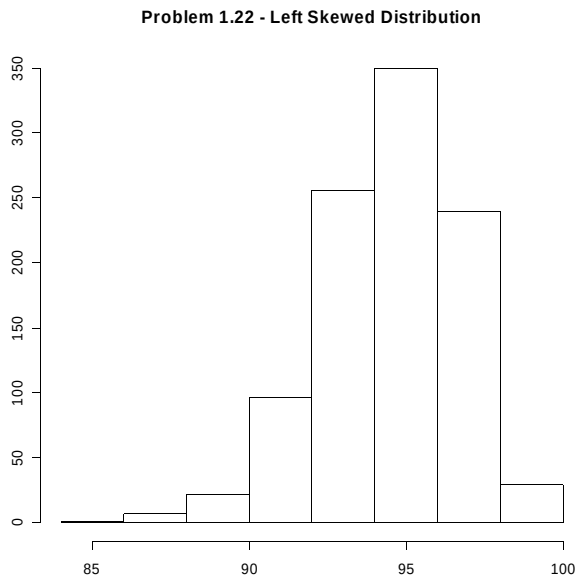
Stat 140 – Fall 2004
Final Solutions for Homework #1 (20 points)

1.22



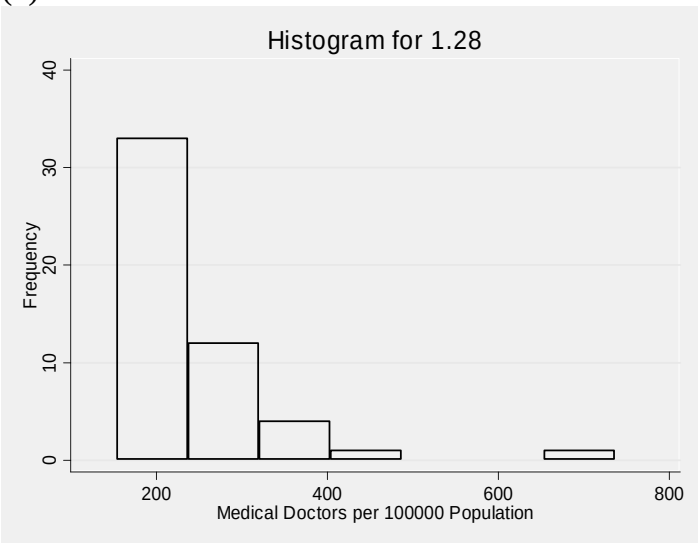
(1 point for a correct graph - any left-skewed histogram is acceptable)

The distribution would be skewed to the left because most of the coins have recent dates (near or less than the current year). However, an old coin will occasionally arise, which would produce a tail extending far to the left. Also, there will never be coins dated in the future. (1 point for reasonable explanation)

1.28

(a) The number of doctors in a state is not a helpful measure because states with larger populations will naturally have a larger number of doctors. However, because of the larger population, doctors may not be as available in larger states. The number of doctors, then, could be misleading. Availability is better described in terms of the number of residents who are to be served by each doctor. (1 point for reasonable explanation)

(b)



Stem-and-leaf plot for Doctors

```
1** | 54,63,67,69,71,73,73,84,90,90,95,98
2** | 00,02,03,03,07,09,11,12,15,18,22,23,24,25,27,30,32,34,35,35, ... (28)
2** | 60,65,91,95
3** | 05,38
3** | 54,74,87
4** | 12
4** |
5** |
5** |
6** |
6** |
7** | 37
```

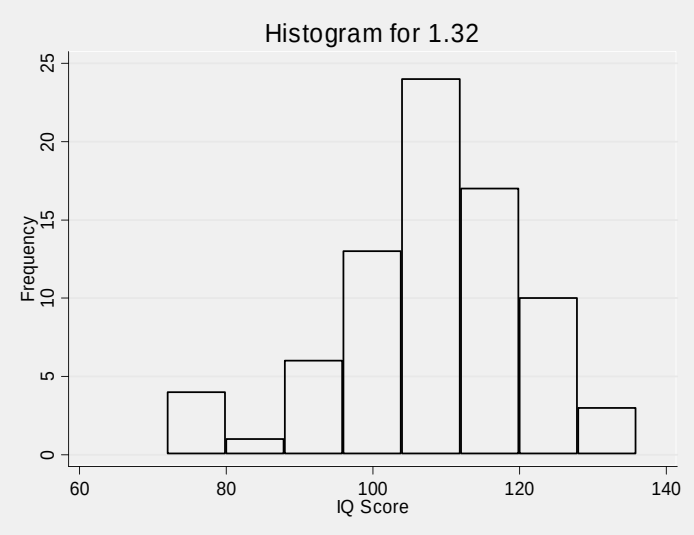
(1 point for histogram or stem and leaf plot)

The distribution is skewed right, with most of the states having less than 400 medical doctors per 100,000 people.

There is one outlier – Washington, D.C., has 737 medical doctors per 100,000 people. This may be due to the large number of people that live in the Washington, D.C., area and work in the city but do not live there. This would cause a large number of doctors to practice there despite the low number of people that actually reside in the city.

(1 point for correct distribution description and for outlier identification and any reasonable explanation).

1.32



Stem-and-leaf plot for iq (IQ)

```
 7*   | 24
 7.   | 79
 8*   |
 8.   | 69
 9*   | 0133
 9.   | 6778
10*   | 0022333344
10.   | 555666777789
11*   | 00001111222233344 44
```

```
11.   | 55688999
12*   | 003344
12.   | 677888
13*   | 02
13.   | 6
```

(1 point for correct histogram or stem and leaf plot)

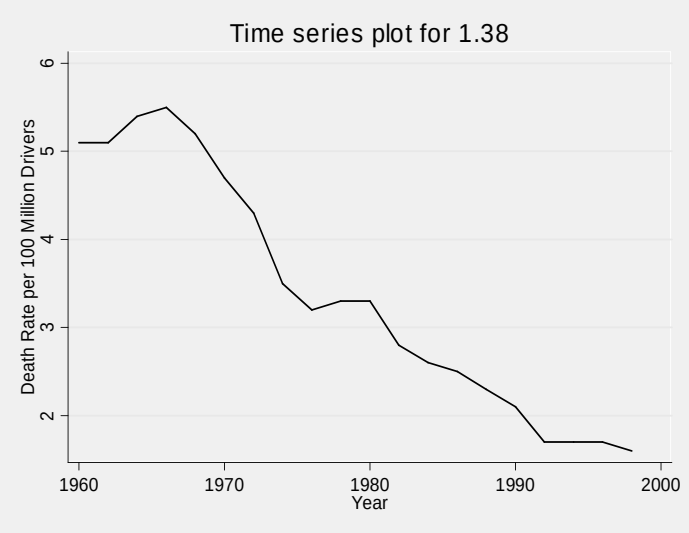
The distribution is slightly skewed left, but mostly symmetric, with a single peak around 110. About half the students seem to have IQ scores less than 110, so the midpoint of the distribution is around 110. The spread is from about 70 to 130. There are a few outliers on the low side, which is what gives the distribution its slight skew.

The midpoint of the distribution is clearly above 100.

(1 point for correct description and midpoint identification)

1.38

(a)



Deaths are generally decreasing as time goes on.

(1 point for correct graph and description)

(b) No, the effect of lower speeds after 1974 and before the mid 1980's is not visible. The decline in deaths is fairly steady over the entire time period. (1 point for identifying steady decline)

(c) A histogram here would not make sense. The time series plot is more helpful because it emphasizes the change over time, which is what we are interested in. Histograms would only tell the distribution of deaths overall. (1 point for identifying the time series plot as more relevant).

1.42

(a) The median score is in the 5496^{th} position. It appears to be just under $40,000. (1 point for correct position and value within $10,000 in either direction)

(b) Q1 would occur in the 2748th position. Q3 would occur in the 8244th position. Q1 appears to be around $30,000, and Q3 appears to be about $50,000 (1 point for correct positions and 1 point for values within $10,000 in either direction).

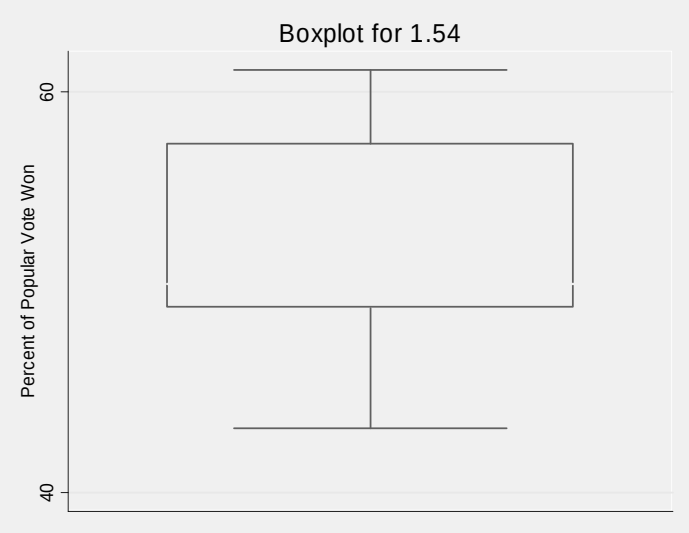
1.44

The median income increases as the highest level of education reached gets higher. The spread and the skewness also seem to go up with highest level of education reached. The skewness for the Higher Degree category is particularly high.

(2 points for correct statements regarding center, skewness and spread)

1.54

(a)



Stem-and-leaf plot for percent (Percent)

percent rounded to nearest multiple of .1
plot in units of .1

```
4** | 32,34
4** |
4** | 79
4** | 92,96,97
5** | 01,07
5** | 39
5** | 51
5** | 74
5** | 88
6** | 07,11
```

The distribution appears to be fairly symmetric, with 2 small values around 43%.

(1 point for correct description and for a boxplot or steam and leaf plot, not a histogram)

(b) The median is 50.4% (1 point)

(c) Q3 is 57.4%. Landslides would be 1956, 1964, 1972, and 1984. (1 point for getting all 4 landslide years)

1.76

(To obtain the 10% trimmed mean, remove the highest and lowest 7 values. To obtain the 20% trimmed mean, remove the highest and lowest 14 values.)

The untrimmed mean is 141.8 and the untrimmed median is 102.5.

The 10% trimmed mean is 118.16.

The 20% trimmed mean is 111.68.

(1 point for calculating the correct means and medians)

Because the data is skewed right, extreme data at the higher end were having a greater effect on the mean than values on the lower end. Thus, trimming the data reduces the mean both times.

(1 point for identifying the correct effect on the mean)

STATA ASSIGNMENT (20 points)

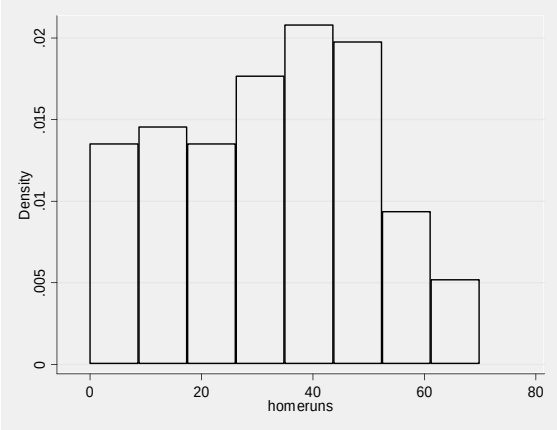
Part 5. (1 point for each correct summary; 3 total points for this part)

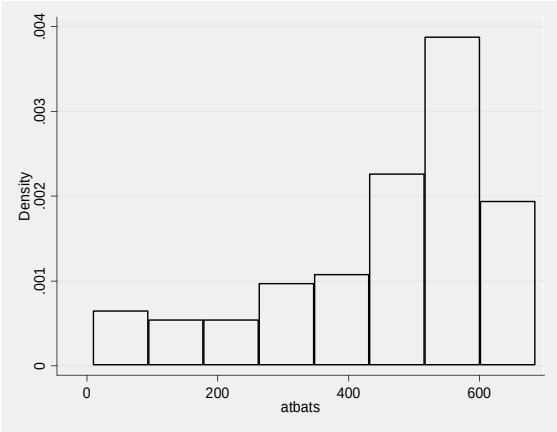
Variable	Obs	Mean	Std. Dev.	Min	Max
player	0				
year	110	1971.236	27.69789	1914	2002
atbats	110	457.9273	162.8404	10	686
homeruns	110	32.65455	17.60183	0	70

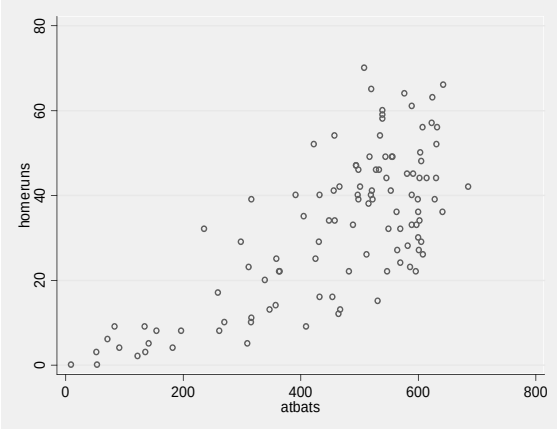
player	Freq.	Percent	Cum.
Aaron	23	20.91	20.91
Griffey	14	12.73	33.64
Maris	12	10.91	44.55
McGwire	16	14.55	59.09
Rodriguez	9	8.18	67.27
Ruth	22	20.00	87.27
Sosa	14	12.73	100.00
Total	110	100.00	

	player	year	atbats	homeruns
1.	Aaron	1954	468	13
2.	Aaron	1955	602	27
3.	Aaron	1956	609	26
4.	Aaron	1957	615	44
5.	Aaron	1958	601	30
6.	Aaron	1959	629	39
7.	Aaron	1960	590	40
8.	Aaron	1961	603	34
9.	Aaron	1962	592	45
10.	Aaron	1963	631	44

6. (1 point for each correct graph; 5 total points for this part)







Homeruns appears to be symmetrically distributed. The distributions of atbats is skewed left. (1 point)

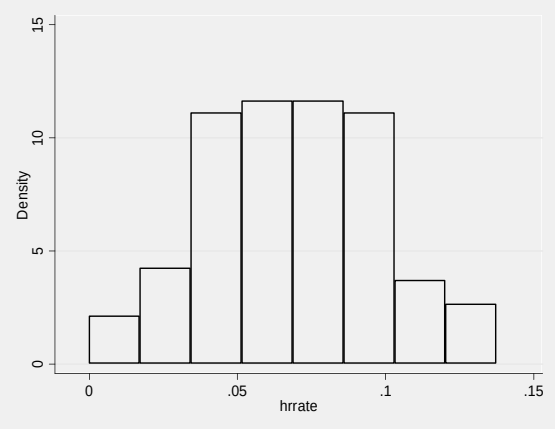
The two have a direct relationship – as one goes up, so does the other one. It is difficult to tell whether the relationship is linear or not from the scatterplot. (1 point)

7. (1 point for correct correlation matrix; 2 total points for this part)

	atbats	homeruns
atbats	1.0000	
homeruns	0.7321	1.0000

The correlation is 0.7321. It is positive because as one goes up, so does the other one. It is less than one because it is not a perfect linear relationship. (1 point)

8. (2 total points for this part)



(1 point for correct graph)

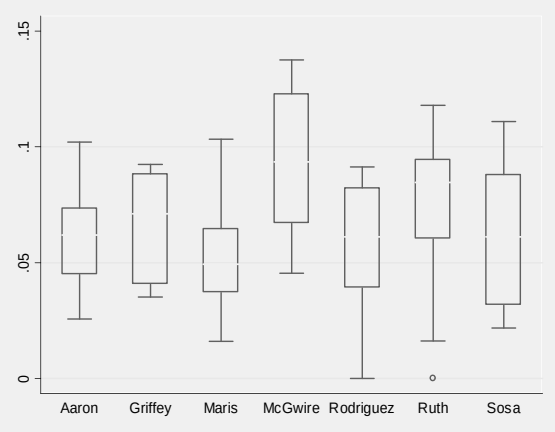
The rates are symmetrically distributed. (1 point)

9. (1 point for each summary and 1 point for graph; 5 total points for this part)

	Summary of homeruns		
player	Mean	Std. Dev.	Freq.
Aaron	32.826087	11.182991	23
Griffey	33.428571	16.046635	14
Maris	22.916667	15.979864	12
McGwire	36.4375	19.646777	16
Rodriguez	33.111111	19.82703	9
Ruth	32.454545	20.210151	22
Sosa	35.642857	21.208152	14
Total	32.654545	17.601833	110

	Summary of hrrate		
player	Mean	Std. Dev.	Freq.
Aaron	.06091769	.01957722	23
Griffey	.06590665	.0221048	14
Maris	.05108443	.0245257	12
McGwire	.09388766	.03003145	16

Rodriguez	.05853504	.02929423	9
Ruth	.07360276	.03200528	22
Sosa	.06511442	.02990548	14
Total	.06815176	.02903871	110



It does not matter whether you look at the rate or homeruns in comparing the players to each other. The boxplot shows the rates, and the centers and spreads vary. McGuire appears to the best and Maris appears to the worst. Rodriguez has the largest variation. (2 points for accurate description of data, mentioning a comparison of their centers and their spreads)

10. (1 point for each correct summary; 3 total points for this part)

Variable	Obs	Mean	Std. Dev.	Min	Max
player	0				
homerun	7	513.1429	186.6167	275	755
atbat	7	7196	2632.81	4382	12364
hrrate	7	.071563	.0138101	.053911	.0942298

	Summary of hrrate		
player	Mean	Std. Dev.	Freq.
Aaron	.06106438	0	1
Griffey	.06769854	0	1
Maris	.053911	0	1
McGuire	.09422984	0	1
Rodriguez	.06800548	0	1
Ruth	.08501012	0	1
Sosa	.07102192	0	1
Total	.07156304	.01381007	7

McGuire is still the best player, and Maris is still the worst, but the order in the middle has changed slightly. (1 point)