Take home assignment 1

Your Name Here

Due September 28, 2015

# Instructions

This take home quiz consists of 28 problems with a total of 28 points. Please work individually: do not collaborate with anyone inside or outside of this class as this assignment is a preparation for the in-class exam. You may use your notes and lab exercises.

Please read through carefully, and complete the R code as instructed. Type in your answers in this document in front of the lines that say "#' Answer:". You can continue your answer on the following lines if it's long, but we're generally expecting brief answers.

If you add allowErrorsInNotebook() at some point, be sure to delete it before you compile and turn in the final version of the notebook or you may miss R errors.

**Important:** if you have problems with the R code, or with compiling your notebook, email your TA or the instructor for help. You are being tested on biostatistical concepts and interpretation, not on your ability to program in R.

## When you are ready to submit

1. Make sure your name is in the "author:" field at the top of the document, (keep the quotes!)
2. Click on File > Compile notebook... to create a word document from your code. Check the word document for completeness
3. If you used allowErrorsInNotebook(), don't forget to remove it before the final run so you'll be sure to catch any errors.
4. Upload the word document (not the R code!) to Blackboard. There will be a folder for take home 1.

This quiz will be graded and factored into your final course grade as indicated in the syllabus.

# Getting set up

library(PubH6002) # the current version is 0.23  
  
babies = read.csv('babies.csv')  
use(babies)

## Note: the variables in this dataset are now visible

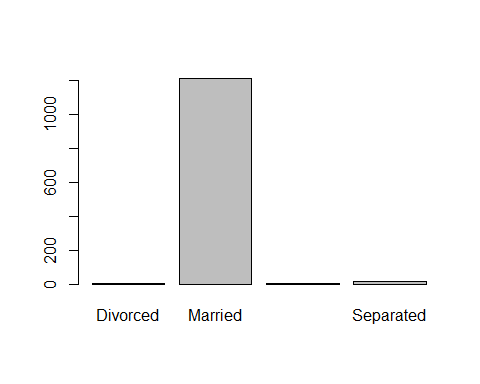
The data are a subset from a comprehensive study of all pregnancies that occurred between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco–East Bay area. The subset is for baby boys born in 1961, of a single birth (i.e. not a twin or triplet) and who survived at least 28 days.

# Marital Status (variable marital)

#### Question 1 (1pt)

The marital status of the mother is recorded in variable marital. Create a barplot of this variable

height <- table(babies$marital)  
barplot(height)



#### Question 2 (1pt)

What kind of variable is this: a) numerical, b) nominal, c) ordinal, or d) binary?

Answer: ordinal

#### Question 3 (1pt)

What percentage of mothers are married? (use the summary function)

summary(marital)

## Divorced Married Never Married Separated NA's   
## 5 1208 6 15 2

Answer:

1208/nrow(babies)

## [1] 0.9773463

#### Question 4 (1pt)

From an analytic point of view, do you see potential problems with this variable?

Answer:

# Infant's Birth Weight (variable wt)

#### Question 5 (1pt)

The variable wt contains the weight of the infants at birth, in ounces.

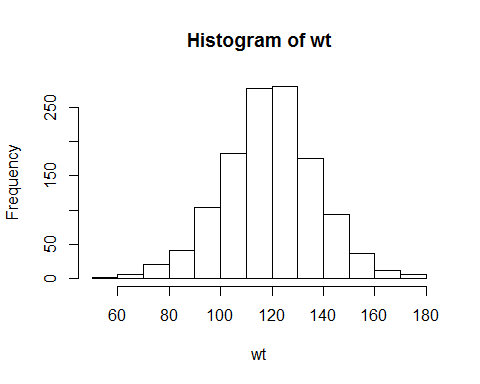
What kind of variable is this: a) numerical, b) nominal, c) ordinal, or d) binary?

Answer: a

#### Question 6 (1pt)

Create histogram and boxplot for variable wt

hist(wt)

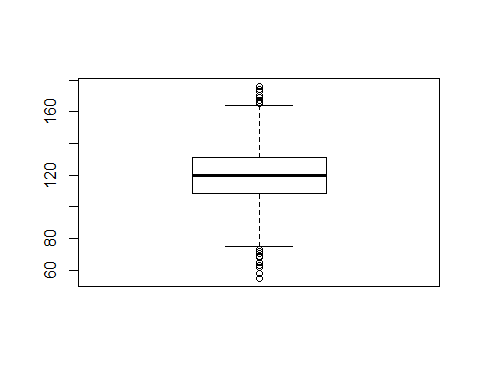


#### Question 7 (1pt)

From the boxplot, what is the median birth weight?

Answer:

boxplot(wt)



#### Question 8 (1pt)

What is the average weight? (you can round to whole ounces if you want)

summary(wt)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 55.0 108.8 120.0 119.6 131.0 176.0

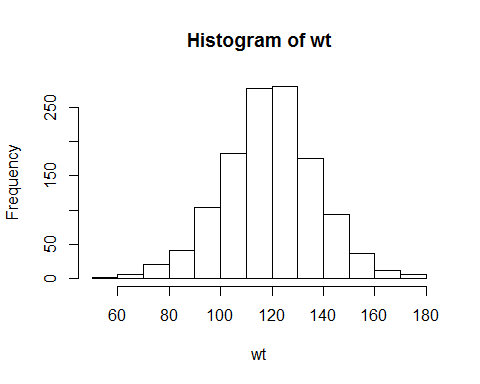
Answer:

#### Question 9 (1pt)

How would you describe the distribution of birth weights: a) more or less symmetric, b) severely skewed to the left, or c) severely skewed to the right?

Answer:

hist(wt)



#looks symmetric

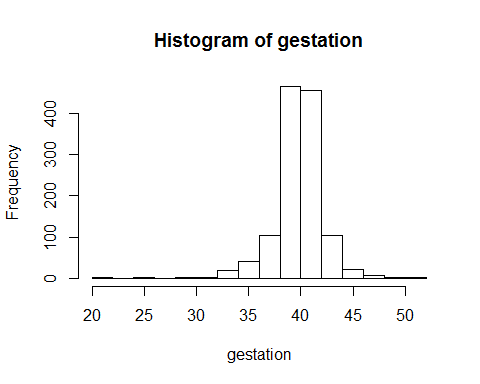
# Infant's Gestational Age (variable gestation)

The gestational age is the duration of the pregnancy in weeks, measured from the first day of the mother's last menstrual cycle to the current date (the infant's birth here). Full term gestation period for a baby is typically 40 weeks. Infants born before 37 weeks are considered preterm or premature. Births after 42 weeks are considered postterm.

#### Question 10 (1pt)

Get a histogram of gestation age (variable gestation)

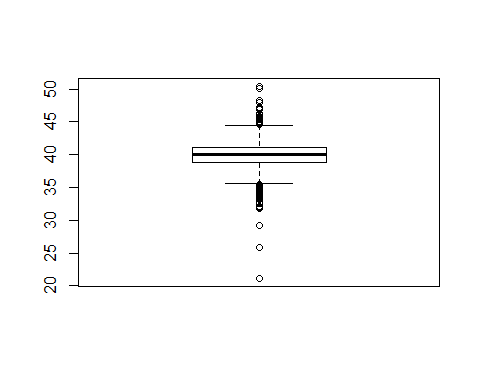
hist(gestation)



#### Question 11 (1pt)

What is the median gestation age?

boxplot(gestation)



median(gestation,na.rm = TRUE)

## [1] 40

Answer:

#### Question 12 (1pt)

Does this distribution have possible outliers? are they on the low end of gestational age, on the high end, or both ends?

Answer: both ends

# Father's age (variable dage)

#### Question 13 (1pt)

The variable dage contains the age of the father in years.

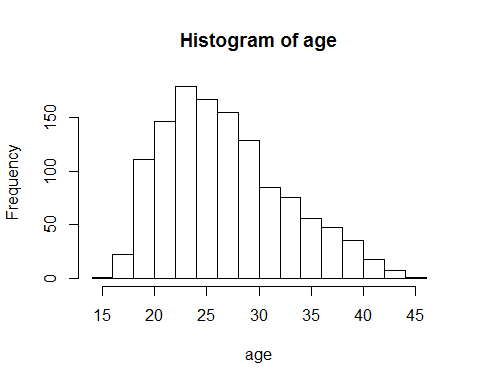
What kind of variable is this? a) numerical b) nominal c) ordinal or d) binary?

Answer: a

#### Question 14 (1pt)

Create histogram and boxplot for variable age

hist(age)



#### Question 15 (1pt)

The distribution of ages is skewed. Is it skewed to the right or to the left? Answer:

#### Question 16 (1pt)

Use the summary function to obtain basic statistics on the variable dage.

summary(dage)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 25.00 29.00 30.74 35.00 99.00

#### Question 17 (1pt)

What is the age of the youngest father in this dataset? Answer:

#### Question 18 (1pt)

Does it look like there are bad data values in this variable? Explain.

Answer:

#### Question 19 (1pt)

The summary function reports several statistics such as the mean, median, etc. Explain why those statistics are not useful here.

Answer:

# Maternal Smoking (variable smoke)

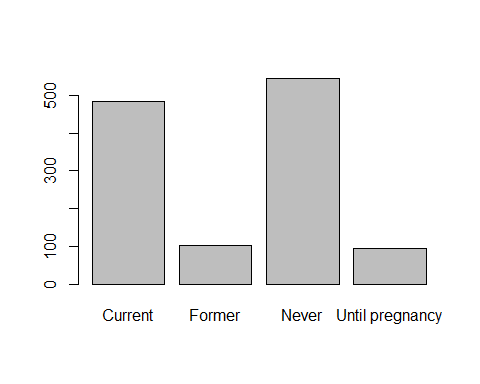
The variable smoke records whether the mother smokes or smoked in the past. The possible answers are

* Never smoked
* Current smoker
* Was a smoker until the current pregnancy (i.e. stopped when she got pregnant)
* Former (i.e. stopped before the current pregnancy)
* Unkown (NA)

#### Question 20 (1pt)

We will consider this variable nominal. Produce a barplot of variable smoke

height <- table(babies$smoke)  
barplot(height)



#### Question 21 (1pt)

What are the two most common levels?

Answer:

#### Question 22 (1pt)

What are the number and proportion of mothers who never smoked in this dataset?

summary(smoke)

## Current Former Never Until pregnancy   
## 484 103 544 95   
## NA's   
## 10

Answer:

#### Question 23 (1pt)

How many missing values do we have for this variable?

Answer:

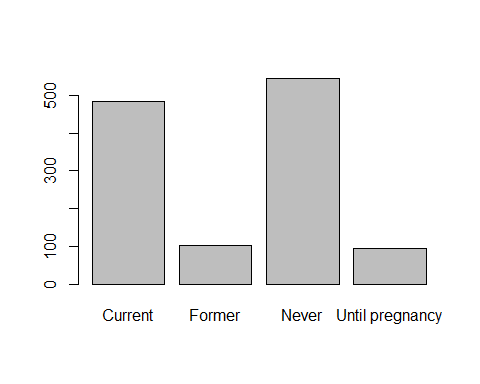
# Number of cigarettes smoked (variable number)

#### Question 24 (1pt)

The variable number contains the average number of cigarettes smoked per day (0 for non-smokers)

Create a barplot

height <- table(babies$smoke)  
barplot(height)



#### Question 25 (1pt)

What kind of variable is this? a) numerical b) nominal c) ordinal or d) binary?

Answer:

#### Question 26 (1pt)

Among smokers (i.e. excluding 0) what is most common level?

Answer:

#### Question 27 (1pt)

How many mothers never smoked based on this variable? Does it agree with the number in the previous section?

summary(smoke)

## Current Former Never Until pregnancy   
## 484 103 544 95   
## NA's   
## 10

Answer:

#### Question 28 (1pt)

From an analytical point of view, are there problems with this variable? Answer: