Professor Notes About the "Univariate EDA - Quant" Homework

- Question 1 asks for a "univariate EDA" – Note that this does not mean to just make a histogram and summary statistics. Performing a univariate EDA means addressing shape, outliers, center, and dispersion from looking at a histogram (or boxplot) and summary statistics. You MUST remember to explain why you chose to use the mean/sd or median/IQR to address shape/center. Your description of the outlier should be specific (i.e., say that "there is an outlier at 71" not that "there is an outlier in the 70-80 range.")
- In question 1 the shape, outliers, center, and dispersion are specifically listed, even if there was no outliers, for both the private and state labs. Also note that I clearly indicated why I chose to use the mean/sd (because of symmetry and no outliers) or median/IQR (occurrence of outliers).
- It would be reasonable to use the median and IQR for both state and private labs in question 1 because of the outlier in the state lab. This would allow a more appropriate comparison between labs in question 2.
- Note how each table and figure is labeled and referred to in the answers.
- You should use no more than one more decimal place than what was recorded for the quantitative variable.

Effluent Sampling Labs

1. The distribution of BOD values for the private lab is approximately symmetric with no obvious outliers (Figure 1). The mean of the private data is 34.64 mg/L with a standard deviation of 10.45 mg/L (Table 1). The mean and standard deviation were used because of the symmetry of the distribution and absence of outliers.

   The distribution of BOD values for the state lab appears to be right-skewed with an outlier at 71 mg/L (Figure 1). The median of the state data is 20 mg/L with an IQR from 9.5 to 33.5 mg/L (Table 1). The median and IQR were used because of the presence of the outlier.

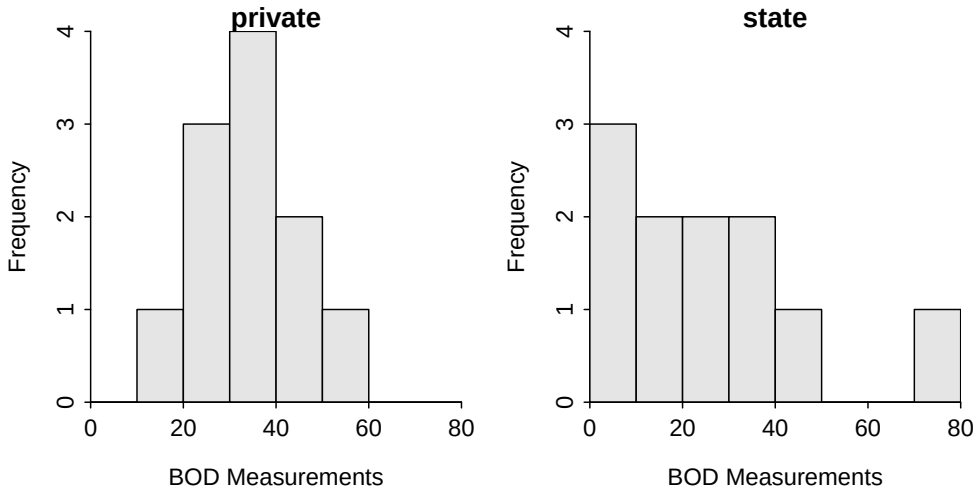

Figure 1. Histograms for BOD measurements for the private (left) and state (right) labs.

Table 1. Descriptive statistics for BOD measurements for the private and state labs.

	lab	n	mean	sd	min	Q1	median	Q3	max
1	private	11	34.6	10.5	15	28.5	35	40.5	54
2	state	11	25.3	19.7	6	9.5	20	33.5	71

2. The two most outstanding differences between the private and state labs are that the state lab BOD values are generally lower (note lower means and medians) and more dispersed (note wider IQR and larger standard deviation).

R Appendix

```
library(NCStats)
setwd('C:/aaaWork/Books/IntroStats/HW/')
d <- read.csv("4_50.csv")
str(d)
hist(bod~lab,data=d,xlab="BOD Measurements")
Summarize(bod~lab,data=d,digits=1)
```