Key for Chem 351: First Exam

To begin, let's load and source data files, scripts, and packages used in this answer key.

```
# loading and sourcing data files and scripts and packages
load("tin.RData")
load("uranium.RData")
load("instruments.RData")
source("shadeArea.R")
library(alr3)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

Problem One

Part (a)

The mean, median, standard deviation, and MAD are shown below; note that we need to include the argument constant = 1 to obtain the MAD value as we defined it in class. All values are in ppm.

```
mean(tin)
```

```
## [1] 256.4416
```

```
median(tin)
```

```
## [1] 255.5
```

```
sd(tin)
```

```
## [1] 64.35412
```

```
mad(tin, constant = 1)
```

```
## [1] 23
```

The mean and the median are similar in value, which suggests the distribution of data is reasonably symmetrical about its central value; if this was not the case, then we would expect to see a clear difference between the two because the mean is more sensitive than the median to skewed data. Because the standard deviation is more sensitive to extreme values than is the MAD, the difference in these values suggests that there are extreme values in the data. Together, these four values suggest that the distribution of results is reasonably symmetrical with extreme values at both ends of the distribution.

Part (b)

A box plot of the full data set, which is shown in Figure 1, indicates that the four labs reporting the smallest values (76.5 ppm, 114 ppm, 155 ppm, and 185 ppm) and the two labs reporting the largest values (395 ppm and 522.09 ppm) are outliers.

```
boxplot(tin, horizontal = TRUE)
```

Removing these six labs from the data set and recalculating the summary statistics (results reported below) suggests the reduced data set has a distribution that is symmetrical around its center, as the mean and the median are similar, and that is less affected by extreme values, as the standard deviation and the MAD are now closer to each other in value. When we wish to characterize an analytical method's accuracy and precision across a large number of laboratories, we need to ensure their work reflects variation across the

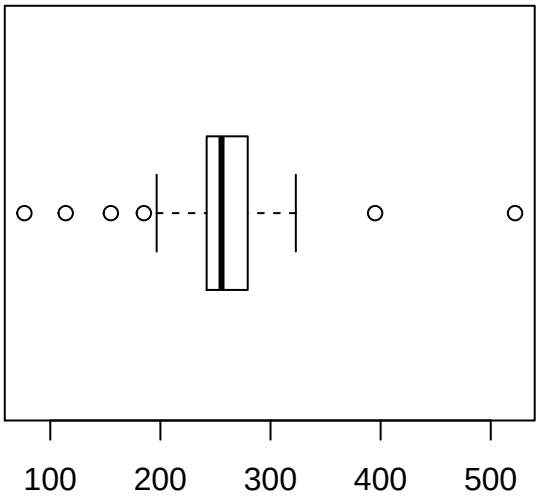

Figure 1: Box plot for the data in problem 1.

same set of determinate errors as differences in systematic errors between the labs widens the distribution and overestimates the method's standard deviation.

```
# note: there are many ways to remove values from a vector; the
# one shown here uses the concatanate function, c(), to idenfity
# the values to delete, places a minus sign in front of it to
# indicate delete and places both inside the bracket, [ ],
# notation used to identify elements in a vector
tin.new = tin[-c(1:4, 45, 46)]
mean(tin.new)
```

```
## [1] 258.7181
```

```
median(tin.new)
```

```
## [1] 258
```

```
sd(tin.new)
```

```
## [1] 26.9509
```

```
mad(tin.new, constant = 1)
```

```
## [1] 15.895
```

Part (c)

To find the probability we use the "pnorm" function and use the mean and the standard deviation for the reduced data set to define the normal distribution. We also use the lower.tail argument, as needed, to control the end of the distribution we need to consider.

```
m = mean(tin.new)
s = sd(tin.new)
# part (a)
pnorm(240, mean = m, sd = s)
```

```
## [1] 0.2436768
```

```
# part (b)
pnorm(280, mean = m, sd = s, lower.tail = FALSE)
```

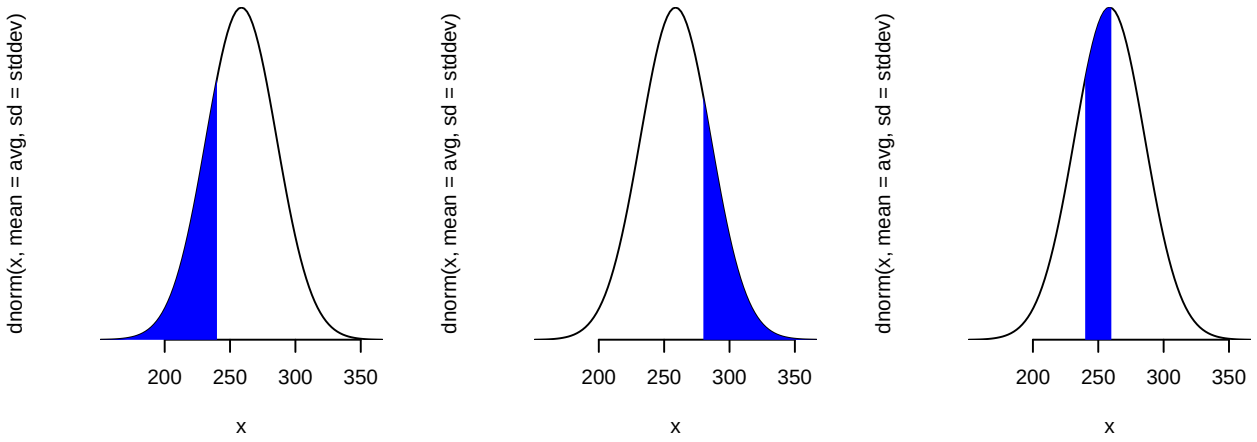

Figure 2: From left-to-right, areas under the normal distribution curve for scenarios i, ii, and iii.

```
## [1] 0.2148642
```

```
#part(c)
pnorm(260, mean = m, sd = s) - pnorm(240, mean = m, sd = s)
```

```
## [1] 0.2752922
```

As shown above, we see that 24.4% of results are less than 240 ppm, that 21.5% of results are greater than 280 ppm, and that 27.5% of results are between 240 ppm and 260 ppm. Figure 2 shows the corresponding areas under the normal distribution curves.

```
old.par = par(mfrow = c(1, 3))
shadeArea(tin.new, 0, 240)
shadeArea(tin.new, 280, 500)
shadeArea(tin.new, 240, 260)
```

```
par(old.par)
```

Part (d)

The result from ICP-IDMS is described as a "standard value," which means we take this as the accepted or theoretical value μ, and use a t-test to evaluate whether there is a significant difference between the mean of the reduced data set and the accepted value. The null hypothesis is that the mean for the reduced data set is the same as the accepted value, and the alternative hypothesis is that the two values are not the same; that is, this is a two-tailed t-test.

```
t.test(x = tin.new, mu = 286.5, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  tin.new
## t = -6.5196, df = 39, p-value = 9.85e-08
## alternative hypothesis: true mean is not equal to 286.5
## 95 percent confidence interval:
##  250.0987 267.3374
## sample estimates:
## mean of x
##  258.7181
```

The p-value from the t-test is less than 0.05; thus, we reject the null hypothesis and accept the alternative

hypothesis that there is a significant difference between the mean of the reduced data set and the accepted value that we cannot explain by the uncertainty in the measurements.

Part (e)

There are a number of ways we can show this, two of which are highlighted here. The first approach is to use a stem-and-leaf plot, using the scale argument to spread out the data. A second approach is to use a histogram and the breaks option to increase the number of bins (but, this requires care as it is easy to manipulate the breaks to give an outcome we "desire"" rather than letting the result emerge from the data itself). The stem-and-leaf plot is shown here

```
stem(tin.new, scale = 2)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   19 | 7
##   20 | 3
##   21 | 8
##   22 | 068
##   23 | 3
##   24 | 234467788
##   25 | 1256
##   26 | 03556
##   27 | 224799
##   28 | 00145
##   29 | 06
##   30 |
##   31 | 0
##   32 | 3
```

and the histogram in Figure 3. Note that both plots suggest we can subdivide the results into three groups, each in the form of a peak in the distribution: a very small set of results centered at approximately 220 ppm, a small, but narrowly distributed set of results centered at approximately 240 ppm, and a larger and more broadly distributed set of results centered at approximately 270 ppm.

```
hist(tin.new, breaks = 10, main = "")
abline(v = 222, lwd = 3, col = "red")
abline(v = 244, lwd = 3, col = "green")
abline(v = 276, lwd = 3, col = "blue")
```

Problem Two

To compare the results from the field and from the lab, we use a t-test. More specifically, because each site is unique and its sample is analyzed both in the field and in the lab, we use a paired t-test. This is critical because the variation in results between sites (samples) is so large that it likely is larger than the variation between field and lab, compromising the comparison. The null hypothesis is that the field and the lab give the same results—that is, the difference between them is zero—and the alternative hypothesis is that their results are different. Because we have no reason to believe that "degradation of samples" must lead to a decrease in concentration, we use a two-tailed t-test.

```
t.test(x = field, y = lab, mu = 0, alternative = "two.sided", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  field and lab
```

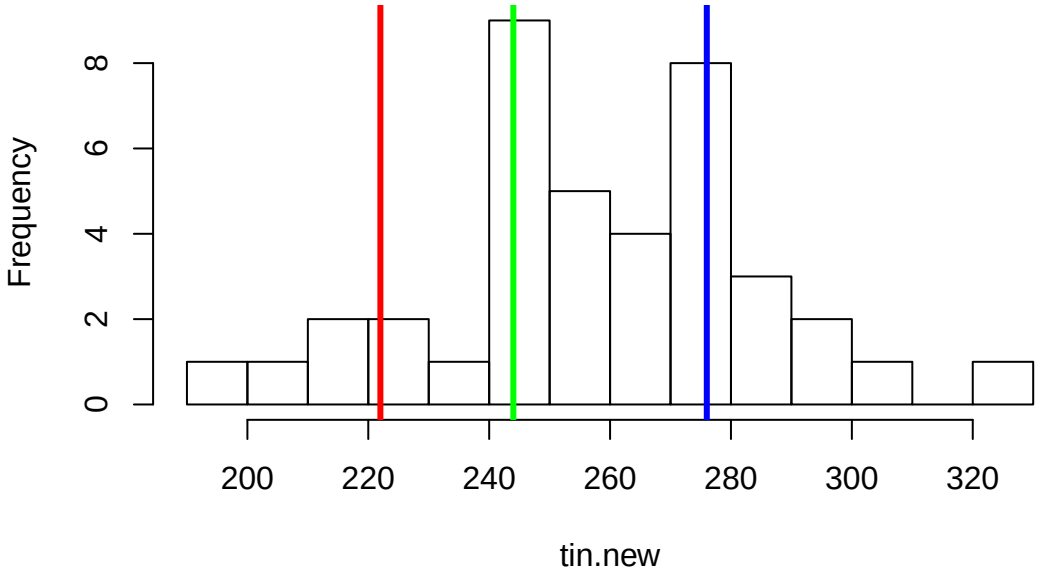

Figure 3: Histogram for the reduced set of laboratories showing evidence of three distinct groups

```
## t = -1, df = 19, p-value = 0.3299
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.351165  2.951165
## sample estimates:
## mean of the differences
##                    -2.7
```

With a p-value of 0.3299, we have no evidence at $\alpha = 0.05$ that we cannot attribute the difference between the results when run in the field and when run in the lab to uncertainty in the measurements.

Problem Three

To determine if there is a difference in the results obtained by these instruments we use an analysis of variance. To do so, we first need to place the data into a two-column data frame consisting of a column for results and column to identify the instrument; note, we do not need the column for days, so we first remove it from the data frame.

```
# note: there are several ways to remove a column from a data
# frame; here using [ , -9] means keep all rows but discard the
# ninth column
inst = instruments[ , -9]
inst = stack(inst)
head(inst)
```

```
##   values   ind
## 1  9.290 inst1
## 2  8.147 inst1
## 3  7.916 inst1
## 4  8.511 inst1
## 5 15.789 inst1
## 6  9.936 inst1
```

Now that the data is properly organized, we complete an analysis of variance using the formula values ~ ind.

The null hypothesis is that the variance between the instruments and the variance within the instruments are equal, and the alternative hypothesis is that the variance between the instruments is significantly larger than the variance within the instruments.

```
inst.aov = aov(values ~ ind, data = inst)
summary(inst.aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ind           7   1818  259.76   75.84 <2e-16 ***
## Residuals   328   1123    3.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of less than 2e-16, we have good evidence that the variance between the instruments is so large that we cannot attribute it to the uncertainty in the results of each instrument. To consider how we might divide the instruments into distinct groups, we use the TukeyHSD test.

```
inst.hsd = TukeyHSD(inst.aov)
inst.hsd
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = values ~ ind, data = inst)
##
## $ind
##                    diff         lwr        upr     p adj
## inst2-inst1 -0.3551667 -1.58716313  0.8768298 0.9877184
## inst3-inst1 -0.1455476 -1.37754408  1.0864488 0.9999621
## inst4-inst1  1.7147143  0.48271783  2.9467107 0.0007398
## inst5-inst1 -4.4930952 -5.72509170 -3.2610988 0.0000000
## inst6-inst1 -5.1136190 -6.34561551 -3.8816226 0.0000000
## inst7-inst1 -1.9849762 -3.21697265 -0.7529797 0.0000383
## inst8-inst1 -3.9469524 -5.17894884 -2.7149559 0.0000000
## inst3-inst2  0.2096190 -1.02237741  1.4416155 0.9995603
## inst4-inst2  2.0698810  0.83788449  3.3018774 0.0000140
## inst5-inst2 -4.1379286 -5.36992503 -2.9059321 0.0000000
## inst6-inst2 -4.7584524 -5.99044884 -3.5264559 0.0000000
## inst7-inst2 -1.6298095 -2.86180598 -0.3978131 0.0017261
## inst8-inst2 -3.5917857 -4.82378217 -2.3597893 0.0000000
## inst4-inst3  1.8602619  0.62826545  3.0922584 0.0001575
## inst5-inst3 -4.3475476 -5.57954408 -3.1155512 0.0000000
## inst6-inst3 -4.9680714 -6.20006789 -3.7360750 0.0000000
## inst7-inst3 -1.8394286 -3.07142503 -0.6074321 0.0001980
## inst8-inst3 -3.8014048 -5.03340122 -2.5694083 0.0000000
## inst5-inst4 -6.2078095 -7.43980598 -4.9758131 0.0000000
## inst6-inst4 -6.8283333 -8.06032979 -5.5963369 0.0000000
## inst7-inst4 -3.6996905 -4.93168693 -2.4676940 0.0000000
## inst8-inst4 -5.6616667 -6.89366313 -4.4296702 0.0000000
## inst6-inst5 -0.6205238 -1.85252027  0.6114726 0.7871073
## inst7-inst5  2.5081190  1.27612259  3.7401155 0.0000000
## inst8-inst5  0.5461429 -0.68585360  1.7781393 0.8777701
## inst7-inst6  3.1286429  1.89664640  4.3606393 0.0000000
## inst8-inst6  1.1666667 -0.06532979  2.3986631 0.0781691
## inst8-inst7 -1.9619762 -3.19397265 -0.7299797 0.0000500
```

Although we can examine this graphically, it is easy to interpret the summary results, which show us that

instruments 1–3 seem to have similar results and that instruments 5, 6, and 8 also seem to have similar results (although the p-value for the difference between instrument 6 and instrument 8 is close to the limit of 0.05); the remaining two instruments—4 and 7—are significantly different from each other and from the other instruments.