STAT 3743 • PROBABILITY & STATISTICS • FALL 2010 • KERNS

Exam I

Name: ANSWER KEY

Note: the questions are randomly generated so these may (not) exactly match those on your paper. The answers below are for *these* and if you have trouble seeing the connection between these and those, ask me.

Directions: SHOW ALL WORK. You may use R for computations, but no other software (and in particular, not the Internet). If you use R to calculate something, then hand write the R code that you typed, together with the numerical answer.

1. For this problem we will study the `USJudgeRatings` data. You can read about them with `?USJudgeRatings`. In particular, let us focus on the variable `INTG`.

   (a) First, store the values of `INTG` in a vector `x`. The quickest way to do this is

   ```
   x <- USJudgeRatings$INTG
   ```

   (There isn't anything to write down for this part).

   (b) Find the IQR of `INTG`.

   (c) Find the Five Number Summary (5NS).

   (d) Use the 5NS to calculate what the width of a boxplot of `INTG` would be.

   (e) Compare your answers (b) and (d). Are they the same? If not, are they close?

   (f) Make a boxplot of `INTG`, and include a sketch of it in your report.

   (g) Are there any potential/suspected outliers? If so, list their values. *Hint:* take a look at `sort(x)`.

   (h) Using the rules discussed in class, classify each of the answers to (g), if any, as *potential* or *suspected* outliers.

Solution:

I will forego showing the first part. We will do the next two parts at once.

```
> IQR(x)

[1] 1

> fivenum(x)

[1] 5.90 7.55 8.10 8.55 9.20
```

The width of the box of a boxplot is the upper hinge minus the lower hinge.

```
> fivenum(x)[4] - fivenum(x)[2]

[1] 1
```

For these data the IQR and the width of the box are identical. For many data sets, however, they are off by a little bit.

Below is a boxplot of the data.

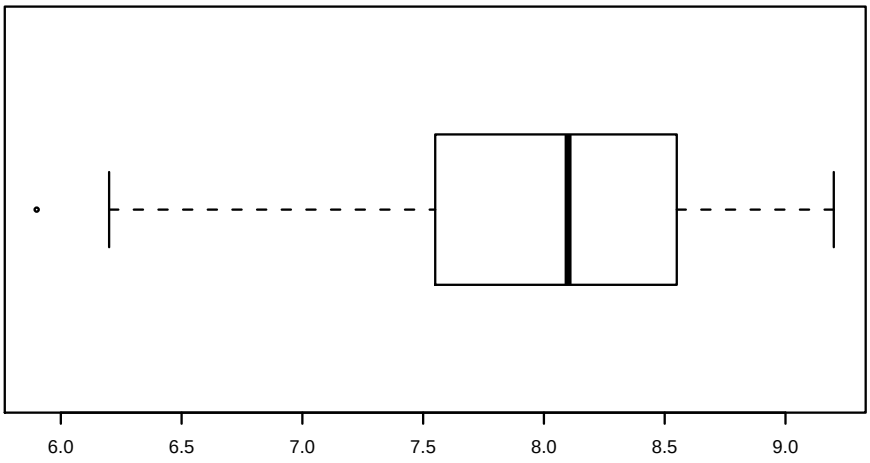

Figure 1: Boxplot of `USJudgeRatings$INTG`

We can see from the boxplot that for these data there is one (1) extreme value, off to the left. The quickest way to get its value is with the `boxplot.stats` function.

```
> boxplot.stats(x)$out

[1] 5.9
```

We could just as easily have sorted the data with `sort(x)` and then looked for the smallest data value (the first entry).

We next use the rules from class to see whether the datum is a potential versus suspected outlier. We do that by calculating the lower and upper fences. Here is a quick way to do that (or you can do it by hand).

```
> w <- fivenum(x)[4] - fivenum(x)[2]
> fivenum(x)[2] - c(3, 1.5) * w

[1] 4.55 6.05

> fivenum(x)[4] + c(1.5, 3) * w

[1] 10.05 11.55
```

Since the datum does not fall outside the outer fence on the left, it is merely a potential outlier.

2. This problem studies the `USJudgeRatings` data. You can read about them with `?USJudgeRatings`. Type `head(USJudgeRatings)` at the command prompt for a quick look at the top of the data set.

   (a) Identify the data type of each of the variables.

   (b) Now type `attach(USJudgeRatings)` at the command prompt which will allow you to simply type variable names without the dollar signs. Try it. Type `ORAL` at the command prompt.

      i. Choose an appropriate visual display for `ORAL` and sketch the graph (just a sketch). You will want to try several choices before you decide on an "appropriate" one.

      ii. Report at least two (2) measures of center for `ORAL`. Based on what you know about the data from above, make a decision about which measure is the better one for these data, and tell me why.

      iii. Report at least two (2) measures of spread for `ORAL`. Again, based on what you know about the data from above, make a decision about which measure is the better one for these data, and tell me why.

      iv. Report at least two (2) measures of shape for `ORAL`. Use the rules-of-thumb we discussed in class to decide if the values you observed are substantially different from zero. *Hint:* don't forget `library(e1071)`.

   (c) Report any other unusual features of `ORAL` that you see.

Solution:

(a) These data are all quantitative, presumably continuous (or at least likely should be taken as continuous).

(b) Descriptive statistics

   i. Visual display. We will put down several to give us a feel for the data.

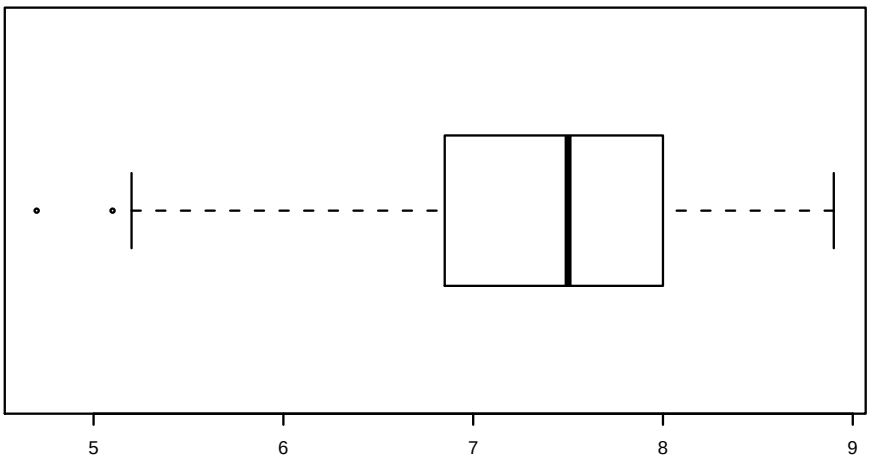

Figure 2: Boxplot of `USJudgeRatings$ORAL`

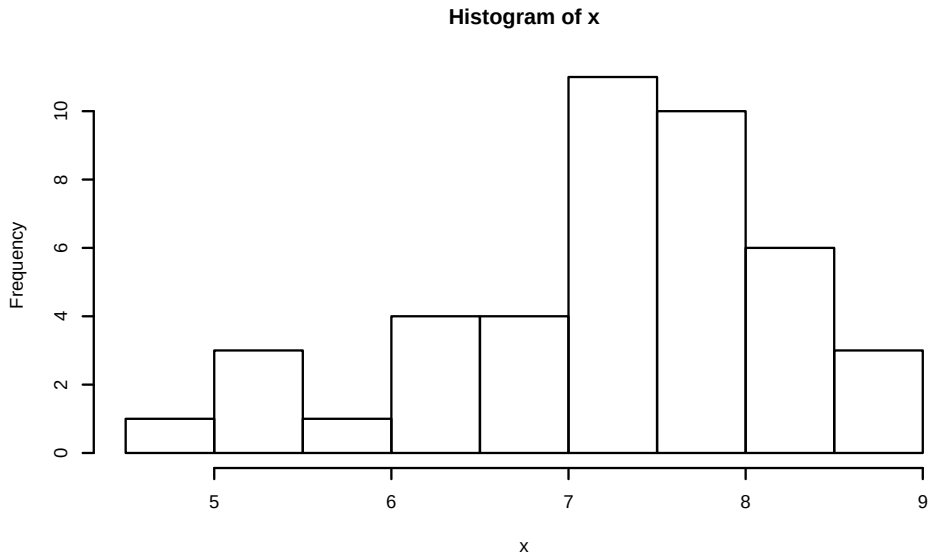

Figure 3: Histogram of `USJudgeRatings$ORAL`

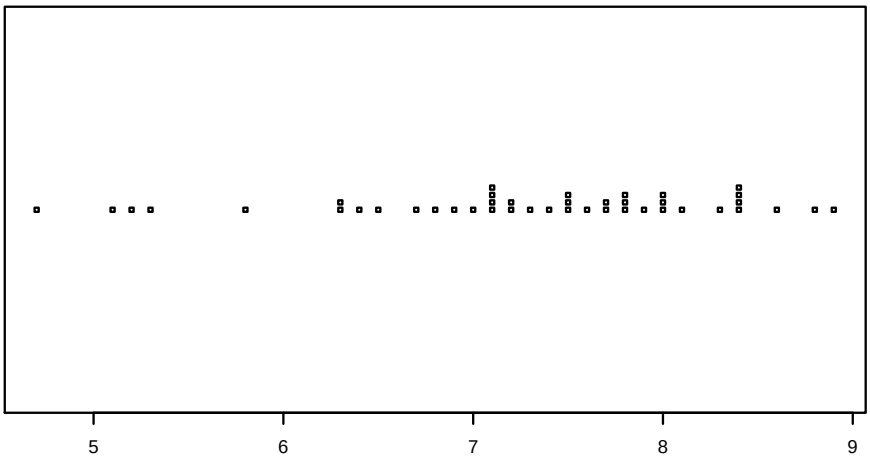

Figure 4: Stripchart of `USJudgeRatings$ORAL`

```
> library(aplpack)
> stem.leaf(x)
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 43
LO: 4.7
    4    5* | 123
```

```
  5    5. | 8
  8    6* | 334
 12    6. | 5789
 21    7* | 011112234
(10)   7. | 5556778889
 12    8* | 000134444
  3    8. | 689
```

For these data it looks like the strip chart conveys the most information about the data (the stemplot does a good job, too). The boxplot really loses most of the information about the shape and granularity of the data, while the histogram falls in the middle, better than the boxplot but not quite as informative as the strip chart.

ii. Measures of center.

```
> mean(x)
[1] 7.293023
> median(x)
[1] 7.5
> mean(x, trim = 0.05)
[1] 7.335897
```

Since these data had potential outlier(s) we should choose a resistant measure of center such as the median or trimmed mean.

iii. Measures of spread.

```
> sd(x)
[1] 1.010044
> IQR(x)
[1] 1.15
> mad(x)
[1] 0.7413
```

Since these data had potential outlier(s) we should choose a resistant measure of spread such as the IQR or MAD.

iv. Measures of shape.

```
> library(e1071)
> skewness(x)
[1] -0.7530389
> kurtosis(x)
[1] 0.01266688
```

To see if these are relatively large we would calculate

```
> n <- length(x)
> sqrt(6/n) * c(2, 4)
[1] 0.7470874 1.4941747
```

So these data are substantially skewed left but not substantially kurtic.

(c) Anything else noteworthy.

We can see from the boxplot that there are two extreme values to the left, and visually we can surmise that these are potential outliers (or we can check by hand).