STA 570 Spring 2011

Lecture 4 *Thursday, Jan 27*

- Descriptive Statistics
  - Graphical
  - Numerical

Homework 3: Due next week in lab.

Corrected Stem and Leaf Plot

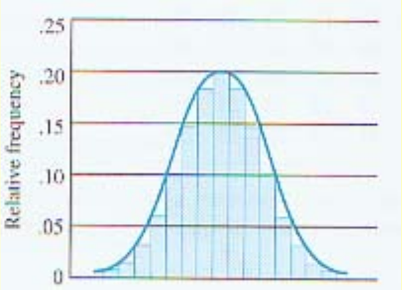
```
Stem Leaf

10 011122
 9 5555566666789
 9 011111222223334444444
 8 55666777778888999
 8 00001122222234
 7 689
 7 01
 6 66
 6 14
 5 8
```

multiply stem.leaf by 10

this is an Example with “split stems”

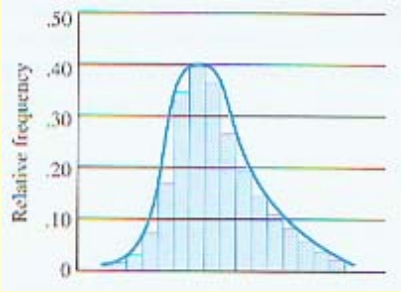
Review: Shapes of Distributions

Symmetric Distribution



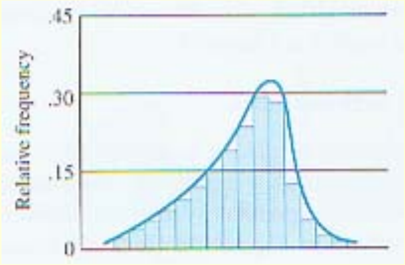
Skewed to the right, positively skewed

Skewed towards the higher values



Skewed to the left, negatively skewed

Skewed towards the lower values



Shape

- Look at the “tails”. If the tails are equal length, then the distribution is symmetric
- If the tail for lower values is longer, the distribution is left skewed
- If the tail for higher values is longer, the distribution is right skewed.
- “Symmetric” gets the benefit of the doubt in describing a distribution. “Roughly symmetric” is fine.

Good Graphics…

- …present large data sets concisely and coherently
- …can replace a thousand words and still be clearly understood and comprehended
- …encourage the viewer to compare two or more variables
- …do not replace substance by form
- …do not distort what the data reveal

Bad Graphics…

- …don’t have a scale on the axis
- …have a misleading caption
- …distort by stretching/shrinking the vertical or horizontal axis
- …use histograms or bar charts with bars of unequal width
- …are more confusing than helpful

Summarizing Data Numerically

- Center of the data
  - Mean: Arithmetic average *(Interval)*
  - Median: Midpoint of the observations when they are arranged in increasing order *(Interval, Ordinal)*
  - Mode: Most frequent value *(Interval, Ordinal, Nominal)*
- Dispersion of the data
  - Variance, Standard deviation
  - Interquartile range
  - Range
- Skewness of the data

Mode

- One statistic mentioned often for categorical data (ordinal or nominal) is the mode, which is the category with the most observations.
- The mode is most meaningful when one of the categories has most of the observations, as in “most faculty at UK have doctoral degrees”
- If the data is spread among many categories, knowing the mode doesn’t provide a full picture.

Central Location for Interval Data

- For interval data, the most common measures of central location are the mean and median.
- The mean is defined as the arithmetic average of the observations. You find this by adding them up and dividing by the total number. If your observations are (2,6,13), the mean is (2+6+13)/3 = 7.

Mean/Median continued

- The median is the “middle” observation of the SORTED data. If your observations are (2,6,13), the median is 6. If your observations are (5,11,0,8,10), the median is 8.
- If there is an even amount of data, average the two middle values. So if the data are (6,10,4,3), the middle values are 4 and 6, and (4+6)/2 = 5. The median is 5.

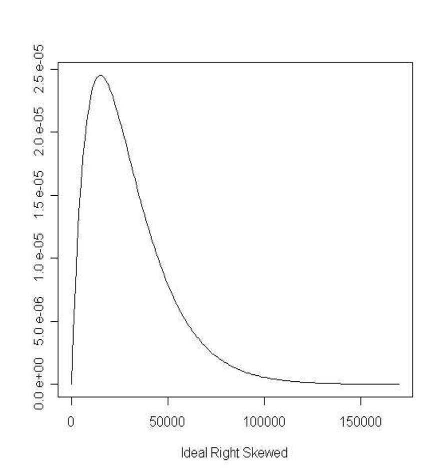
Differences between the mean and median

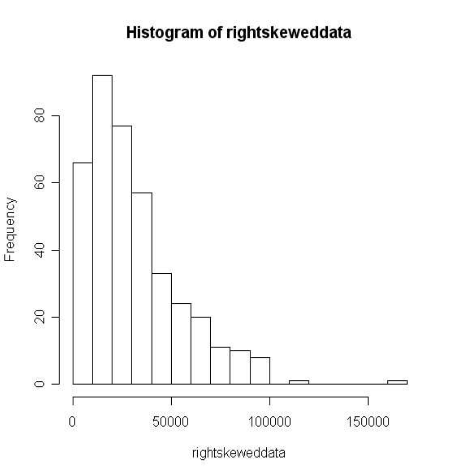
- The median is robust, which means that outliers do not affect it. The mean is not.
- Suppose we have data (1,4,6,10,12). The mean is 33/5 = 6.6 while the median is 6.
- Suppose we change the 12 to 14000. The median is still 6, but the mean changes to 14021/5 = 2804.2. Note also that the median is still close to most of the data, but the mean is nowhere close to any data point.

Mean

- The mean is highly influenced by outliers. That is, data points that are far from the rest of the data.
- Right skewed distribution:

The mean is pulled to the right.





Mean (Average)

- The mean requires numerical values. Only appropriate for quantitative data.
- It does not make sense to compute the mean for nominal variables.
- Example "Nationality" (nominal):

  Germany = 1, Brazil = 2,

  U.S. = 3, China = 4, India = 5
- Mean nationality = 2.8???

Mean

- Sometimes, the mean is calculated for ordinal variables, but this does not always make sense.
- Example “average health” (on an ordinal scale):

  excellent=1, good=2, fair=3, poor=4
- Mean (average) health=2.1

- Another example: “GPA = 3.8” is also a mean of observations measured on an ordinal scale

Mean

- Assume that each measurement has the same “weight”
- Then, the mean is the center of gravity for the set of observations
- This is because the sum of the distances to the mean is the same for the observations above the mean as for the observations below the mean

Median

- The median is the measurement that falls in the middle of the ordered sample
- When the sample size n is odd, there is a middle value
- It has the ordered index $(n+1)/2$
- Example: 1.1, 2.3, 4.6, 7.9, 8.1

  $n=5$, $(n+1)/2=6/2=3$, Index =3,

  Median = 3rd smallest observation = 4.6

Median

- When the sample size *n* is even, average the two middle values
- Example: 3, $\underline{7}$, $\underline{8}$, 9, *n=4,*

  (n+1)/2=5/2=2.5, Index =2.5

  Median = midpoint between 2nd and 3rd smallest observation = (7+8)/2 =7.5

Median

- The median can be used for interval data and for ordinal data
- The median can not be used for nominal data because the observations can not be ordered on a scale

- How can the median be found from a stem and leaf plot?

Mean versus Median

- Mean: Interval data with an approximately symmetric distribution
- Median: Interval or ordinal data

- The mean is sensitive to outliers,
  the median is not

Mean vs. Median

Observations	Median	Mean
1, 2, 3, 4, 5	3	3
1, 2, 3, 4, 100		
3, 3, 3, 3, 3		
1, 2, 3, 100, 100		

Mean vs. Median

- If the distribution is symmetric, then Mean=Median
- If the distribution is skewed, then the mean lies more toward the direction of skew
- Mean and Median Online Applet

Percentiles

- The *p*th percentile is a number such that *p* % of the observations take values below it, and (100-*p*)% take values above it
- 50th percentile = median
- 25th percentile = lower quartile
- 75th percentile = upper quartile

Quartiles

- 25th percentile

  = lower quartile

  = median of the observations below the median

- 75th percentile

  = upper quartile

  = median of the observations above the median

- Median and Quartiles can be found from a stem and leaf plot
- Example: Murder Rate Data

```
Stem Leaf                 #
 20 3                  1
 19
 18
 17
 16
 15
 14
 13 135                  3
 12 7                  1
 11 334469                  6
 10 2234                 4
  9 08                 2
  8 03469                 5
  7 5                  1
  6 034689                  6
  5 0238                 4
  4 46                 2
  3 0144468999                10
  2 039                  3
  1 67                 2
    ----+----+----+----+
```

A quarter of the states
has murder rate above…

The median murder rate is…

A quarter of the states
has murder rate below…

- Median and Quartiles can be found from a stem and leaf plot
- Example: Murder Rate Data

```
Stem Leaf                #
  20 3                 1
  19
  18
  17
  16
  15
  14
  13 135                 3
  12 7                 1
  11 334469                6
  10 2234                4
   9 08                2
   8 03469               5
   7 5                 1
   6 034689              6
   5 0238                4
   4 46                2
   3 0144468999              10
   2 039                 3
   1 67                2
    ----+----+----+----+
```

A quarter of the states has murder rate above…

The median murder rate is…

A quarter of the states has murder rate below…

- Median and Quartiles can be found from a stem and leaf plot
- Example: Murder Rate Data

```
Stem Leaf                #
  20 3                  1
  19
  18
  17
  16
  15
  14
  13 135                 3
  12 7                  1
  11 334469               6
  10 2234                4
   9 08                 2
   8 03469               5
   7 5                 1
   6 034689              6
   5 0238               4
   4 46                 2
   3 0144468999           10
   2 039                 3
   1 67                 2
    ----+----+----+----+
```

A quarter of the states has murder rate above…

The median murder rate is…

A quarter of the states has murder rate below…

Five-Number Summary

- Maximum, Upper Quartile, Median, Lower Quartile, Minimum
- SAS output (Murder Rate Data)

```
      Quantile      Estimate

100% Max          20.30
75% Q3           10.30
50% Median         6.70
25% Q1            3.90
0% Min           1.60
```

Five-Number Summary

- Maximum, Upper Quartile, Median, Lower Quartile, Minimum
- Example: The five-number summary for faculty salaries (in $1000) in the Mathematics Department (2006) is minimum=50, Q1=63, median=74, Q3=92, maximum=112.
- What does this suggest about the shape of the distribution?
- Also, the mean salary is reported as $77,000. Is this consistent with the shape of the distribution?

Summarizing Data Numerically

- Measures of Location / Central Tendency
  - Where is the data located?
  - Where is the “middle” of the data?
  - Mean, Median (Mode, Percentiles)
- **Measures of Variation**
  - **How variable are the data?**
  - **How spread out about the “middle” are the data?**
  - **Range, Variance, Standard Deviation, Interquartile Range**
- Measures of Skewness

Sample and Population Measures of Variation

- Range: maximum - minimum
- Variance (sample / population):

$$s^2 = \frac{\sum (Y_i - \overline{Y})^2}{n-1} \qquad \sigma^2 = \frac{\sum (Y_i - \mu)^2}{N}$$

- Standard Deviation (sample / population):

$$s = \sqrt{\frac{\sum (Y_i - \overline{Y})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\sum (Y_i - \mu)^2}{N}}$$

- Interquartile Range: Q3-Q1

Range

- Range: Difference between the largest and smallest observation
- Very much affected by outliers (a misrecorded observation may lead to an outlier, and affect the range)
- The range does not always reveal different variation about the mean

Sample Variance

$$s^2 = \frac{\sum (Y_i - \overline{Y})^2}{n-1}$$

The variance of *n* observations is the sum of the squared deviations, divided by *n-1.*

Variance: Interpretation

- The variance is about the average of the squared deviations
- “average squared distance from the mean”
- Unit: square of the unit for the original data
- Difficult to interpret
- Solution:
  - Take the square root of the variance
  - Then, the unit is the same as for the original data

Sample Standard Deviation

- The sample standard deviation *s* is the positive square root of the sample variance

$$s = \sqrt{\frac{\sum (Y_i - \overline{Y})^2}{n-1}}$$

Standard Deviation: Properties

- $s \geq 0$ always
- *s=0* only when all observations are the same
- If data is collected for the whole population instead of a sample, then *n-1* is replaced by *N*
- *s* is sensitive to outliers

Standard Deviation Interpretation: Empirical Rule

- If the histogram of the data is approximately symmetric and bell-shaped, then
  - About **68%** of the data are within **one** standard deviation from the mean
  - About **95%** of the data are within **two** standard deviations from the mean
  - About **99.7%** of the data are within **three** standard deviations from the mean

More precisely: 68.27% – 95.45% – 99.73%

Example

- Distribution of (old) class scores was approximately bell-shaped with mean 88 and standard deviation 9
- About 68% of the 81 scores are between ________
- About 95% are between _____________
- If you have a score above 95, you are in the top _______%

Yet Another Example

- "Number of people you have known personally who have committed suicide in the last 12 months"

Response	Frequency	Percentage
0	1344	88.8
1	133	8.8
2	25	1.7
3	11	0.7
4	1	0.1

•Mean = 0.15

•Standard Deviation = 0.46

•Are 68% of the observations between –0.31 and 0.61?

Standard Deviation (*s*) vs. Interquartile Range (*IQR*)

- Standard Deviation is affected by outliers
- Interquartile Range is not affected by outliers
- *NB: The Range is most affected by outliers*

- Whenever you use the Median instead of the Mean, you should also use the IQR instead of the Standard Deviation

Example Data Sets

- One Variable Statistical Calculator
- Modify the data sets and see how mean and median, as well as standard deviation and interquartile range change
- Look at the histograms and stem-and-leaf plots – does the empirical rule apply?
- Make yourself familiar with the standard deviation
- Interpreting the standard deviation takes experience