Stat1600 Solution to Midterm #1 Form C

1. The areas (rounded to nearest ten square miles) of counties in New Jersey are shown in the stem and leaf plot below:

```
-------------
Leaf Unit = 10

0  5
1  039
2  2337
3  11366
4  278
5  037
6  4
7
8  2
-------------
(So, for example, the first observation under stem 0 is 50
 and the observation under stem 8 is 820)
```

(a) (**5 points**) After examining the Stem-and-Leaf Plot, what seems to be the shape of this data?

Answer: Slightly Right skewed

Reason: Long right tail

(b) (**5 points**) In addition to the Stem-and-Leaf Plot, what other type(s) of data presentation would be most appropriate for this data? Explain your answer. Your choices are: (Bar Chart, Pie chart, Histogram, Dotplot, Box-and-Whisker Plot)

Answer:

Histogram, Dotplot, and Box-and-Whisker Plot can also be used for such (numerical) data.

(c) (**15 points**) Give the five-number summary then construct a boxplot (with whiskers extending to the two extremes, MIN and MAX). (**Hint**: How do you construct a sorted list of the data from the stem-and-leaf plot above?)

MIN and MAX: MIN=50, MAX=820

Median: (show your work)

$.5(n+1) = .5 \times 22 = 11$ and hence

$$MED = \text{(11th ordered value)} = 330.$$

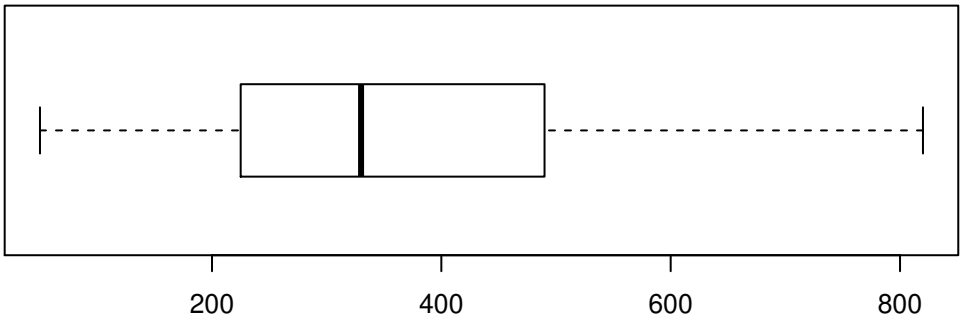
Quartiles, Q_1 and Q_3: (show your work)

$.25(n+1) = .25 \times 22 = 5.5$ and hence

$$Q_1 = \frac{\text{(5th ordered value) + (6th ordered value)}}{2} = \frac{220+230}{2} = 225$$

$$Q_3 = \frac{\text{(5th largest value) + (6th largest value)}}{2} = \frac{500+480}{2} = 490$$

boxplot:



2. Groups of dolphins were observed off the coast of Iceland near Keflavik in 1998 (adapted from Source: *Activities of Dolphin Groups* (http://www.statsci.org/data/general/dolpacti.html) from OzDASL). Fifteen groups were observed during noon time having one of three main activities— travelling (**`Travel`**), feeding (**`Feed`**) or socializing (**`Social`**):

`Social Travel Social Travel Travel Feed Travel Social`

`Travel Feed Feed Feed Travel Travel Feed`

(a) (**5 points**) What type of data is being analyzed here (Nominal, Ordinal, Interval, or

Ratio? Explain your answer.

The data values are of nominal data type since there exists no sense of ordering.

(b) (**5 points**) What type(s) of data presentation would be most appropriate for this data? Explain your answer. Your choices are: (Bar Chart, Pie chart, Stem-and-Leaf Plot, Histogram, Dotplot, Box-and-Whisker Plot)

ANSWER:

Choices that can be used are bar chart and pie chart. Others can not be used since the data set is categorical.

(c) (**15 points**) Construct a frequency-relative frequency table of this data.

	Frequency	Rel.Freq.
Travel	7	46.7
Feed	5	33.3
Social	3	20.0
TOTAL	15	100.0

3. The Bay Area Air Quality Management District (BAAQMD) made 9 independent measurements (in ppm (parts per million)) of the carbon monoxide over the period from September 11, 1990 to March 30, 1993 from one of the stacks of an oil refinery northest of San Francisco (adapted from *Measuring Air Pollution Story* from DASL):

15 20 12.5 20 5 165 20 15 25

(a) (**7 points**) Calculate the mean, $\overline{X}$ of this data.
ANSWER: 33.06

CALCULATION/REASON: (show your work)

$$\overline{X} = \frac{15 + 20 + 12.5 + 20 + 5 + 165 + 20 + 15 + 25}{9} = 33.06$$

(b) (**8 points**) Let's compare the summary statistics of carbon monoxide measurements for two other stacks of the oil refinery:

Statistics	Stack A	Stack B
Mean	38	36
Standard Deviation	20	40

Based on the information provided in the table, would you expect one of these stacks to have more extremes in their carbon monoxide measurements than the other?

ANSWER: Yes

REASON:

Stack B has more extremes than Stack A since the former has a much larger standard deviation.

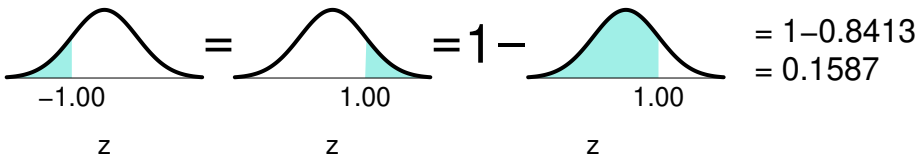
4. The chest sizes, measured in inches, of Scottish militiamen in the early 19th century were recorded (adapted from DASL: *Chest sizes of Militiamen*). A histogram of chest sizes shows an approximately normal curve. It is known that the mean chest size is 39.9 inches with a standard deviation of 2.3 inches. Given this information, use the Z-Table provided to answer the following questions:

   (a) (**15 points**) What percentage of soldiers having chest size under 37.6 inches?

Note that the z-score of 37.6 is

$$z = \frac{37.6 - 39.9}{2.3} = -1.00$$

Hence, the chance that a soldier having chest size under 37.6 inches can be computed as the area under the Z curve to the left of -1.00 which equals the area to the right of 1.00 and is calculated as $1-0.8413$ =0.1587 (see figures below). That is, 15.87%.



   (b) (**10 points**) What percentage of soldiers having chest size between 35.3 and 44.5 inches?

Note that the z-score for 35.4 is

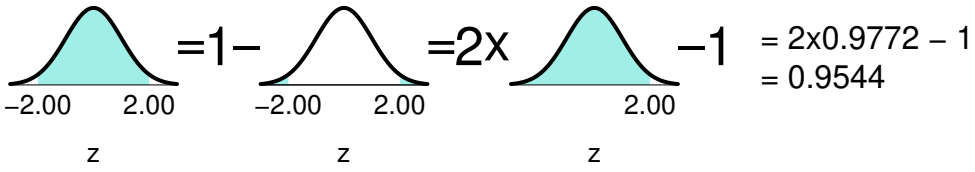
$$z = \frac{35.3 - 39.9}{2.3} = -2.00$$

and that the z-score for 44.5 is

$$z = \frac{44.5 - 39.9}{2.3} = 2.00.$$

Hence the percentage of men soldiers having chest size between 35.3 and 44.5 inches is the area (in percentage) under the Z curve bounded by -2.00 and 2.00 which is approximately 95% according to the empirical rules.

Using Z Table, the probability is 0.9544 or 95.44%:



(c) (**10 points**) A soldier had chest size measured at least as large as 80% of all Scottish militiamen. What was his chest size?

The soldier was at 80th or higher percentile. The 79.95th z-percentile is 0.84 (note: the area under Z curve to the left of 0.84 is 0.7995 and the area under Z curve to the left of 0.85 is 0.8023). Hence, the 80.23th N(mean=39.9,SD=2.3)-percentile is

$$39.9 + (0.85 \times 2.3) = 39.9 + 1.955 = 41.855 \approx 41.9$$

His chest size was 41.9.

5. (**5 points**) Extra Credit Problem
Assume we have a data set that includes the area codes of our customers' cellphones. Some examples in the data set include 269, 610, 484 etc. Clearly these are numbers, but are considered nominal (categorical) data. Discuss or prove why these values are actually nominal (categorical) data and should not be analyzed as numeric data.

Although telephone area codes come as number, there exists no sense of ordering in the area codes. That is telephone area codes are nominal data values and should by no means be analyzed as numeric data. An average of, say 269, 616, 989, is 624.67. An area code of 624.67? That is located in the 'average' place? That is absurd!