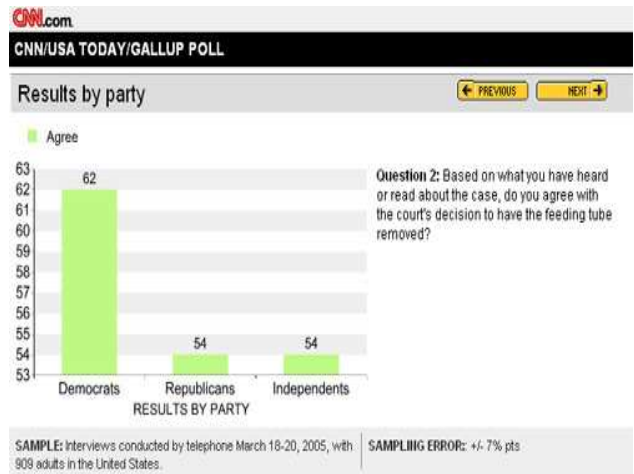


Things you'll know (or know better to watch out for!) when you leave in December:

1. What you can and cannot infer from graphs.
2. How to construct (in your head!) and interpret confidence intervals.
3. How to conduct tests on population parameters within a population and between/across populations. These parameters include means, variances, odds ratios, and others.
4. How and when to carry out a linear regression analysis and how to interpret the results.

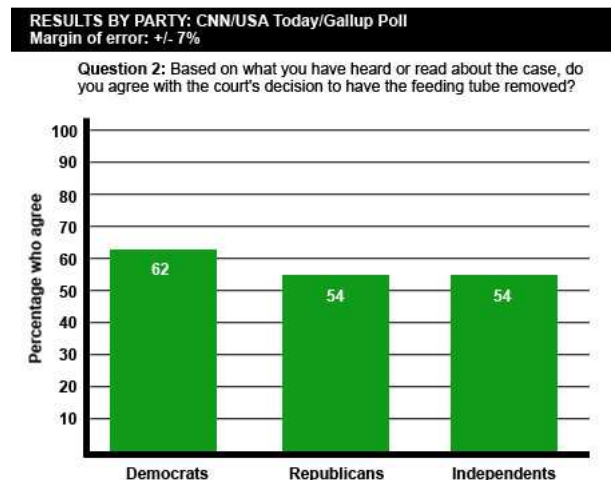


mediamatters.org claimed that this was a misleading graph:

"In presenting the results of a CNN/USA Today/Gallup poll, CNN.com used a visually distorted graph that falsely conveyed the impression that Democrats far outnumber Republicans and Independents in thinking the Florida state court was right to order Terri Schiavo's feeding tube removed.

According to the poll, when asked if they "agree[d] with the court's decision to have the feeding tube removed," 62 percent of Democratic respondents agreed, compared to 54 percent of Republicans, and 54 percent of Independents. But these results were displayed along a very narrow scale of 10 percentage points, and thus appeared to show a large gap between Democrats and Republicans/Independents:

Presented in this manner, the graph suggests that the gap between the two groups is overwhelming, rather than only 8 percentage points, within the poll's margin of error of +/- 7 percentage points. "



When constructing graphs, do the following, and hope (or maybe insist) that others do the same !

1. Clearly label x and y axes.
2. Use relevant scales.
3. Indicate exactly which subset of the data are being represented and how. Were continuous measurements changed to binary or integer for plotting purposes ? Were other transformations done ?
4. BE VERY CAUTIOUS about drawing general conclusions from graphs that do not involve error bars (or some way of representing error). More on this later in the semester.

Responses to questions in the Breast Cancer Consortium Questionnaire fall into these categories. Examples are below.

Nominal Data: Questions 1, 2, 5, 16...

Ordinal Data: Question 4, 9, 17, (disregard never and not sure)...

Discrete Data: How many sisters have been diagnosed with breast cancer ? (Question 7b rephrased)

Continuous Data: Questions 14 (weight could be measured in arbitrarily small units - can take on any value in some range - "discreteness" of measurements only limited by the measuring device)

Data can be classified into a few basic types.

Nominal Data: Numeric values that represent classes or categories - the categories are not ordered. Magnitude of numerical value is not important.

Ordinal Data: Numeric values that represent classes or categories - the categories are ordered. Magnitude of numerical value is not important.

Ranked Data*: Numeric values that represent the order of ranked observations. By assigning ranks, information about the magnitude of the values and their differences is lost; however, ranks still retain much useful information. *Convention is to rank from lowest to highest and then assign numeric values to each ranked observation starting with the lowest. Pagano and Gavreau (page 10) have this switched around.

Discrete Data: Numbers that represent measurable quantities (as opposed to just labels). Magnitude is important. Discrete data takes on specified values that differ by fixed amounts; intermediate values are not possible.

Continuous Data: Numbers that represent measurable quantities taking on any value in some range. Again, magnitude is important. The difference between any two values can be arbitrarily small.

Absolute and Relative Frequencies of Mammograms for 5,447,140 mammograms recorded in the Breast Cancer Surveillance Consortium (BCSC) study from 1996 - 2004 (inclusive).

(<http://breastscreening.cancer.gov/statistics>)

Race	Number of Mammograms	RF (%)
White	3,785,762	69.5
Black	283,251	5.2
Hispanic	397,641	7.3
Asian	266,910	4.9
American Indian	59,919	1.1
Other	653,657	12.0

For nominal (shown here) and ordinal data, a *frequency distribution* consists of the set of classes or categories along with the numerical counts in each. A *relative frequency distribution* shows the proportion of counts that fall into each class or category. A relative frequency (RF) value for any category is obtained by dividing the number of observations in that category by the total number of observations. This can be reported as a percentage (as shown) by multiplying the resulting fraction by 100.

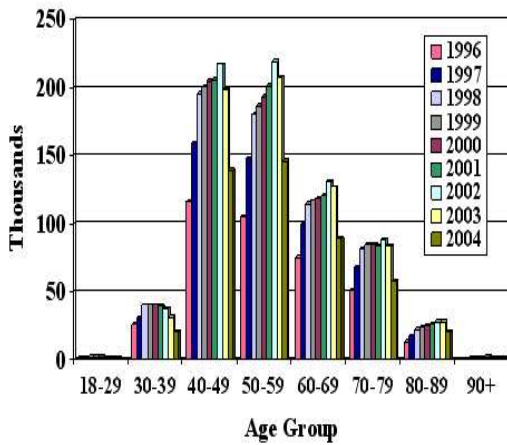
This table is a bit misleading since we can't tell which populations (if any) are under (or over) represented. This data does not consider the proportions of each race in the general population.

Number of mammograms taken in 1999 grouped by patient's age

Age	Number of Mammograms	RF (%)	CRF (%)
18-29	2,000	0.4	0.4
30-39	27,000	5.7	6.1
40-49	141,000	29.6	35.7
50-59	135,000	28.3	64.0
60-69	87,000	18.2	82.2
70-79	65,000	13.6	95.8
80-89	19,000	4.0	99.8
90-over	1,000	0.2	100.0

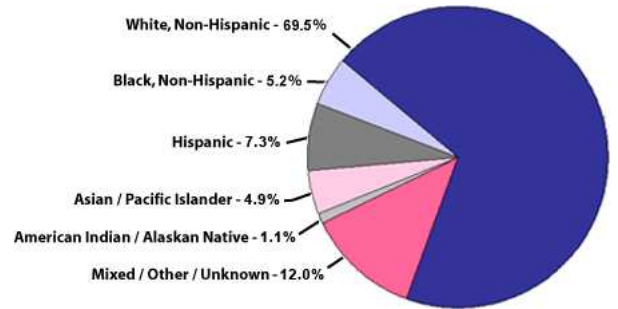
For discrete or continuous data, we must break down the range of values into a series of distinct non-overlapping intervals. If there are too many intervals, not much of a summary is obtained; if there are too few, information can be lost. Although it is not necessary (and is not done in this BCSC study), intervals are often constructed so that they have equal widths. The *cumulative relative frequency* (CRF) for an interval is the proportion of the total number of observations that have a value less than or equal to the upper limit of the interval. This too can be expressed as a percentage. In the table above, we see that 35.7% of mammograms are performed on women at or under the age of 49.

Bar chart showing number of mammograms in each age group for years 1996 - 2004.



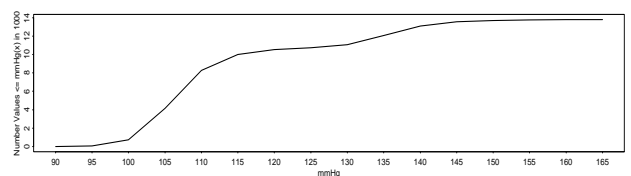
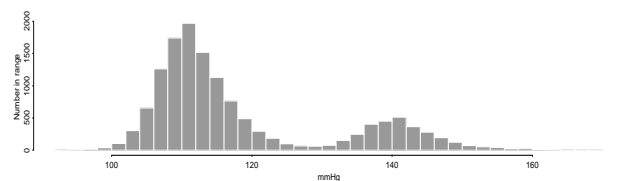
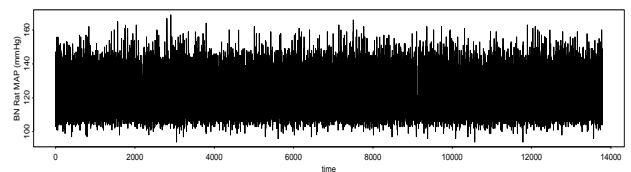
Graphical Summaries

Percentage of Mammograms by Race and Ethnicity. This pie chart shows the racial distribution of 5,447,140 mammograms recorded by the Breast Cancer Surveillance Consortium from 1996 - 2004 (inclusive). Again, we can't tell which populations (if any) are under (or over) represented. This data does not consider the proportions of each race in the general population.



Rat Mean Arterial Pressure

Histograms are used to display a frequency distribution for discrete or continuous data. If relative frequencies (proportions) are displayed, the histogram is often called a *probability histogram*. In this case, the heights sum to one. Again, note that we must break down the range of values into a series of distinct non-overlapping intervals. If there are too many intervals, not much of a summary is obtained; if there are too few, information can be lost.



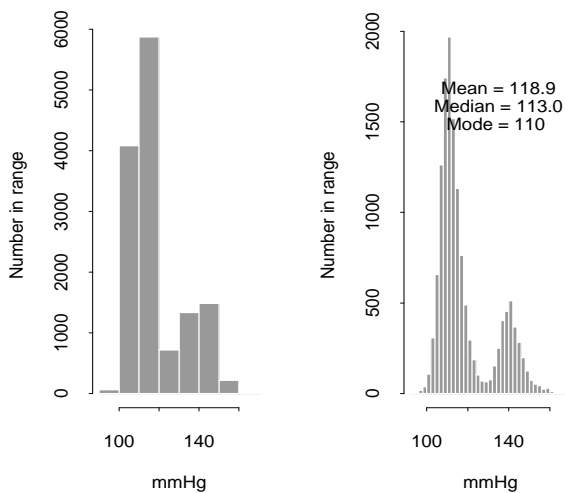
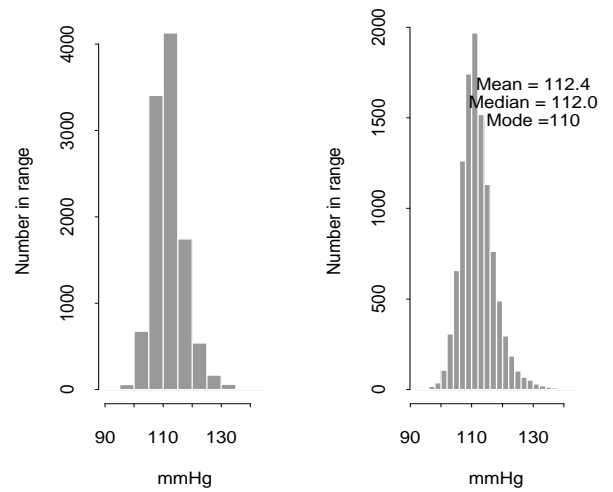
Numerical Summaries of Data

Arithmetic Mean: Sum of data values divided by the total number of values.

Median: Value which separates data into two halves: half the data values are greater than the median, half are smaller than the median. The median is less sensitive to outliers than the mean.

Mode: Value (or set of values) that occurs most frequently.

More precise definitions will be given soon.



Numerical Summaries of Data

Let n data values be denoted by x_1, x_2, \dots, x_n

Mean: Sum of data values divided by the total number of values.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Mode: Value (or set of values) that occurs most frequently.

Median: Value which separates data into two halves: half the data values are greater than the median, half are smaller than the median. The median is less sensitive to outliers than the mean.

Quantiles or percentiles: The n^{th} Quantile (percentile) is the smallest value which is greater than or equal to n percent of the data. For example, the 95th percentile is the value that is greater than or equal to 95% of the data and less than or equal to the remaining 5 %.

Quartiles: The 25th and 75th quantiles are called quartiles.

Range: Difference between the largest and the smallest data values.

Interquartile Range: Difference between the 75th and 25th percentiles. Consequently, it contains the middle 50% of the observations.

A bit of detail on Quantiles

Intuitively, the p^{th} quantile (percentile) is the smallest value V_p such that p percent of the sample points are less than or equal to V_p . The median, being the 50th percentile, is a special case of a quantile. Quartiles are also special cases of quantiles.

Note that this is not a precise definition. For example, if you have a data set with $n = 20$ values, what would the median be ?

Precise Definition of a Quantile

For a data set of size n , the p^{th} quantile is defined by

1. The $(k + 1)^{th}$ largest sample point if $\frac{np}{100}$ is not an integer. Here k is the largest integer less than $\frac{np}{100}$.
2. The average of the $(\frac{np}{100})^{th}$ and $(\frac{np}{100} + 1)^{th}$ largest observations if $\frac{np}{100}$ is an integer.

...More Numerical Summaries of Data

The (sample) *variance* of the data set is defined by

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

The rationale for using $(n-1)$ in the denominator (as opposed to n) will be given in a few weeks.

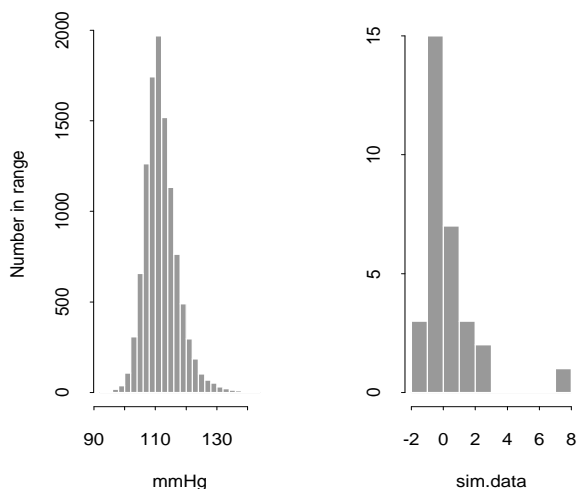
The *coefficient of variation* is the standard deviation divided by the mean.

$$CV = \frac{s}{\bar{x}}$$

It is often multiplied by 100 % to give a %.

If the 40th and 60th percentiles lie an equal distance from the midpoint, and the same is true for the 30th and 70th, the 20th and 80th, and all other pairs of percentiles that sum to 100, the data are *symmetric*.

A *symmetric* distribution has the same shape on each side of the 50th percentile. Shown below are histograms of MAP (left) data and simulated (right) data.



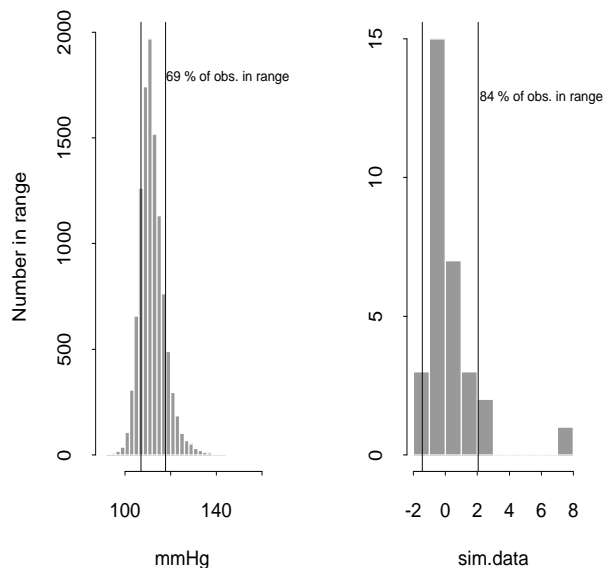
Summarizing the Distribution

The mean and standard deviation of a data set can be used to summarize characteristics of the entire distribution.

Empirical Rule: If the data are symmetric and unimodal, then approximately 67 % of the observations lie within the interval $\bar{x} \pm 1 \cdot \sigma$; approximately 95 % of the observations lie within $\bar{x} \pm 2 \cdot \sigma$.

There is a more precise version of this that we will see later this semester.

Vertical lines are drawn at $\bar{x} \pm 1 \cdot \sigma$ for the MAP (left) and simulated (right) data. 69 % of the observations fall between the lines for the MAP data; 84 % for the simulated data! The *empirical rule* does not work if the data are not symmetric and unimodal.



Chebychev's Inequality

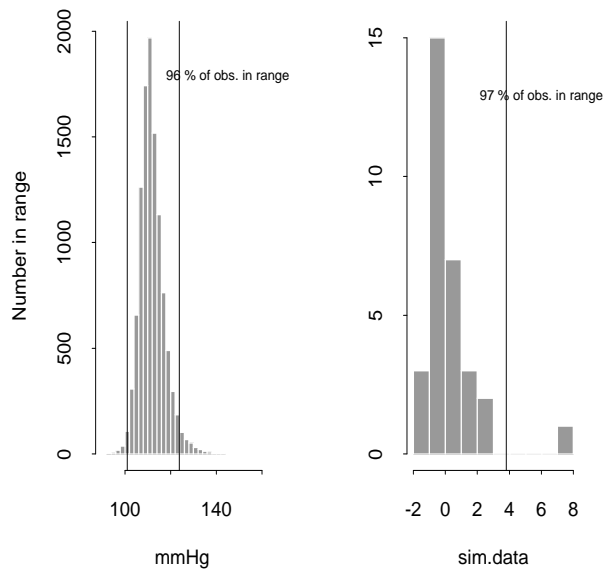
Chebychev's Inequality will work even when the data are not symmetric or unimodal.

Chebychev's Inequality: For any number k that is greater than 1, at least $1 - \left(\frac{1}{k}\right)^2$ lie within k standard deviations of their mean.

So, for $k = 2$, Chebychev's inequality tells us that at least

$$1 - \left(\frac{1}{2}\right)^2 = \frac{3}{4}$$

of the values lie within 2 standard deviations of the mean.



Box Plots

A *boxplot* is used for discrete or continuous data. The lower bound of the box is the 25th quartile of the data; the upper end is the 75th quartile. The 50th percentile (median) is indicated. Horizontal lines are drawn at the most extreme values outside the box that are not more than $1.5 \times$ Interquartile Range beyond either of the bounding quartiles.

Generating a Box Plot

To generate a box plot,

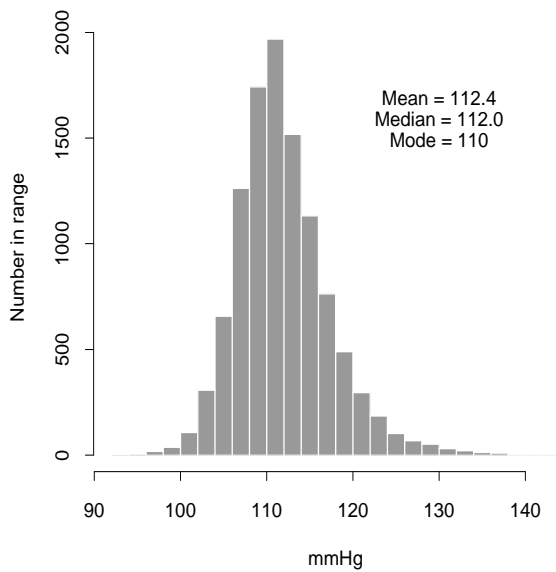
Plot the box. Upper bound is 75th quartile (75Q), lower bound is 25th quartile (25Q). Draw a line at the median. Note that 25Q and 75Q are NOT standard notation.

Calculate the Inter Quartile Range (75Q-25Q).

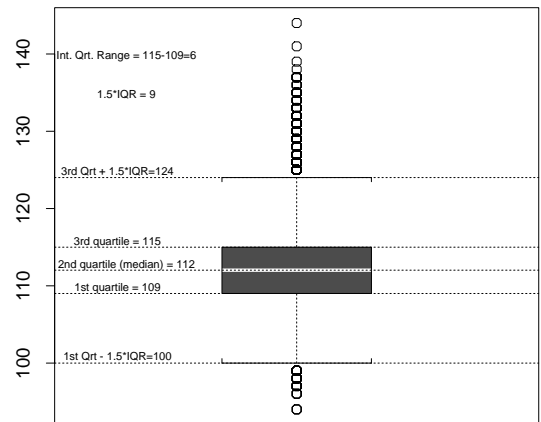
Draw horizontal lines at most extreme points closest to without going outside $75Q + 1.5 \times IQR$ and $25Q - 1.5 \times IQR$.

Draw in remaining points that fall outside the horizontal lines.

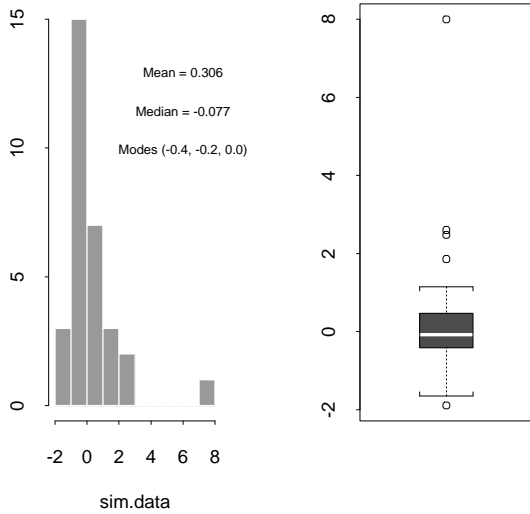
Rat Mean Arterial Pressure (MAP) Data



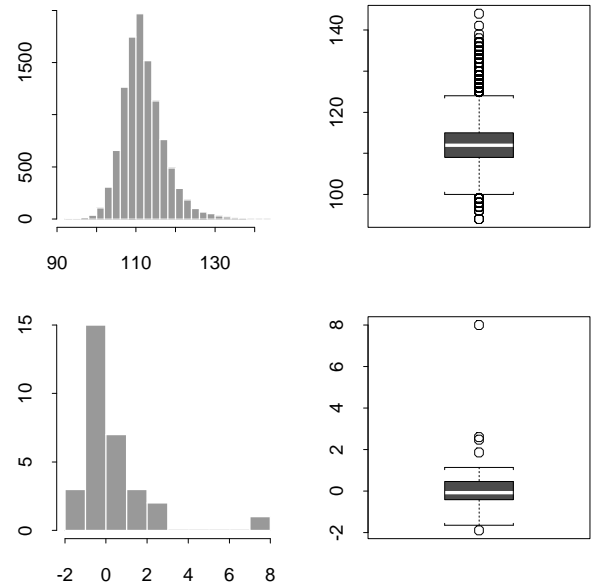
A box plot for the MAP pressure data shown earlier is given below.



Note that the horizontal lines need not be the same distance from the box. Again, the horizontal lines are drawn at the most extreme values outside the box that are NOT more than 1.5 * Interquartile Range beyond either of the bounding quartiles.



Histograms and Boxplots of MAP data (upper) and Simulated data (lower)

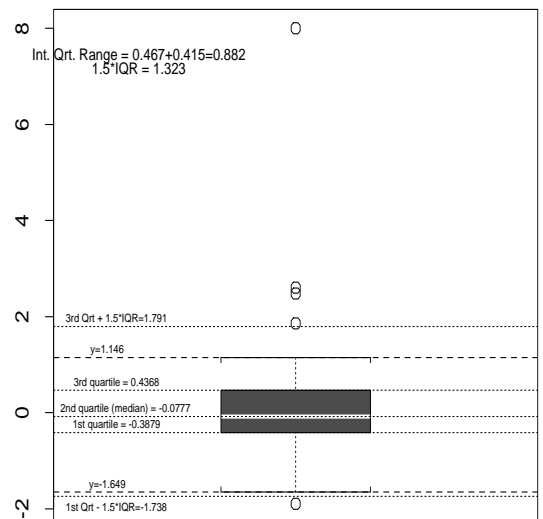


Simulated data from previous slide

-1.892 -1.649 -1.328 -0.913 -0.839 -0.724 -0.442 -0.415 -0.361 -0.300
 -0.282 -0.224 -0.177 -0.173 -0.145 -0.078 -0.049 -0.012 0.006 0.124
 0.159 0.266 0.406 0.467 0.905 1.072 1.146 1.858 2.478 2.602 8.000

A boxplot function will calculate quartiles. It might interpolate between values or impose the restriction that the quartiles be one of the data values. For now, let's impose the restriction that the quartiles be one of the data values. Find the median and the 25th and 75th quartiles. Where would the horizontal lines for the box plot be drawn ?

(hint: $25Q - 1.5 * IQR = -1.738$ and $75Q + 1.5 * IQR = 1.791$).



(From Stat 541 exam): Three data sets were generated. The box plots and histograms of each data set are shown. Match the box plot to the histogram generated by the same data.

(a) → (b) → (c) →

