**SEIS 631 Foundations of Data Analysis**

**Assignment 6**

In 2004, the state of North Carolina released a large data set containing information on births recorded in the state. This data set is useful to researchers studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of observations from this data set.

**Exploratory analysis:** Load the nc data set into our workspace which is in the **nc.RData** file. We have observations on 13 different variables, some categorical and some numerical. The meaning of each variable is as follows.

| **variable** | **description** |
| --- | --- |
| fage | father’s age in years. |
| mage | mother’s age in years. |
| mature | maturity status of mother. |
| weeks | length of pregnancy in weeks. |
| premie | whether the birth was classified as premature (premie) or full-term. |
| visits | number of hospital visits during pregnancy. |
| marital | whether mother is married or not married at birth. |
| gained | weight gained by mother during pregnancy in pounds. |
| weight | weight of the baby at birth in pounds. |
| lowbirthweight | whether baby was classified as low birthweight (low) or not (not low). |
| gender | gender of the baby, female or male. |
| habit | status of the mother as a nonsmoker or a smoker. |
| whitemom | whether mom is white or not white. |

|  |
| --- |
| **Q1) What does each case (observation) in this data represent? How many cases do we have?** |

As a first step in the analysis, we should consider summaries of the data. This can be done using the summary command:

**summary(nc)**

As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren’t sure or want to take a closer look at the data, make a graph.

Consider the possible relationship between a mother’s smoking habit and the weight of her baby. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

We can compare the means of the distributions using the following function to split the weight variable into the habit groups, then take the mean of each using the mean function.

**by(nc$weight, nc$habit, mean)**

|  |
| --- |
| **Q2) What mean weight do you get for smokers and non-smokers?** |

There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test.

|  |
| --- |
| **Q3) Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.** |

Next, we introduce a new function, inference, that we will use for conducting hypothesis tests and constructing confidence intervals.

**inference(y = nc$weight, x = nc$habit, est = "mean", type = "ht", null = 0,**

**alternative = "twosided", method = "theoretical")**

The first argument is y, which is the response variable that we are interested in: nc$weight. The second argument is the explanatory variable, x, which is the variable that splits the data into two groups, smokers and non-smokers: nc$habit. The third argument, est, is the parameter we’re interested in: "mean" (other options are "median", or "proportion".) Next we decide on the type of inference we want: a hypothesis test ("ht") or a confidence interval ("ci"). When performing a hypothesis test, we also need to supply the null value, which in this case is 0, since the null hypothesis sets the two population means equal to each other. The alternative hypothesis can be "less", "greater", or "twosided". Lastly, the method of inference can be "theoretical" or "simulation" based.

|  |
| --- |
| **Q4) What is the point estimate for this hypothesis test?** |

|  |
| --- |
| **Q5) What is the p-value for this hypothesis test?** |

Change the type argument to "ci" to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

|  |
| --- |
| **Q6) What Confidence Interval do you get?** |

|  |
| --- |
| **Q7) Interpret the Confidence Interval in words.** |

By default the function reports an interval for (*μnonsmoker* − *μsmoker*). We can easily change this order by using the order argument:

inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,

alternative = "twosided", method = "theoretical",

order = c("smoker","nonsmoker"))

|  |
| --- |
| **Q8) What Confidence Interval do you get?** |

**Submission:**

Submit the following two files (no need to compress into a folder):

* **Pdf or word document** with answers to all the questions asked in the assignment.
* **.R file** with relevant code. It is useful to add comments within the code with a pound sign.