Section 8.1: Frequency Distributions

In this section, we look at ways to organize data in order to make it more user friendly. It is difficult to obtain any meaningful information from the data as presented in the two data sets below. We will look at ways to organize and present this data in ways from which a meaningful summary of the data can be derived at a glance. There is a whole discipline devoted to this sort of problem. A classic book in the field still today is E.R. Tufte, *The Visual Display of Quantitative Information.* Cheshire, Connecticut: Graphics Press. (1983)

Data Set 1 A random sample of 20 students at the University of Notde Same were asked to estimate the average number of hours they spent per week studying outside of class. Also their eye color and the number of pets they owned was recorded. The results are given below.

Student	# Hours Studying	Eye Color	# Pets
Student 1	10	blue	1
Student 2	7	brown	0
Student 3	15	brown	3
Student 4	20	green	1
Student 5	40	blue	2
Student 6	25	green	1
Student 7	22	hazel	0
Student 8	13	brown	5
Student 9	12	gray	4
Student 10	21	hazel	3
Student 11	16	blue	1
Student 12	22	green	1
Student 13	25	brown	1
Student 14	30	green	2
Student 15	29	brown	0
Student 16	25	green	4
Student 17	27	gray	0
Student 18	15	hazel	1
Student 19	14	blue	2
Student 20	17	brown	2

Data Set 2: EPAGAS The Environmental Protection Agency (EPA) perform extensive tests on all new car models to determine their mileage ratings. The 25 measurements given below represent the results of the test on a sample of size 25 of a new car model.

EPA mileage ratings on 25 cars

36.3	41.0	36.9	37.1	44.9
40.5	36.5	37.6	33.9	40.2
38.5	39.0	35.5	34.8	38.6
41.0	31.8	37.3	33.1	37.0
37.1	40.3	36.7	37.0	33.9

Frequency Table or Frequency Distribution

To construct a frequency table, we divide the observations into **classes or categories**. The number of observations in each category is called the **frequency** of that category. A **Frequency Table** or

Frequency Distribution is a table showing the categories next to their frequencies. When dealing with **Quantitative data** (data that is numerical in nature), the categories into which we group the data may be defined as a range or an interval of numbers, such as $0 - 10$ or they may be single outcomes (depending on the nature of the data). When dealing with **Qualitative data**(non numerical data), the categories may be single outcomes or groups of outcomes. When grouping the data in categories, make sure that they are disjoint (to ensure that observations do not fall into more than category) and that every observation falls into one of the categories.

Example: Data Set 1 We create frequency distributions for the data on eye color and the number of pets owned below. Note that we lose some information from our original data set by separating the data. There are simple methods of presenting paired data which we do not have time to study in this course.

Eye Color (Category)	**# of Students** (Frequency)
Blue	
Brown	
Gray	
Hazel	
Green	
Total	

# Pets (Category)	**# of Students** (Frequency)
0	
1	
2	
3	
4	
5	
Total	

Note that the sum of the frequencies equals the total number of observations, in this case the number of students in our sample.

Eye Color (Category)	**# of Students** (Frequency)
Blue	4
Brown	6
Gray	2
Hazel	5
Green	3
Total	20

# Pets (Category)	**# of Students** (Frequency)
0	4
1	7
2	4
3	2
4	2
5	1
Total	20

The **relative frequency** of a category is the frequency of that category (the number of observations that fall into the category) divided by the total number of observations:

$$\text{Relative Frequency of Category i} = \frac{\text{frequency of category i}}{\text{total number of observations}}.$$

We may wish to also/only record the **relative frequency** of the classes (or outcomes) in our table.

Eye Color (Category)	**Proportion of Students** (Rel. Frequency)
Blue	
Brown	
Gray	
Hazel	
Green	
Total	

# Pets (Category)	**Proportion of Students** (Rel. Frequency)
0	
1	
2	
3	
4	
5	
Total	

Eye Color (Category)	**Proportion of Students** (Rel. Frequency)
Blue	0.20
Brown	0.30
Gray	0.10
Hazel	0.25
Green	0.15
Total	20

# Pets (Category)	**Proportion of Students** (Rel. Frequency)
0	0.20
1	0.35
2	0.20
3	0.10
4	0.10
5	0.05
Total	20

Choosing Categories

- When choosing categories, **the categories should cover the entire range of observations, but should not overlap.** If the categories chosen are intervals one should specify what happens to data at the end points of the intervals.
  - For example if the categories are the intervals 0-10, 10-20, 20-30, 30-40, 40-50. One should specify which interval 10 goes into, which interval 20 goes into etc... Mathematicians normally use different brackets in interval notation to indicate whether the endpoint is included or not. The notation $[0, 10)$ denotes the interval from 0 to 10 where 0 is included in the interval but 10 is not.

- Common sense should be used in forming categories. **Somewhere between 5 and 15 categories gives a meaningful picture that is easily processed.** However if there are only 3 candidates for a presidential election and you conduct a poll to determine who those polled will vote for, then it is natural to choose 3 categories.

- **To choose intervals as categories** with quantitative data, one might subtract the smallest observation from the largest and divide by the desired number of intervals. This gives a rough idea of interval length. You should then adjust it to a simpler (larger) number which is relatively close to it. Then make intervals of the desired length where the first starts at a natural point lower than the minimum observation and the last ends at a natural point greater than the maximum observation.

  – For example, if you data ranged from 1 to 29, and you wanted to create 6 categories as intervals of equal length. The length of each should be approximately $\frac{29-1}{6} \approx 4.667$. It is natural to use 6 intervals of length 5 in this case, with the first starting at 0 and the last ending at 30. If we decide to include the right end point and exclude the left end point for each interval, our intervals are :$(0, 5],\ (5, 10],\ (10, 15],\ (15, 20],\ (20, 25],\ (25, 30]$.

Example: Data set 2 Make a frequency distribution (table) for the data on mileage ratings using 5 intervals of equal length. Include the left end point of each interval and omit the right end point.

Mileage (Category)	**# of cars** (Frequency)
[,)	
[,)	
[,)	
[,)	
[,)	
Total	

EPA mileage ratings on 25 cars

36.3	41.0	36.9	37.1	44.9
40.5	36.5	37.6	33.9	40.2
38.5	39.0	35.5	34.8	38.6
41.0	31.8	37.3	33.1	37.0
37.1	40.3	36.7	37.0	33.9

We are told to divide the data into 5 intervals of equal length. The smallest value in the data is 33.9 and the largest is 44.9 and $\frac{44.9 - 33.9}{5} = 2.2$. If we start at 30.0 and use intervals of length 3, 5 intervals later will end at 45.0 so we cover the data.

Mileage (Category)	**# of cars** (Frequency)
[30, 33)	1
[33, 36)	5
[36, 39)	12
[39, 42)	6
[42, 45)	1
Total	25

The value 39.0 goes in the interval $[39, 42)$ NOT the interval $[36, 39)$.

Example: Data set 1 Make a frequency distribution (table) for the data on the estimated average number of hours spent studying in data set 1, using 7 intervals of equal length. Include the left end point of each interval and omit the right end point.

Hours Studying	**# of students**
(Category)	(Frequency)
[,)	
[,)	
[,)	
[,)	
[,)	
[,)	
[,)	
Total	

Student	# Hours Studying
Student 1	10
Student 2	7
Student 3	15
Student 4	20
Student 5	40
Student 6	25
Student 7	22
Student 8	13
Student 9	12
Student 10	21
Student 11	16
Student 12	22
Student 13	25
Student 14	30
Student 15	29
Student 16	25
Student 17	27
Student 18	15
Student 19	14
Student 20	17

We are told to divide the data into 7 intervals of equal length. The smallest value in the data is 7 and the largest is 40. Since $\frac{40-7}{7} = 4.71428571428571$ so lets use intervals of length 5 starting at 5, we will end at 40. Since we have a value of 40 and we have agreed to put end point values to the right hand interval, this does not quite work. If we start with 6 we will be OK.

Hours Studying	**# of students**
(Category)	(Frequency)
[6, 11)	2
[11, 16)	5
[16, 21)	3
[21, 26)	6
[26, 31)	3
[31, 36)	0
[36, 41)	1
Total	20

Representing Qualitative data graphically

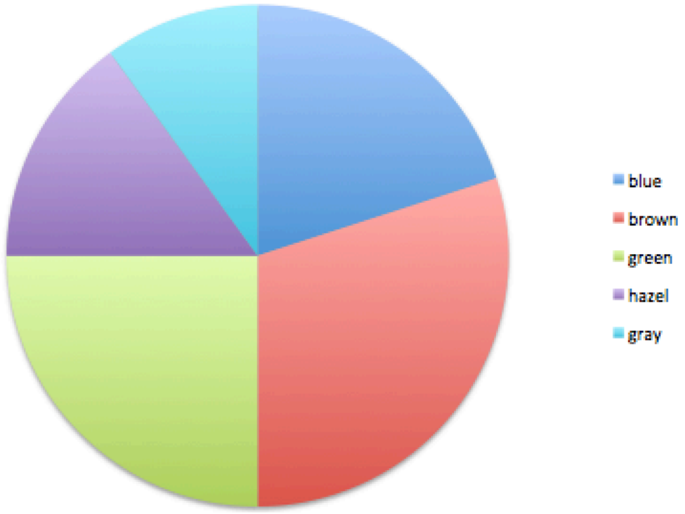
Pie Chart One way to present our qualitative data graphically is using a **Pie Chart**. The pie is represented by a circle (Spanning 360^0). The size of the pie slice representing each category is proportional to the relative frequency of the category. The angle that the slice makes at the center is also proportional to the relative frequency of the category; in fact the angle for a given category is given by:

$$\text{category angle at the center} =$$

$$\text{relative frequency category} \times 360^0.$$

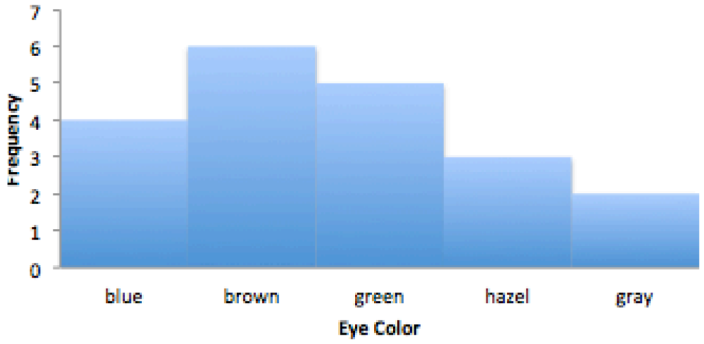
The pie chart always adheres to the **area principle**. That is the proportion of the area of the pie devoted to any category is the same as the proportion of the data that lies in that category.This principle is commonly violated to alter perception and subtly promote a particular point of view (see end of lecture).

Example 1 Below we present the data on eye color from data set 1 in a pie chart on the left.



Bar Graphs We can also represent our data graphically on a **Bar Chart** or **Bar Graph**. Here the categories of the qualitative variable are represented by bars, where the height of each bar is either the category frequency, category relative frequency, or category percentage.

The bases of all bars should be equal in width. Having equal bases ensures that the bar graph adheres to the **area principle**, which in this case means that the proportion of the total area of the bars devoted to a category(= area of the bar above a category divided by the sum of the areas of all bars) should be the same as the proportion of the data in the category. This principle is often violated to promote a particular point of view (see end of lecture).

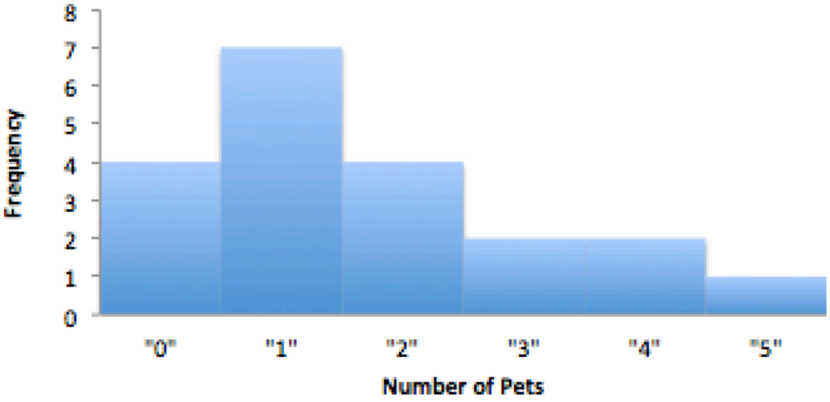


Representing Quantitative data using a Histogram

Histograms A **histogram** is a bar chart in which each bar represents a category and its height represents either the frequency, relative frequency(proportion) or percentage in that category.

If a variable can only take on a finite number of values (or the values can be listed in an infinite sequence) the variable is said to be **discrete**.

For example the number of pets in Data set 1 was a discrete variable and each value formed a category of its own. In this case, each bar in the histogram is centered over the number corresponding to the category and all bars have equal width of 1 unit. (see below).



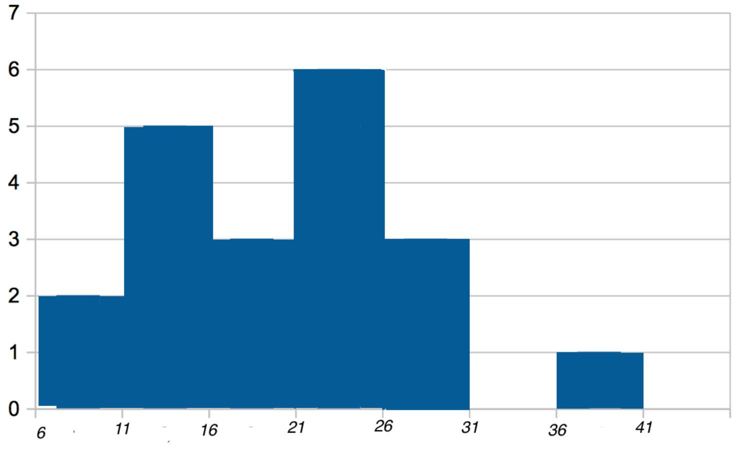
If a variable can take all values in some interval, it is called a **continuous** variable. If our data consists of observations of a continuous variable, such as that in data set 2, the categories used for our histogram should be intervals of equal length (to adhere to the area principle) formed in a manner similar to that described above for frequency tables. The bases of the bars in our histogram are comprised of these categories of equal length and their heights represent either the frequency, relative frequency or percentage in each category. Because it is difficult to tell from the histogram alone which endpoints are included in the categories, we adopt the convention that the categories(intervals) include the left endpoint but not the right endpoint.

Example Construct a histogram for the data in data set 2 on EPA mileage ratings, using the categories used above in the frequency table. Use the frequency of observations in each category to define the height of the bars.

Mileage (Category)	**# of cars** (Frequency)
[,)	
[,)	
[,)	
[,)	
[,)	
Total	

On the left is the frequency data from above.

Hours Studying (Category)	**# of students** (Frequency)
[6, 11)	2
[11, 16)	5
[16, 21)	3
[21, 26)	6
[26, 31)	3
[31, 36)	0
[36, 41)	1
Total	20

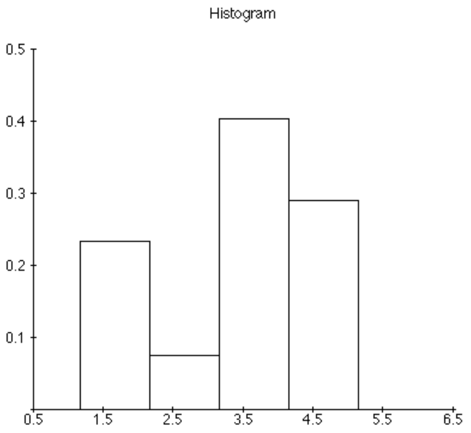


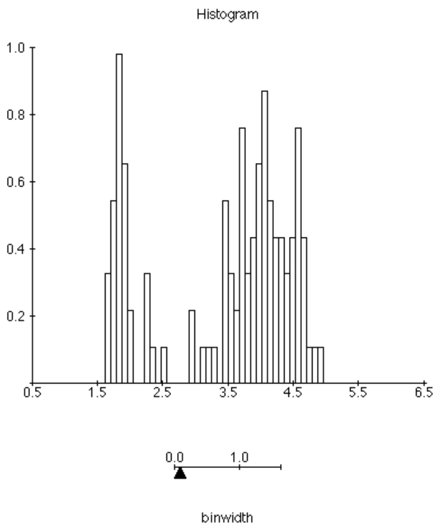
Decreasing the width of the categories for histograms

For large data sets one can get a finer description of the data, by decreasing the width of the class intervals on the histogram. The following Histograms are for the same set of data, recording the duration (in minutes) of eruptions of the Old Faithful Geyser in Yellowstone National Park. They show the histogram for the same set of data, with two different class interval lengths. The applet on the website

`Duration of Eruptions for Old Faithful`

allows you to change the width of the class intervals yourself.





Stem and Leaf Display Another graphical display presenting a compact picture of the data is given by a stem and leaf plot.

To construct a Stem and Leaf plot

- Separate each measurement into a stem and a leaf – generally the leaf consists of exactly one digit (the last one) and the stem consists of 1 or more digits.

e.g.: 734 stem = 73, leaf=4

2.345 stem = 2.34, leaf=5.

Sometimes the decimal is left out of the stem but **a note is added on how to read each value**. For the 2.345 example we would state that 234|5 should be read as 2.345.

Sometimes, **when the observed values have many digits,** it may be helpful either to round the numbers (round 2.345 to 2.35, with stem=2.3, leaf=5) or truncate (or dropping) digits (truncate 2.345 to 2.34).

- Write out the stems in order increasing vertically (from top to bottom) and draw a line to the right of the stems.
- Attach each leaf to the appropriate stem.
- Arrange the leaves in increasing order (from left to right).

Example Make a Stem and Leaf Plot for the data on the average number of hours spent studying per week given in Data Set 1.

10, 5, 15, 20, 40, 25, 22, 13, 12, 21

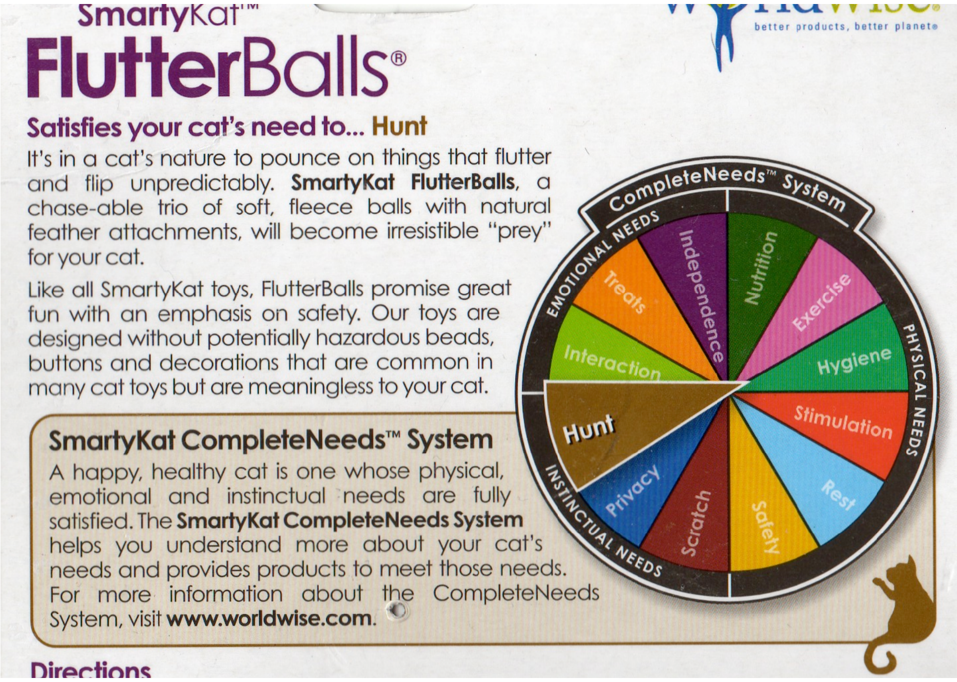
16, 22, 25, 30, 29, 25, 27, 15, 14, 17

All are data points are 2 digit integers and the tens digit goes from 0 to 4.

0	5								
1	0	2	3	4	5	5	6	7	
2	0	2	3	4	5	5	6	7	9
3	0								
4	0								

Extras : How to Lie with statistics

Example This (faux) pie chart, shows the needs of a cat, and comes from a box containing a cat toy. Note that the "categories" are not distinct and they use an exploding slice to distort the are for Hunting, which is the need of your cat that this particular toy is supposed to fulfill.



SmartyKat™

FlutterBalls®

Satisfies your cat's need to... Hunt

It's in a cat's nature to pounce on things that flutter and flip unpredictably. **SmartyKat FlutterBalls**, a chase-able trio of soft, fleece balls with natural feather attachments, will become irresistible "prey" for your cat.

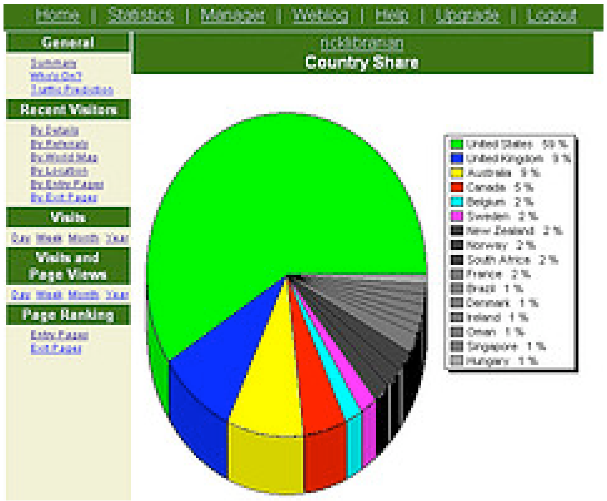
Like all SmartyKat toys, FlutterBalls promise great fun with an emphasis on safety. Our toys are designed without potentially hazardous beads, buttons and decorations that are common in many cat toys but are meaningless to your cat.

SmartyKat CompleteNeeds™ System

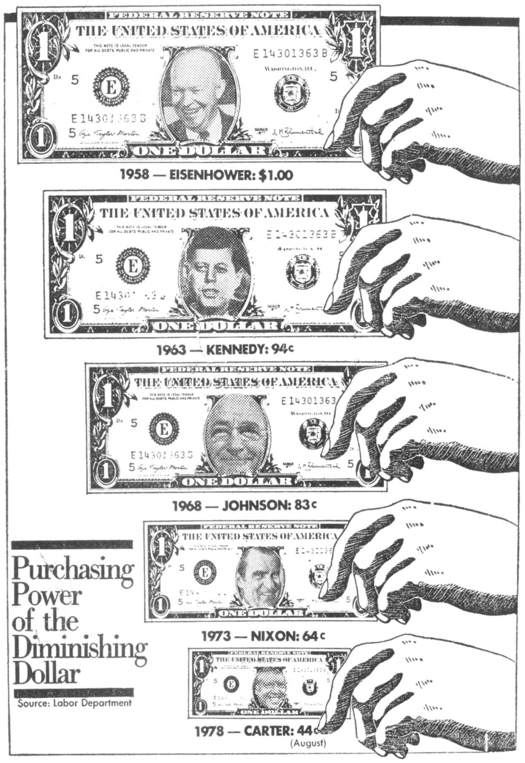
A happy, healthy cat is one whose physical, emotional and instinctual needs are fully satisfied. The **SmartyKat CompleteNeeds System** helps you understand more about your cat's needs and provides products to meet those needs. For more information about the CompleteNeeds System, visit **www.worldwise.com.**

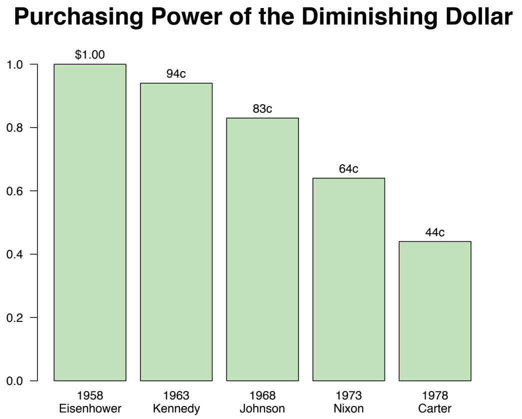
Directions

A subtle way to lie with statistics is to violate the area rule. The pie chart below is distorted to make the areas of regions devoted to some categories proportionally larger than they should be by stretching the pie into an oval shape and adding a third dimension.

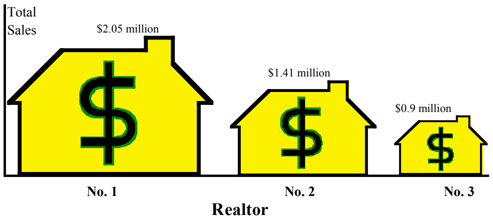


Example Both of the following graphs represent the same information. The graph on the left violates the area principle by making the base of the bars (banknotes) of unequal width.





Example All of the following graphs violate the area principle by replacing the bars by irregular objects in addition to making the bases of unequal length.



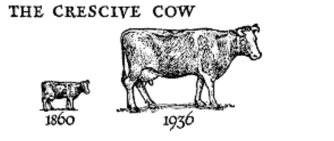
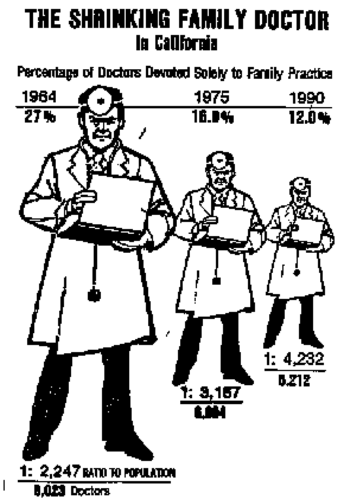


Figure 1: Graphical distortion of data
SOURCE: Darrell Huff. 1993. *How to Lie with Statistics* WW Norton & Co, 72.


THE SHRINKING FAMILY DOCTOR
In California
Percentage of Doctors Devoted Solely to Family Practice
1: 4,232
1: 3,167
1: 2,247 RATIO TO POPULATION
8,023 Doctors