

Project #1: Data Summary

The data which is stored in the file **nc2005birth1000.xls** is a random sample of 1000 birth records taken by the North Carolina State Center for Health and Environmental Statistics in 2005. Of particular interest will be incidents of low infant birth weight. Low birth weight is commonly defined as less than 2500 grams. Over the course of the semester we will investigate the relationship of several variables with low birth weight and each other. The goal of this assignment will be to summarize the variables in this data set both graphically and numerically. The variables in this study are:

Variable Label	Description
PLURALITY	Number of children born of the pregnancy
SEX	Sex of child (1=Male, 2=Female)
FAGE	Age of father (years)
MAGE	Age of mother (years)
WEEKS	Completed Weeks of Gestation (weeks)
VISITS	Number of prenatal visits
MARITAL	Marital status (1=married, 2=not married)
RACEMOM	Race of Mother (0=Other Non-white, 1=White, 2=Black 3=American Indian, 4=Chinese, 5=Japanese, 6=Hawaiian, 7=Filipino, 8=Other Asian or Pacific Islander)
HISPMOM	Mother of Hispanic origin (C=Cuban, M=Mexican, N=Non-Hispanic, O=Other and Unknown Hispanic, P=Puerto Rican, S=Central/South American, U=Not Classifiable)
GAINED	Weight gained during pregnancy (pounds)
LOWBW	0=infant was not low birth weight 1=infant was low birth weight
TPOUNDS	Weight of child (pounds)
SMOKE	0=mother did not smoke during pregnancy 1=mother did smoke during pregnancy
MATURE	0=Mother is age 34 or younger 1=Mother is 35 or older
PREMIE	0=infant was not premature 1=infant was premature premature defined at 36 weeks or sooner

The goal of this assignment is to obtain summary statistics for the variables in the data set. This is an important activity of most statistical studies. In your report, clearly label all tables and when appropriate give the units of measure. The components of the assignment are given below. Be sure your

presentation is clear and organized. The use of tables is required.

1. Provide the appropriate numerical summary for each of the variables. This entails determining if you will create a frequency table or a table with the mean, standard deviation, min, Q1, median, Q3, and maximum.
2. Create a histogram for the variables **FAGE**, **MAGE**, **WEEKS**, **GAINED**, and **TPOUNDS**. Describe the shape of the distribution. Is the mean or the median a better measure of center for each of these variables?
3. Construct side-by-side boxplots for the variable of **TPOUNDS** for the two groups of smokers and nonsmokers. Interpret your graph. Does the boxplot indicate a difference in the distribution of baby weight for smoking and nonsmoking mothers? Calculate the mean, median, and standard deviation of pounds for the smoking and nonsmoking mothers. What do these statistics indicate?
4. Suppose a friend of yours has given birth to a 10.3 pound baby. Would you consider this baby “heavy”? Why or why not? In writing explain your reasoning. Another friend of yours had a baby and gained approximately 30 pounds during her pregnancy. Explain to her why she should not be too depressed over this occurrence. Comment on the percentage of women who reported smoking and drinking during pregnancy and the reliability of the responses.
5. Lastly, propose three other variables you would like to investigate in regard to weight of the infant. Give three explicit questions you would ask the mother prior to delivery and explain why you want to know that information.