Histogram

- Most commonly used tool in descriptive statistics.

- Histogram for discrete data:
  - Determine the frequency and relative frequency of each x value.
  - Mark possible x values on a horizontal scale.
  - Above each value, draw a rectangle whose height is the relative frequency (or the frequency) of that value.

- Histogram for continuous data:
  - Divide the range of the data into classes (5-10) of *equal width*. (It can also be unequal.)
  - Determine the frequency and relative frequency for each class.
  - Mark the class boundaries on a horizontal measurement axis.
  - Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

Constructing histogram

- Example: The maximum daily temperature in degrees Fahrenheit measured from May to September 1973 at La Guardia Airport. (154 observations)

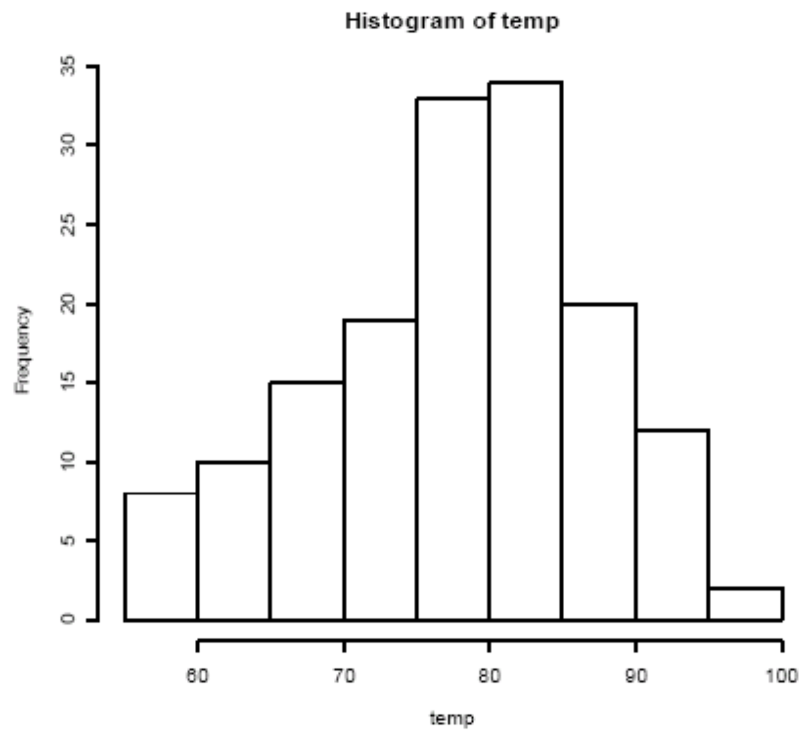
  Data

  {67 72 74 62 56 66 65 59 61 69 74 69 66 68 58 64 66 57 68 62 59 73 61 61 57 58 57 67 81 79 76 78 74 67 84 85 79 82 87 90 87 93 92 82 80 79 77 72 65 73 76 77 76 76 76 75 78 73 80 77 83 84 85 81 84 83 83 88 92 92 89 82 73 81 91 80 81 82 84 87 85 74 81 82 86 85 82 86 88 86 83 81 81 81 82 86 85 87 89 90 90 92 86 86 82 80 79 77 79 76 78 78 77 72 75 79 81 86 88 97 94 96 94 91 92 93 93 87 84 80 78 75 73 81 76 77 71 71 78 67 76 68 82 64 71 81 69 63 70 77 75 76 68}

  Draw a histogram.

Example cont.

Class	Count	Percent
55-59.9	8	5.2
60-64.9	10	6.5
65-69.9	15	9.8
65-74.9	19	12.4
75-79.9	33	21.6
80-84.9	34	22.2
85-89.9	20	13.1
90-94.9	12	7.9
95-99.9	2	1.3



- R demo. `>hist(x)` (option: `breaks=`...)

Examining distributions

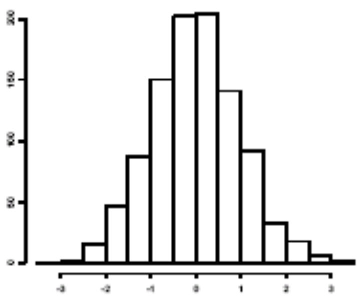
- When examining a distribution, look at its shape, center and spread. Look for clear deviations from the overall shape.

- We are interested in whether it is symmetric or skewed, as well as the number of modes.

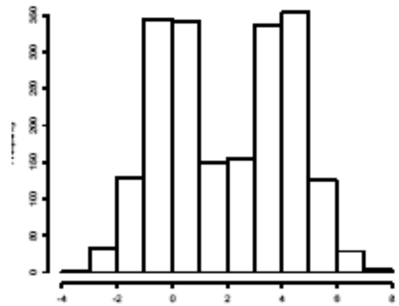
- Outliers are observations that lie outside of the overall pattern of a distribution.

Examining distributions

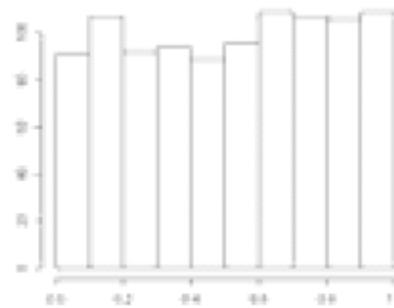
(a) Symmetric, unimodal



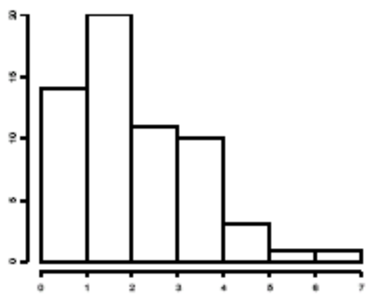
(b) bimodal



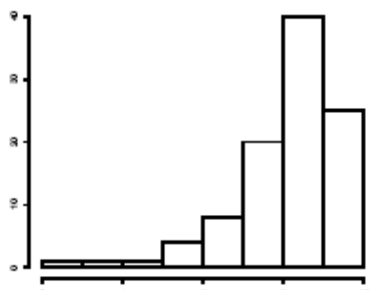
(c) Uniform



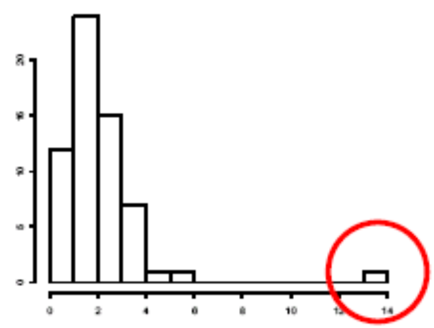
(d) right skewed



(e) left skewed



(f) Outlier



Examining a new data set

1. Examine each variable by itself.
2. Study the relationship between variables.

For both steps 1 and 2 we want to:

- Display the data graphically.
- Summarize the data numerically (Statistics).
- Construct a mathematical model.

Describing distributions numerically

- For single variables, We are interested in summaries that provide information about the center and spread of the distribution.

- A statistic is a numerical summary of data.

- The two most common measures of center are the mean and median.

- “generous” vs. “selfish”.

Mean

- If we have *n* ,observations, their mean is defined by,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

or

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

Ex. Calculate the mean of the data set: {1,2,3,4,5}.

$$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

Ex. Calculate the mean of the data set: {1,2,3,4,30}.

$$\bar{x} = \frac{1+2+3+4+30}{5} = \frac{40}{5} = 8$$

Mean cont.

- The mean is non-resistant, meaning that it is influenced by very large or very small data points that are extreme values for the data set.

Median

The median, written as M, is defined as the middle value of a data set.

1. List all *n* observations in order of size.
2. If *n* is odd, the median is the center value of the ordered list.
3. If *n* is even, the median is the average of the two center observations.

Median Cont.

<u>Ex.</u> Calculate the median of {6,2,5,19,12,10}.

2 5 6 10 12 19

M is the average of 6 and 10, hence M=8.

<u>Ex.</u> Calculate the median of {1,2,3,4,5} and {1,2,3,4,30}.

1 2 3 4 5 M=3.

1 2 3 4 30 M=3.

Median cont.

- The median is resistant (robust) to the extremes in the data set. Extremely large or small values do NOT influence the median.

Measures of variability

- Mean and median provide measures of location (center).
- One also needs some measures of variability to further describe the spread of the data set.
- Commonly used numerical values that can summarize the spread of a distribution.
  - Range
  - Interquartile Range (IQR)
  - Standard deviation

Quartiles

- The median divides the data into two groups of equal size.
- The quartiles divide the data into four groups of equal size.

Quartiles cont.

To find the quartiles:

1. Find the median.
2. Find the first quartile (Q1, or the *lower fourth*) by finding the median of the lower half of the data.
3. Find the third quartile (Q3, or the *upper fourth*) by finding the median of the upper half of the data.

   (*When n is odd include the median in both halves in steps 2 and 3.*)

Ex. Find the quartiles for the data set {2,4,6,8,12,14,18,19,41}.

2 4 6 8 12 14 18 19 41

2 4 6 8 12

12 14 18 19 41

M = 12 Q1 = 6 Q3=18

IQR

- The Interquartile Range, IQR, is the distance between the first and third quartiles,

$$IQR = Q3 - Q1.$$

- The IQR measures the spread of the middle 50% of the data.

- An observation is a suspected outlier if it falls more than 1.5*IQR from the closest fourth. An outlier is extreme if it is more than 3*IQR from the nearest fourth, and it is mild otherwise.

Ex. Can any of the observations in the data set {2,4,6,8,12,14,18,19,41} be considered outliers?

Recall we had M = 12, Q1=6, Q3=18. Therefore, IQR = 18 – 6 = 12.

1.5*IQR = 1.5*12 = 18. Q3+18 = 36, Q1-18 = -12. Since 41 > 36, 41 is classified as a potential outlier.

Boxplot

- A five number summary lists, in order, the minimum, Q1, the median, Q3, and the maximum.

- A boxplot is a graphical representation using a five number summary.
    1. Draw a vertical (horizontal) measurement scale.
    2. Place a rectangle to the right of (above) this axis; the lower (left) edge of the rectangle is at the lower fourth, and the upper (right) edge is at the upper fourth.
    3. Place a horizontal (vertical) line segment inside the rectangle at the location of the median.
    4. Draw "whiskers" out from either end of the rectangle to the smallest and largest observations that are NOT outliers.
    5. Using dots to represent outliers.

- R demo. `>boxplot(x)`

Standard deviation

- The variance and standard deviation are measures of spread that indicate how far values in the data set are from the mean, on average.

- Consider the observations $x_1, x_2, x_3, \ldots, x_n$.

- The deviations $(x_i - \bar{x})$ display the spread of x_i about their mean $\bar{x}$.

- The sum of the deviations is always 0, as some of the deviations are positive and others are negative.

- Squaring the deviations makes them all positive. Observations far from the mean will have large positive squared deviations.

- The variance is the ‘average’ squared deviation.

Standard deviation

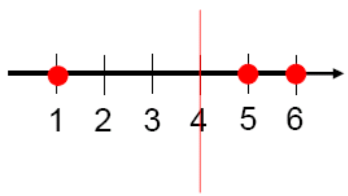
- If we have *n* observations $x_1, x_2, x_3, \ldots, x_n$. The variance is defined as

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- The standard deviation, *s*, is the square root of the variance.
  1. *s* is a measure of spread about the mean and should be used when the mean is used as the measure of center.
  2. If *s*=0, then all the values in the data set are exactly the same (no spread). Why?
  3. The more spread out the data, the greater the standard deviation.
  4. *s* is always positive.
  5. *s* has the same unit of measurement as the original data

Standard deviation

<u>Ex.</u> Let $x_1 = 1, x_2 = 5, x_3 = 6$



$$\bar{x} = 4$$

Calculate the deviations:

$$(x_1 - \bar{x}) = -3$$

$$(x_2 - \bar{x}) = 1$$

$$(x_3 - \bar{x}) = 2$$

Note that the deviations sum to 0.

Calculate the squared deviations:

$$(x_1 - \bar{x})^2 = 9$$

$$(x_2 - \bar{x})^2 = 1$$

$$(x_3 - \bar{x})^2 = 4$$

Calculate the ‘average’ squared deviation:

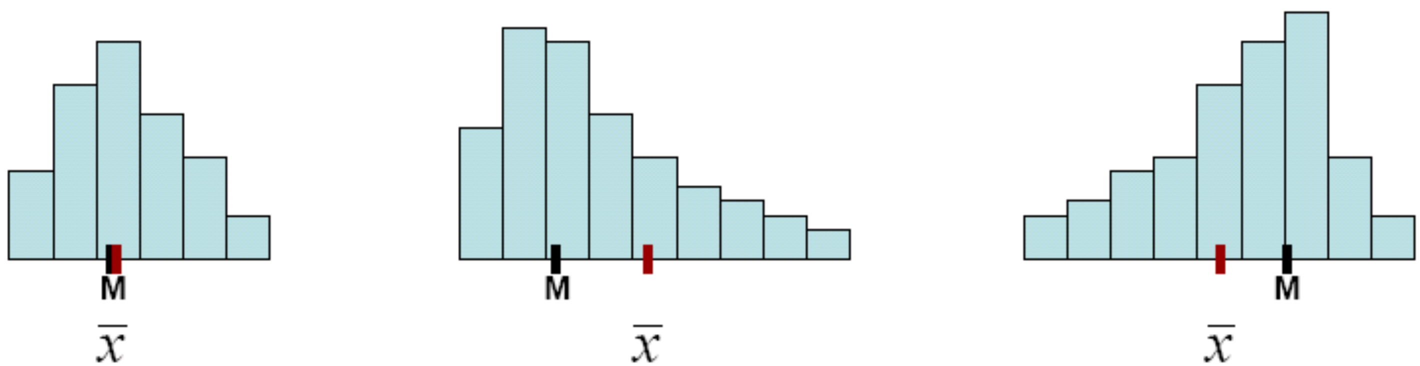
$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{2}(9+1+4) = \frac{14}{2} = 7$$

Degree of freedom

- As the sum of the deviations are always zero, the last deviation can be found once we know the other n-1.
- Only n-1 of the squared deviations can vary freely, so we average by dividing the total by n-1.
- n-1 are the degrees of freedom of the variance and standard deviation.

Measures of center and spread

- If the distribution is:
  1. symmetric, then $\bar{x} = M$ and both are located exactly in the middle of the distribution.
  2. skewed right, then $\bar{x} > M$.
  3. skewed left, then $\bar{x} < M$.



- As a rule of thumb: if a data set is reasonably symmetric use the mean and standard deviation, if it is highly skewed use the five-number summary.