

Sample First Midterm Examination
(with solutions)

1. Baseball pitcher Nolan Ryan played in 20 games or more in the 24 seasons from 1968 through 1992. Here are the numbers of games he played in during these years.

21 39 39 35 30 33
 25 41 37 21 35 32
 27 42 31 35 30 30
 30 28 34 29 34 27

(a) Make a stemplot of these data, ordering the leaves on each stem

Note: Use 4 or more stems

Five stems is probably the best choice here, using the first digit twice.

```

2* | 11
2. | 57789
3* | 000012344
3. | 555799
4* | 12
    
```

1* | 1 represents 11 Leaf digit unit = 1

The next larger size is 10 stems, using the digit 5 times.

```

2* | 11
2t |
2f | 5
2s | 77
2. | 89
3* | 00001
3t | 23
3f | 44555
3s | 7
3. | 99
4* | 1
4t | 2
    
```

This is really too many stems for a data set with n = 24. It's hard to make out the shape

1* | 1 represents 11 Leaf digit unit = 1

Note: When there are two stems per first digit, the lower one gets leaves 0, 1, 2, 3 and 4 and the upper one gets leaves 5, 6, 7, 8 and 9. When there are 5 stems per first digit, the first gets leaves 0 and 1, the second 2 and 3, and so on. Thus neither of the following is a proper stemplot.

```

2* | 115
2. | 7789
3* | 000012344555
3. | 799
4* | 12
    
```

```

2* | 1157789
3. | 000012344
3* | 555799
4. | 12
    
```

(b) Find the quartiles of these data (first or lower quartile, third or upper quartile, and median). Here is some MacAnova output:

```
Cmd> ryan <- vector(21,39,39,35,30,33,25,41,37,21,35,32,\
  27,42,31,35,30,30,30,28,34,29,34,27) # enter data

Cmd> sort(ryan)
(1)      21      21      25      27      27
(6)      28      29      30      30      30
(11)     30      31      32      33      34
(16)     34      35      35      35      37
(21)     39      39      41      42
```

Having the data ordered makes this pretty easy.

$n = 24$ is even, so the median is half way between 31 and 32, the 12th and 13th values in order of size. Thus $M = 31.5$.

Q_1 is the median of the lower half (lower 12 values), that is half way between 28 and 29, the 6th and 7th values. $Q_1 = 28.5$.

Q_3 is the median of the upper half (upper 12 values), that is half way between 35 and 35, the 6th and 7th values counting from the maximum. $Q_3 = 35$

More MacAnova output

```
Cmd> n <- 24

Cmd> sum(ryan)
(1)      765

Cmd> write(sum((ryan - sum(ryan)/n)^2))
NUMBER:
(1)      722.625
```

(c) Use these values to calculate the sample mean \bar{x} and standard deviation s . Do not enter the data in your calculator. Show your work.

$$\bar{x} = \underline{31.875} \qquad s = \underline{5.6052}$$

$$\bar{x} = \sum x_i / n = 765/24 = 31.875$$

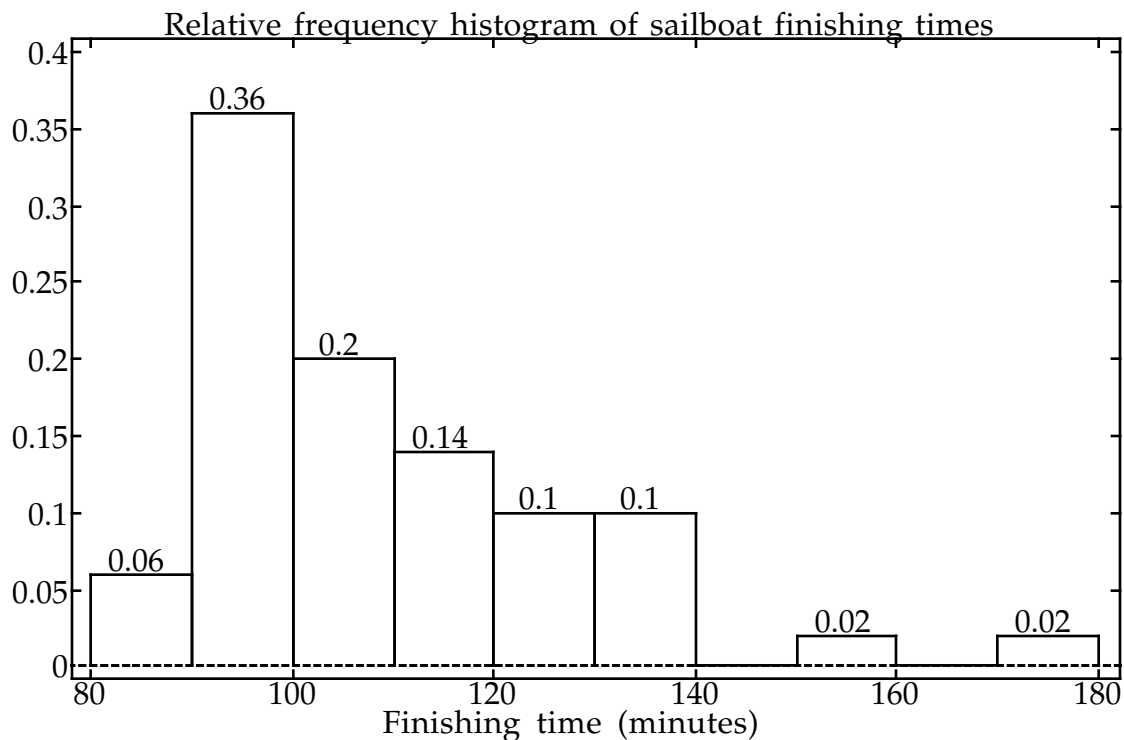
$$s = \sqrt{\sum (x_i - \bar{x})^2 / (n - 1)} = \sqrt{722.625 / 23} = \sqrt{31.418} = 5.6052$$

2. Fifty boats, all of the same design, competed in a six mile sailboat race on Lake Superior.

(a) Suppose you are told (not necessarily correctly) that the distribution of finishing times (times from start to finish) of these boats was approximately normal, with mean 110 minutes and standard deviation 20 minutes. Approximately what proportion of the finishes would you guess were between 90 and 130 minutes?

$110 - \sigma = 90$ and $110 + \sigma = 130$ so you are asked to guess the proportion of boats withing one standard deviation of the mean. By the 68-95-99.7 rule, about 68% of the data should be withing σ of the mean. More precisely, if the sailboat finishing times were exactly normal, Table A gives the proportion as $2(.8413 - .5) = .6826 = 68.26\%$ or $.8413 - .1587$

Here is a relative frequency histogram of the actual finishing times of the 50 boats. The number above each bar is the relative frequency of sailboats in that time class.



(b) Describe in no more than two sentences the shape of the distribution of finishing times. Would you expect the median to be larger than the mean, smaller than the mean, or about the same?

The data are strongly positively skewed and positively skewed; possibly there are two outliers.

You would expect the median to be smaller than the mean. This is characteristic of positively skewed data.

Sample Statistics 5021 First Midterm Examination with solutions

(c) What percent of the sail boats actually finished between 90 and 130 minutes after the start.

Summing the relative frequencies of the corresponding bars, the proportion is
 $.36 + .20 + .14 + .10 = .80$ or 80%

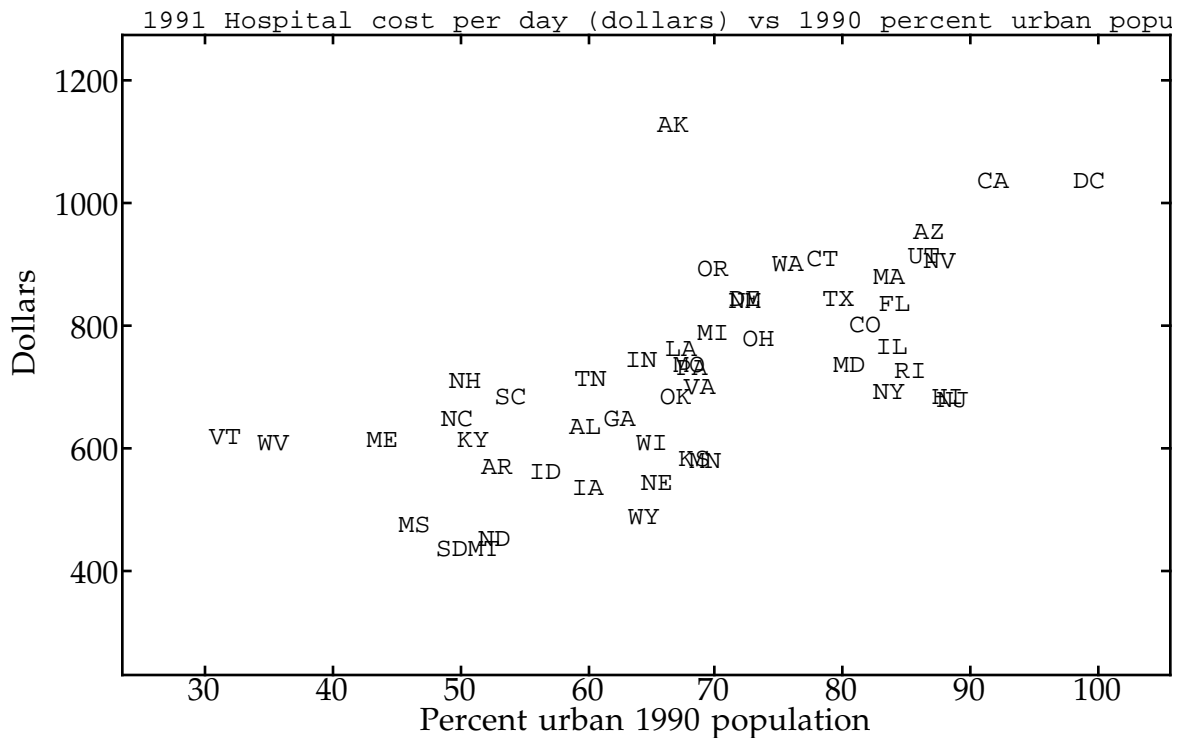
This is substantially more than the 68% predicted by the 68-95-99.7 rule.

3. The following table lists data for the 50 states and the District of Columbia from the 1993-1994 Statistical Abstract of the United states. There are two variables, the percent of the population living in urban areas in 1990 (% urban) and the average 1991 hospital cost per day (Hosp \$/day). Also included are the two letter abbreviations for the states.

	% urban	Hosp \$/day		% urban	Hosp \$/day		% urban	Hosp \$/day		% urban	Hosp \$/day
AK	67.5	1130	ID	57.4	565	MT	52.5	437	RI	86.0	730
AL	60.4	637	IL	84.6	770	NC	50.4	651	SC	54.6	684
AR	53.5	571	IN	64.9	745	ND	53.3	454	SD	50.0	436
AZ	87.5	955	KS	69.1	585	NE	66.1	546	TN	60.9	716
CA	92.6	1037	KY	51.8	616	NH	51.0	713	TX	80.3	846
CO	82.4	801	LA	68.1	764	NJ	89.4	680	UT	87.0	915
CT	79.1	910	MA	84.3	880	NM	73.0	840	VA	69.4	701
DC	100.0	1038	MD	81.3	740	NV	88.3	907	VT	32.2	621
DE	73.0	845	ME	44.6	617	NY	84.3	694	WA	76.4	904
FL	84.8	836	MI	70.5	792	OH	74.1	782	WI	65.7	611
GA	63.2	652	MN	69.9	582	OK	67.7	684	WV	36.1	612
HI	89.0	686	MO	68.7	737	OR	70.5	896	WY	65.0	489
IA	60.6	538	MS	47.1	479	PA	68.9	732			

Sample Statistics 5021 First Midterm Examination with solutions

Here is a scatterplot of these data.



(a) Describe in about two sentences any regular pattern to the data and any apparent deviations from the pattern.

There is a positive relationship, somewhat linear. There is one striking outlier in the y direction, namely Alaska where hospital costs are well out of line with other states.

(b) Alaska (AK) has the highest hospital cost of any state. In a regression of hospital cost on percent rural population, is Alaska likely to be an influential case? Why or why not.

No, Alaska is not likely to be an influential case. Although it appears to be an outlier in the y-direction, it is not an x-outlier. With $x = 67.5$, and $\bar{x} = 68.8$ its percent urban population is very close to the overall mean.

(c) Here are various summary numbers ($x = \% \text{ urban}$, $y = \text{Hosp } \$/\text{day}$).

$n = 51$	$\sum x = 3509.0$	$\sum y = 36,789$
$\sum (x - \bar{x})^2 = 11,539.94$	$\sum (y - \bar{y})^2 = 1,272,643.7$	$\sum (x - \bar{x})(y - \bar{y}) = 80,509.43$
$s_x = 15.192$	$s_y = 159.54$	$r_{xy} = 0.6643$

Find the intercept and slope of the least squares regression line. Use the summary numbers given. Do not key in the data in your calculator.

$$b = r_{xy} s_y / s_x = 0.6643 \cdot 159.54 / 15.192 = 6.976$$

$$\text{or } b = S_{xy} / S_{xx} = \sum (x - \bar{x})(y - \bar{y}) / \sum (x - \bar{x})^2 = 80,509.43 / 11,539.94 = 6.9766$$

$$a = \bar{y} - b\bar{x} \text{ Since } \bar{y} = 36,789 / 51 = 721.35 \text{ and } \bar{x} = 3509.0 / 51 = \mathbf{68.804},$$

$$a = 721.35 - (6.9766) \times (68.804) = \mathbf{241.33}$$

Sample Statistics 5021 First Midterm Examination with solutions

4. Here are counts from a study of all gun related deaths in Milwaukee, between 1990 and 1994.

	Handgun	Shotgun	Rifle	Unknown	Totals
Homicides	468	28	15	13	524
Suicides	124	22	24	5	175
Totals	592	50	39	18	699

(a) Find the conditional distribution of gun type given cause of death.

This would have been more precise if it had asked to find the *conditional distributions*, since there are 2, the distribution of gun type given homicides and the distribution of gun type given suicides.

	Handgun	Shotgun	Rifle	Unknown	Total
Homicides	$468/524 = .893$	$28/524 = .053$	$15/524 = .029$	$13/524 = .025$	1.000
	Handgun	Shotgun	Rifle	Unknown	Total
Suicides	$124/175 = .709$	$22/175 = .126$	$24/175 = .137$	$5/175 = .029$	1.001

(b) On the basis of these distributions, in not more than 2 or 3 sentences, describe the differences in types of weapons used in gun related homicides and suicides.

It appears that proportion of suicides committed with a handgun is about 79% of the proportion of murders committed with a handgun. Conversely, shotguns were more than twice as popular and rifles more than 4 times as popular suicides than among murderers.