STAT 614 - HW 1

Problem 1

1. How many individuals are in the data set?

There are 124 individuals although not every individual has values for every variable.

2. How many variables are in this data set?

There are 40 variables.

3. Can you tell if any of the variables are categorical (i.e. qualitative)? Identify specific ones.

Yes, exposure GROUP is categorical. As is a person's biological sex. The area in which they lived is also categorical. (Others *may be* categorical; I just haven't checked!)

Two important variables that were studied were (1) MAXFT = the number of finger-wrist taps in the dominant hand (a measure of neurological function) and (2) IQF = the Wechsler full-scale IQ score. You will explore the relationship of lead exposure to one of these two outcome variables.

4. Is this an observational study or a randomized experiment? Explain why.

This is an observational study. Exposure levels were not randomized

5. How many individuals have MAXFT scores measured? How many have IQF scores measured?

All 124 individuals have IQF scores, but 25 individuals are missing MAXFT scores so 124 – 25 = 99 have MAXFT scores.

6. **Pick one of** `MAXFT` or `IQF` of interest to you. We are primarily interested in comparing the distribution of the outcome of interest (MAXFT or IQF) for the two different groups of children (GROUPS 1 and 2, those children with elevated blood-lead levels $\geq$ 40 μg/ml and those with lower levels, < 40 μg/ml, respectively.)

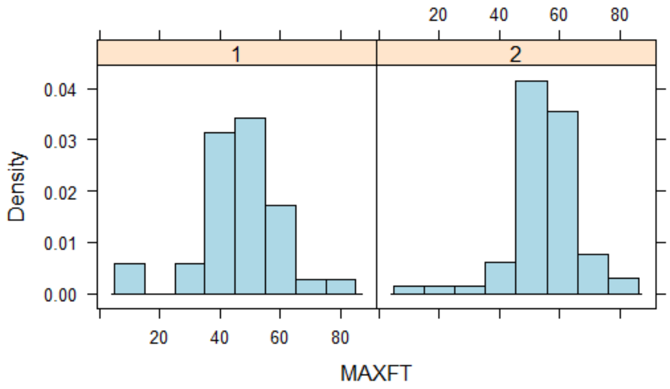
For MAXFT:

a. What are the mean and median of the outcome of interest MAXFT for each GROUP?

The mean finger wrist-tap measurement of the high blood-lead level group is 47.43 taps and the mean of the low group is 54.44 taps. The median finger wrist-tap measurement of the two groups is 48 and 53.5 taps, respectively. These give estimates for the typical values of finger wrist-tap measurements of the two groups.

b. Describe the shape of the distribution (i.e. histogram) of the outcome for each GROUP.

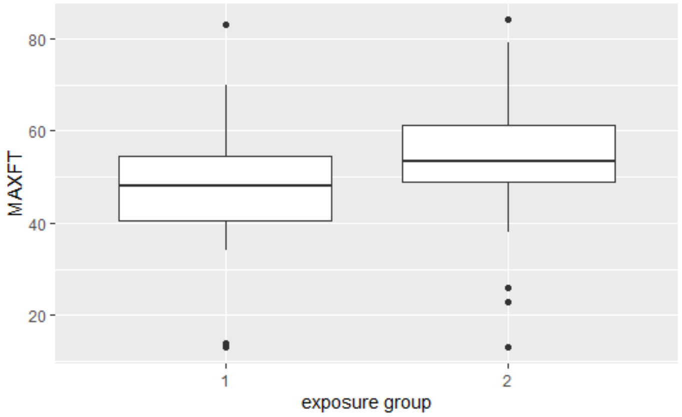
The MAXFT variable for each group as unimodal and roughly symmetric (about the respective means or medians given in part (a) above). The high blood-lead level group has outliers on the lower and upper ends and is slightly more variable – as demonstrated by the standard deviations, which for the high exposure group is 13.16 taps and for the low group is 12.06 taps.



c. What information can we get from the Boxplot of the outcome for each GROUP?

The boxplot illustrates the five-number summary and highlights the outliers. It can also be used to assess the symmetry of the distributions. There is some slight skewness displayed in the low exposure level group (GROUP=2).

```
factor(GROUP) min   Q1 median    Q3 max    n missing
            1  13 40.5   48.0 54.50  83   35      11
            2  13 49.0   53.5 61.25  84   64      14
```



d. Based on these summaries, what is your assessment of the differences between the two groups of children on the outcome of interest? Discuss the role of randomization in this study.

These results suggest that, on average, the low blood-lead level group had higher mean finger wrist-tap scores and hence, higher neurological function than the high blood-lead level group.

(Note: There is a great deal of overlap in the two distributions, but you can see an association/difference between the groups. Now, whether this is a meaningful difference requires expert knowledge of MAXFT! We will use a formal test to assess whether observed differences are “statistically different”, i.e. expected by chance or unexpected by chance and hence potentially “an effect.” Statistically different is not the same as meaningful or practical difference!)

The participants were not randomly selected to be in the study and therefore we cannot or should hesitate to generalize these results to the population of all children. Furthermore, the treatment (high vs low exposure) was not randomly assigned to the subjects and therefore we cannot conclude definitively that it is the lead exposure that is causing the differences in neurologic function There could be other variables which were not measured that actually explain the differences between the two groups. (Where people live might be related to their socioeconomic status, which could have in impact on educational opportunities or health care individuals receive and these could be explaining the differences in development seen here.)

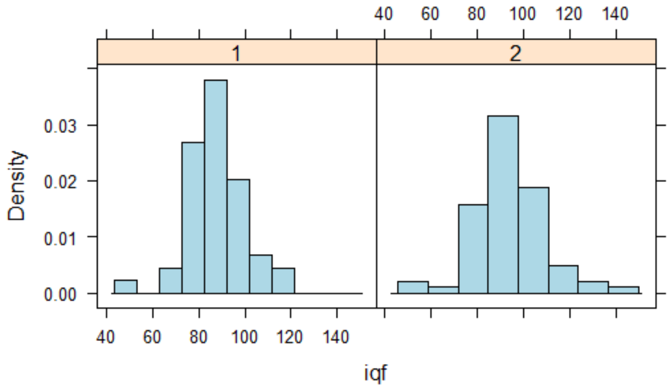
For IQF:

a. What are the mean and median of the outcome of interest IQF for each GROUP?

The mean IQF of the high lead exposure group is 88.02 units with a median of 88 and the mean IQF of low lead exposure group is 92.88 units with a median of 94.

b. Describe the shape of the distribution (i.e. histogram) of the outcome for each GROUP.

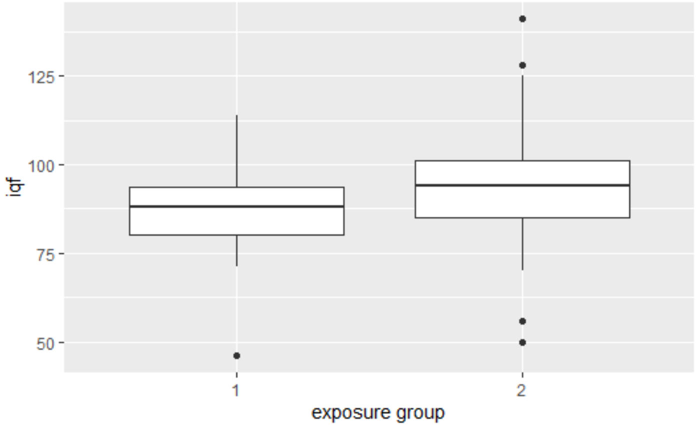
The distribution of IQF is mostly symmetric and unimodal with some outliers. The high exposure group has one low IQF outlier while the low exposure group has two high and two low IQF outliers (as displayed in the boxplot below). These plots also illustrate that there is a high amount of overlap in the two distributions.



c. What information can we get from the Boxplot of the outcome for each GROUP?

The boxplot illustrates the five-number summary and highlights the outliers. It can also be used to assess the symmetry of the distributions. These look reasonably symmetric.

```
 factor(GROUP) min Q1 median     Q3 max  n missing
             1  46 80     88  93.75 114 46       0
             2  50 85     94 101.00 141 78       0
```



d. Based on these summaries, what is your assessment of the differences between the two groups of children on the outcome of interest? Discuss the role of randomization in this study.

These results suggest that, on average, the low blood-lead level group had higher mean IQF scores and hence, higher neurological function than the high blood-lead level group.

(Note: There is a great deal of overlap in the two distributions, but you can see an association/difference between the groups. Now, whether this is a meaningful difference requires expert knowledge of IQF! We will use a formal test to assess whether observed differences are “statistically different”, i.e. expected by chance or unexpected by chance and hence potentially “an effect.” Statistically different is not the same as meaningful or practical difference!)

The participants were not randomly selected to be in the study and therefore we cannot or should hesitate to generalize these results to the population of all children. Furthermore, the treatment (high vs low exposure) was not randomly assigned to the subjects and therefore we cannot conclude definitively that it is the lead exposure that is causing the differences in neurologic function There could be other variables which were not measured that actually explain the differences between the two groups. (Where people live might be related to their socioeconomic status, which could have in impact on educational opportunities or health care individuals receive and these could be explaining the differences in development seen here.)

Problem 2

Re-visit the study design for Case Study 1.1.2 Sex Discrimination in Employment from Chapter 1 of the textbook or our class notes. Briefly contrast the case study design with that of the study described in the *New York Times* article "Bias Persists for Women of Science, a Study Finds'' and in the manuscript *Science Faculty's Subtle Gender Biases Favor Male Students* (both are given in Blackboard under the *Homework* tab.) Briefly describe the overall goals of each study and give the strengths and weakness of each. Which study do you find has more compelling evidence for the hypotheses of interest?

Answers will vary but the crux is that the *Science Faculty* study involved a randomized experiment. If you get this point, you get full credit! There are many items that can be critiqued about a study design so if what you say is reasonable, then that is fine. Ask me if you have questions.