Lecture 2: Visual Display of Data

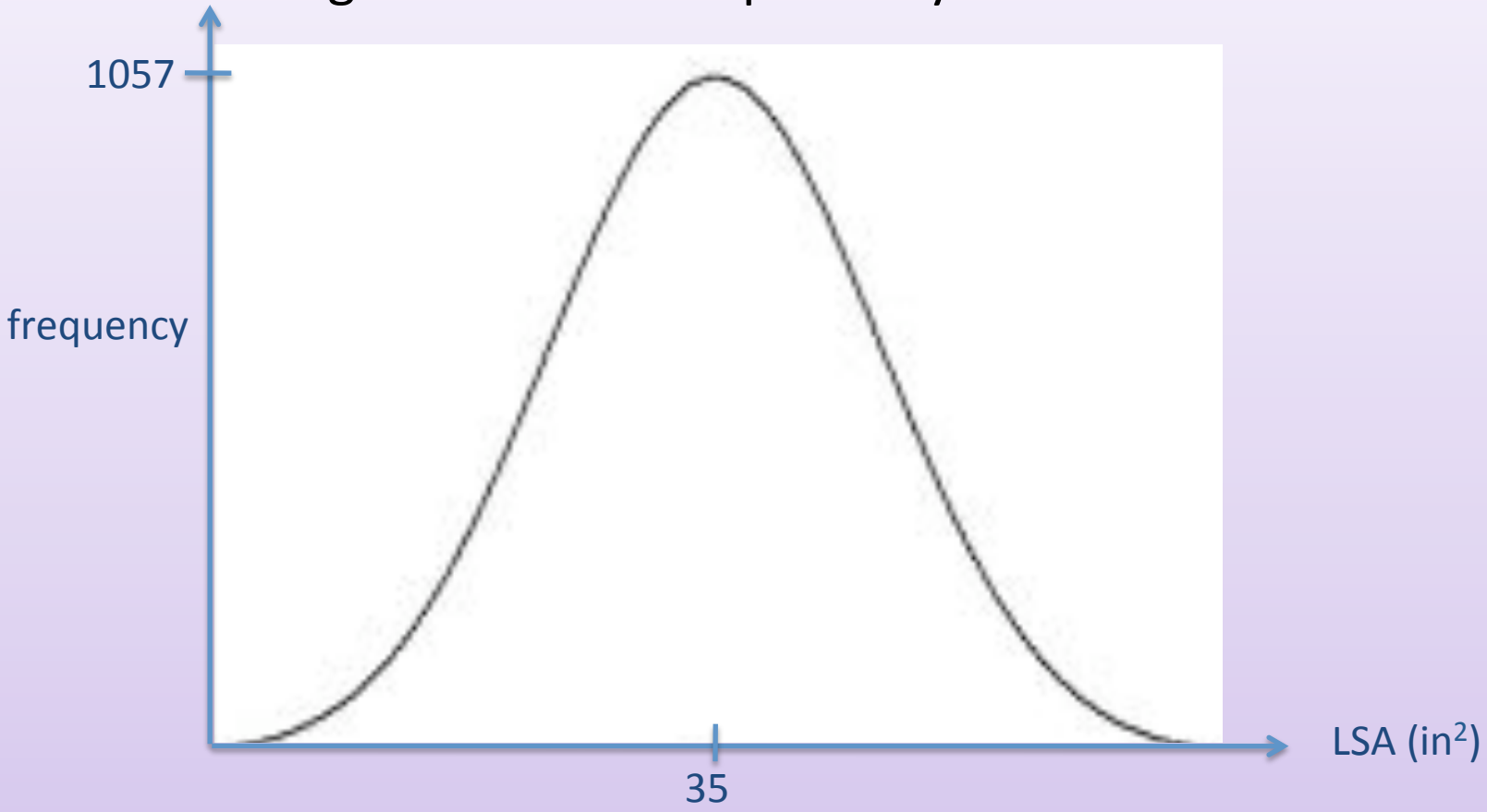
- Overview: In addition to considering various summary statistics, it is also common to consider some visual display of the data
- Outline:
  1. Histograms
  2. Scatter Plots
  3. Assignment

Frequency Distributions

- Consider the experiment of measuring the leaf surface area (LSA) of a certain type of plant in a certain location
- It's *possible* that we could make tens of thousands of measurements and it's possible that we obtain several "repeat" measurements for various values
- For example, we may find that 1057 of the plants we observed had LSA 35 in^2 and 239 of the plants we observed had LSA 13 in^2; and so on
- If we were to plot the LSA vs. the frequency that number was observed, we would obtain a **frequency distribution**

Normal Distribution

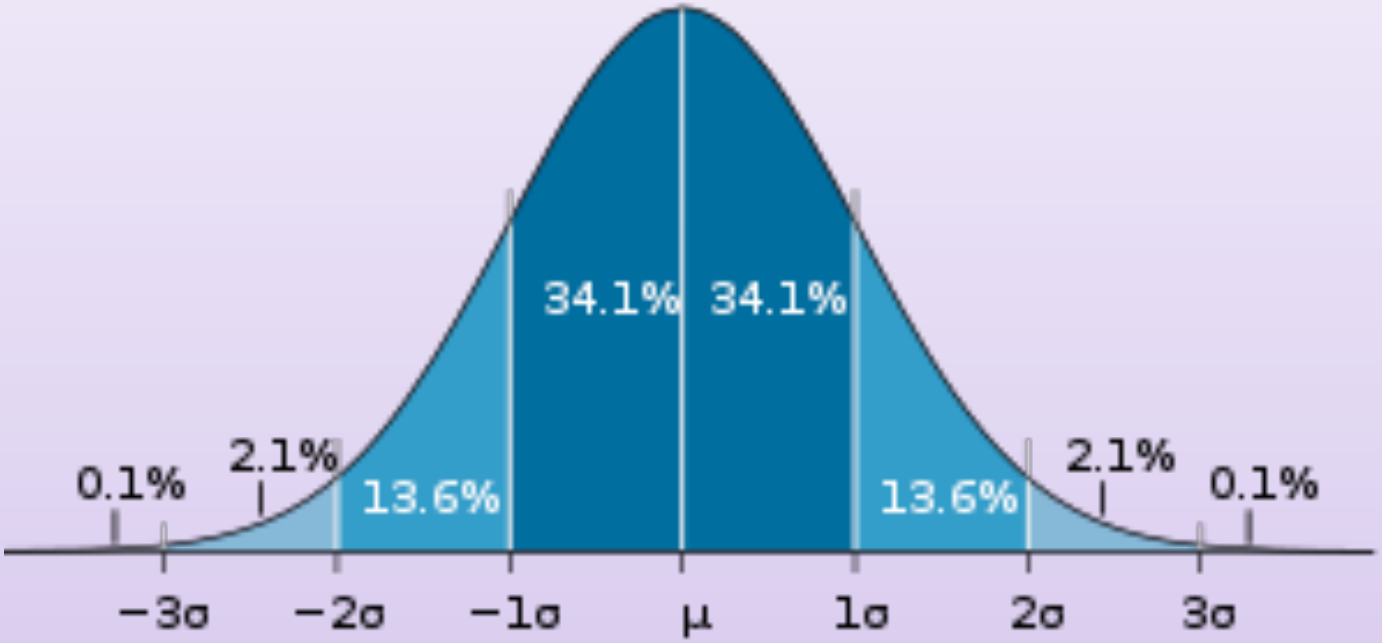
- There's a good chance our plot may look like this:



- This is the familiar "bell-curve" – a normal distribution

Normal Distribution

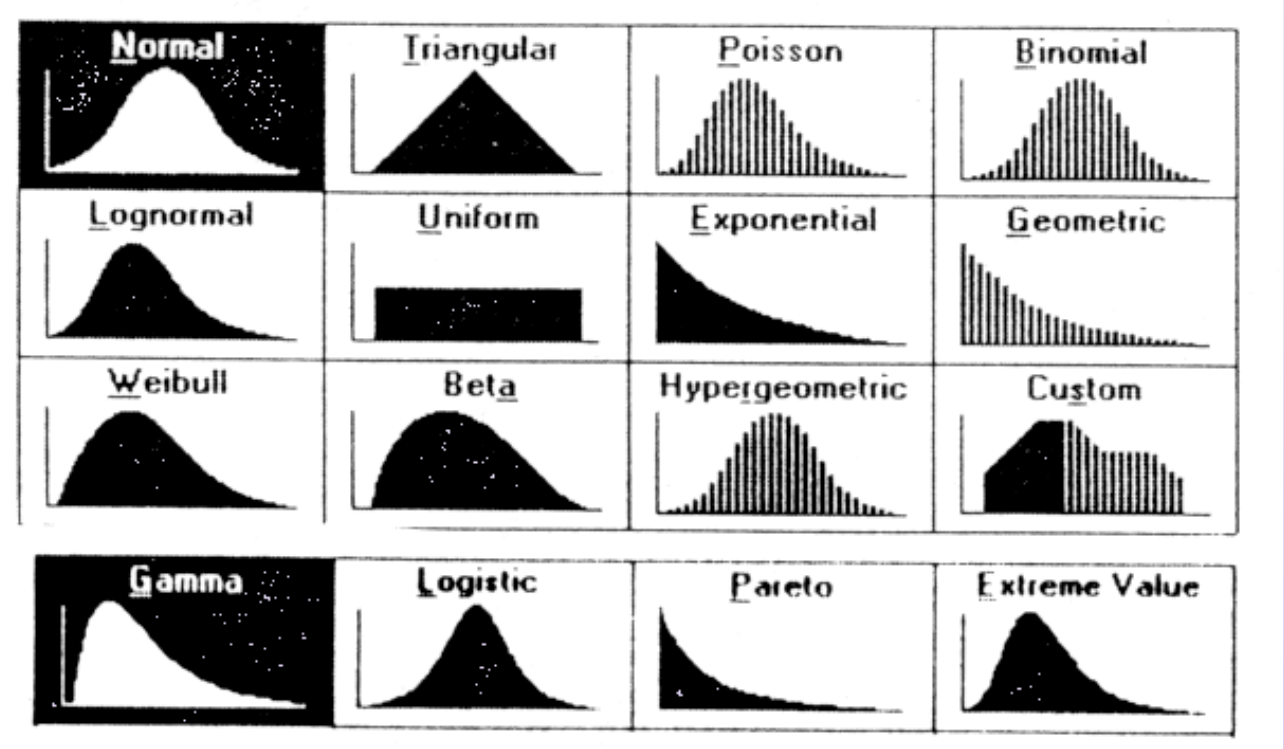
- If the data are normally distributed then, ≈68.3% of the values are within 1 standard deviation of the mean; ≈95.4% within 2 SDs; ≈99.7% within 3 SDs



- In any distribution, smaller σ implies data are closer together and larger σ implies data are spread further apart

Probability Distributions

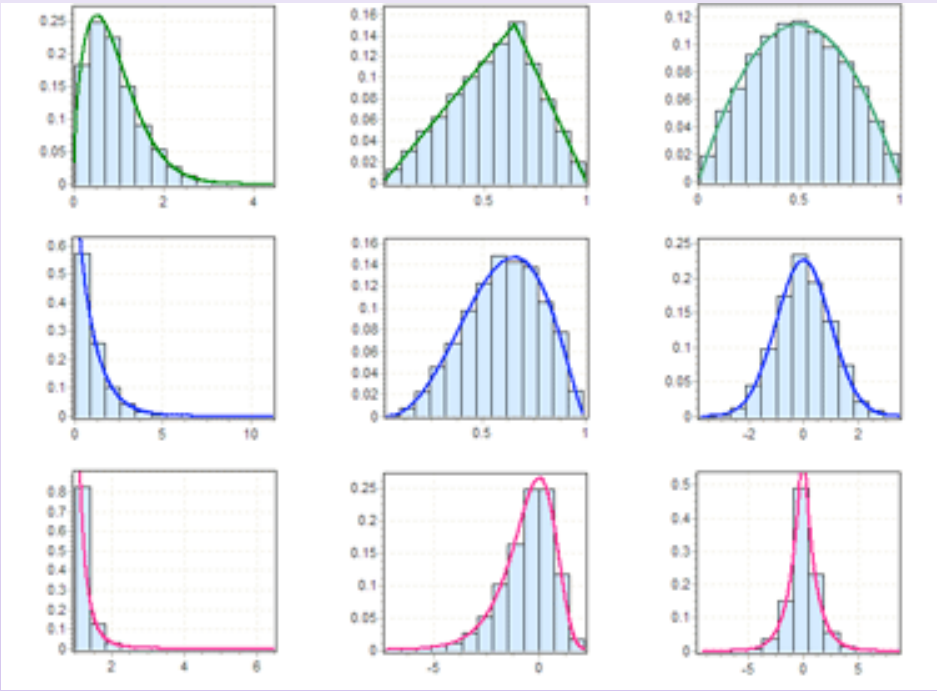
- There are many different probability distributions:



- As can be seen here, the type of underlying variable- discrete or continuous- determines the type of distribution function- probability mass function or probability distribution function

Probability Distributions & Histograms

- (Back to our experiment) Realistically, we won't obtain enough measurements for our distribution to be a nice, smooth curve. A **histogram** is an estimate of the probability distribution of a continuous variable:



Constructing a Histogram

1. Decide on the number of classes (bars) for the histogram
   - Typically 10-20, but for a small dataset, 4-6
2. Determine the width of the classes (CW)
   - CW = (Range) / (# classes) rounded **up** to precision of data
3. Select the smallest observed value as the lower limit of the first class and add multiples of the CW from step (2) to obtain the *lower* (not upper!) limits for each class
4. To find the upper class limits, add the CW to the lower class limits and subtract from this the precision of the data
   - So if the CW is 0.3 and the lower limit of the first class is 3.7, then the upper class limit for the first class is 3.7 + .3 -.1 = 3.9
   - There is a gap between the upper limit of one class and the lower limit of the next; the gap width is the precision of the data
5. The class *boundaries* will be the midpoint of these gaps

Example 2.2 Histogram for Black Bear Data

- In 2002, Katie Settlage and her field research team collected the following weights for female black bears in the southern Appalachians:

60	85	95	85	115
75	140	145	120	110
90	115	75	125	80
80	80	80	110	75
120	150	38	118	

where the weight is given in pounds. Let's go through the steps of constructing a histogram for this set of data.

Example 2.2 Histogram for Black Bear Data

60	85	95	85	115
75	140	145	120	110
90	115	75	125	80
80	80	80	110	75
120	150	38	118	

1. Decide on the number of classes (bars) for the histogram:
   This is small data set- 24 points- so we choose 6 classes

2. Determine the width of the classes (CW):
   CW = (Range) / (# classes) rounded **up** to precision of data
   = (150 – 38) / 6 = 18.66… => 19

Example 2.2 Histogram for Black Bear Data

3. Select the smallest observed value as the lower limit of the first class and add multiples of the CW from step (2) to obtain the *lower* (not upper!) limits for each class

Class	Lower				
1	38				
2					
3					
4					
5					
6					

Example 2.2 Histogram for Black Bear Data

3. Select the smallest observed value as the lower limit of the first class and add multiples of the CW from step (2) to obtain the *lower* (not upper!) limits for each class

Class	Lower				
1	38				
2	57				
3					
4					
5					
6					

Example 2.2 Histogram for Black Bear Data

3. Select the smallest observed value as the lower limit of the first class and add multiples of the CW from step (2) to obtain the *lower* (not upper!) limits for each class

Class	Lower				
1	38				
2	57				
3	76				
4	95				
5	114				
6	133				

Example 2.2 Histogram for Black Bear Data

4. To find the upper class limits, add the CW to the lower class limits and subtract from this the precision of the data

Class	Lower	Upper			
1	38	56			
2	57	75			
3	76	94			
4	95	113			
5	114	132			
6	133	151			

Example 2.2 Histogram for Black Bear Data

5. The class *boundaries* will be the midpoint of the gaps between the upper and lower boundaries

Class	Lower	Upper	Lower Boundary	Upper Boundary	
1	38	56	37.5	56.5	
2	57	75	56.5	75.5	
3	76	94	75.5	94.5	
4	95	113	94.5	113.5	
5	114	132	113.5	132.5	
6	133	151	132.5	151.5	

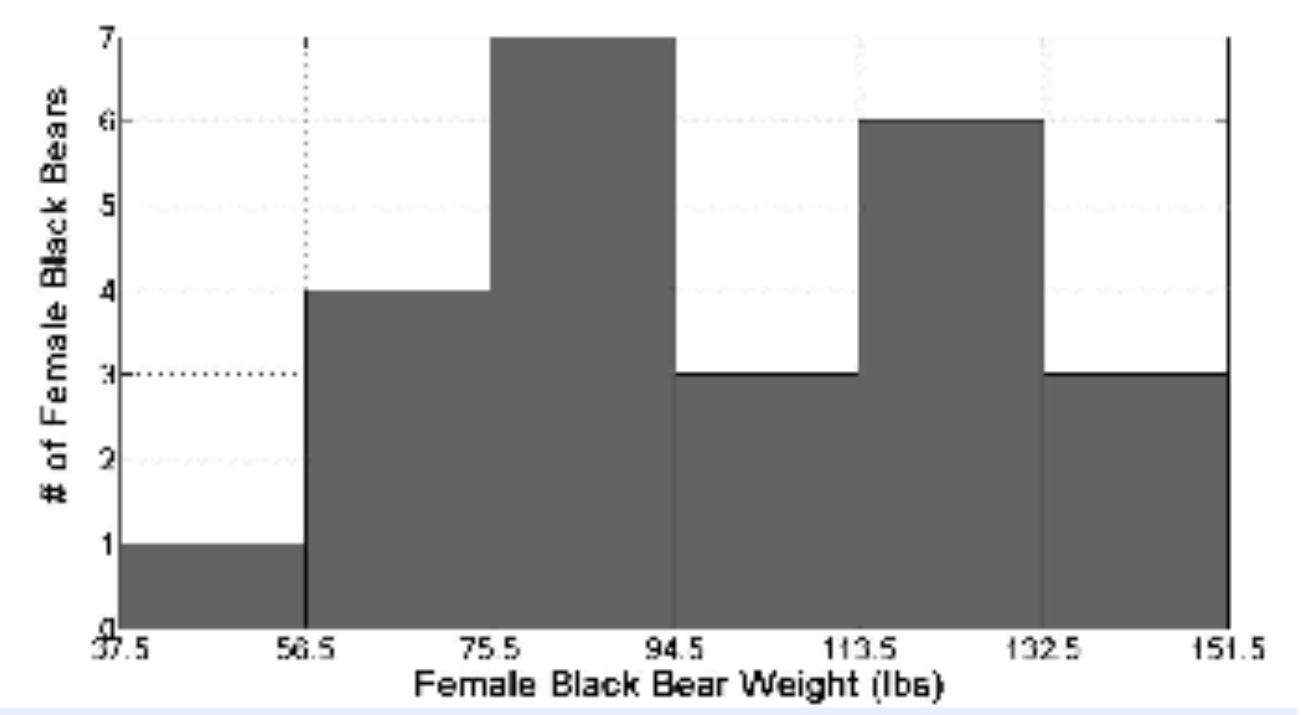
Example 2.2 Histogram for Black Bear Data

6. We need to count the frequencies for each class

Class	Lower	Upper	Lower Boundary	Upper Boundary	Frequency
1	38	56	37.5	56.5	1
2	57	75	56.5	75.5	4
3	76	94	75.5	94.5	7
4	95	113	94.5	113.5	3
5	114	132	113.5	132.5	6
6	133	151	132.5	151.5	3

Example 2.2 Histogram for Black Bear Data

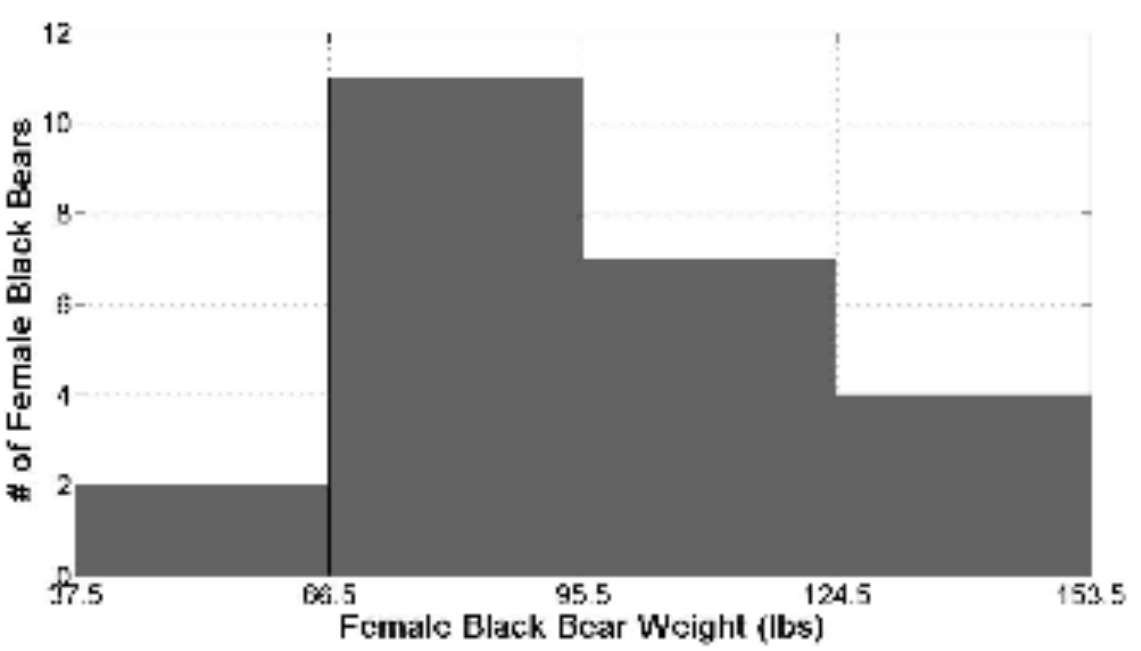
7. Draw the histogram:



The histogram seems to have two peaks- we say it is *bimodal*

Example 2.2 Histogram for Black Bear Data

However, when classifying data as bimodal, we must be careful. Suppose we had chosen only four classes. We would have produced this histogram:



This histogram appears to be unimodal.

Example 2.2 Histogram for Black Bear Data

8. Finally, we may also consider the *relative* frequencies for each class:

Class	Lower	Upper	Lower Boundary	Upper Boundary	Frequency	Relative Frequency
1	38	56	37.5	56.5	1	1/24
2	57	75	56.5	75.5	4	1/6
3	76	94	75.5	94.5	7	7/24
4	95	113	94.5	113.5	3	1/8
5	114	132	113.5	132.5	6	1/4
6	133	151	132.5	151.5	3	1/8

Histograms vs Scatter Plots

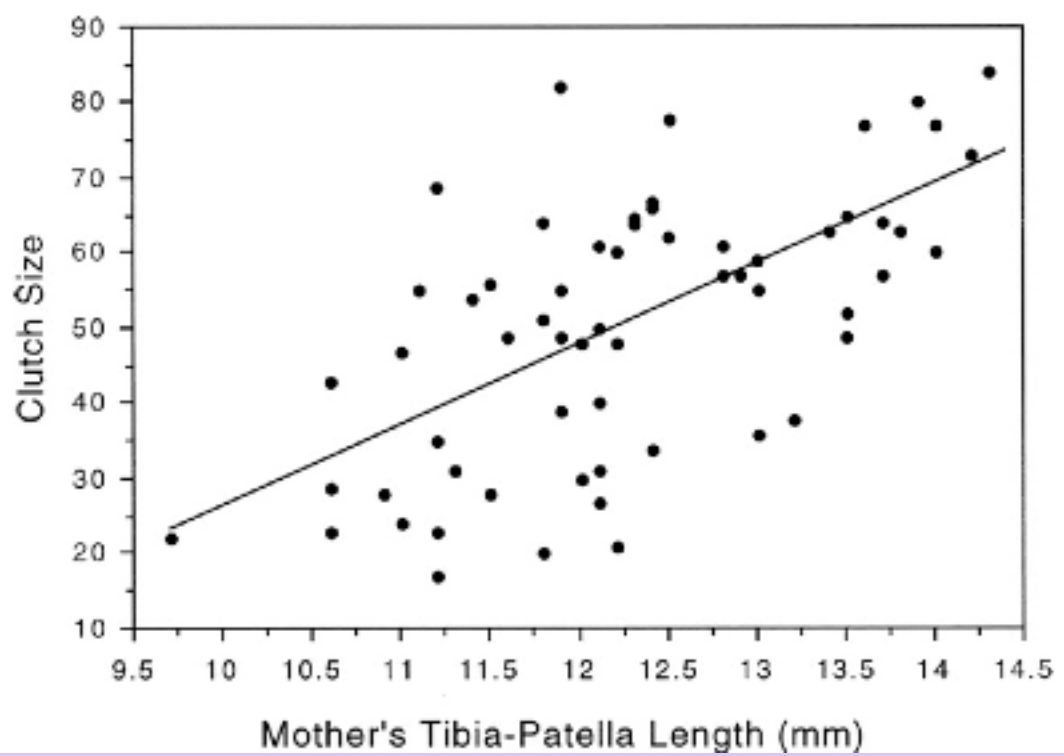
- Histograms are used to visually display *single* sets of measurements
- In many cases however, the value obtained from a measurement is related to other characteristics or factors affecting the same object being measured; for example:
  - a child's birth weight is related to the mother's birth weight
  - leaf photosynthesis rate is affected by the light level
  - the number of eggs laid by a spider may be related to the size of the mother
- The objective of a scatter plot is to determine visually whether there *appears* to be any relationship between two different observations or characteristics of a single biological object

Creating Scatter Plots

- Creating a scatter plot is easy once it is decided which characteristics are to be included
  - Each axis is chosen to cover the range of measurements for that observation set
  - One measurement is chosen for the horizontal axis and the other for the vertical- which is chosen for which axis which does not matter unless you have some reason to believe that one measurement *causes* changes in the other, in which case it is conventional to use the horizontal axis for the measurement that is viewed as the causative factor
- Then for each object measured, a single point is plotted on the graph at the appropriate coordinates

Creating Scatter Plots

- For example, suppose we plot the number of eggs laid by vs the sizes (length in mm) of 38 different spiders:



2. Assignment

- Exercises 2.2, 2.3, 2.4, 2.7, 2.9, 2.10

Homework

2.3 (a), (e)

The length of time an outpatient must wait for treatment is a variable that plays an important part in the design of outpatient clinics. The waiting times (in minutes) for 50 patients at a pediatric clinic are as follows:

35	22	63	6	49	19	15	83	46	19
16	31	24	29	26	68	42	57	64	8
23	47	21	51	7	40	19	46	16	32
108	33	55	32	22	36	25	27	37	58
39	10	42	72	13	51	45	77	16	28

Construct a histogram by hand using 5 classes.

Solution: First, Determine the width of the classes (CW):

CW = (Range) / (# classes) rounded **up** to precision of data

= (108 – 6) / 5 = 20.4… => 21

Homework

2.3 (a), (e)

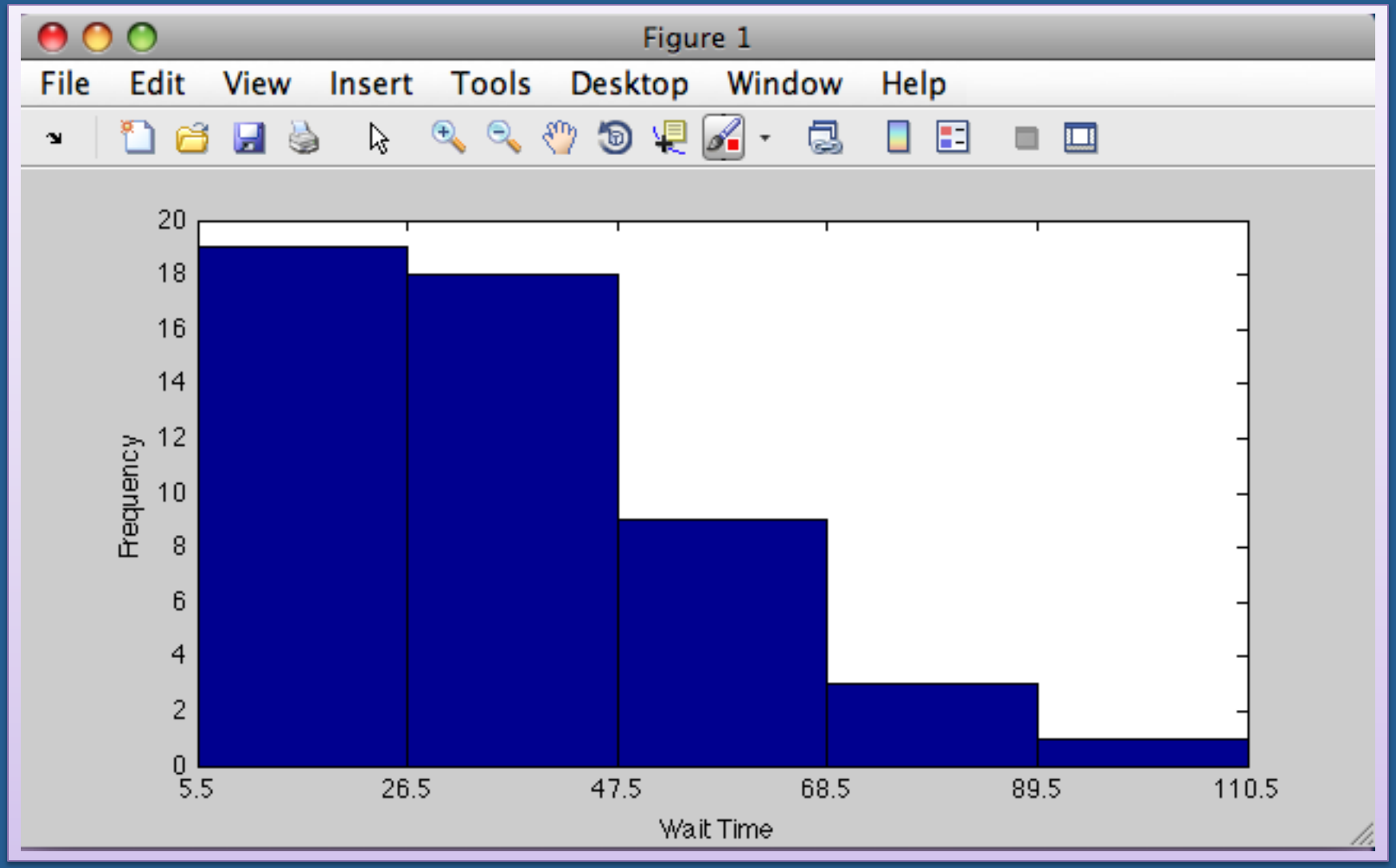
Next, we fill in the table:

Class	Lower	Upper	Lower Boundary	Upper Boundary	Frequency
1	6	26	5.5	26.5	19
2	27	47	26.5	47.5	18
3	48	68	47.5	68.5	9
4	69	89	68.5	89.5	3
5	90	110	89.5	110.5	1

Finally, draw the histogram:

Homework

2.3 (a), (e)



Homework

2.4 (b)

The regulations of the Board of Health specify that the fluoride level in drinking water not exceed 1.5 parts per million (ppm). Each of the 11 measurements below represent average fluoride levels over 15 days in ppm.

0.75 0.86 0.84 0.97 0.94 0.89 0.88 0.78 0.77 0.76 0.82

(b) Construct a relative frequency histogram for this data using an appropriate number of classes (by hand).

Solution: First, Determine the width of the classes (CW):

CW = (Range) / (# classes) rounded **up** to precision of data

= (0.97 – 0.75) / 4 = 0.055… => 0.06

Homework

2.4

Next, we fill in the table:

Class	Lower	Upper	Lower Boundary	Upper Boundary	Frequency	Relative Frequency
1	0.75	0.80	0.745	0.805	4	4/11
2	0.81	0.86	0.805	0.865	3	3/11
3	0.87	0.92	0.865	0.925	2	2/11
4	0.93	0.98	0.925	0.985	2	2/11

Homework

2.4

Finally, draw the histogram: