Math 203
Topics for the Chapter 8 quiz:
Data and patterns

To do:

Build graphical representations of data sets, using stem-and-leaf plots, histograms, bar charts, line graphs, and pie charts.

Build graphical representations of more than one data set, for side-by-side comparison, using double stem-and-leaf plots, comparison histograms, multiple bar charts, multiple line graphs, and multiple pie charts.

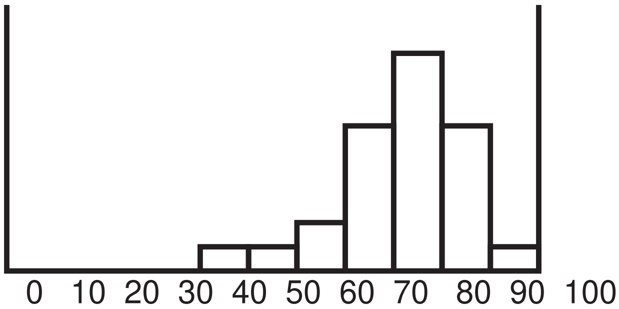
Describe patterns in data revealed by these graphical representations: clutering, gaps, outliers, trends.

Describe some of the ways that a visual presentation of data can lead to misleading or distorted conclusions.

Data sets, collections of numbers that are the result of an experiment, a survey, a historical record, or one of any number of other sources, contain a wealth of information; how can we find patterns in the data? One common approach is to graphically represent the data.

Stem-and leaf plot: list the data in increasing order (for convenience). List the data between consecutive multiples of 10 on successive lines, listing their last digits only. Example: some exam scores.

10	0
9	2,2,3,6,7,8
8	0,3,4,6,6,7,8,9,9
7	1,4,7,7,7,8
6	5,6
5	7
4	1

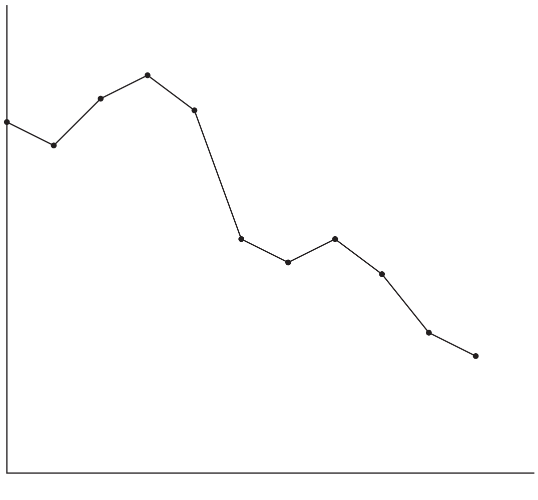


Histogram: group data points together according to size in a collection of "bins" (or "measurement classes"). Plot only the number of points in each bin, as a rectangle ("mast") with height given by the number in each bin.

Stem-and-leaf is sort of a crude histogram, although it has the advantage that the entire data set can be recovered from it. Both allow us to get a sense of the distribution of the data, to locate clusters (tall masts close together), gaps, and outliers (data points lying away from the main cluster(s)). The bin size of the histogram can be changed to reveal finer detail (small bins).

Bar graphs are like histograms, but the bins can be anything that distinguishes points in the data set: male/female, level of education, etc. The heights of the bars represent the *frequency* of the events, how many times the event occured in the data set. The information of the chart can be expressed as a *frequency table*, listing events and frequencies side-by-side. The advantage of the chart is that it allows us to use our innate ability to estimate relative sizes to visualize the differences among frequencies. "Outliers" are represented by the shortest bars.

Line graphs: if the data represents a quantity that changes over time (population, annual income, etc.), then graphing the data as points of the form (date,quantity) and connecting consecutive dots by straight ine segments constructs a line graph for the data. The slopes of the lines visually indicate when the quantity has gone up or down over time. This allows us to visually spot *trends*, sequences of line segments all (or mostly all) either sloping upward or sloping downward.



Pie charts are like bar graphs in that they visually represent a collection of events; each event is alloted a sector of a full circle, the angle given to an event = (360)(the fraction of the total that the event represents) degrees. Essentially, we see each event as a fraction of the whole. This is useful when the actual number of occurances is less important, and we want to see which event account for most of the total. Outliers are represented by the narrowest slices.

Charting for comparison:

If we have several data sets (representing, perhaps, surveys taken on different dates), or we wish to look at how the distribution of a data set differs on certain sub-populations of the population that the data set was built on (male/female, different age groups), comparative charts can help.

A double stem-and-leaf plot is essentially the combination of stem-and-leaf plots for two data sets (quiz grades for two classes?); we build the leaves for one plot to the right of the stem (100,90,80, etc.) and the other to the left of the stem. This is a form (still containing everything from the two data sets) of *comarison histogram*; we use the same bins on two or more data sets, and place the masts for the different data sets counting the numbers in each bin side-by-side. This allows for direct side-by-side comparison of the clusters, gaps, and outliers of the several data sets at once, allowing us to visually spot the differences (or similarities) across data sets. One potential difficulty is that different sets might reveal the most useful information using different bin sizes, but for a comparison histogram the same bins must be used for all data sets.

If the data is time dependent (enrollment totals for several schools by year), a multiple line graph, building line graphs for each data set, using the same axes, can allow us to find correlations among the trends contained in each data set. (E.g., population trends in predator species usually follow, but lag behind, the population changes of their prey.)

Multiple bar charts, which like histograms place the frequency in a bin for each data set side-by side (grouping by bins), allow us to make comparisons across data sets; we can see that some populations land in certain bins more often than other populations. A multiple pie chart is essentially a pie chart for each data set; again, the bins used must be the same across all data sets. By comparing pie charts we can spot differences in relative frequencies among the various populations.

Sources of distortion, misunderstanding, misuse:

Visually representing data can be extremely useful in detecting patterns in data, letting us understand at a glance what the basic distribution of the data is, and spotting trends. But it is possible to **mis**represent data, and to distort the underlying patterns (intentionally or unintentionally), when representing data graphically.

Perhaps the most common way to alter someone's impression of data is to change the vertical scale used. Displying only a smaller range of values enhances the differences between data points. This can be either good or bad; a difference that is very small, and therefore insignificant, can be made to appear quite significant. Worse, the vertical scale can be omitted, rendering the comparison all but useless. Scales needn't start at 0 to be honest; some data (SAT scores, credit ratings) only take values in a restrictive range.

Sometimes data is presented using no scale at all; the bars get higher/lower, but not in proportion to the values indicated. The visual mipact is of a trend; the varying rates of the trend are lost. Trends can be enhanced by using 2- or 3-dimensional objects for their representation; doubling the side length of a cube increases its volume by a factor of 8. The eye sees the much increased volume and interprets the rise as being much larger than it is.

We tend to assume that the bin sizes in a histogram are all the same; manipulating this can make clustering appear where it might not really be. The same varying of the horizontal scale in line graph can enhance or downplay trends in the data.

3-dimensional effects can obscure the message in data. Putting the bars of a bar chart on a uniformly sized "pedestal" diminishes the differences between them. Adding shadows can have the same effect. Putting larger objects in the background makes them appear smaller; putting them in the forground makes them look disproportionally larger.

Using maps to represent data by state or country can be misleading; if the data represents information about the population, the fact that population is not proportional to size can lead to distortion.

And this is just the tip of the iceberg...