Biostatistics and Informatics

BTM6000

Jochen G. Raimann

Session 2: Descriptive Statistics

Slide credit to
Prof. Levi Waldron
CUNY SPH

Material to cover for session 2

1. Review session #1
2. Frequency tables, bar plots, histograms, box plots, dot plots
3. Measuring center: Mean, Median, Mode
4. Measuring spread: Standard Deviation, Variance, Interquartile Range, Range

Vocabulary review

- **Population**: the entire group of interest
- **Sample**: a subset of the population
- **Record or observation:** data of an individual
- **Descriptive statistics**: summarization, visualization
- **Inferential statistics**: learning about the population from the sample

Vocabulary review

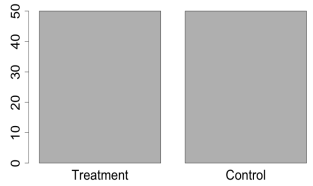
- **Variable**: any characteristic that is recorded for the subjects in a study
  - **Nominal** (**binary**, **categorical**)
    - treatment group, marital status, race
  - **Ordinal**
    - birth order among siblings, ranking
    - Sometimes nominal + ordinal lumped as categorical
  - **Quantitative** (**discrete**, **continuous**)
    - number of children
    - height, weight, age

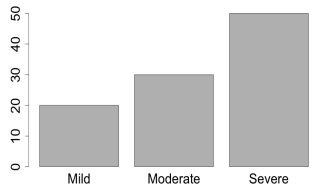
Nominal / ordinal data summary

- Frequency tables / contingency tables
- Barplot

Treatment	Control
50	50

Mild	Moderate	Severe
20	30	50





(one variable)

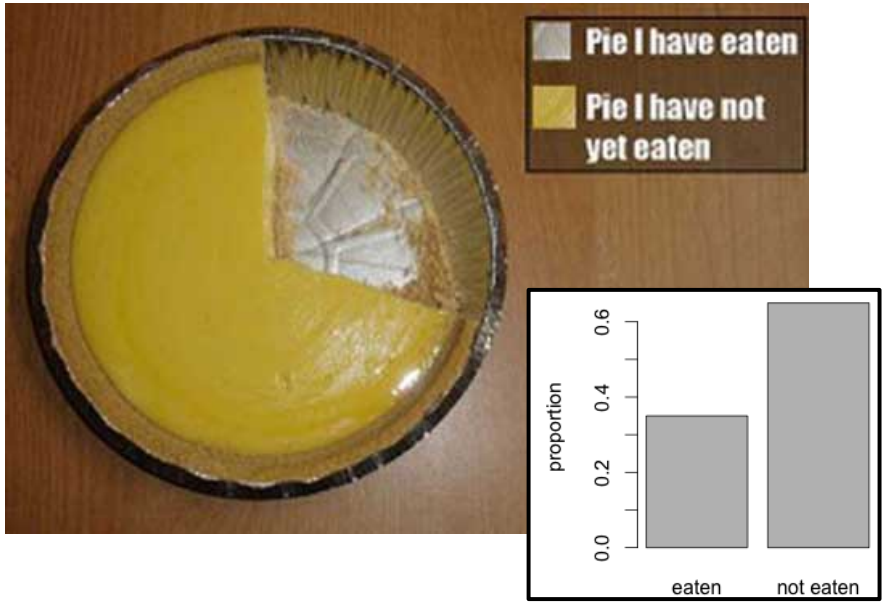
Nominal / ordinal data summary

- Pie charts are another alternative:

Generally disliked by statisticians.

Why? Can you tell what proportion of the pie I have eaten? I can't.

Anything a pie chart can show, a bar chart can show better.



Nominal / ordinal data summary: Frequency, Proportion, & Percentage

- If 4 students received an “A” out of 40 students, then:
  - the **frequency** is 4
  - the **proportion** or **relative frequency** is 0.10 =4/40
  - the **percentage** is 10% (.10*100=10%)
- Any of these can be used for a frequency table or barplot.

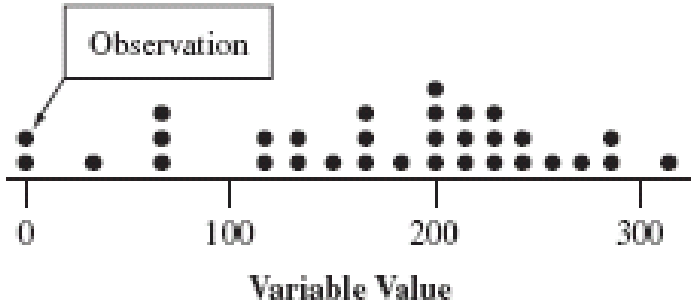
Quantitative variables: graphical summary

- **Dot plot:** shows a dot or dash for each observation
- **Histogram:** categorizes the data then shows it as a barplot
- **Box plot:** qualitative and quantitative measures together

- These plots help identify central tendency, spread, skew, and outliers

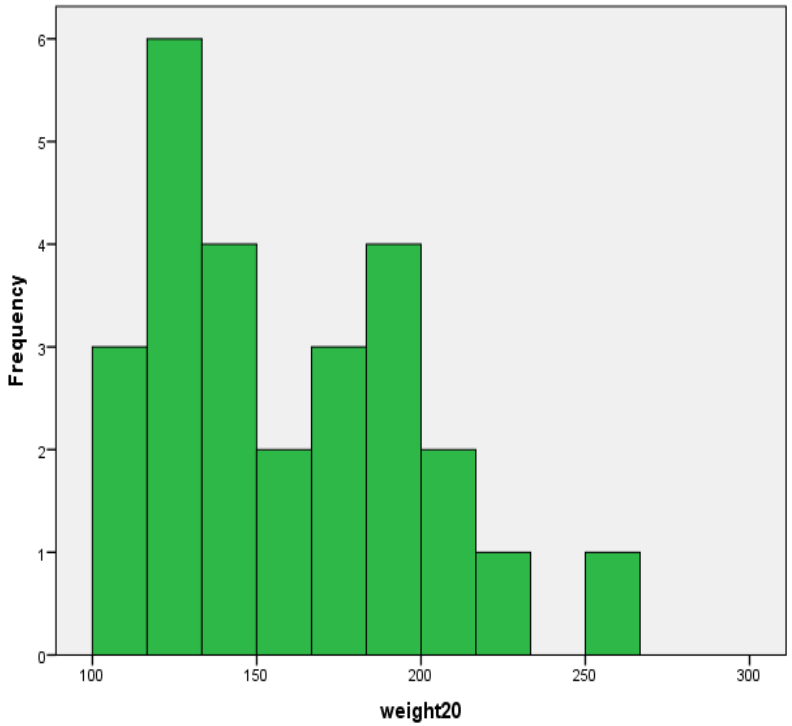
Dot Plots

Dot plots show every individual observation:



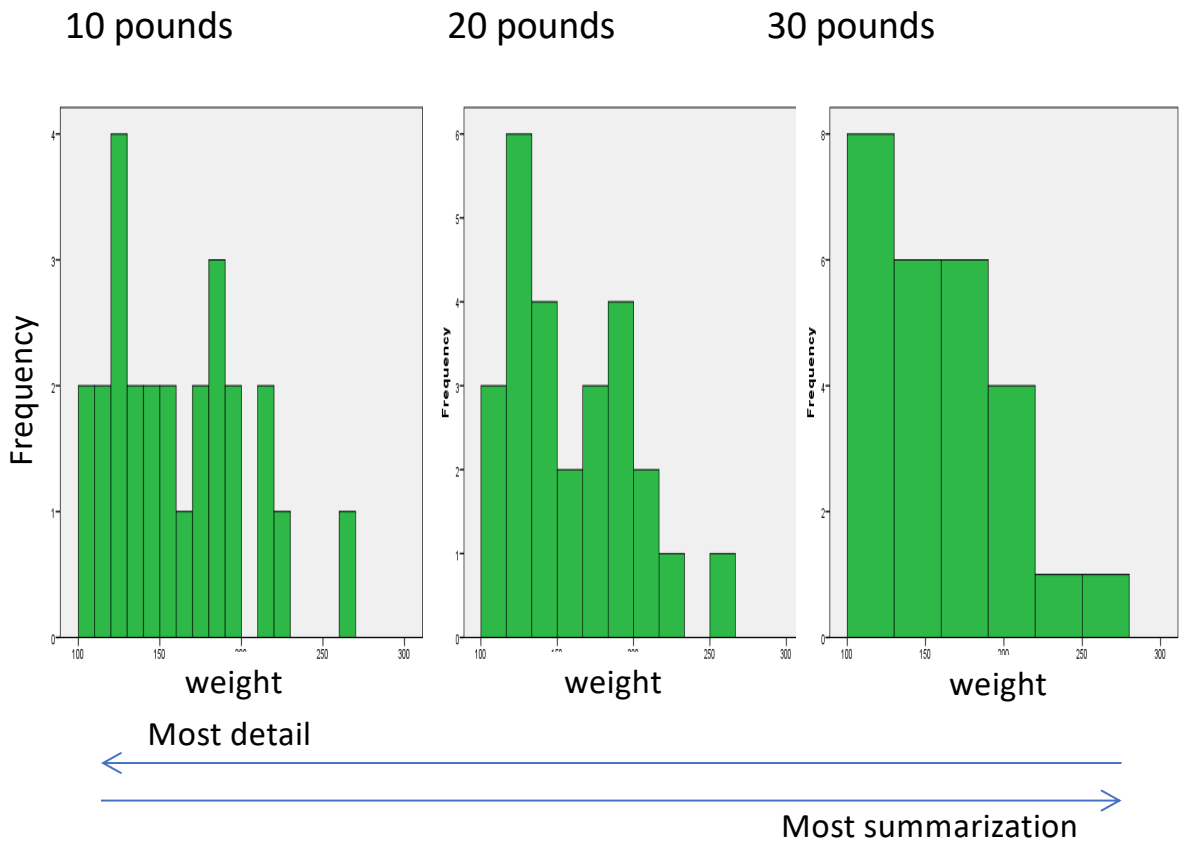
Above, some “binning” has been done, making this like a histogram

Histograms



- A **Histogram** is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable

Histogram: effect of the bin size



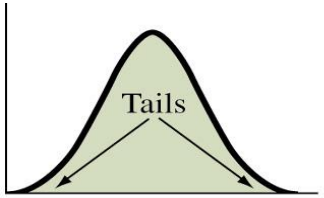
Interpreting Histograms

- Look for **center, spread, and shape:**
  - Assess where a distribution is **centered** by finding the median (half of the data points are above the median, half are below)
  - Assess the **spread** of a distribution
  - Assess the **shape** of a distribution:
    - roughly symmetric?
    - skewed to the right?
    - skewed to the left?

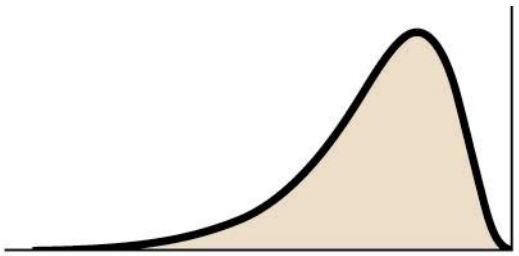
Shape: skew

- *Symmetric Distributions*:

left and right sides of the histogram are mirror images

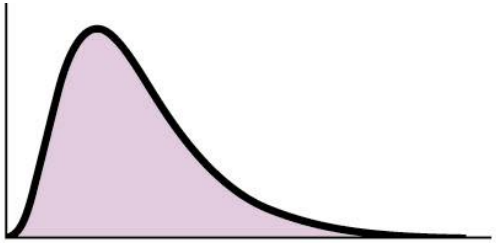


Symmetric



Skewed to the left

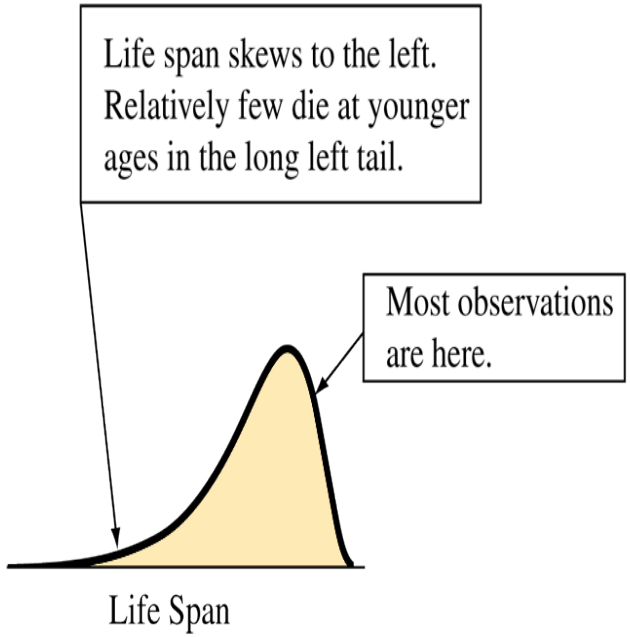
A distribution is skewed to the left if the left tail is longer than the right tail

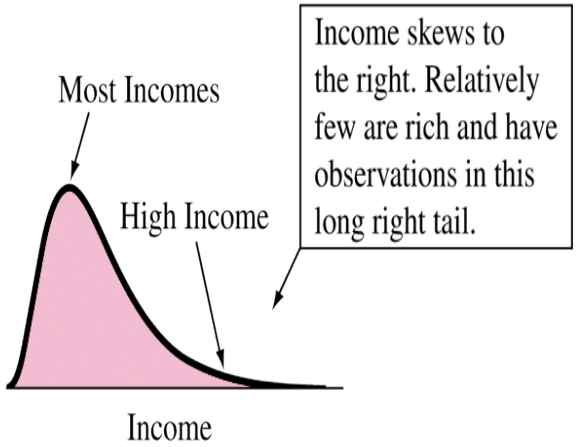


Skewed to the right

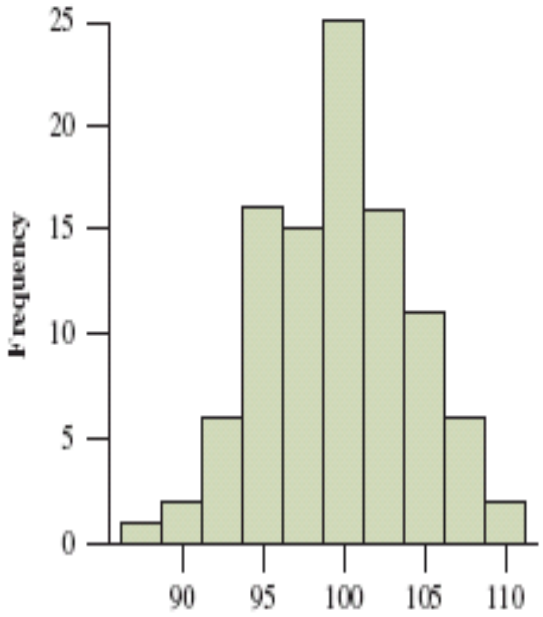
A distribution is skewed to the right if the right tail is longer than the left

Examples of Skewness

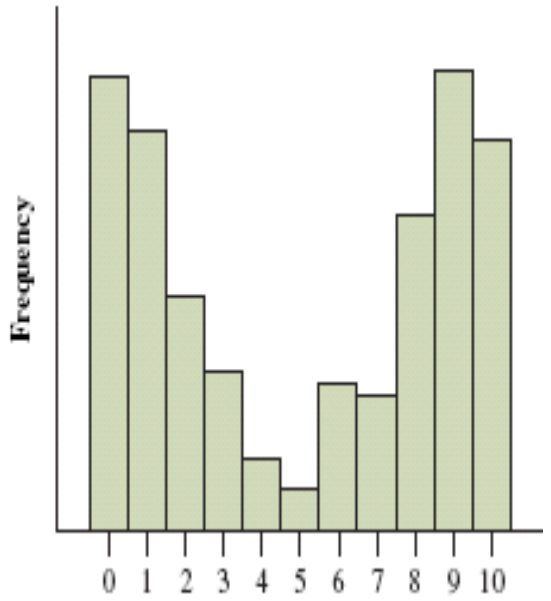




Shape: Modality



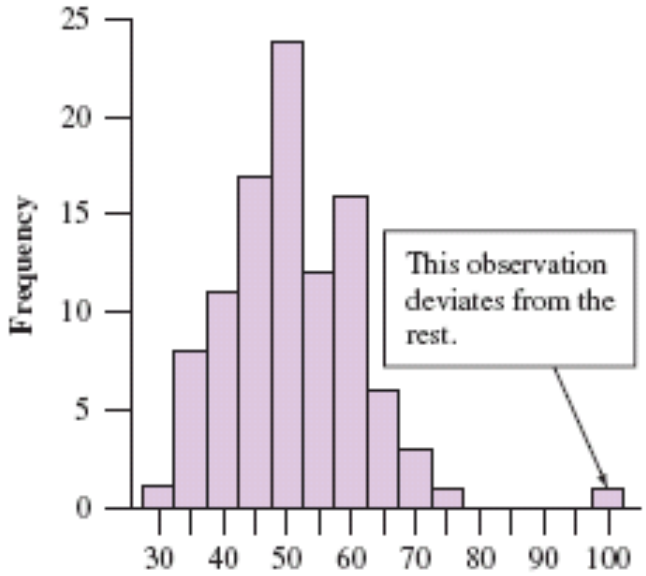
Unimodal



Bimodal

Outliers

- An **Outlier** falls far from the rest of the data:



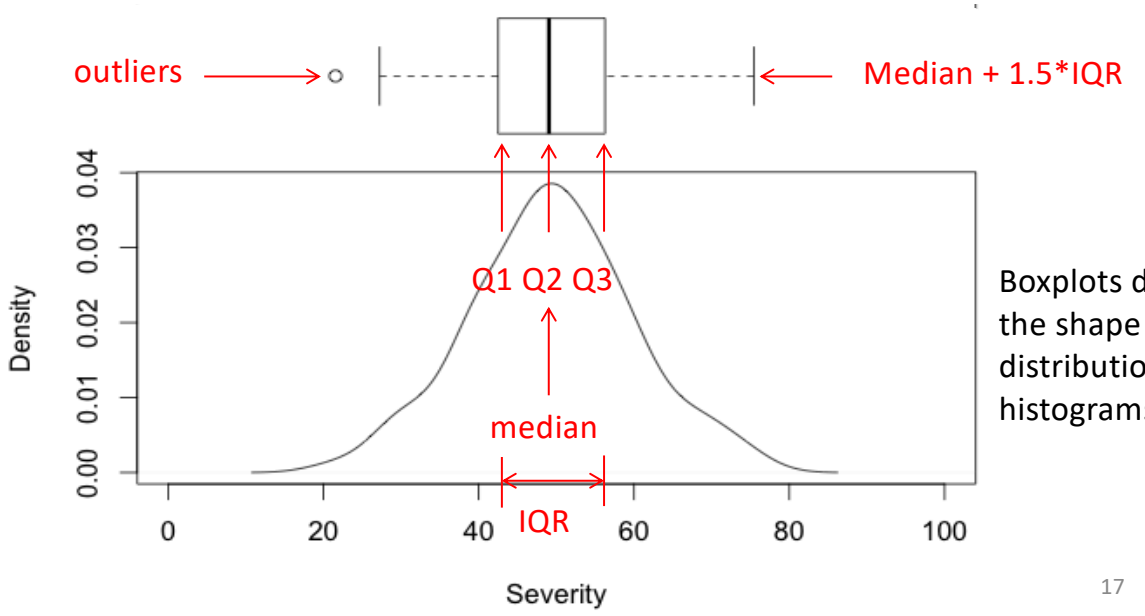
Outliers deserve special attention

It may or may not be justifiable to remove outliers

In general we are hesitant to remove any data, unless there is a very good reason

Boxplot for a single distribution

Percentile = % of data below a given value
Quartile = 25, 50, 75th percentile
Median = 50th percentile
IQR = Interquartile Range = 25th – 75th percentile



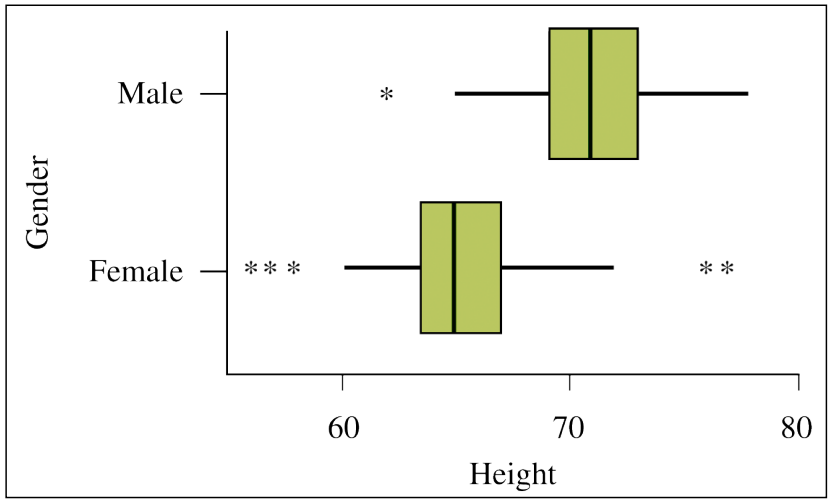
Boxplots do not display the shape of the distribution as clearly as histograms

17

Boxplot for Comparing Distributions

However boxplots are useful for making graphical comparisons of two or more distributions.

They are useful for identifying potential outliers



Numerical summaries of quantitative variables

- Mean, median, mode
- Quartiles, percentiles
- Range
- Standard deviation

Mean

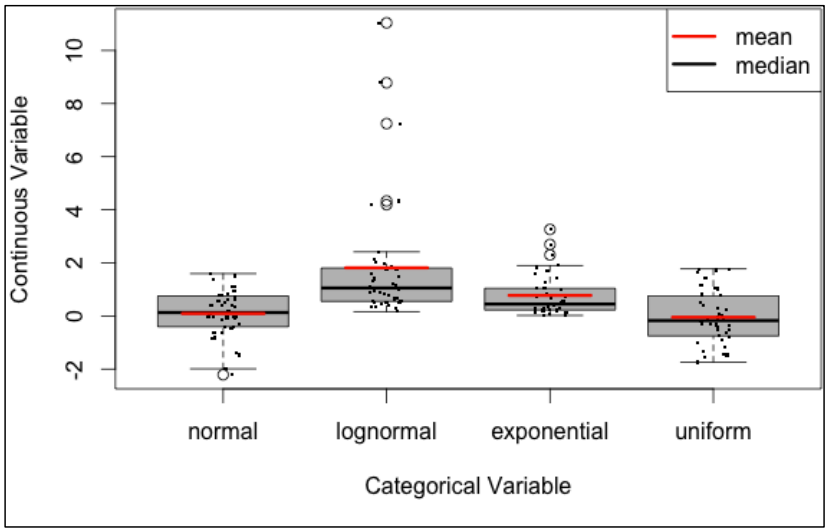
- The mean is the sum of the observations divided by the number of observations
- E.g., add up how much each person paid for lunch, divide by the number of people

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

(capital sigma notation
or summation notation)

Mean II

- The mean is not ***resisitant***.
- *It is strongly influenced by outlying values*



Median

- The **median** is the midpoint of the observations when they are ordered
- If the number of observations is:
  - Odd, then the median is the middle observation
  - Even, then the median is the average of the two middle observations
- Median is the **50^{th} percentile**

Median

- The median is the midpoint of the observations when they are ordered from the smallest to the largest (or from the largest to smallest)
- 50^{th} percentile, 2^{nd} quartile
- Examples:
  - 1, 2, **3**, 4, 5
  - 1, 2, **3**, 4, 1000000000
  - 1, 2, 3, 4: median = **2.5**

Median height of the 7 Dwarfs

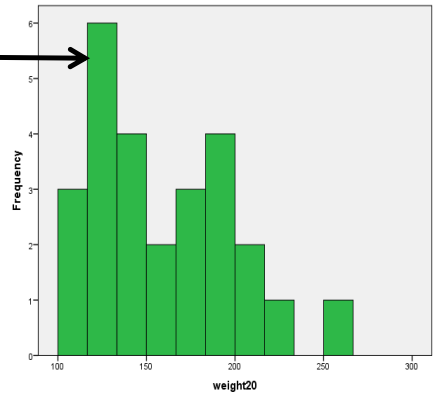
when the 7 dwarfs are arranged on order of height,
the median dwarf is in the middle (#4)

Median *is* resistant to outliers

Mode

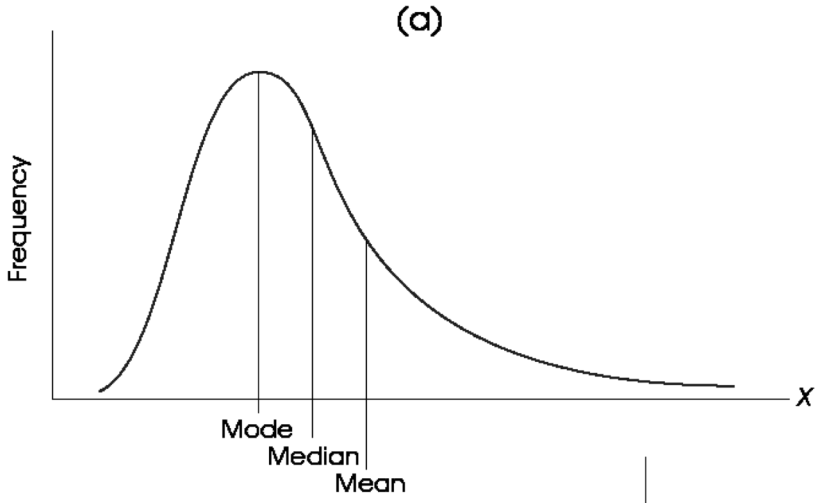
- Mode is the value that occurs most often
  - The mode is most often used with categorical data
  - The mode is the highest bar in the histogram or barchart

mode

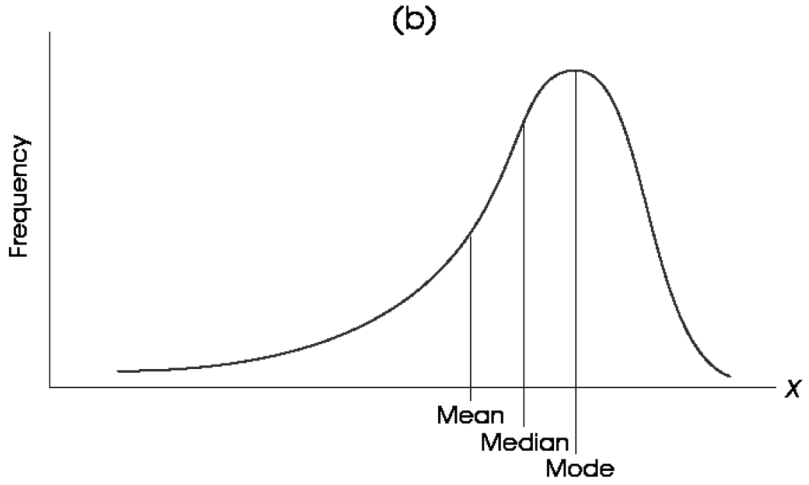


Note that for a histogram, the mode changes with the width of the bins

Central Tendency & Skewness

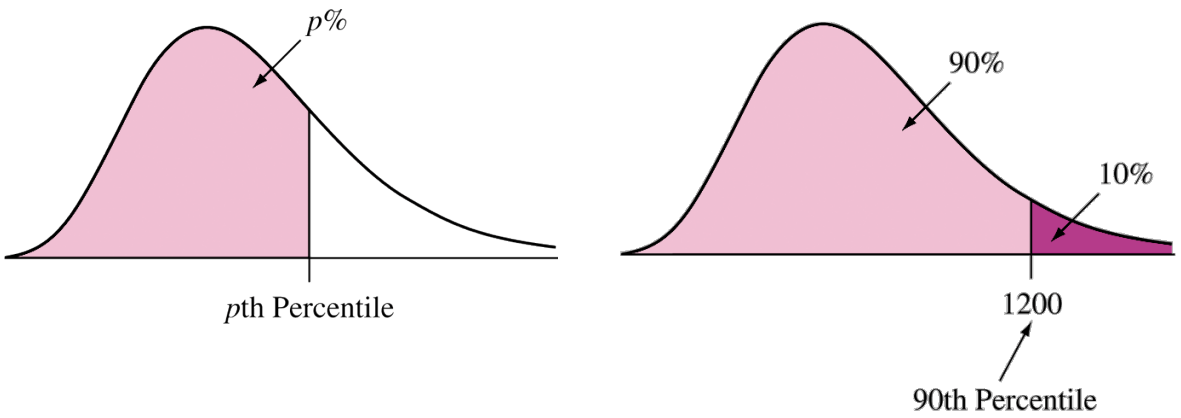


Can infer skewness from differences between mean and median



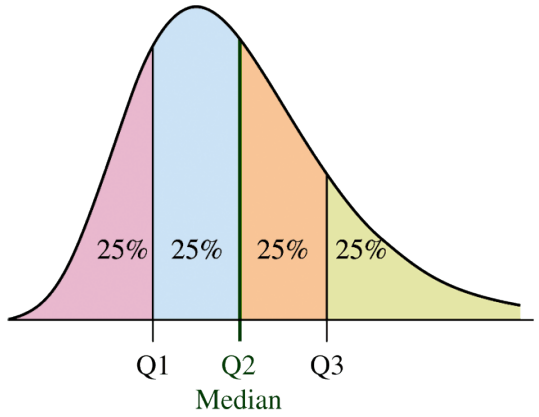
Percentile

- The **p^{th} percentile** is a value such that p percent of the observations fall <u>below or at that value</u>



Quartiles

- Quartiles split the data into four equal parts
  - The median is the second quartile, Q2. 50% of observations are below, 50% are above Q2.
  - The first quartile, Q1, is the median of the lower half of the observations. 25% of observations are below Q1.
  - The third quartile, Q3, is the median of the upper half of the observations. 75% are below.

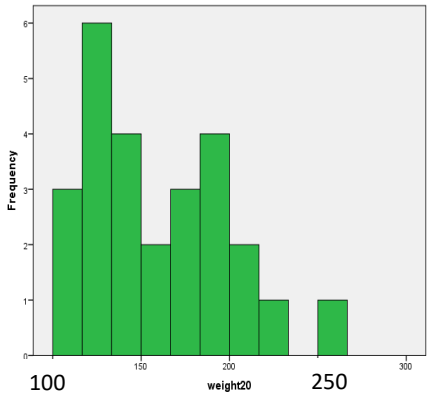


Range

- **Range** is one way to measure the spread of a quantitative variable
- The range is the difference between the largest and smallest values in the data set;

Range = max – min

Range ~ 250 – 100 = 150



Note that range is strongly affected by outliers

Variance and Standard Deviation (SD)

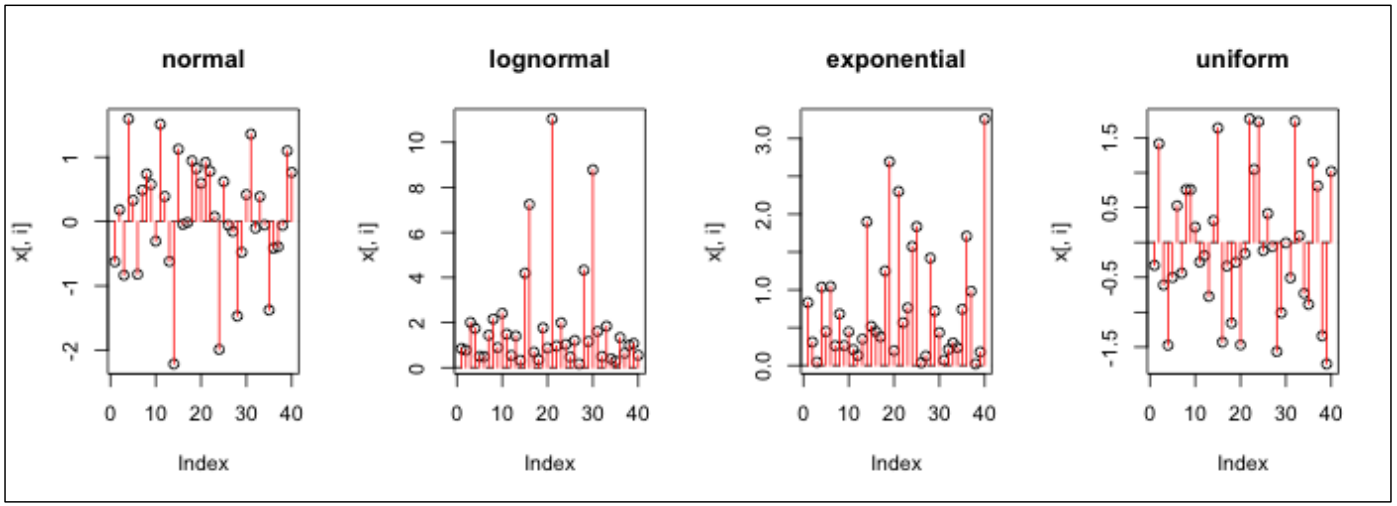
- Gives a measure of variation by:
  - Calculating squared *deviations* of each observation from the mean
  - Calculating the *adjusted mean* of squared deviations

- Variance is the square of standard deviation

n-1 for a *sample*
n for a *population*

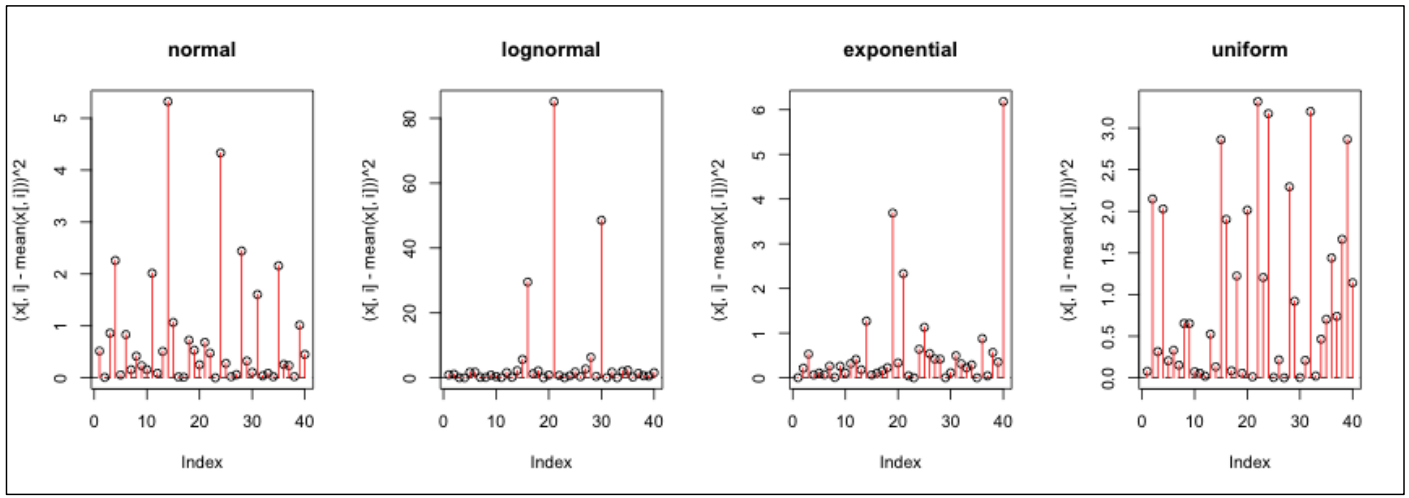
$$\text{var} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Deviations from mean: $x - \bar{x}$



$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Squared deviations: $\left(x - \bar{x}\right)^2$

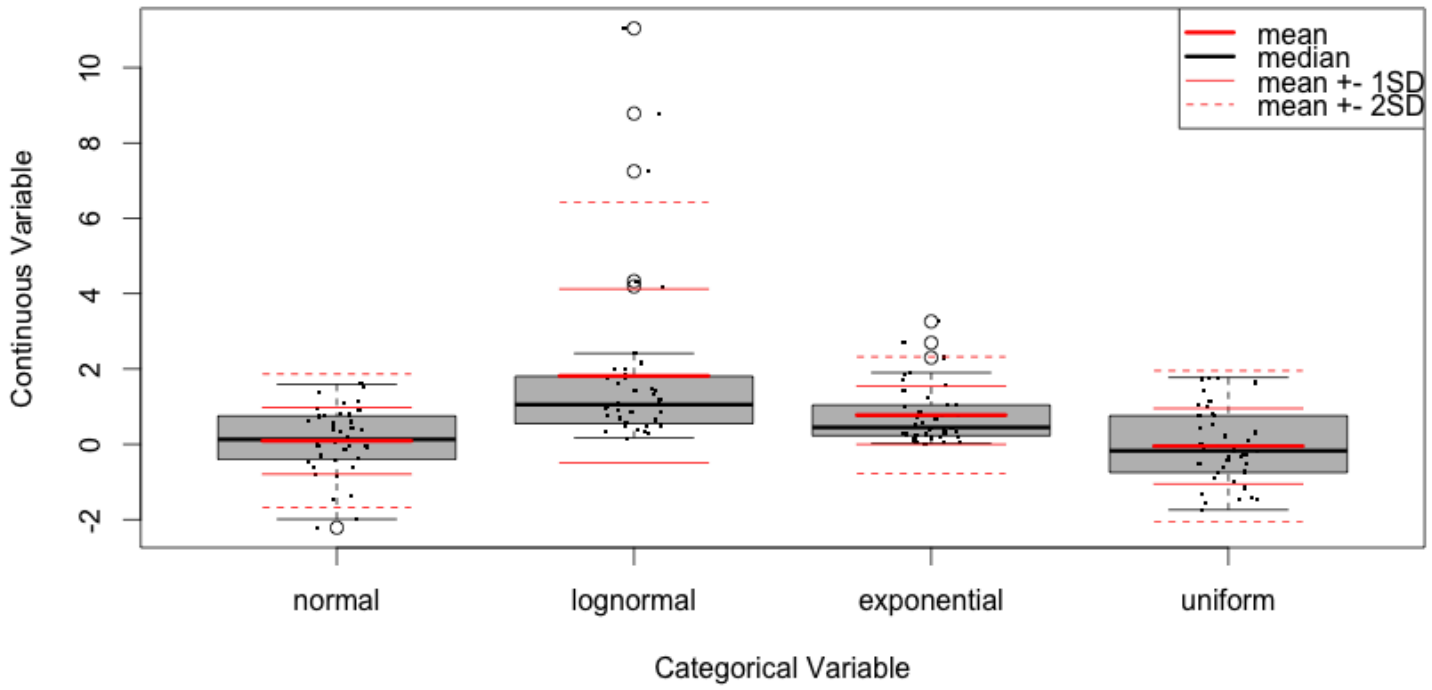


- SD not robust to outliers/skew, but convenient for describing normal distributions

Standard Deviation

- Has the same units as the observed variable
- s measures the spread of the data
- s = 0 only when all observations have the same value, otherwise s > 0. As the spread of the data increases, s gets larger.
- s is not *resistant*. Strong skewness or a few outliers can greatly increase s.

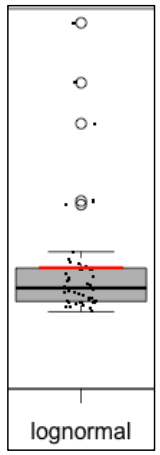
Standard Deviation



Alternative measures of spread

* Interquartile Range (IQR):
  - Q3 – Q1 (75^{th} percentile minus 25^{th} percentile)
* Median Absolute Deviation (MAD):
  - median(x – median(x))
- Range:
  - max(x) – min(x)

***robust to outliers and skew**



Vocabulary Summary

- Dot plot, histogram, box plot
- Mean, Median, Mode
- Percentiles, quartiles, Q1, Q2, Q3
- Standard deviation, IQR, MAD, range
- Capital-Sigma (summation) notation
- Unimodal, bimodal distributions
- Skew, left skew, right skew
- Outliers

Learning Summary

- Define and calculate Mean, Median, Mode, Range, Standard Deviation
- Create histograms, boxplots, barplots, frequency tables
- Identify skewed and symmetric distributions
- Identify outlier data points