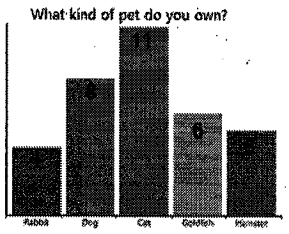
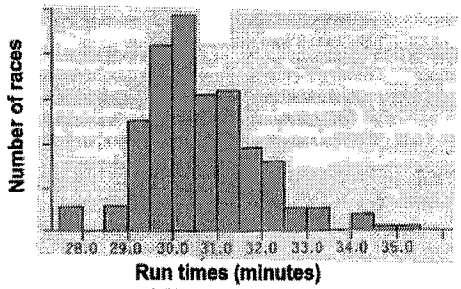
AP Statistics – Lesson Notes - Chapter 4: Displaying Quantitative Data

Graphing Frequency (count) with Numerical (Quantitative) data

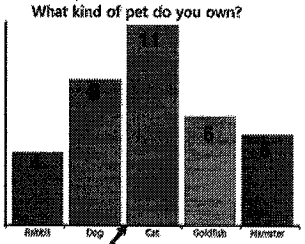
In this chapter we are (mainly) still considering the frequency (count), but instead of counts of category amounts, the variable is continuous, so if we want something like a bar chart, we must divide the possible variable values into something like categories - called 'bins':



Bar chart
(Categorical data)



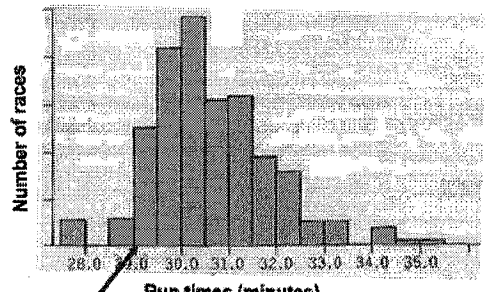
Histogram
(Quantitative data)



Bar chart

gaps between bars
implying the order is arbitrary

order doesn't matter

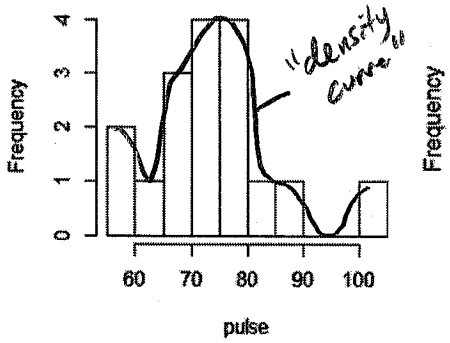


Histogram

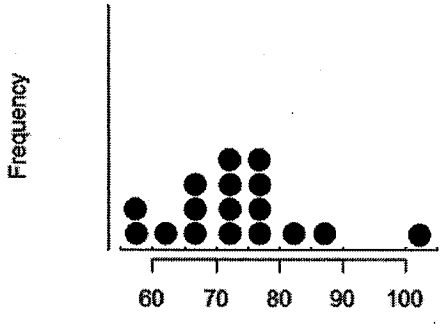
no gaps between bars because graph
accounts for all possible values in a range

order matters
(increasing, left to right)

Pulse rates of women: 56, 60, 68, 72, 76, 80, 88, 64, 68, 72, 76, 80, 68, 105, 72, 84, 72



Histogram



Dotplot

(Note: Histograms, Dotplots, and Stem-and-Leaf displays obey the area principle)

Pulse rates of women: 56, 60, 68, 72, 76, 80, 88, 64, 68, 72, 76, 80, 68, 72, 84, 72, 105

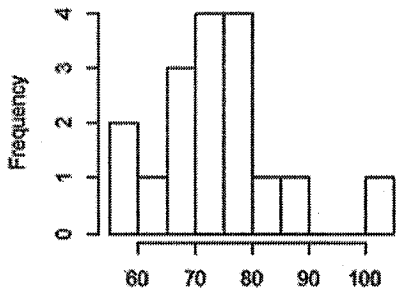
Stem-and-Leaf

10	5
9	
8	0 8 0 4
7	2 6 2 6 2 2
6	0 8 4 8 8
5	6

10	5
10	
9	
9	
8	8
8	0 0 4
7	6 6
7	2 2 2 2
6	8 8
6	0 4
5	6
5	

← "split stem"

(compare to histogram)



Stem-and-leaf diagrams show shape, outliers, and preserve the original data values. (can also be used to compare center & spread)

always include a key:

Describing Shape, Center, and Spread

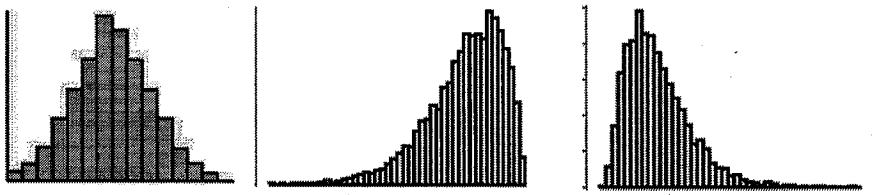
Because the data is numerical and shown in order, the shape of the distribution is meaningful and can be described.

SOCS

Shape **Outliers** **Center** **Spread**

Shape: Symmetric or skewed
Outliers: Gaps, unusual features, IQR rule of thumb
Center: Median or mean
Spread: Range, IQR, Variance, or Standard Deviation

Shape: Symmetric or skewed



Symmetrical | **Skewed left (tail toward lower values)** | **Skewed right (tail toward higher values)**

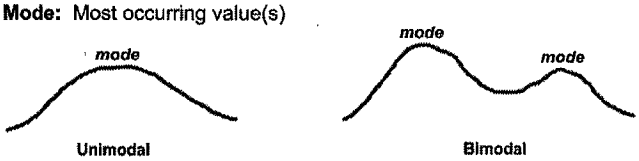
A **'statistic'** is a number that represents a data set in some way.

Center: Median or mean

Median: Middle value (or average of two middle values) position of median = $\frac{n+1}{2}$

Mean: $\bar{x} = \frac{\sum x}{n}$

Mode: Most occurring value(s)



Unimodal **Bimodal**

How mean, median, and skew are related

What is the mean and median of this data set?

6 7 8 9 10

$$\overline{X} = \frac{6+7+8+9+10}{5} = 8$$

$$Median = 8$$

If we move the '10' farther away from the mean, how does mean and median change?

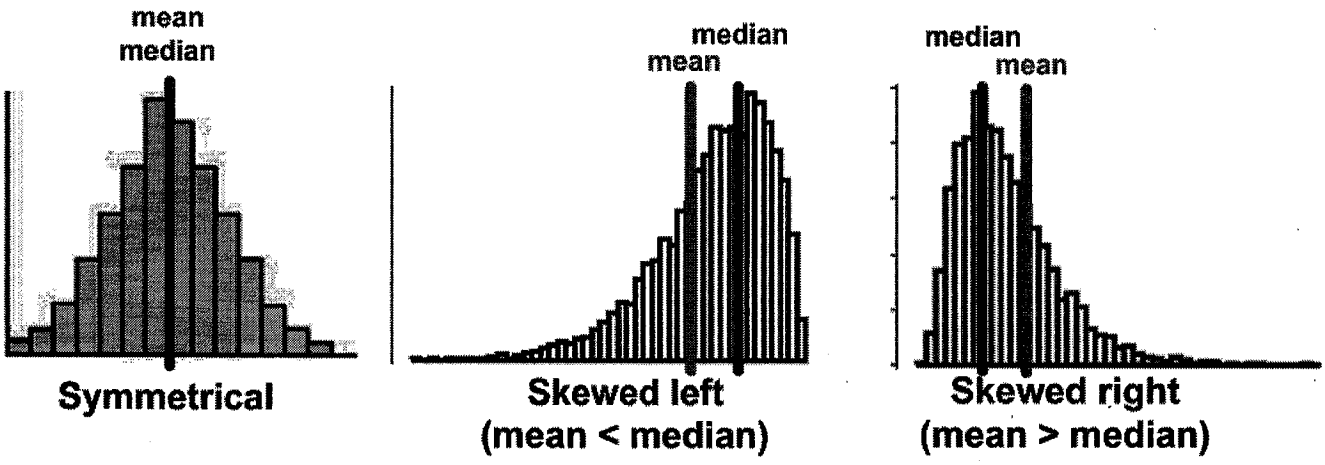
6 7 8 9 20

$$\overline{X} = \frac{6+7+8+9+20}{5} = 10$$

$$Median = 8$$

The median is unaffected, but the outlier 'pulls the mean towards itself'

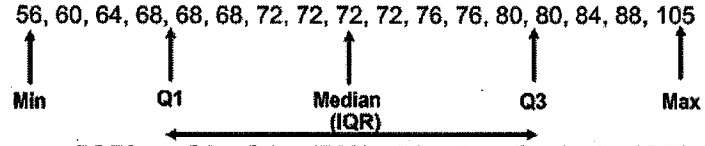
In a perfectly symmetric distribution, the mean and median are aligned, but in skewed distributions, the mean is pulled in the direction of the tail:



Spread: Range, IQR, Variance, or Standard Deviation

Range: Maximum value - Minimum value

Quartiles:



Interquartile Range (IQR): Q3 - Q1 (50% of the data lies in the IQR)

Variance: $s^2 = \frac{\sum\left(x-\overline{x}\right)^2}{n-1}$

Standard Deviation: $s = \sqrt{s^2} = \sqrt{\frac{\sum\left(x-\overline{x}\right)^2}{n-1}}$

Conceptual meaning of variance and standard deviation

women's pulse (beats per minute)

56, 60, 64, 68, 68, 68, 72, 72, 72, 72, 76, 76, 80, 80, 84, 88, 105

How far away from mean is this value?

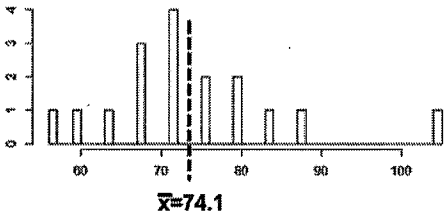
68 - 74.1 = -6.1 bpm

What is the 'typical' or 'average' value of the distances away from the mean?

If we just added the distances for each, some would be positive, some negative, and they would cancel each other out. So we square the difference to make them all positive before we add. Then we divide by the number of distances^2 to get average distance^2.

This is **variance:**

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$



$\bar{x}$=74.1

But we really want distances (spread) in terms of the original units (bpm, not bpm^2) so we take the square root of the result.

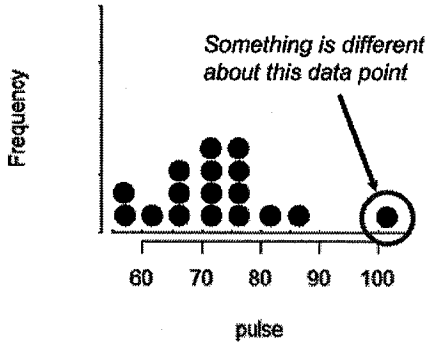
This the **standard deviation:**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

divide by n-1 because we are using the mean in the calculation. (more complete explanation posted on www.mrfelling.com)

Standard deviation is 'typical' or 'average' distance of all data points from the mean.

Outliers: Gaps, unusual features, IQR rule of thumb



Outlier rule-of-thumb

A data point is an outlier if its value is:

$$< Q1 - 1.5IQR$$

or

$$> Q3 + 1.5IQR$$

A.K.A. the "upper and lower fences"

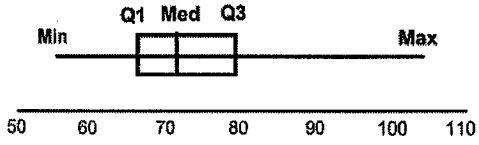
Summarizing information about a dataset

Pulse rates of women: 56, 60, 68, 72, 76, 80, 88, 64, 68, 72, 76, 80, 68, 72, 84, 72, 105

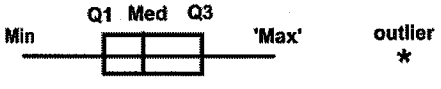
5 number summary

Min: 56
Q1: 68
Median: 72
Q3: 80
Max: 105

Box & Whisker Plot



Sometimes, outliers are excluded and plotted separately:



Is 105 an outlier?

$$IQR = Q3 - Q1 = 80 - 68 = 12$$

$$upper\ fence = Q3 + 1.5IQR = 80 + 1.5(12) = 98$$

105 > 98, so, yes, 105 is an outlier.

Calculator statistics and plots

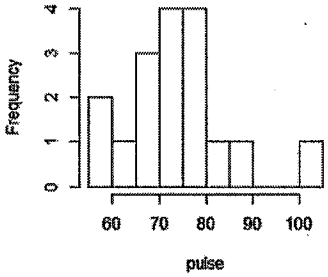
If you have complete data set: Use 1 list and 1-Var-Stats

Pulse rates of women: 56, 60, 68, 72, 76, 80, 88, 64, 68, 72, 76, 80, 68, 72, 84, 72, 105

- Enter data in L1 (Stats, Edit)
- Stats, Calc, 1-Var Stats

If you have a histogram: Use 2 lists and 1-Var-Stats

- Enter representative bar values data values in L1 (Stats, Edit)
- Enter data counts (height of bars) in L2
- Stats, Calc, 1-Var Stats w/FreqList set to L2



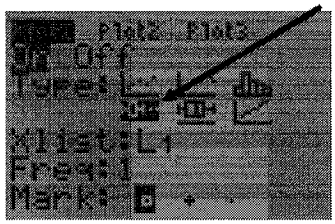
To display a histogram:

- Enter data in L1 (Stats, Edit) or L1 and frequency in L2
- "Y=" and clear out any equations
- 2nd "Y=" to enter set up for Statistics Plot 1.
- Set to match the screen on the right:
- Zoom: 9 (ZoomStat)

'WINDOW' change Xscl 'GRAPH' to change bin width

To display a boxplot:

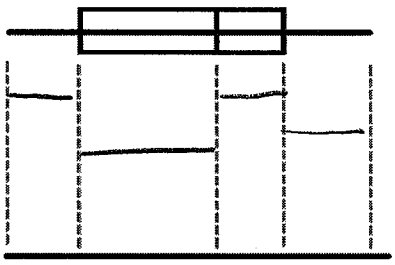
- Enter data in L1 (Stats, Edit) or L1 and frequency in L2
- "Y=" and clear out any equations
- 2nd "Y=" to enter set up for Statistics Plot 1.
- Set to match the screen on the right:
- Zoom: 9 (ZoomStat)

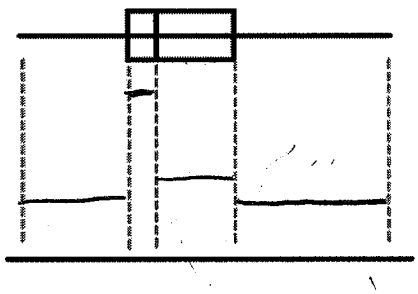


always use the one with the dots so you'll see outliers

Boxplot to Histogram (approximate) - the 'Aquarium'

The distribution is broken into quarters at the Q1, Median, and Q3 marks and each quarter contains about 25% of the 'water' (distribution):





'Resistant' Statistics

Statistics are ***resistant*** if they are not overly affected by the presence of outliers.

Measures of Center	Measures of Spread
Median: resistant	IQR: resistant
Mean: not resistant	Variance: not resistant
	Std deviation: not resistant

Conclusions: If the plot (box, histogram, stem & leaf, dot) of the distribution shows that it is...

...symmetric:

- Can use mean or median to measure center.
- Can use standard deviation or IQR to measure spread.

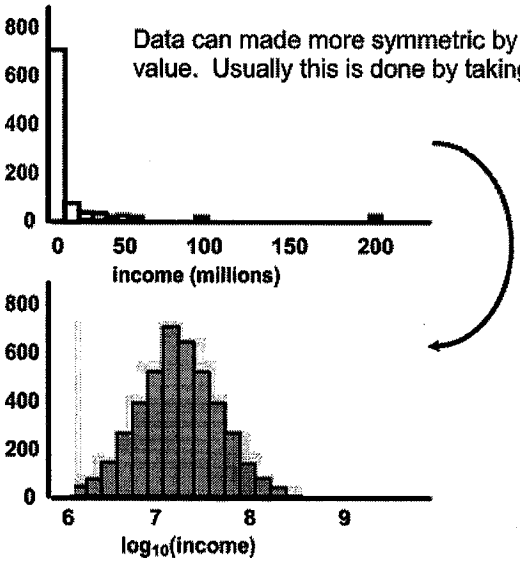
...skewed:

- Must use median to measure center.
- Must use IQR to measure spread.

Re-expressing Skewed Data (Transforms)

Sometimes, data is so skewed that it is difficult to conclude things even with a picture:

Example: CEOs asked, 'what is your annual income?'

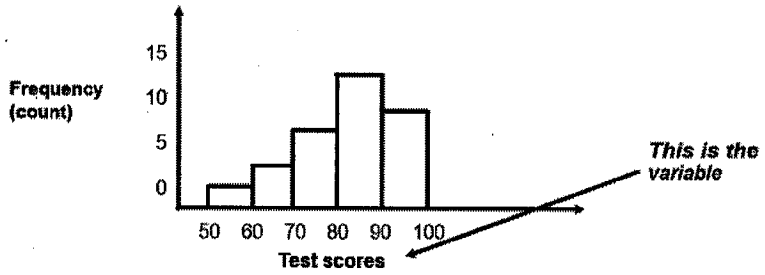


Data can made more symmetric by re-expressing, or transforming each data value. Usually this is done by taking a logarithm (base 10) or square root.

original income value	transformed income value $\log_{10}$
1,500,000	6.2
15,000,000	7.2
150,000,000	8.2

This chapter is mostly about *single variable* data

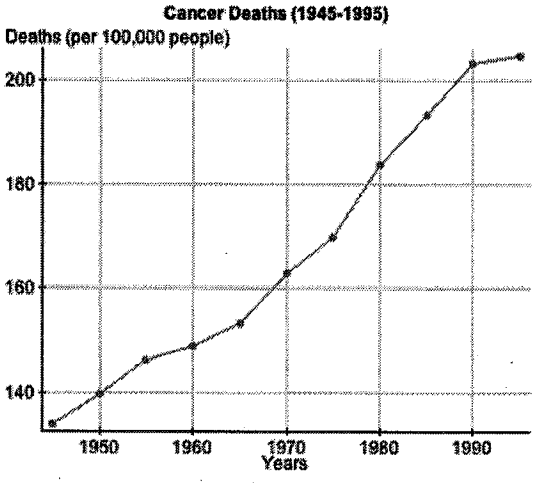
For most of the data in this chapter, there is only **one** variable. If we are looking at a histogram, the values of the variable are shown on the horizontal axis and the vertical axis isn't another variable...it is the frequency or count of the number of data with a particular value:



Timeplots

This chapter is mostly about displaying information about the frequency (counts) for various 'bin' values of a variable. The book mentions one other type of plot called a 'Timeplot' which is fundamentally different (more like later chapters).

A ***timeplot*** is simply a 2-D plot of actual data values (not counts) vs time.



'Variable' vs. 'Count'