STA 321 Spring 2014

Lecture 3 Tues*day, January 28*

- Sampling and Measurement
  - Sampling Plans
  - Sampling and Nonsampling Error
- Descriptive Statistics
  - Graphical
  - Numerical

Sampling Plans

- Simple Random Sampling (SRS)
- **Stratified Random Sampling**
- **Cluster Sampling**
- Systematic Sampling

Stratified Sampling

- Suppose the population can be divided into separate, non-overlapping groups ("***strata***") according to some criterion.
- Select a simple random sample independently from each group.

Why could stratification be useful?

- We may want to draw inference about population parameters for each subgroup
- Sometimes, ("proportional stratified sample") estimators from stratified random samples are more precise than those from simple random samples

Proportional Stratification

- The proportions of the different strata are the same in the sample as in the population
- Mathematically:

Population size N, subpopulation sizes N_i

Sample size n, subsample sizes n_i

$$\frac{n_i}{n} = \frac{N_i}{N}$$

Proportional Stratification

- Example:
  - Total population of the US: 304 Million
  - Population of Kentucky: 4 Million (1.3%)
  - Suppose you take a sample of size n=304 of people living in the US.
  - If stratification is proportional, then 4 people in the sample need to be from Kentucky
  - Suppose you take a sample of size n=1000. If you want it to be proportional, then 13 people (1.3%) need to be from Kentucky.

Cluster Sampling

- The population can be divided into a set of non-overlapping subgroups (the clusters)
- The clusters are then selected at random, and all individuals in the selected clusters are included in the sample
- Cluster Sampling is usually less precision than SRS and Stratified sampling.
- Cluster sampling cost usually less and more convenient.

Summary of Important Sampling Plans

- **Simple Random Sampling (SRS)**
  - Each possible sample has the same probability of being selected.
- **Stratified Random Sampling**
  - Non-overlapping subgroups (strata)
  - SRSs are drawn from each strata
- **Cluster Sampling**
  - Non-overlapping subgroups (clusters)
  - Clusters selected at random
  - All individuals in the selected clusters are included in the sample
- **Systematic Sampling**
  - Useful when the population consists as a list
  - A value *K* is specified. Then one of the first *K* individuals is selected at random, after which every *Kth* observation is included in the sample

Types of Bias

- **Selection Bias**
  – Selection of the sample systematically excludes some part of the population of interest
- **Measurement/Response Bias**
  – Method of observation tends to produce values that systematically differ from the true value
- **Nonresponse Bias**
  – Occurs when responses are not actually obtained from all individuals selected for inclusion in the sample

Bias?

- Pittsburgh is known to have a very good medical center. However, in “America’s Most Liveable cities”, Pittsburgh was marked down on health care.
- The variable used as a proxy for healthcare was “mortality rate in hospitals”.
- Why would a good medical center perform poorly on mortality rate?

Sampling and Nonsampling Error

- Assume you take a random sample of 100 UK students and ask them about their political affiliation (Democrat, Republican, Independent)
- Now take another random sample of 100 UK students
- Will you get the same percentages?

Sampling Error

- No, because of sampling variability.
- Also, the result will not be exactly the same as the population percentage, unless you take a “sample” consisting of the whole population of 25,000 students (this would be called a “census”)

  or if you are very lucky

Sampling Error

- **Sampling Error** is the error that occurs when a statistic based on a sample estimates or predicts the value of a population parameter.
- In random samples, the sampling error can usually be quantified.
- In ***nonrandom*** samples, there is also sampling variability, but its extent is ***not predictable***.

Nonsampling Error

- Everything that could also happen in a census, that is when you ask the whole population
- Examples: Bias due to question wording, question order, nonresponse (people refuse to answer), wrong answers (especially to delicate questions)

Graphic Descriptive Statistics

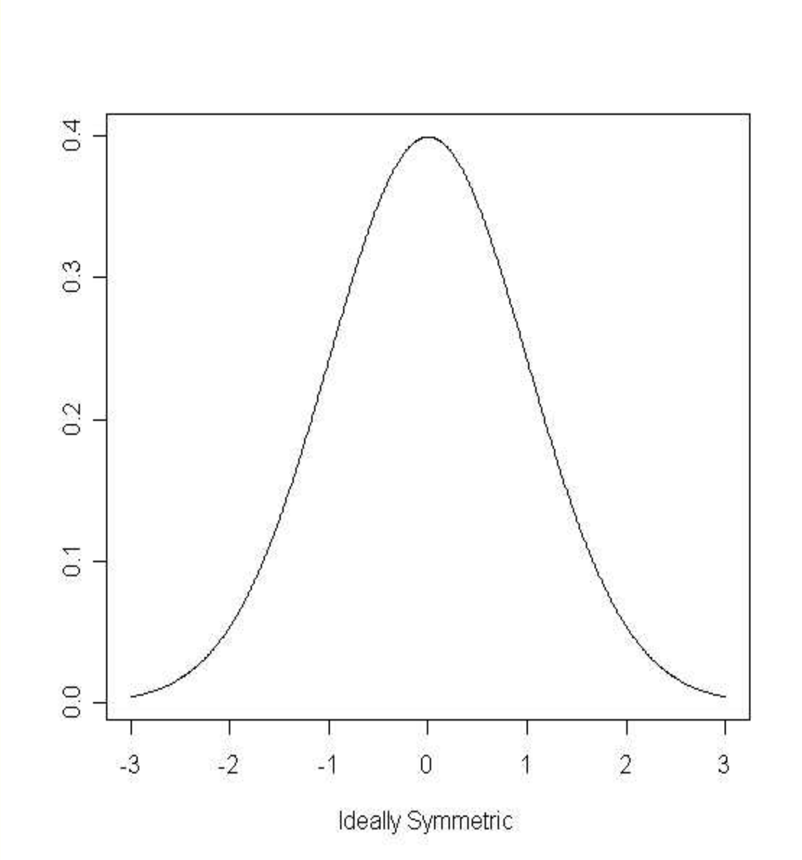
- Summarize data graphically
- Use graphs, tables, and numbers
- Condense the information from the dataset

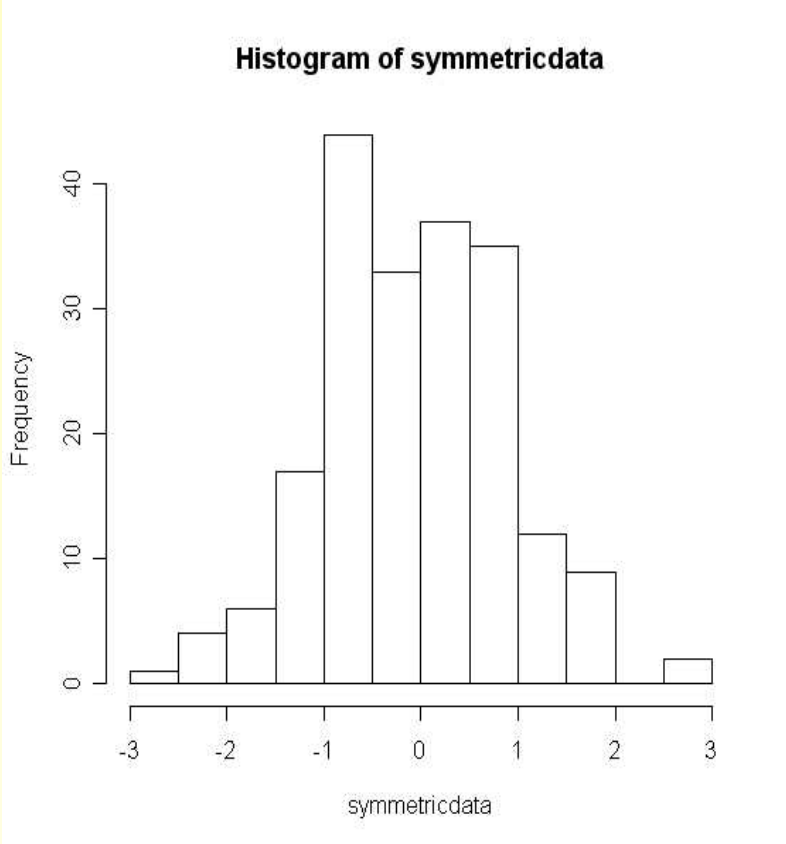
- Graphs for Interval data:
  Histogram, Stem and Leaf Plot,
  Box Plot (later)
- Graphs for Nominal/Ordinal data:
  Bar graph, Pie chart

The four features of distributions

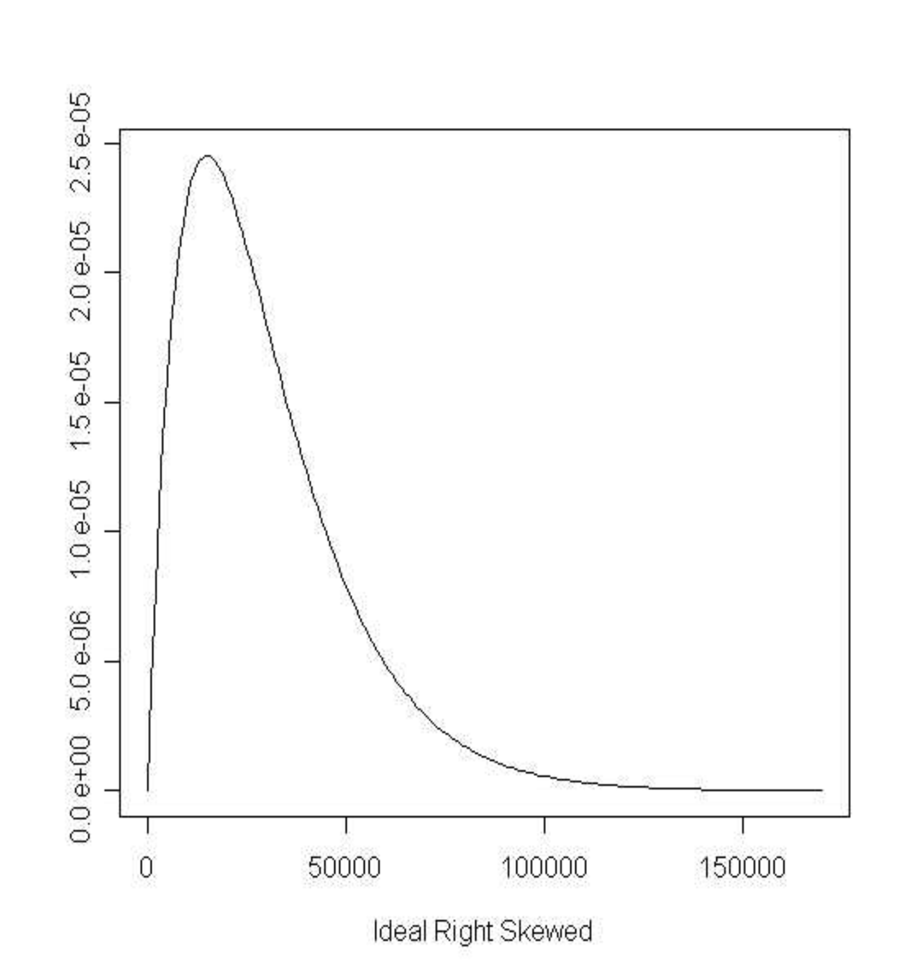
- Central Location – where are most of the observations?
- Spread – how far apart are the observations?
- Shape – Symmetric or skewed?
- Outliers – are any observations very far from the rest?

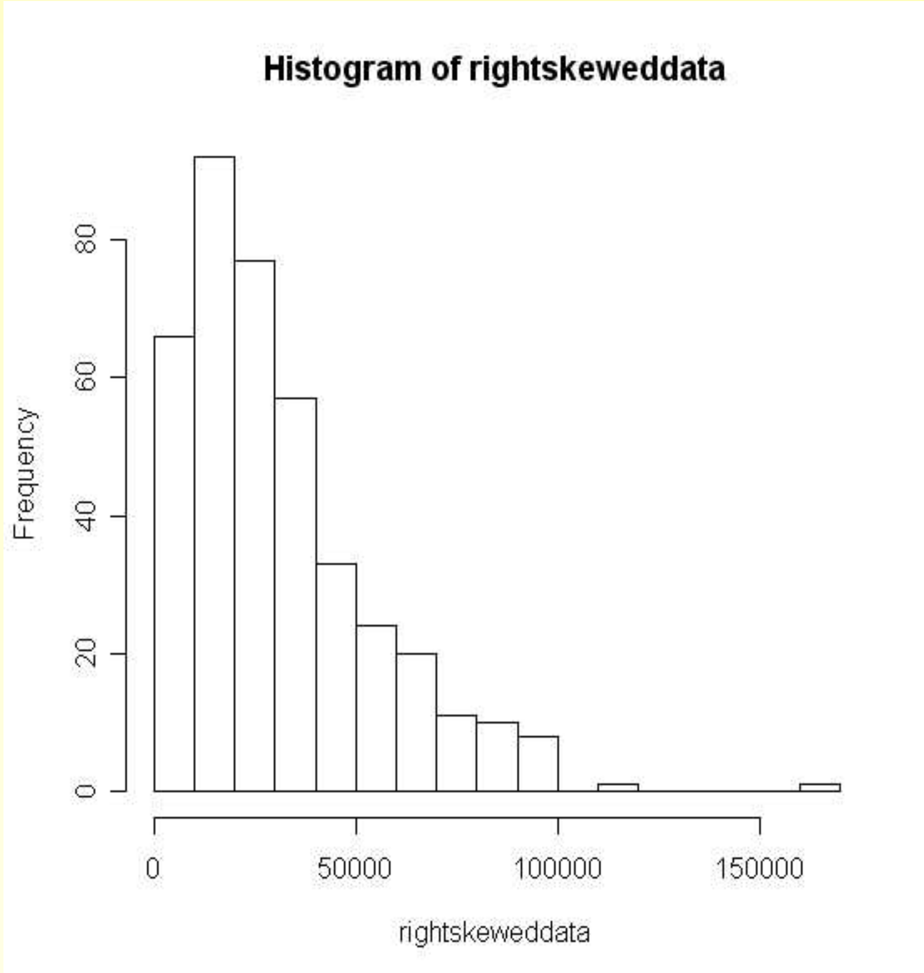
Symmetric Data – Ideally and Practically



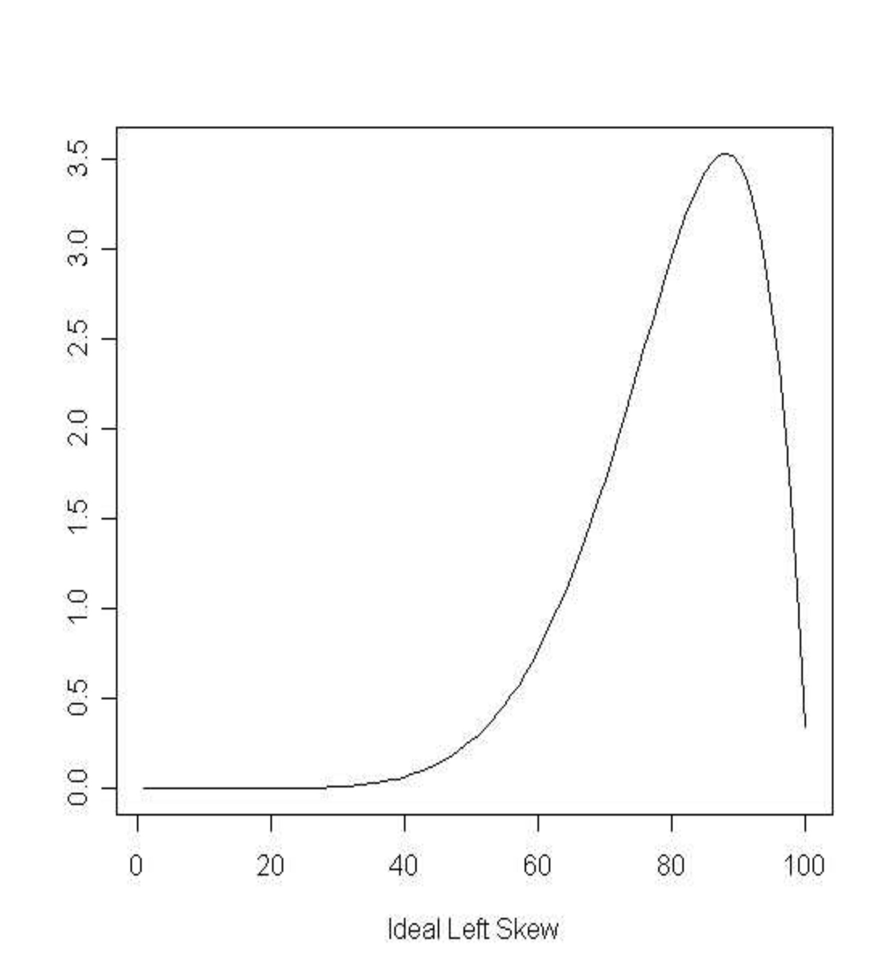


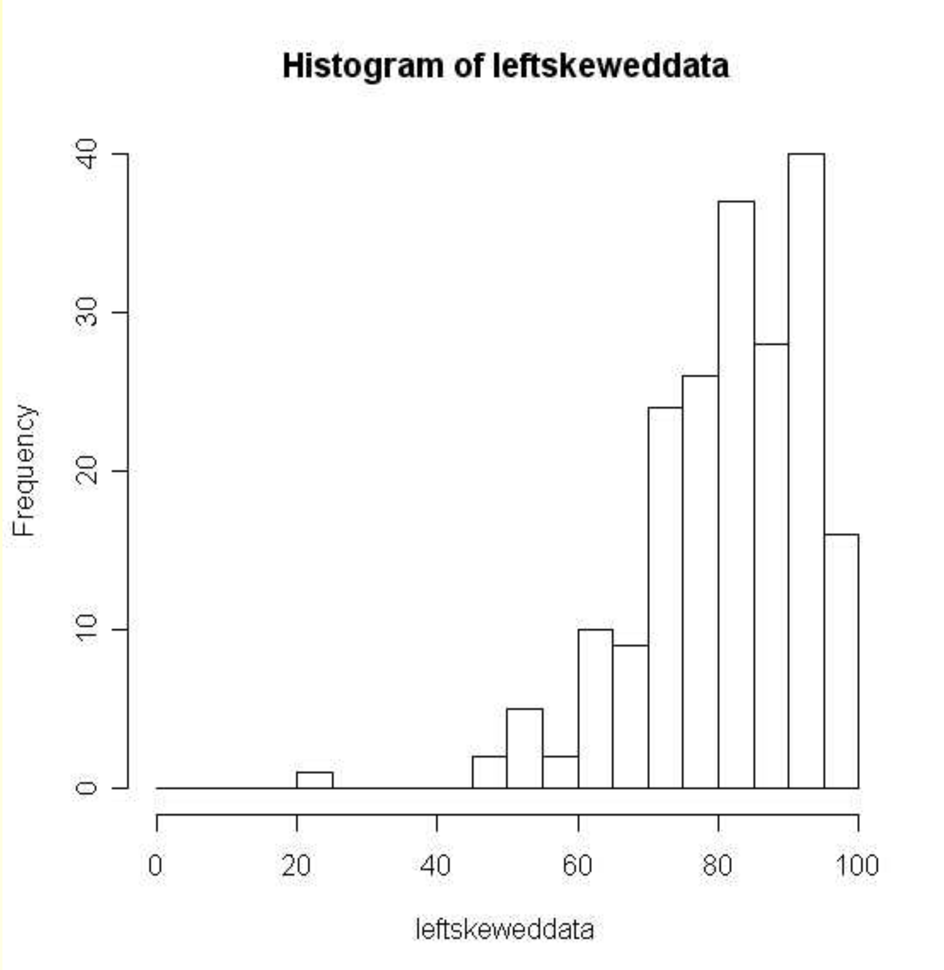
Right skewed data – ideally and practically





Left skewed data – ideally and practically





Histogram (Interval Data)

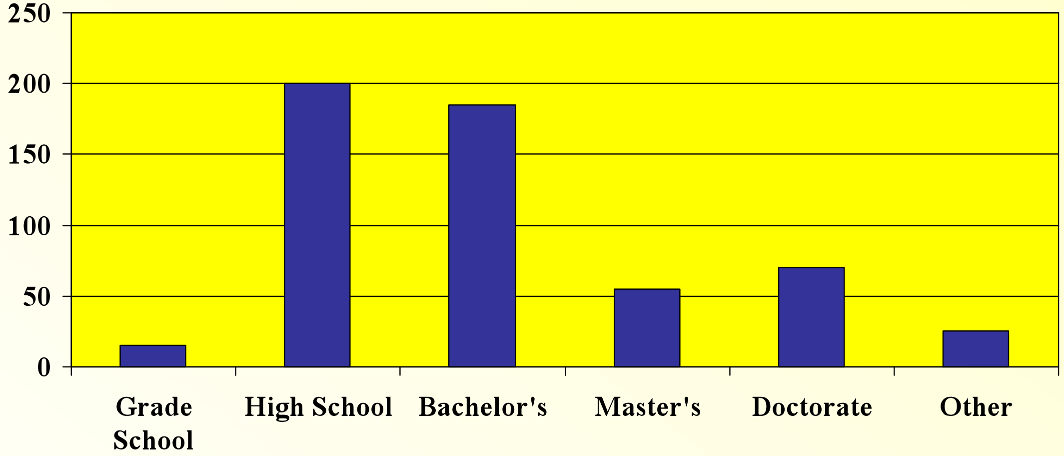
- Find frequencies for each interval
- Draw a bar over each interval, the height of the bar represents the relative frequency for that interval
- Bars should be touching; I.e., equally extend the width of the bar at the upper and lower limits so that the bars are touching.

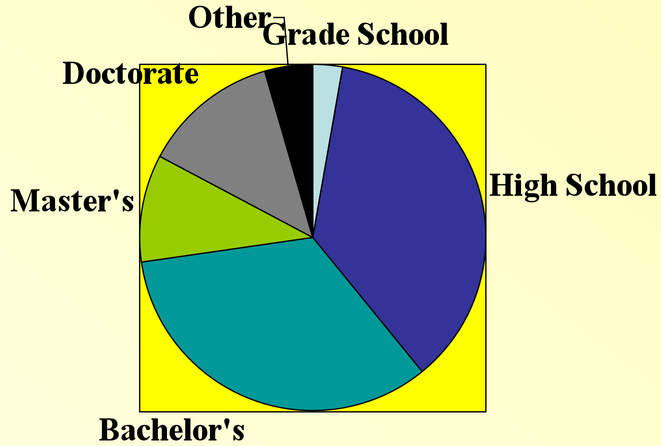
Bar Graph (Nominal/Ordinal Data)

- Histogram: for *interval (quantitative) data*
- Bar graph is almost the same, but it is for *qualitative data*
- Difference:
  - The bars are usually separated to emphasize that the variable is categorical rather than quantitative
  - For nominal variables (no natural ordering), order the bars by frequency, except possibly for a category “other” that is always last

Bar Graph and Pie Chart for “Highest Degree Achieved”

If the data is ordinal, classes are presented in the natural ordering, except that “Other” is usually at the end.





Stem and Leaf Plot

- Write the observations ordered from smallest to largest
- Each observation is represented by a stem (leading digit(s)) and a leaf (final digit)
- Looks like a histogram sideways
- Contains more information than a histogram, because every single measurement can be recovered

Stem and Leaf Plot (Interval Data)

Stem Leaf

100 011122
90 5555566666789
90 0111112222233344444444
80 55666777778888999
80 000011222222234
70 689
70 01
60 66
60 14
50 8

this is an Example with "split stems"

Stem and Leaf Plot

- Useful for small data sets (<100 observations)
- Practical problem:
  - What if the variable is measured on a continuous scale, with measurements like

    1267.298, 1987.208, 2098.089, 1199.082 etc.
  - Use common sense when choosing “stem” and “leaf”

Stem and Leaf Plot

- Can also be used to compare groups:
  Back-to-Back Stem and Leaf Plots, using the same stems for both groups.
- Murder Rate Data from U.S. and Canada

- By the way, it doesn’t really matter whether the smallest stem is at top or bottom of the table

Example

- Data set:

  5.5, 18.5, 6.0, 5.5, 5.3,

  5.8, 11.0, 6.1, 7.0, 14.5,

  10.4, 7.6, 4.3, 7.2, 10.5,

  6.5, 3.3, 2.0, 4.1, 6.2
- Create a stem and leaf plot

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.

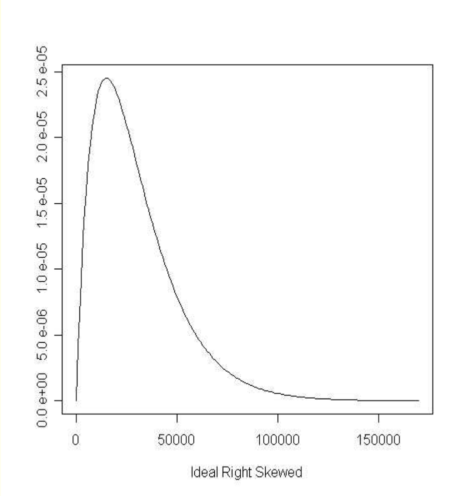
Differences between the mean and median

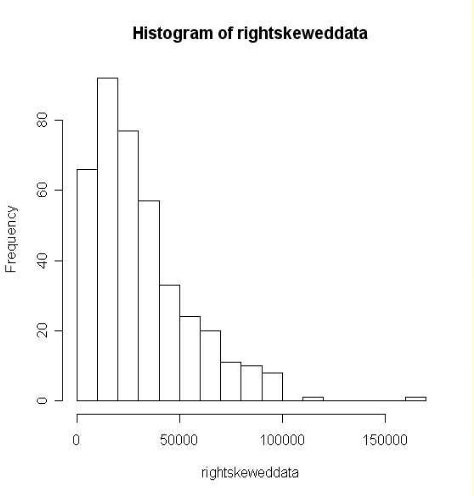
- The median is robust, which means that outliers do not affect it. The mean is not.
- Suppose we have data (1,4,6,10,12). The mean is 33/5 = 6.6 while the median is 6.
- Suppose we change the 12 to 14000. The median is still 6, but the mean changes to 14021/5 = 2804.2. Note also that the median is still 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 Springs 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, <u>4.6</u>, 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, <u>7</u>, <u>8</u>, 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 <u>not</u> 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?

Grouped or Ordinal Data (Mean, Median, Mode)

- "How often do you read the Kernel?"

Response	Frequency
every day	969
a few times a week	452
once a week	261
less than once a week	196
Never	76

- Identify the mode
- Identify the median response, if possible
- Find the mean, if possible

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