Statistics 1, Fall 2008 (Norris)
Solutions to Evens for HW #2

Chapter 2

2.22 Since nothing is known about the shape of the data distribution, you must use Tchebysheff's Theorem to describe the data.

a The interval from 60 to 90 represents $\mu \pm 3\sigma$ which will contain at least 8/9 of the measurements.

b The interval from 65 to 85 represents $\mu \pm 2\sigma$ which will contain at least 3/4 of the measurements.

c The value $x = 65$ lies two standard deviations below the mean. Since at least 3/4 of the measurements are within two standard deviation range, *at most* 1/4 can lie outside this range, which means that at most 1/4 can be less than 65.

2.24 **a** The stem and leaf plot generated by ***Minitab*** shows that the data is roughly mound-shaped. Note however the gap in the center of the distribution and the two measurements in the upper tail.

Stem-and-Leaf Display: Weight

```
Stem-and-leaf of Weight  N  = 27
Leaf Unit = 0.010

 1   7   5
 2   8   3
 6   8   7999
 8   9   23
 13  9   66789
 13  10
(3)  10  688
 11  11  2244
 7   11  788
 4   12  4
 3   12  8
 2   13
 2   13  8
 1   14  1
```

b Calculate $\sum x_i = 28.41$ and $\sum x_i^2 = 30.6071$, the sample mean is

$$\overline{x} = \frac{\sum x_i}{n} = \frac{28.41}{27} = 1.052$$

and the standard deviation of the sample is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{30.6071 - \frac{(28.41)^2}{27}}{26}} = 0.166$$

c The following table gives the actual percentage of measurements falling in the intervals $\overline{x} \pm ks$ for $k = 1, 2, 3$.

k	$\overline{x} \pm ks$	Interval	Number in Interval	Percentage
1	1.052 ± 0.166	0.866 to 1.218	21	78%
2	1.052 ± 0.332	0.720 to 1.384	26	96%
3	1.052 ± 0.498	0.554 to 1.550	27	100%

d The percentages in part **c** do not agree too closely with those given by the Empirical Rule, especially in the one standard deviation range. This is caused by the lack of mounding (indicated by the gap) in the center of the distribution.

e The lack of any one-pound packages is probably a marketing technique intentionally used by the supermarket. People who buy slightly less than one-pound would be drawn by the slightly lower price, while those who need exactly one-pound of meat for their recipe might tend to opt for the larger package, increasing the store's profit.

2.42 The ordered data are:

0, 1, 3, 4, 4, 5, 6, 6, 7, 7, 8

a With $n = 12$, the median is in position $0.5(n+1) = 6.5$, or halfway between the 6th and 7th observations. The lower quartile is in position $0.25(n+1) = 3.25$ (one-fourth of the way between the 3rd and 4th observations) and the upper quartile is in position $0.75(n+1) = 9.75$ (three-fourths of the way between the 9th and 10th observations). Hence, $m = (5+6)/2 = 5.5$, $Q_1 = 3 + 0.25(4-3) = 3.25$ and $Q_3 = 6 + 0.75(7-6) = 6.75$. Then the five-number summary is

Min	Q_1	Median	Q_3	Max
0	3.25	5.5	6.75	8

and

$$IQR = Q_3 - Q_1 = 6.75 - 3.25 = 3.50$$

b Calculate $n = 12$, $\sum x_i = 57$ and $\sum x_i^2 = 337$. Then $\bar{x} = \dfrac{\sum x_i}{n} = \dfrac{57}{12} = 4.75$ and the sample standard deviation is

$$s = \sqrt{\frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{337 - \dfrac{(57)^2}{12}}{11}} = \sqrt{6.022727} = 2.454$$

c For the smaller observation, $x = 0$,

$$z\text{-score} = \frac{x - \bar{x}}{s} = \frac{0 - 4.75}{2.454} = -1.94$$

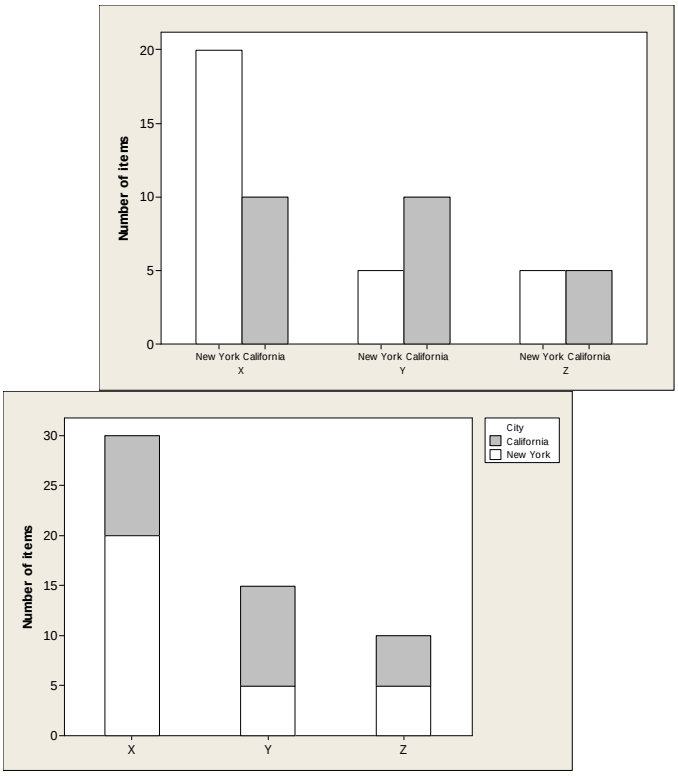
and for the largest observation, $x = 8$,

$$z\text{-score} = \frac{x - \bar{x}}{s} = \frac{8 - 4.75}{2.454} = 1.32$$

Since neither *z*-score exceeds 2 in absolute value, none of the observations are unusually small or large.

Chapter 3

3.2 **a** The side-by-side comparative bar chart (next page) shows the two states (each shaded differently) in groups of three, each corresponding to one of the three categories. The vertical axis measures the number of items in the category.



a **b**

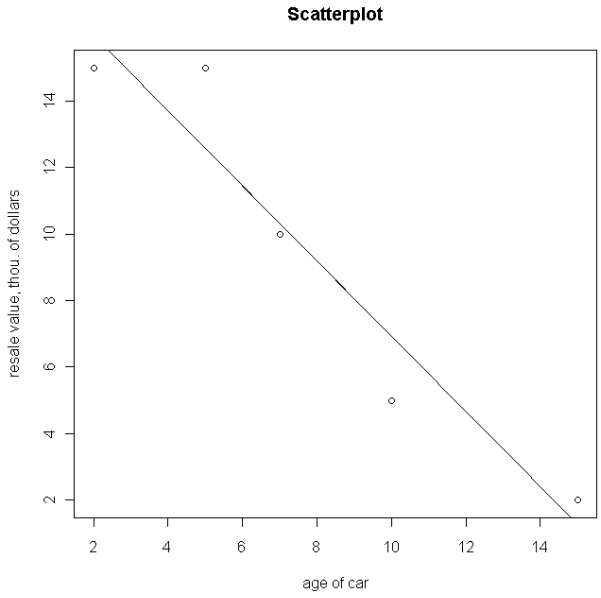
b The stacked bar chart (figure **b**) shows each of the three categories, with the number of items for New York and California (each shaded differently) stacked on top of each other.

c The side-by-side bar chart is easier to understand, since you can directly compare the two states. The stacked bar chart requires the reader to decipher the difference between the two pieces of the bar.

d A line chart is another possible graphical method which could be used.

Additional Handout Problems

1. a)



b) I expect the linear correlation coefficient to be negative since the older the car is, the less money it will be worth. In other words, as the age of the car increases, the resale value decreases.
c) Using the calculations shown in the table below:

$$s_{xy} = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{n-1} = \frac{255 - \frac{1}{5}(39)(47)}{5-1} = \frac{-111.6}{4} = -27.9$$

Using a calculator to obtain s_x and s_y:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-27.9}{(4.97)(5.86)} = -.96$$

	x	y	xy
	2	15	30
	10	5	50
	7	10	70
	5	15	75
	15	2	30
Total	39	47	255

d) $b = \frac{s_{xy}}{s_x^2} = \frac{-27.9}{(4.97)^2} = -1.13$

$$a = \bar{y} - b\bar{x} = \frac{47}{5} - (-1.13)\frac{39}{5} = 18.21$$

$y = 18.21 - 1.13x$

e) The best fit line is shown on the scatterplot in part (a).
f) For every one year increase in the car's age, the average resale value decreases by about $1100 dollars.
g) y=18.21-1.13(8)=9.17 about $9000 dollars.

2) Extra Credit. No, multiplying each x value will not change the correlation coefficient. However, it will change the equation of the best fit line. This can be proven algebraically, but you need to know some things about summation notation. Alternately, if you try it on a small data set, you will get the same results. However, trying it on a data set is not a proof that the correlation coefficient will always be the same when x is multiplied by a constant. Here is the algebraic proof:

Note that $\sum_{i=1}^{n} kx_i = k\sum_{i=1}^{n} x_i$ and letting $\tilde{x}_i = kx_i$

st.dev of $\tilde{x}_i = s_{\tilde{x}} = \sqrt{\frac{\sum (kx_i - k\bar{x})^2}{n-1}} = \sqrt{\frac{\sum k^2(x_i - \bar{x})^2}{n-1}} = \sqrt{k^2 \cdot \frac{\sum (x_i - \bar{x})^2}{n-1}}$

$= \sqrt{k^2 s_x^2} = |k| s_x$

since mean of $\tilde{x}_i = \frac{\sum_{i=1}^{n} kx_i}{n} = \frac{k\sum x_i}{n} = k\bar{x}$

Then, $r_{new} = \frac{s_{\tilde{x}y}}{s_{\tilde{x}} s_y} = \frac{\left(\frac{\sum (kx_i) y_i - \frac{1}{n}(\sum kx_i)(\sum y_i)}{n-1}\right)}{|k| s_x \cdot s_y}$

$$= \frac{\left(\frac{k\sum x_i y_i - \frac{1}{n} \cdot k \sum x_i \sum y_i}{n-1}\right)}{|k| s_x s_y}$$

$$= \frac{\left(\frac{k\left(\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i\right)}{n-1}\right)}{|k| s_x s_y}$$

$$= \left(\frac{k}{|k|}\right)\frac{\left(\frac{\sum xy - \frac{1}{n}\sum x \sum y}{n-1}\right)}{s_x s_y} = \frac{k}{|k|} \cdot r_{old} = r_{old} \text{ if } k>0$$

3) a) negative. The more units the student carries, the more homework he/she has so the less sleep he/she is likely to get.

b) positive. As temperature increases, ice cream sales are also likely to increase.
c) zero. There not likely to be any relationship between last digit of SSN and mathematical aptitude.
d) negative. The more kilometers the person trains per week, the faster he/she is likely to run. Hence, the time in 5 km race is likely decrease with increased number of km run per week.

new slope $= \frac{s_{\tilde{x}y}}{s_{\tilde{x}^2}} = \frac{\left(\frac{k\left(\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i\right)}{n-1}\right)}{k^2 s_x^2}$

$$= \frac{1}{k}\left(\frac{s_{xy}}{s_{x^2}}\right) = \frac{1}{k}(\text{old slope})$$

So the equation of the best fit line does change, but r stays the same for $k>0$.