3 Graphical Displays of Data

Reading: SW Chapter 2, Sections 1-6

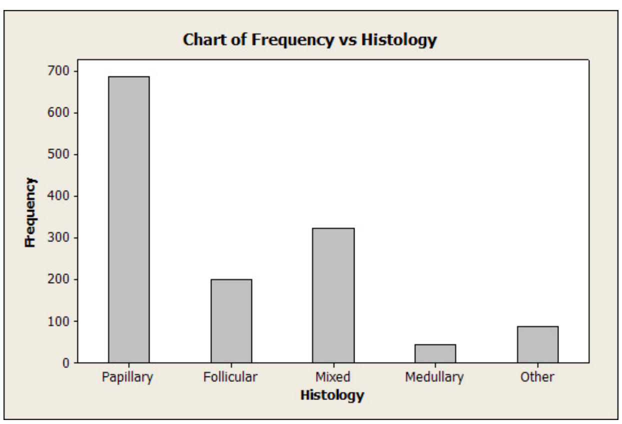
Summarizing and Displaying Qualitative Data

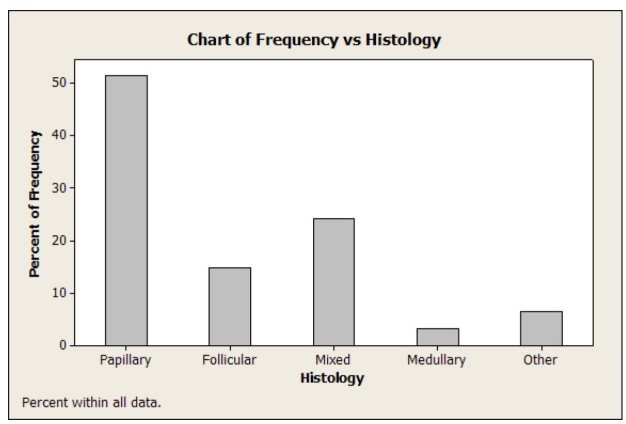
The data below are from a study of thyroid cancer, using NMTR data. The investigators looked at all thyroid cancer cases diagnosed among NM residents between 1/1/69 and 12/31/91. A small percentage of cases were omitted (those that weren't first primary; those without more than 60 days of follow-up without another diagnosis of cancer), leaving 1338 cases of thyroid cancer.

A **frequency distribution** for a categorical variable gives the counts or frequency with which the values occur in the various categories. The frequency distribution for histologic type is given below. The **relative** frequency distribution gives the proportion (i.e number of cases divided by sample size) or percentage (proportion times 100%) of cases in each histologic category.

Histology	Frequency	Relative Frequency	Percentage
Papillary	687	$687/1338 = 0.51$	51%
Follicular	199	$199/1338 = 0.15$	15%
Mixed	323	$323/1338 = 0.24$	24%
Medullary	43	$43/1338 = 0.03$	3%
Other	86	$86/1338 = 0.06$	6%
Total	1338	0.99(1.00)	99% (100%)

The frequency distribution is usually summarized graphically via a **bar graph**, sometimes called a **bar chart**. The next page give frequency and relative frequency distributions generated by **Minitab**. Erik will show you how to do this in LAB.





The information conveyed is the same in both graphs. The graph of percentages has real advantages when comparing two groups with much different sample sizes, however.

Example: SW pages 12, 14 - colors of Poinsettia.

Graphical Summaries of Numerical Data

There are four (actually, there a many more) graphical summaries of primary interest: the **histogram**, the **dotplot**, the **stem and leaf** display, and the **boxplot**. Each of these is easy to generate in **Minitab**. Our goal with a graphical summary is to see patterns in the data. We want to see what values are typical, how spread out are the values, where do the values tend to cluster, and what (if any) big deviations from the overall patterns are present. Sometimes one summary is better than another for a particular data set.

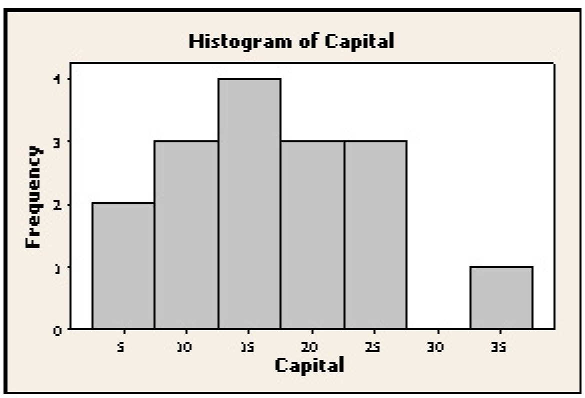
Histogram

The **histogram** breaks the range of data into several equal width intervals, and counts the number (or proportion, or percentage) of observations in each interval. The histogram can be viewed as a **grouped frequency distribution** for a continuous variable. Here is the "help" entry from Minitab describing histograms:

Histograms

Graph > Histogram

Use to examine the shape and spread of sample data. Histograms divide sample values into many intervals called bins. Bars represent the number of observations falling within each bin (its frequency). In the histogram below, for example, there are two observations with values between 2.5 and 7.5, three observations with values between 7.5 and 12.5, and so on.



Observations that fall exactly on an interval boundary are included in the interval to the right (or left, if the last bin).

Why is it reasonable to group measurements whereas with categorical data we computed the number of observations with each distinct data value?

Most texts, including SW, discuss the choice of intervals. We will use **Minitab** for our calculations, which usually does quite a good job of choosing the intervals for us. We already saw histograms of MAO levels in the previous section.

The real strength of histograms is showing where data values tend to cluster. Their real weakness is that the choice of intervals (bins) can be arbitrary, and the apparent clustering can depend considerably on the choice of bins. Histograms work pretty well with larger data sets, where the choice of bins usually has little effect; for smaller data sets, dotplots or stem and leaf displays usually are a much better choice.

Dotplot

Where histograms try to condense the data into relatively few bins, dotplots present a similar picture but emphasize the distinct values. Dotplots are particularly good at comparing different data sets, especially smaller data sets. One big advantage is that you usually see all the data, so no information is lost in the dotplot. The biggest disadvantage is that it gets pretty "noisy" for large data sets.

Here is the "help" entry from Minitab describing dotplots:

Dotplots

Graph > Dotplot

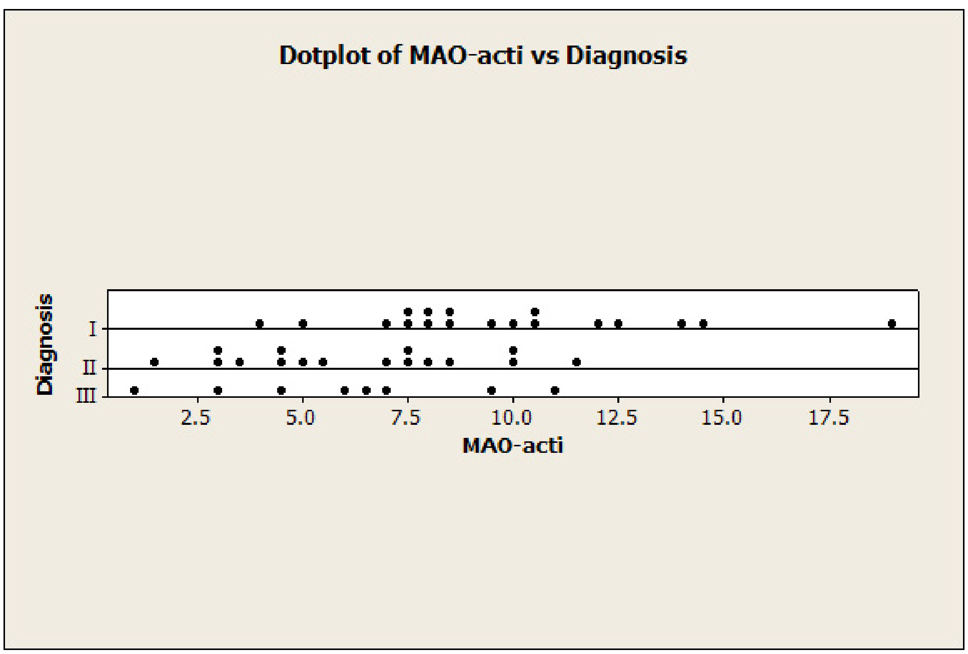
Use to assess and compare distributions by plotting the values along a number line. Dotplots are especially useful for comparing distributions.

The x-axis for a dotplot is divided into many small intervals, or bins. Data values falling within each bin are represented by dots.

If possible, Minitab displays a dot for each observation. Otherwise, a dot represents multiple observations with a footnote indicating the maximum number of observations represented by each dot.

Note Dotplots can only be brushed when dots represent individual observations.

Earlier we looked at histograms of MAO activity levels for schizophrenic patients of three different diagnoses. The dotplots for the three data sets make comparisons quite easy. Isn't it a lot easier to see the nature of differences here than using the three histograms in the previous section?



Stem and Leaf Display

A **stem and leaf** display defines intervals for a grouped frequency distribution using the base 10 number system. Intervals are generated by selecting an appropriate number of lead digits for the data values to be the stem. The remaining digits comprise the leaf. Following is Minitab's "help" entry for the Stem and Leaf:

Stem-and-Leaf

Graph > Stem-and-Leaf
Stat > EDA > Stem-and-Leaf
***Character Graphs* > Stem-and-Leaf**

Use to examine the shape and spread of sample data. Minitab displays a stem-and-leaf plot in the Session window. The plot is similar to a histogram on its side, however, instead of bars, digits from the actual data values indicate the frequency of each bin (row).

Below is a stem-and-leaf plot for a data set with the following five values: 3, 4, 8, 8, and 10.

```
Stem-and-leaf of C1  N  = 5
Leaf Unit = 1.0

 1   0  3
 2   0  4
 2   0
(2)  0  88
 1   1  0
```

The display has three columns:

- The leaves (right) – Each value in the leaf column represents a digit from one observation. The "leaf unit" (declared above the plot) specifies which digit is used. In the example, the leaf unit is 1.0. Thus, the leaf value for an observation of 8 is 8 while the leaf value for an observation of 10 is 0.
- The stem (middle) – The stem value represents the digit immediately to the left of the leaf digit. In the example, the stem value of 0 indicates that the leaves in that row are from observations with values greater than or equal to zero, but less than 10. The stem value of 1 indicates observations greater than or equal to 10, but less than 20.
- Counts (left) – If the median value for the sample is included in a row, the count for that row is enclosed in parentheses. The values for rows above and below the median are cumulative. The count for a row above the median represents the total count for that row and the rows above it. The value for a row below the median represents the total count for that row and the rows below it.

  In the example, the median for the sample is 8, so the count for the fourth row is enclosed in parentheses. The count for the second row represents the total number of observations in the first two rows.

Look carefully at the display – how would the example above change if the numbers were 30, 40, 80, 80 and 100 instead of 3, 4, 8, 8, and 10. Try it and confirm the display looks the same with one important difference. Following is the stem and leaf for the MAO activity levels of Diagnosis I patients.

```
Stem-and-Leaf Display: MAO-acti

Stem-and-leaf of MAO-acti  group = 1    N  = 18
Leaf Unit = 1.0


  2   0  45
  7   0  67777
 (4)  0  8899
  7   1  001
  4   1  2
  3   1  44
  1   1
  1   1  8
```

Let's examine this display, and make sure we can pick out what the actual numbers are. Look at the original values (from SW). Is Minitab rounding numbers or just truncating excess digits? SW would have you put larger numbers on top. That would seem more conventional, except stem and leaf displays almost always are done Minitab's way with the larger numbers on the bottom. There is a good reason for this – if you turn the graph 90 degrees counterclockwise, you end up with a regular histogram (what are the bins?)

The stem and leaf was invaluable for "paper and pencil" data analysis. It is very quick to do by hand, and it has the advantage of keeping the original data right on the display. It also sorts the data (puts them in order), which allows quick calculation of medians and quartiles. I find the dotplot a better tool, often, when summarizing small to moderate-sized data sets on the computer. The stem and leaf is harder to use for comparing several groups, but still is more common in practice than dotplots.

Erik will show you how to generate stem and leaf displays in **Minitab**, and a few of the options.

Example

Two stem and leaf displays for a data set on age at death for SIDS cases in Washington state are given below. The first is for the data recorded in days, the second for the data recorded in weeks. Note that the maximum value is 307 days, or 43.9 weeks.

```
Stem-and-Leaf Display: SIDS days

Stem-and-leaf of SIDS days  N  = 78
Leaf Unit = 10


  9    0  222222333
 18    0  444444555
 31    0  6666666667777
(16)   0  8888888889999999
 31    1  00000111111
 20    1  22333
 15    1  4455
 11    1  6777
  7    1  88
  5    2  0
  4    2  23
  2    2
  2    2  7
  1    2
  1    3  0

Stem-and-Leaf Display: SIDS weeks

Stem-and-leaf of SIDS weeks  N  = 78
Leaf Unit = 1.0


  8    0  33334444
 27    0  5666666778889999999
(22)   1  0111111112222223333444
 29    1  55556666677889
 15    2  0112344
  8    2  5669
  4    3  23
  2    3  9
  1    4  3
```

The structure of the two stem and leaf displays is slightly different. In particular, the days display corresponds to a histogram with intervals of width 20 (confirm this!). The weeks display corresponds to a histogram with intervals of width 5 (confirm!). Minitab does give you some control over interval widths, but usually makes the right choice by default.

Boxplots

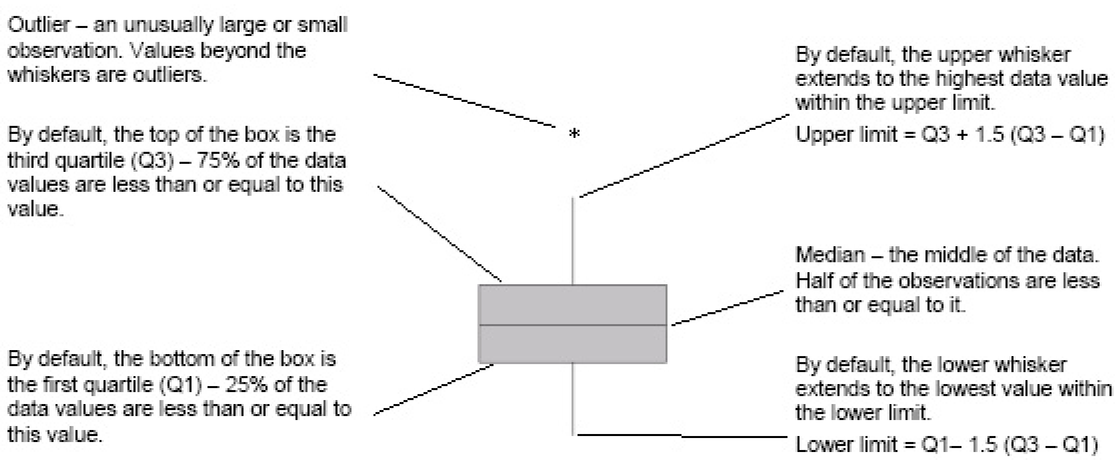
Boxplots have become probably the most useful of all the graphical displays of numerical data. I can go weeks without computing histograms, dotplots, or stem and leaf displays, but I usually compute several boxplots per week. They succinctly summarize central location (average), spread and shape of the data, and highlight outliers while permitting simple comparison of many data sets at once. Following is the Minitab "help" description of boxplots.

Boxplots

Graph > Boxplot

Stat > EDA > Boxplot

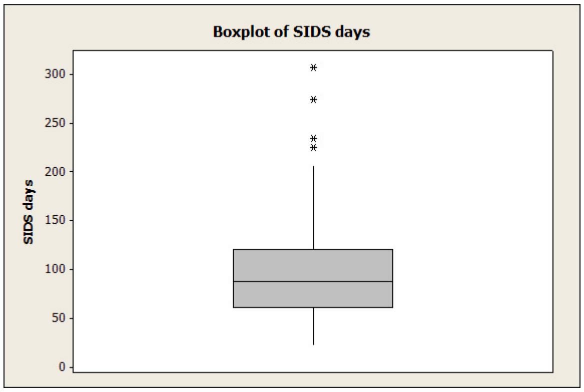
Use boxplots (also called box-and-whisker plots) to assess and compare sample distributions. The figure below illustrates the components of a default boxplot.

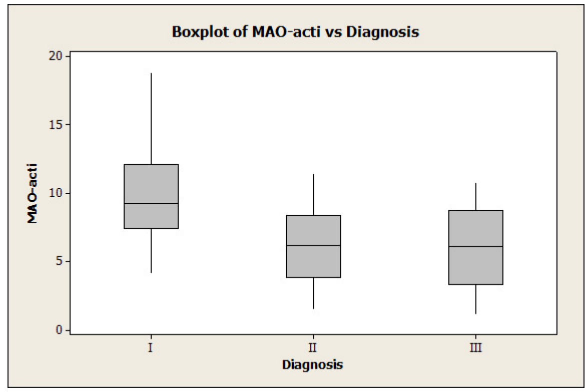


Note By default, Minitab uses the quartile method for calculating box endpoints. To change the method for a specific graph to hinge or percentile, use Editor > Edit Interquartile Range Box > Options. To change the method for all future boxplots, use Tools > Options > Individual Graphs > Boxplots.

Lots of elementary texts make the boxplots simpler by connecting the whiskers to the extremes of the data; this keeps them from highlighting outliers and, in my opinion, erases substantial utility of the boxplot. Minitab will allow you to compute those neutered boxplots, but you should not. The box part of the boxplot is Q_1, M, and Q_3, a range containing half the data. The whiskers connect the box to the extremes of "normal" looking data, and anything more extreme is plotted separately (and importantly) as an outlier. Relative distance of the quartiles from the median, and relative length of the whiskers tells us a lot about the shape of the data (we will explore that below). Several packages, including Minitab, allow you to clutter the boxplot with a lot of other features, but I usually prefer not to.

Boxplots of the SIDS and MAO data sets are below. Let's pick out important features.





Interpretation of Graphical Displays for Numerical Data

In many studies, the data are viewed as a subset or **sample** from a larger collection of observations or individuals under study, called the **population**. A primary goal of many statistical analyses is to generalize the information in the sample to **infer** something about the population. For this generalization to be possible, the sample must reflect the basic patterns of the population. There are several ways to collect data to ensure that the sample reflects the basic properties of the population, but the simplest approach, by far, is to take a random or "representative" sample from the population. A **random sample** has the property that every possible sample of a given size has the same chance of being the sample eventually selected. Random sampling eliminates any systematic biases associated with the selected observations, so the information in the sample should accurately reflect features of the population. The process of sampling introduces random variation or random errors associated with summaries. Statistical tools are used to calibrate the size of the errors.

Whether we are looking at a histogram (or stem and leaf, or dotplot) from a sample, or are conceptualizing the histogram generated by the population data, we can imagine approximating the "envelope" around the display with a smooth curve. The smooth curve that approximates the population histogram is called the **population frequency curve**. Statistical methods for inference about a population usually make assumptions about the shape of the population frequency curve. A common assumption is that the population has a normal frequency curve. In practice, the observed data are used to assess the reasonableness of this assumption. In particular, a sample display should resemble a population display, provided the collected data are a random or representative sample from the population. Several common shapes for frequency distributions are given below, along with the statistical terms used to describe them.

The first display is **unimodal** (one peak), **symmetric** and **bell-shaped**. This is the prototypical normal curve. The boxplot (laid on its side for this display) shows strong evidence of symmetry: the median is about halfway between the first and third quartiles, and the tail lengths are roughly equal. The boxplot is calibrated in such a way that 7 of every 1000 observations are outliers (more than $1.5(Q_3 - Q_1)$ from the quartiles) in samples from a population with a normal frequency curve. Only 2 out of every 1 million observations are extreme outliers (more than $3(Q_3 - Q_1)$ from the quartiles). We do not have any outliers here out of 250 observations, but we certainly could have