

Unit 1 - Summarizing Data
Practice Problems

Solutions

#1.

- a. Qualitative - ordinal
- b. Qualitative - nominal
- c. Quantitative - ratio
- d. Qualitative - nominal
- e. Quantitative - ratio
- f. Quantitative - interval

#2a. By hand, here is the stem and leaf diagram I constructed. Other groupings for the stem are okay.

Stem	Leaf
0	1 1 1 1
0	2 3 3 3 3 3 3
0	4 4 4 4 4 5 5 5 5 5
0	7 7 7 7 7 7 7
0	8 8 8 8 8 8
1	0 0 0 1 1
1	2 2 3 3
1	
1	7 7

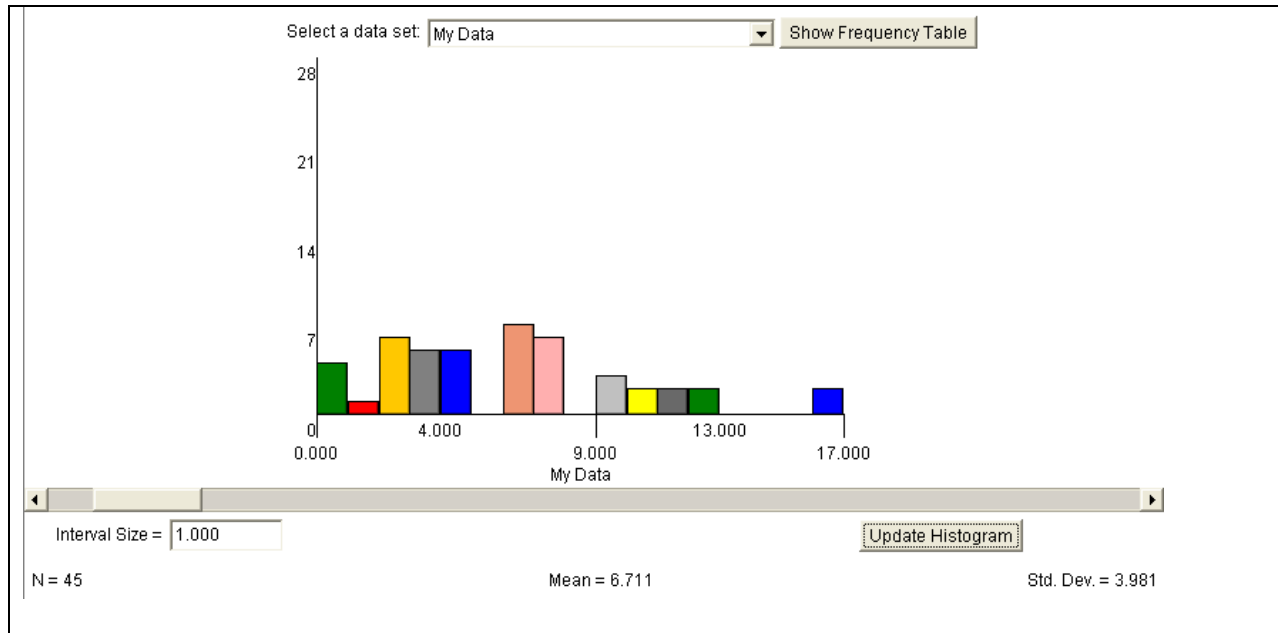
#2b. By hand, this is what I produced. Other class intervals are okay.

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Rel. Frequency
0-1	4	.0889	4	.0889
2-3	7	.1556	11	.2444
4-5	10	.2222	21	.4667
6-7	7	.1556	28	.6222
8-9	6	.1333	34	.7556
10-11	5	.1111	39	.8667
12-13	4	.0889	43	.9556
14-15	0	0	43	.9556
16-17	2	.0444	45	1.0000
TOTAL	45	1.0000		

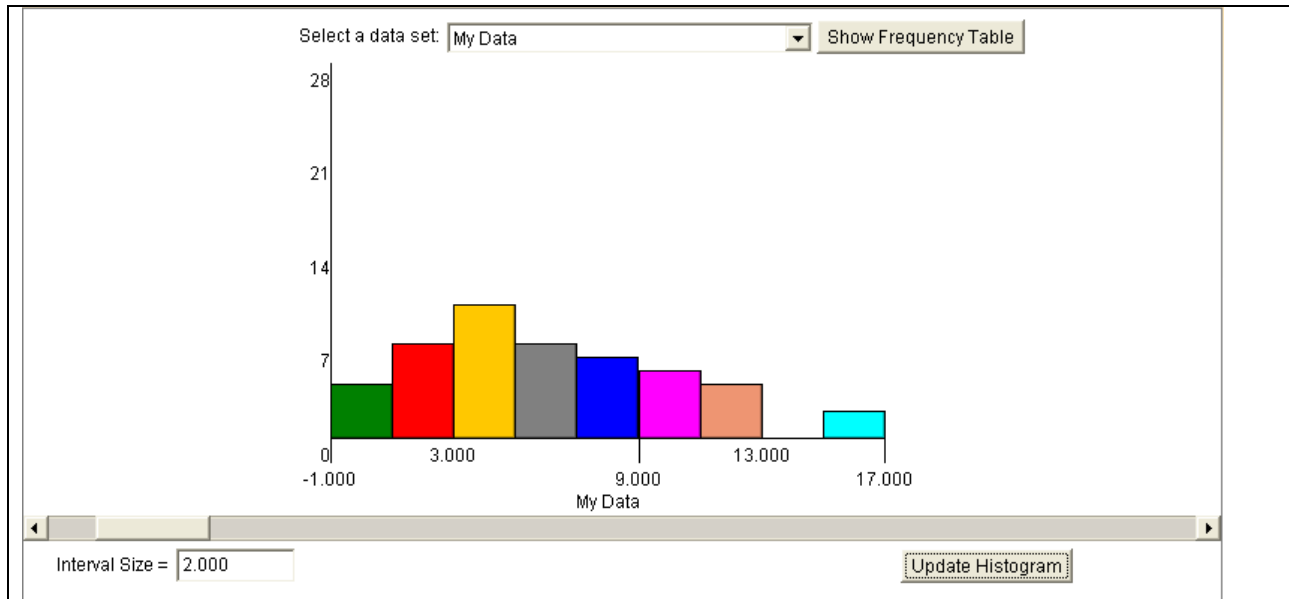
#2c. I did not construct a histogram by hand, I used the Shodor applet that can be found under

<http://www.shodor.org/interactivate/activities/Histogram/>

At the dialogue box “select a data set”, I scrolled down to choose **My Data**. Next, I scrolled down until I found a data entry box. I entered my data, selected **interval size = 1.00** and then clicked on **Update Histogram**. Here is what I got.



If you like, you can play with different choices of interval size. For example, interval size=2.00 yields the following.

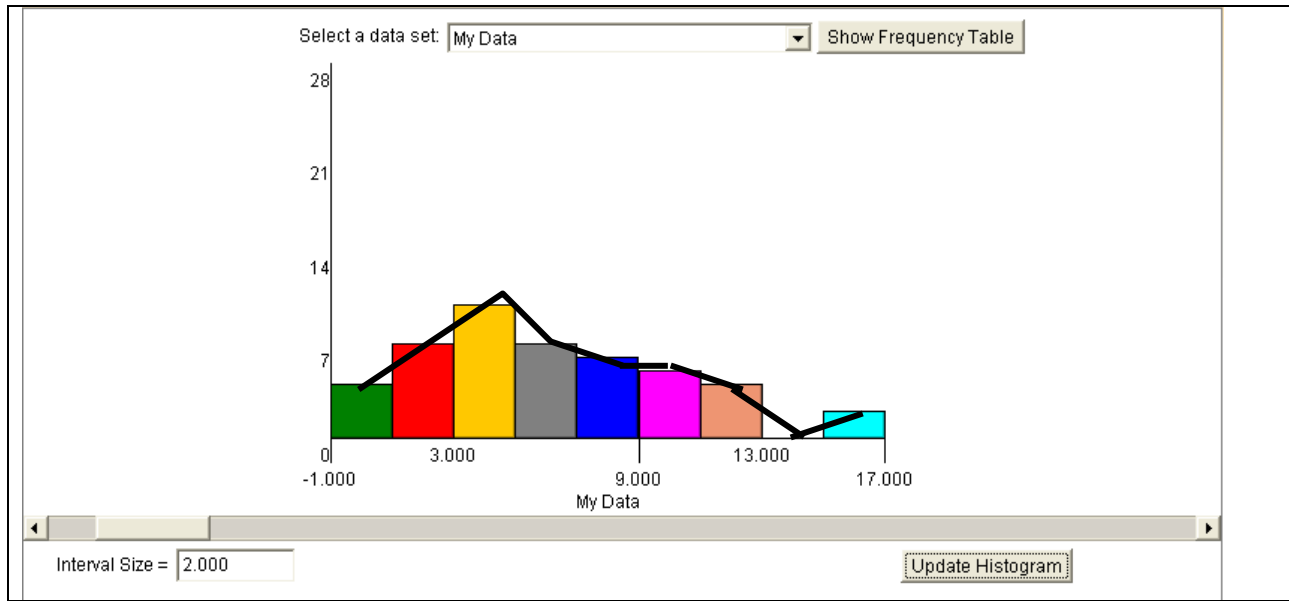


If you would like to try generating a histogram in SAS or Stata or Minitab visit the Summarizing Data topic page in the course resource website. Go to

<http://www-unix.oit.umass.edu/~biep540w/webpages/summarizing.htm>

#2d. I did not do this by hand, either. A frequency polygon plot is similar to a histogram. The first step is to choose class intervals. Next, note for each class interval the frequency (or relative frequency of data values in the interval). Plotted on the x-axis is the midpoint of the interval. Plotted on the y-axis is the frequency (or relative frequency).

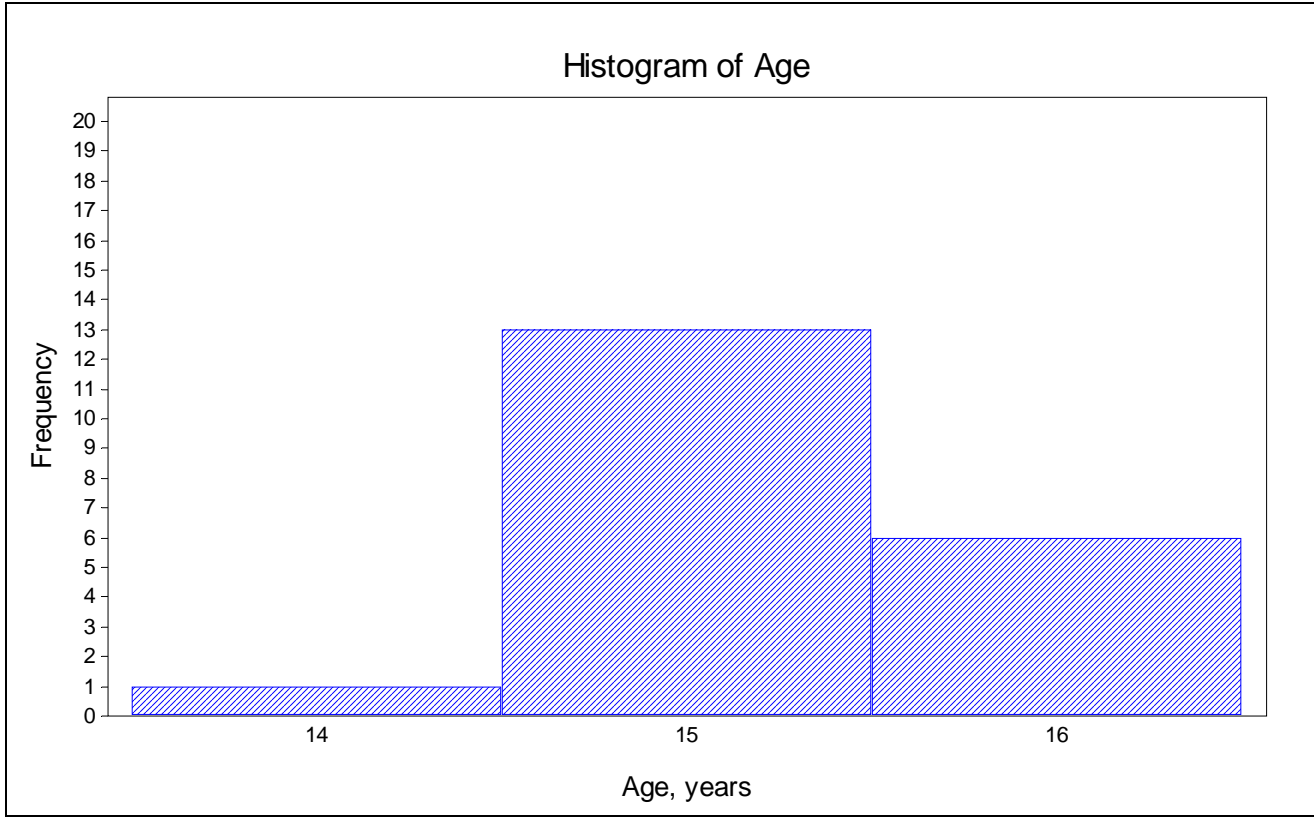
Thus, a frequency polygon can be appreciated as an overlay of the histogram (and therefore communicating the same summarization as the histogram). With a little bit of artistic license (MS word doesn't allow lots of precision), the frequency polygon is the graphed line below in bold black.



#3a.

Age	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Rel. Frequency
14	1	.05	1	.05
15	13	.65	14	.70
16	6	.30	20	1.00
TOTAL	20	1.00		

#3b.



#3c. Males tend to be taller than females

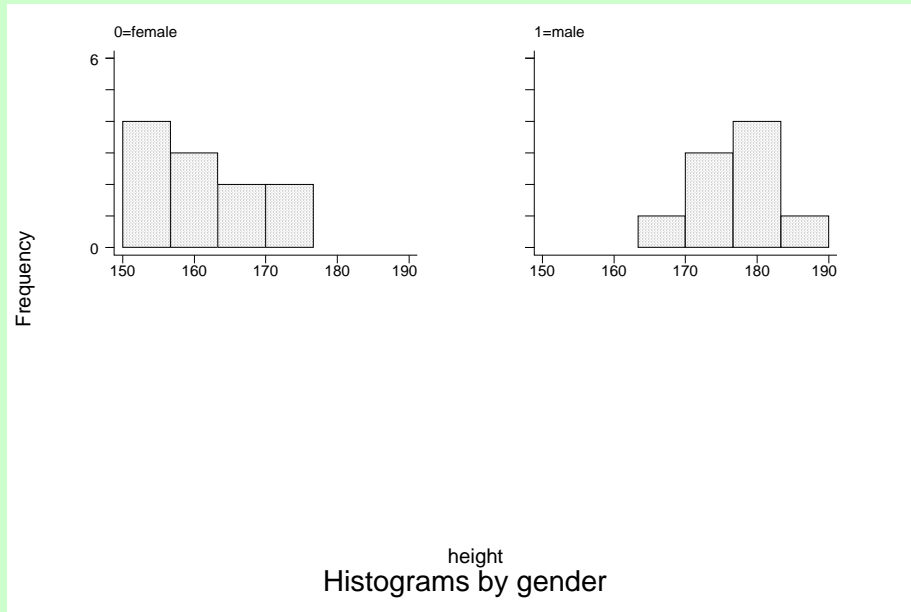
Females	Stem	Males
1 3 4	15	
6 9	15	
0 3	16	
6 7	16	7
0 1	17	4 3 3
	17	8 7
	18	3 3
	18	5

#3d.

Class Interval	<u>FEMALES</u>		<u>MALES</u>	
	Freq.	Re. Freq.	Freq.	Rel. Freq.
150-159	5	.45	0	0
160-169	4	.36	1	.11
170-179	2	.18	5	.56
180-189	0	0	3	.33

I then used STATA to produce the histogram.

```
. graph height, by(gender) bin(6) freq ytick(0,1,2,3,4,5,6) xlabel
```



#4a.

$$\begin{aligned}(X_1 + X_2 + X_3 + X_4)^2 &= \left[\sum_{i=1}^4 X_i \right]^2 \\ &= (3 + 1 + 4 + 6)^2 \\ &= 14^2 \\ &= 196.\end{aligned}$$

#4b.

$$\begin{aligned}X_1^2 + X_2^2 + X_3^2 + X_4^2 &= \sum_{i=1}^4 X_i^2 \\ &= 3^2 + 1^2 + 4^2 + 6^2 \\ &= 9 + 1 + 16 + 36 \\ &= 62.\end{aligned}$$

#4c.

$$\begin{aligned}\sum_{i=1}^4 (X_i - 1)^2 &= (3-1)^2 + (1-1)^2 + (4-1)^2 + (6-1)^2 \\ &= 2^2 + 0^2 + 3^2 + 5^2 \\ &= 4 + 0 + 9 + 25 \\ &= 38.\end{aligned}$$

Note:

$$\begin{aligned}\sum_{i=1}^4 (X_i - 1)^2 &= \sum_{i=1}^4 [X_i^2 - 2X_i + 1] \\ &= \sum_{i=1}^4 X_i^2 - 2 \sum_{i=1}^4 X_i + 1 \sum_{i=1}^4 1 \\ &= 62 - (2)(14) + (1)(4) \\ &= 38.\end{aligned}$$

#4d.

$$\begin{aligned}\sum_{i=1}^4 3X_i &= 3 \sum_{i=1}^4 X_i \\ &= 3(14) \\ &= 42\end{aligned}$$

5A. A stem and leaf diagram might come in handy. Stems are shaded, leaves are not.

3	68851865	→	3	1 5 5 6 6 8 8 8
4	50165165310		4	0 0 1 1 1 3 5 5 5 6 6
5	39113		5	1 1 3 3 9
6	90		6	0 9

MEAN $\bar{x} = \frac{1}{n} \sum_{i=1}^{26} X_i$

$$= \frac{1}{n}(1156) = 44.46 \quad \text{so } \bar{x} = 44.5$$

MEDIAN First solve $\left(\frac{n+1}{2}\right) = \left(\frac{26+1}{2}\right) = 13.5$

Median is midpoint of 13th and 14th observation.

$$\tilde{x} = \frac{1}{2}(41 + 43) \quad \text{so } \tilde{x} = 42$$

MODE This sample is tri - modal 38,41,45

RANGE Maximum - Minimum

$$= 69 - 31 \quad \text{so range} = 38$$

VARIANCE Let's save ourselves the trouble of a very long brute force formula by using the formula for grouped data.

Let j index the unique values. There are 14 unique values.

j	X_j	f_j	$(x_j - \bar{x})^2$	$f_j(x_j - \bar{x})^2$
1	31	1	182.25	182.25
2	35	2	90.25	180.50
3	36	2	72.25	144.50
4	38	3	42.25	126.75
5	40	2	20.25	40.50
6	41	3	12.25	36.75
7	43	1	2.25	2.25
8	45	3	0.25	0.75
9	46	2	2.25	4.50
10	51	2	42.25	84.50
11	53	2	72.25	144.50
12	59	1	210.25	210.25
13	60	1	240.25	240.25
14	69	1	600.25	600.25
TOTALS		26		1998.50

$$S^2 = \frac{\sum_{j=1}^{14} f_j(x_j - \bar{x})^2}{\left(\sum_{j=1}^{14} f_j\right) - 1} = \frac{1998.50}{25} \quad \text{So } S^2 = 79.94$$

$$\text{Standard deviation} \quad S = \sqrt{S^2} \quad \text{So } S = 8.94$$

25th Percentile

First solve $(.25)(n) = (.25)(26) = 6.5$

So 25th percentile is the 7th observation $P_{25} = 38$

75th Percentile

First solve $(.75)(n) = (.75)(26) = 19.5$

So 75th percentile is the 20th observation $P_{75} = 51$

5B.

2	5 5 5 5 5 5 5 5
2	6 6 6 6 6
2	8 8 8
3	0 1
3	4 4

$$MEAN \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{21} X_i = \frac{1}{21}(568) = 27.04 \quad \text{So } \bar{x} = 27.0$$

$$MEDIAN \quad \text{Solving } \left(\frac{n+1}{2}\right) = \left(\frac{21+1}{2}\right) = 11$$

Median is the 11th observation. $\text{So } \tilde{x} = 26$

$$MODE \quad \text{mode} = 25$$

$$RANGE \quad \text{Maximum} - \text{Minimum} \\ = 34 - 25. \quad \text{So} \quad \text{Range} = 9$$

Variance There are 6 unique values.

j	X _j	f _j	(x _j - \bar{x}) ²	f _j (x _j - \bar{x}) ²
1	25	9	4	36
2	26	5	1	5
3	28	3	1	3
4	30	1	9	9
5	31	1	16	16
6	34	2	49	98
TOTALS		21	49	167

$$S^2 = \frac{\sum_{j=1}^6 f_j (x_j - \bar{x})^2}{\left(\sum_{j=1}^6 f_j\right) - 1} = \frac{167}{20} \quad \text{So} \quad S^2 = 8.35$$

Standard deviation $S = \sqrt{S^2} = \sqrt{8.35}$ So $S = 2.89$

25th Percentile

Solving (.25) (n) = (.25) (21) = 5.25

So 25th percentile is 6th observation

$$P_{25} = 25$$

Note - I get this by noticing from the table above that the smallest value (=25) occurs with a frequency of 9 times in the sample.

75th Percentile

Solving (.75) (n) = (.75) (21) = 15.75

So 75th percentile is 16th observation

$$P_{75} = 28$$

Note – I get this by noticing in the table that the value = 28 occurs with a frequency of 3 times in the sample and comes after the first 9 observations all equal to 25 and after the next 5 observations all equal to 26, so that the value of 28 is the 15th, 16th and 17th observations in the ordered sample.

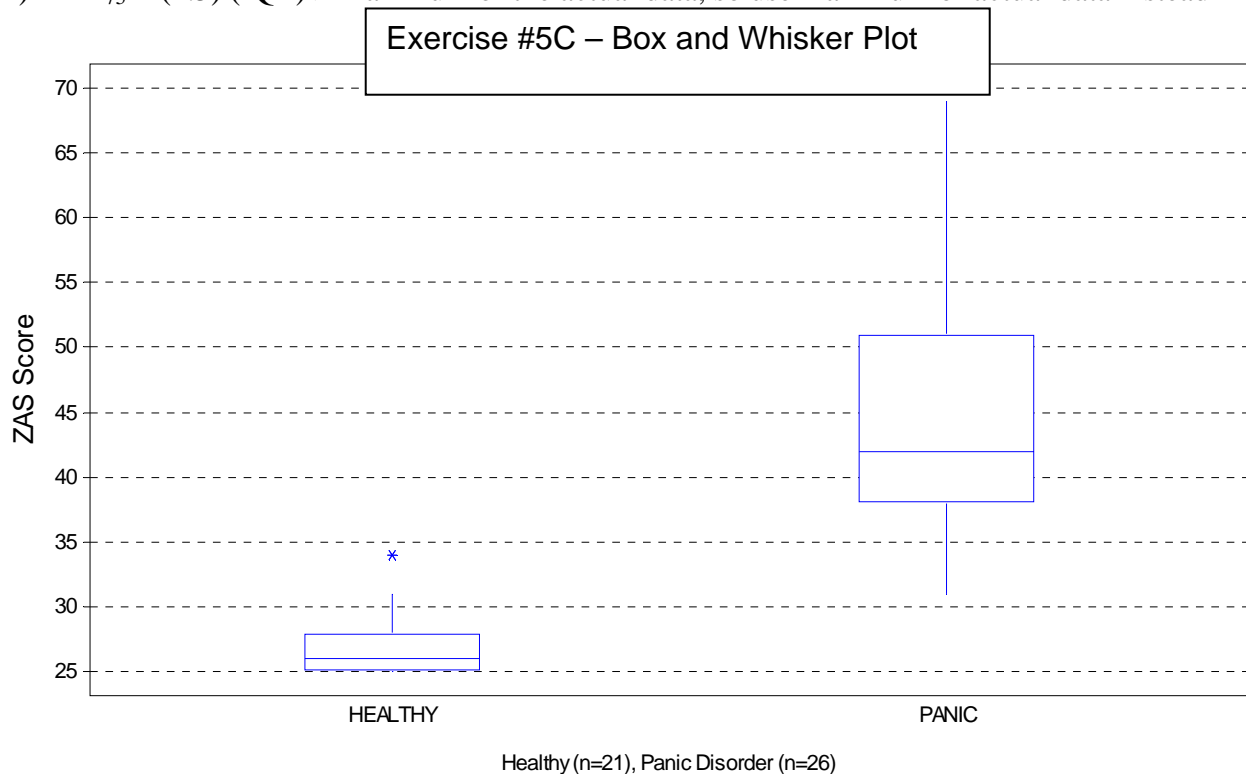
5C. REMINDER - Use the same scale when comparing two groups.

	Group	
	Patients	Controls
Mean	44.5	27.0
Median	42	26
P25	38	25
P75	51	28
Interquartile Range (IQR)	13	3
P25-(1.5)(IQR)	18.5	20.5
P75+(1.5)(IQR)	70.5	32.5*
Min	31*	25*
Max	69*	34

*=Whisker

Notes on Whiskers

- 1) IF $P_{25} - (1.5)(IQR) < \text{minimum of the actual data}$, so use minimum of actual data instead
- 2) IF $P_{75} + (1.5)(IQR) > \text{maximum of the actual data}$, so use maximum of actual data instead



6A.

Class Endpoints	Class Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Freq.
5-14.99	10	5	.067	5	.067
15-24.99	20	10	.133	15	.200
25-34.99	30	20	.267	35	.467
35-44.99	40	22	.293	57	.760
45-54.99	50	13	.173	70	.933
55-64.99	60	5	.067	75	1.000
TOTALS			1.000		

6B.

A cumulative relative frequency polygon for grouped data is, unfortunately, not straightforward in SAS or Stata.

Solution using Excel.

Step 1: Enter your “x” and “y” points into your worksheet such that

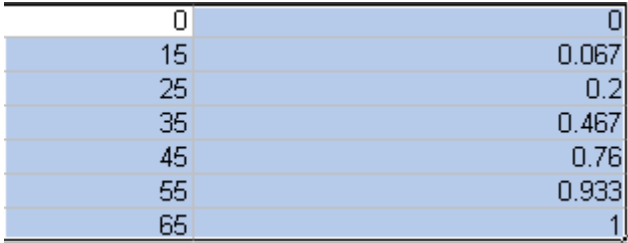

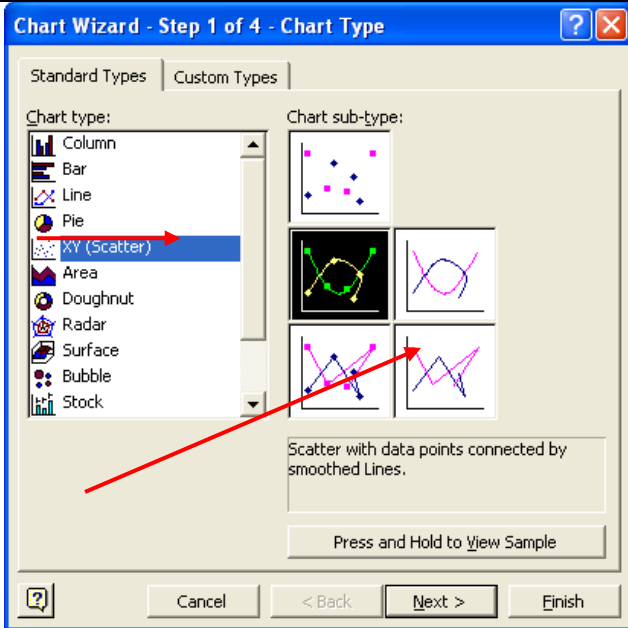
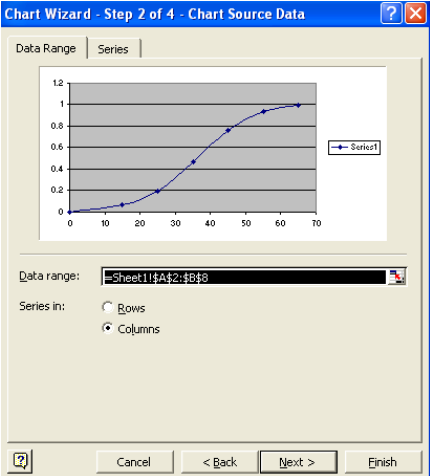
“x” = Endpoint of class interval

“y” = Cumulative relative frequency for the interval

note – Be sure to include an $(x,y) = (0,0)$

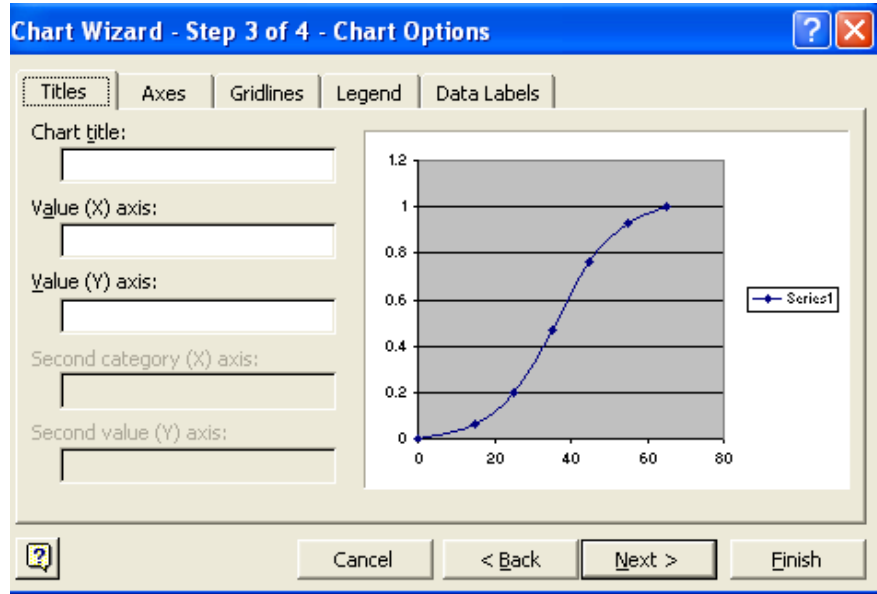
<u>x=age</u>	<u>y=cumulative relative frequency</u>
0	0
15	0.067
25	0.2
35	0.467
45	0.76
55	0.933
65	1

Step 2: Use the chart wizard in excel as follows.

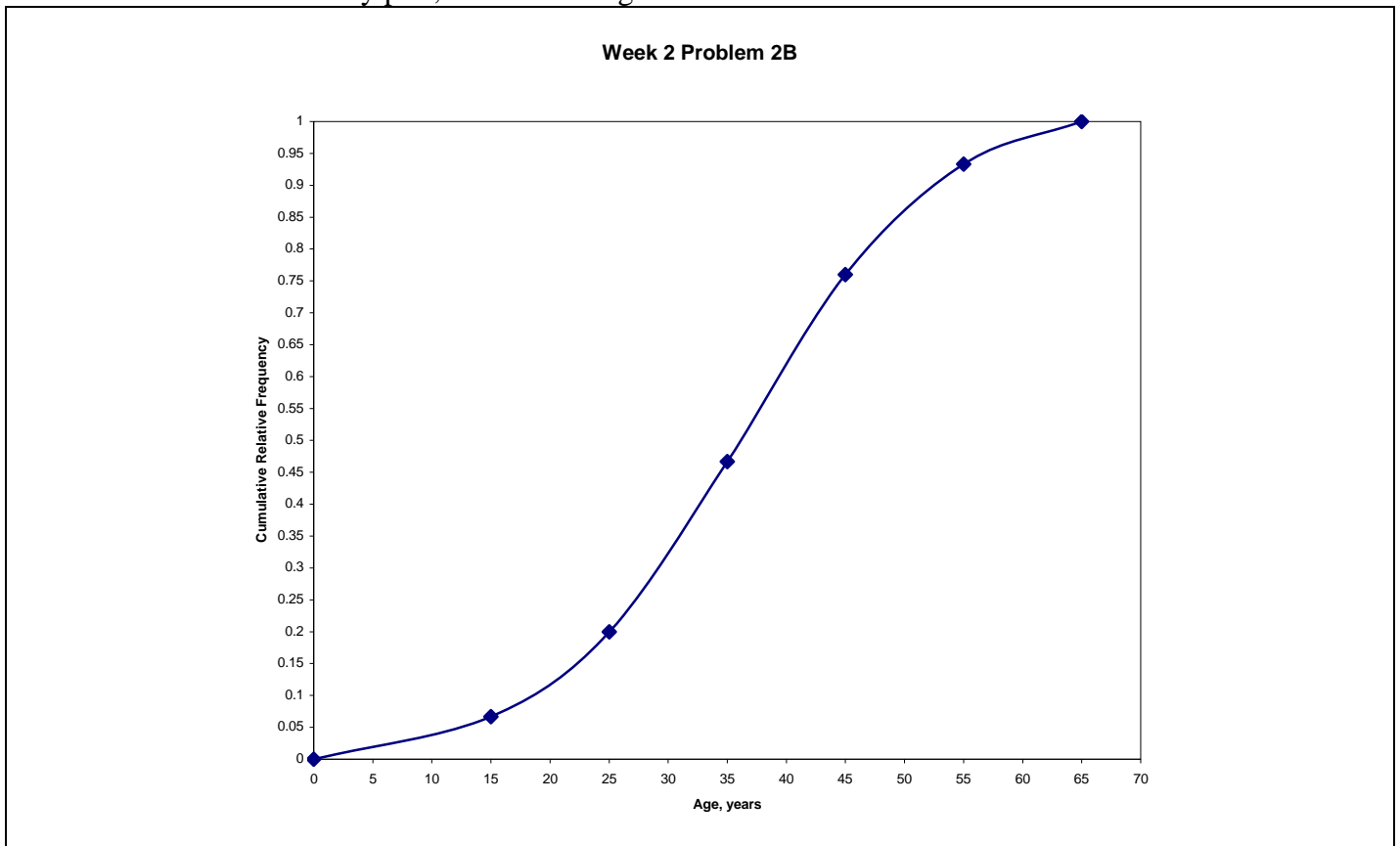
<p>Highlight the data you want to plot</p>	
<p>Click on the chart wizard from the upper toolbar</p>	
<p>Under Chart Type: - Select XY (Scatter) (Scatter)</p> <p>Under Chart sub-type - Highlight the plot with the dots connected</p> <p>Click Next</p>	
<p>You should see the following</p> <p>Click Next</p>	

You will then see a menu that lets you add legends and titles, etc.

And, if you like, you can change such things as shading, tick marks, etc.



After some aesthetics on my part, this is what I got:



Estimates are $P_{10} = 17$ $P_{50} = 36$ $P_{25} = 26$ $P_{75} = 44.5$

6C.

Midpoint X_j	Frequency f_j	$X_j f_j$	$(x_j - \bar{x})$	$f_j(x_j - \bar{x})^2$
10	5	50	-25.7	3302.45
20	10	200	-15.7	2464.90
30	20	600	-5.7	649.80
40	22	880	4.3	406.78
50	13	650	14.3	2658.37
60	5	300	24.3	2952.45
Total	75	2680		12434.75

$$MEAN \quad \bar{x} = \frac{\sum_{j=1}^6 f_j x_j}{\sum_{j=1}^6 f_j} = \frac{2680}{75} \quad \text{So } \bar{x} = 35.7$$

MEDIAN *Note to reader* – I've consulted a number of texts on this. There is no single correct answer. With interval data, whatever median you calculate is an approximation. Here is what is suggested in Think and Explain with Statistics (Lincoln E. Moses, page 64)

$$\text{First solve } \frac{n+1}{2} = \frac{75+1}{2} = 38^{th} \text{ observation}$$

Examination of the table reveals that the 38th observation is in the interval 35 to 44.99

Set the following quantities:

The letter l = lower limit of interval = 35

The letter u = upper limit of interval = 44.99

R = cumulative frequency up to the lower limit of interval = 35

M = # observations contained in interval = 22

N = total # observations = 75

An approximate solution for the median is calculated as

$$\tilde{x} = l + \left[\frac{N/2 - R}{M} \right] (u - l) = 35 + \left[\frac{75/2 - 35}{22} \right] (44.99 - 35) = 36.135 \text{ or } \mathbf{37}$$

VARIANCE

$$S^2 = \frac{\sum_{j=1}^6 f_j (x_j - \bar{x})^2}{\left(\sum_{j=1}^6 f_j\right) - 1} = \frac{12434.75}{74} \text{ so } S^2 = 168.04$$

Standard deviation $S = \sqrt{S^2}$ so $S = 13.0$

7A. Remember to use the same scale.

	1	2	3
median	8	8	7
mean	13.6	6.4	8.2
left box edge = P ₂₅	5	4	5
right box edge = P ₇₅	11	11	12
IQR = P ₇₅ -P ₂₅	6	7	7
P ₂₅ -(1.5)(IQR)	-4	-6.5	-5.5
P ₇₅ +(1.5)(IQR)	20	21.5	22.5
left whisker	5	1	4
right whisker	20	13	14

7B.

When data are skewed by extreme values, medians and quartiles give a better feel for the bulk of the data than do means and standard deviations. This example also illustrates that, as sample size increases, the range can only increase. Notice that the extreme value of 40 occurred in the sample with the largest sample size.