INTRODUCTION TO STATISTICAL REASONING

36-201 **Lab #3 -Partial Solutions**

Question #1-2 The stem-and-leaf plot is given below. The distribution is unimodal and looks pretty symmetric. The center of the distribution is in the 70's. There is a gap in the lower tail suggesting that the value, 48, might be an outlier. The range is 46.

```
4|8
5|
5|8
6|124
6|8
7|001114
7|678899
8|002
8|56
9|34
```

Question #3 $ML = \frac{(25+1)}{2} = 13$, so the median is 76. $QL_1 = \frac{(13+1)}{2} = 7$, so $Q_1 = 70$. $QL_3 = \frac{(25+13)}{2} = 19$, so $Q_3 = 80$. $IQR = 80 - 70 = 10$.

Question #4 The exam score equal to $\overline{x} - s$ is 63.6. The exam score equal to $\overline{x} + s$ is 84.8. 17 out of 25 or 68% of the observations fall between these two values.

Question #5 The exam score equal to $\overline{x} - (2 \cdot s)$ is 53. The exam score equal to $\overline{x} + (2 \cdot s)$ is 95.4. 24 out of 25 or 96% of the observations fall between these two values.

Question #6 The exam score equal to $\overline{x} - (3 \cdot s)$ is 42.4. The exam score equal to $\overline{x} + (3 \cdot s)$ is 106. 25 out of 25 or 100% of the observations fall between these two values.

Question #7 Yes.

Question #8 The standard deviation rule says that 68% of the distribution of exam scores should be between 63.6 and 84.8. This leaves 32% of the distribution below 63.6 and above 84.8. Because the distribution is symmetric, one would expect half of these or $\boxed{16\%}$ to be below $63.6 \sim 64$. In the actual data, 4 out of 25 or 16% of the observations are less than 64.

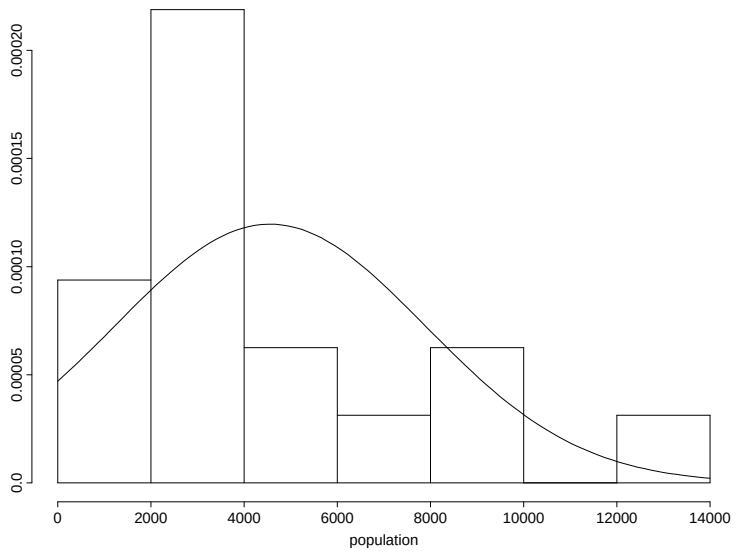

Figure 1: Histogram and overlaid normal curve for the population data.

Question #9. Here are the numerical descriptive statistics for the population densities (people per square mile) as reported in the 1980 U.S. Census for 16 cities in the U.S. The information required in the lab questions can be gotten from here. In particular, the IQR is $Q3 - Q1 = 7235 - 2082 = 5153$, and the Range is $Max - Min = 12413 - 637 = 11776$.

```
                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
C1             16     4556     3284     4274     3335      834

              MIN      MAX       Q1       Q3
C1            637    12413     2082     7235
```

Question #10. Here is a crude copy of the boxplot:

```
                 --------------------
            -----I    +             I----------------------
                 --------------------
        +---------+---------+---------+---------+---------+------C1
        0      2500      5000      7500     10000     12500
```

Question #11. Based on the boxplot, and the histogram in Figure 1, I would say this distribution is

- unimodal
- skewed right
- has no outliers

Question #12. Figure 1 shows a histogram of the population data, overlaid with a normal curve. The histogram does not appear to follow the overlaid normal curve all that well. The most obvious problem is that the mode, or high point, of the normal curve does not match up at all well with the mode of the histogram. You can also see that the normal curve is symmetric about its mode, whereas the histogram is not at all symmetric.

Question #13. There is lots of evidence against using the 68–95–99.7% rule here (either of the following two answers is enough):

- The boxplot and the histogram overlaid with a normal curve each indicate that the distribution of the data is not at all symmetric and unimodal, let alone actually normally distributed.
- We can use the sorted data in C1 together with the computed mean and standard deviation to see how well the rule actually works. The sorted data is

```
C1
    637    1634    1920    2051    2173    2689    2962    3244    3323
   3481    4684    5984    7652    8561    9480   12413
```

and from the descriptive statistics for Question #9 we know the mean is 4556 and the SD is 3335; so we can calculate:

68–95–99.7 Interval		Number of Data Points in Interval	Percent of Data Points in Interval	Expected Percent
$Mean - 1 \cdot SD$	$Mean + 1 \cdot SD$	12	75%	68%
$Mean - 2 \cdot SD$	$Mean + 2 \cdot SD$	15	94%	95%
$Mean - 3 \cdot SD$	$Mean + 3 \cdot SD$	16	100%	99.7%

The 95% and 99.7% parts of the rule seem to be working OK, but the 68% part of the rule isn't working at all well.

Question 14. Figure 2 shows boxplots for each of the transformations

- LET C2 = C1**(.667)
- LET C3 = C1**(.500)
- LET C4 = C1**(.333)
- LET C5 = LOGE(C1)

As the power goes down the right tail gets shorter and the left tail gets longer, until finally the boxplot for C5=LOGE(C1) is actually skewed left instead of skewed right. The plot with the most symmetric tails is probably C4=C1**(1/3), but here there's still some skewing "in the box" (note that the median is not exactly centered between the quartiles Q1 and Q3).

Note: If you chose a different plot, you will not be penalized. In a case like this the evidence is somewhat ambiguous and there are arguments for and against each plot.

Question #15. Here is a crude copy of the boxplot for C4=C1**(1/3).

```
                    ---------------------
        --------------I      +            I--------------
                    ---------------------
   ------+---------+---------+---------+---------+---------+C4
       9.0      12.0      15.0      18.0      21.0      24.0
```

By looking at it and the histogram in Figure 3, I would say that this distribution is

- More nearly symmetric, though the middle 50% is still a bit skewed right
- Unimodal, but not clearly so. An argument can be made for bimodal here too.
- No outliers

Note: Perhaps there is an even better transformation, maybe some power between 1/3 and 0, that would make the data follow a normal curve even more closely...

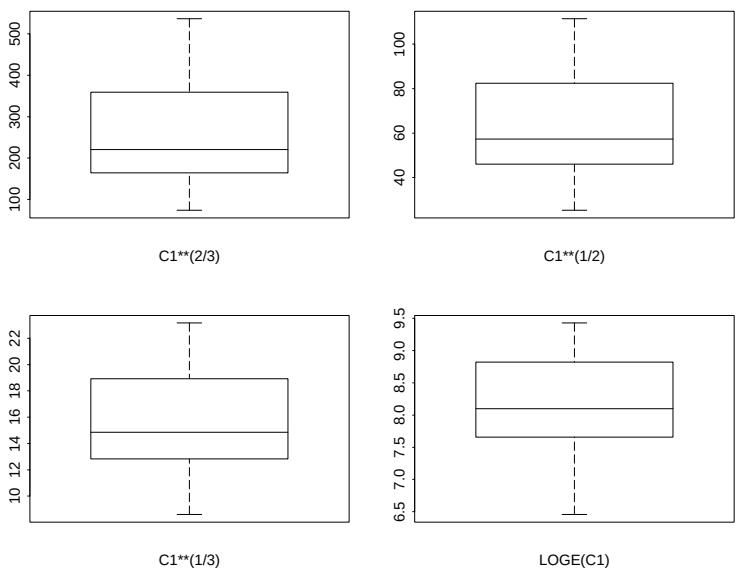

Figure 2: Boxplots for transformations of the population data.

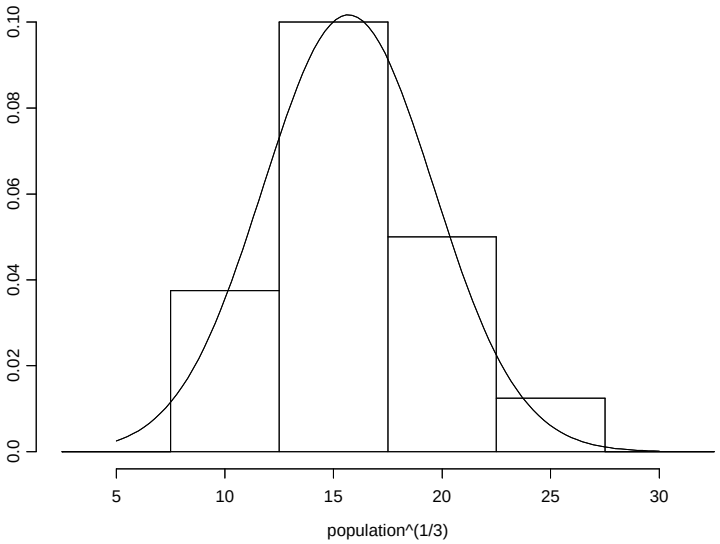

Figure 3: Histogram for the best transformation of the population data.