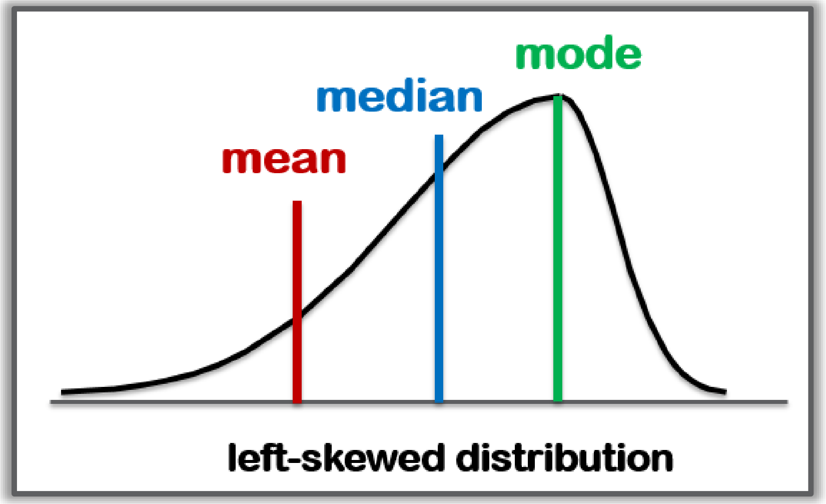
Univariate and Bivariate EDA: Central Tendency



- Central tendency (location measures):
- Means: arithmetic/harmonic/geometric mean; trimmed mean; weighted mean
- Medians: median, weighted median
- Wilcoxon rank-sum test to compare medians vs. t-test to compare means

Prof. Nagiza F. Samatova
samatova@csc.ncsu.edu
Department of Computer Science
North Carolina State University

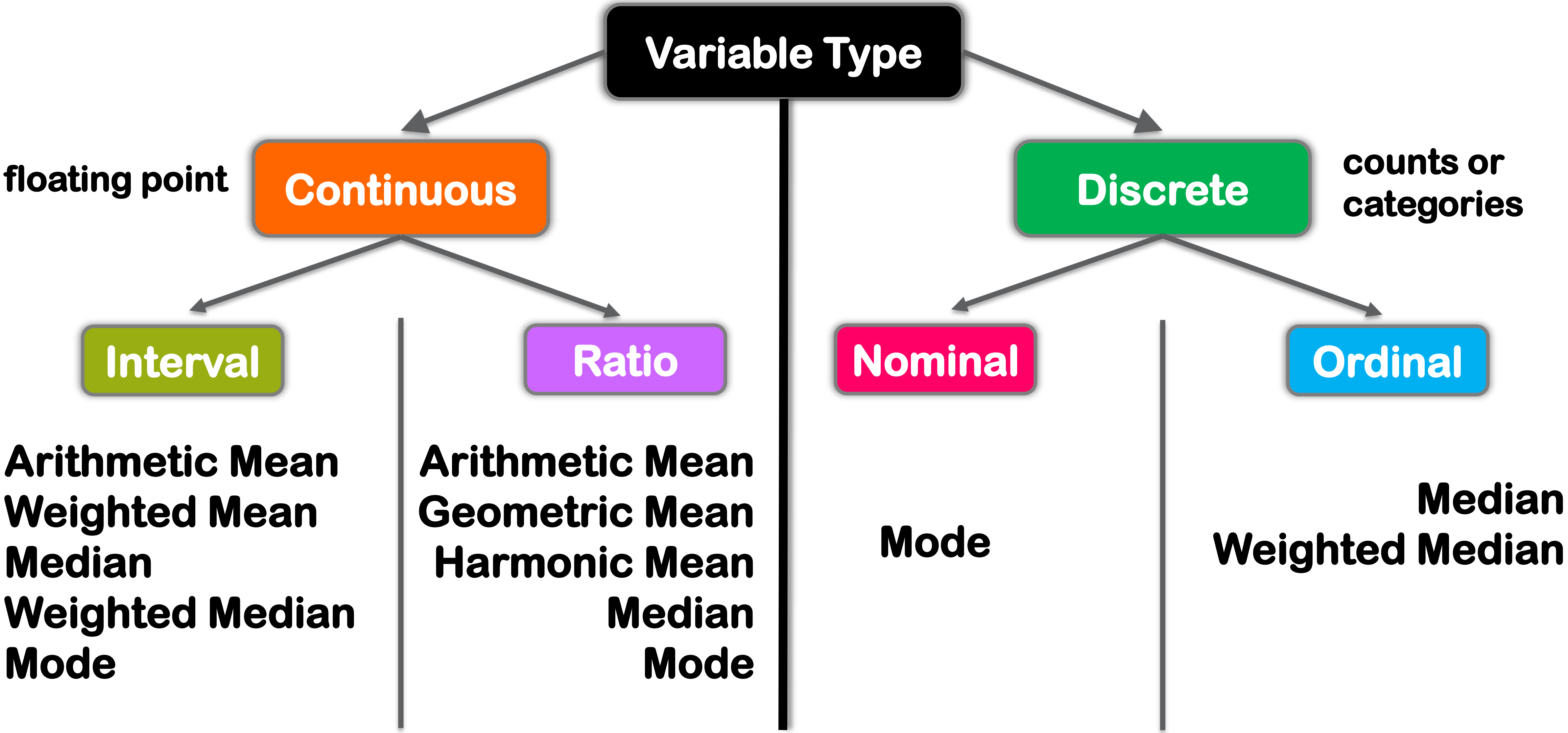
NC STATE Executive Education

February 2019

Quantitative vs. Binary / Categorical

Feature	Description	Example	Statistical Operation	Discrete vs. Continuous
Nominal	values are different names: provide enough info to distinguish one object from another (=, ≠)	zip codes, employee ID, eye colors, sex:{male, female}	**mode**, contingency, entropy, χ^2-test	discrete
Ordinal	values provide enough info to order objects (<, >)	grades (A, A-, B, B+) size (small, medium, large)	**median**, percentiles, rank correlation, run tests, sign tests	discrete
Interval	the differences between values are meaningful: allow ordering and subtraction but not other arithmetic operations	calendar dates, time, temperature in Celsius	**median, mean**, standard deviation, Pearson's correlation, t- and F-tests	both
Ratio	both differences and ratios are meaningful (*, /)	monetary quantities, counts, age, length, temperature	**mean, median, geometric mean, harmonic mean**, percent variation	continuous

Centrality: Continuous vs. Discrete Variable



Centrality Tendency: Location Measures

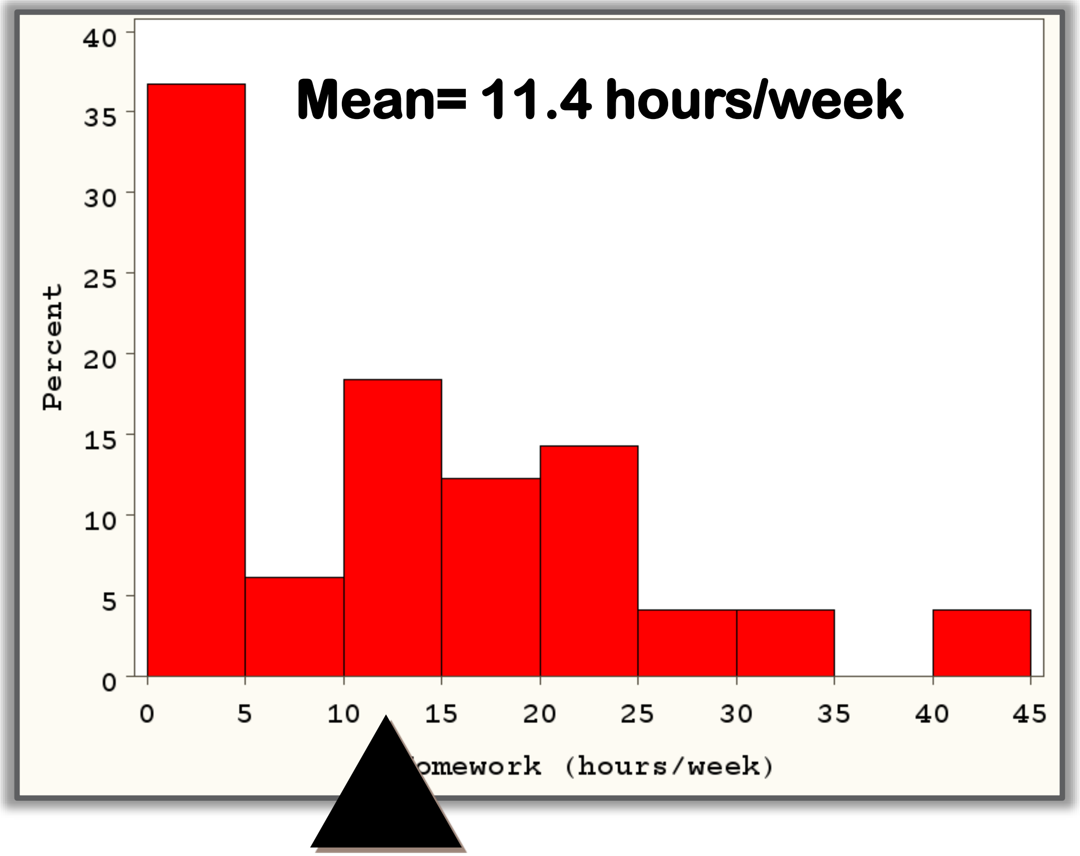
Measure	Description	Synonyms
Mean	sum of all values divided by the number of values	arithmetic mean, average
Weighted Mean	sum of all values times a weight divided by the sum of the weights	weighted average
Median	value such that one-half of the data lies above and the other-half lies below	50th percentile
Weighted Median	value such that the sum of the weights is equal for the lower and upper halves of the sorted list of data values	
Trimmed Mean	average of all values after dropping a fixed number of extreme values	truncated mean
Robust	not sensitive to extreme values, or outliers	resistant
Outlier	data value that is very different from most of the data	extreme value
Mode	the most frequently observed value in the data	
Geometric Mean	characteristic of the average growth rate between positive values	
Harmonic Mean	characteristic of an average rate	

Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

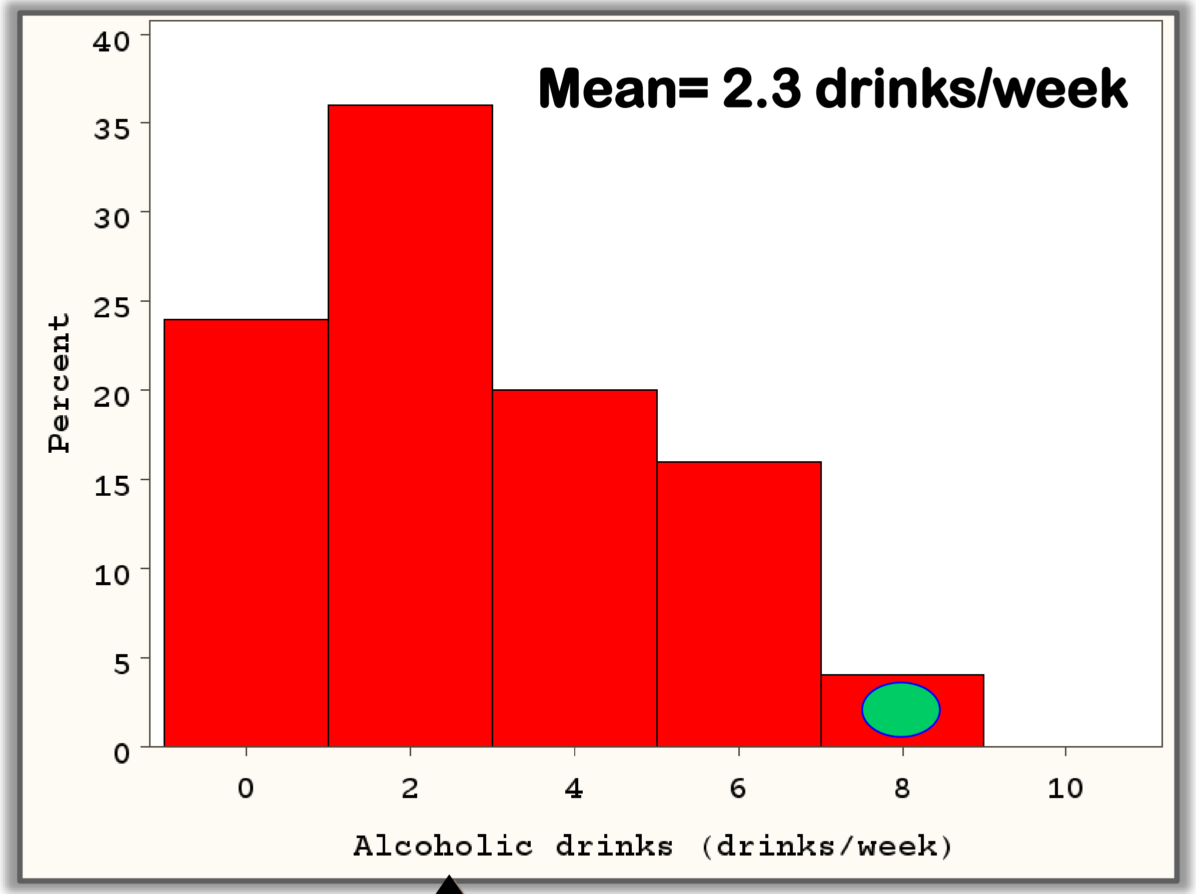
- **Mean** – **the arithmetic average; the balancing point**
  - *calculation:* the sum of values divided by the sample size
- **Example: Participant Age**
  - Data: 17 19 21 22 23 23 23 38

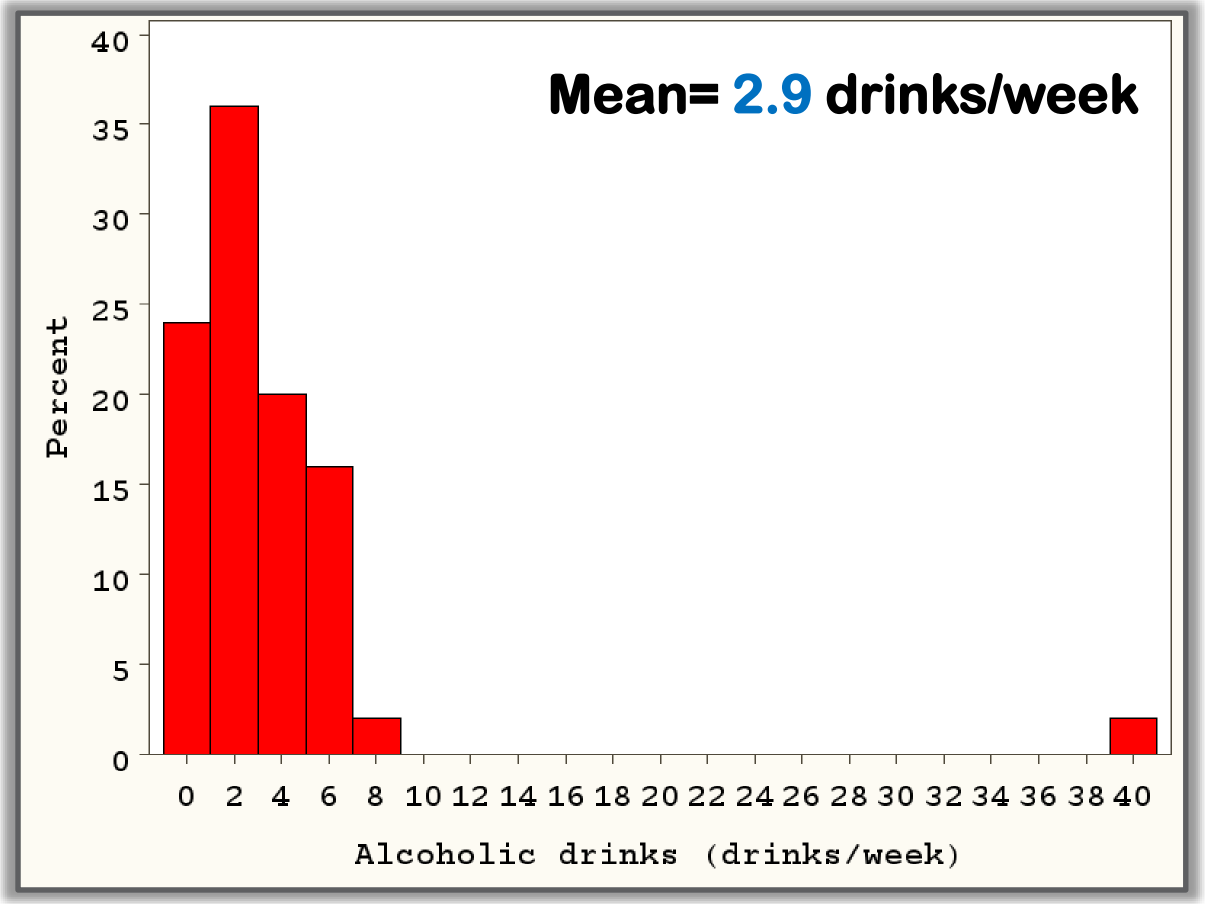
$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{17+19+21+22+23+23+23+38}{8} = 23.25$$

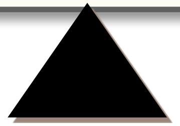


The balancing point

Mean: Affected by Extreme Values…







The balancing point

Mean vs. Trimmed Mean

- **Trimmed mean**
  - eliminates the influence of extreme values
  - widely used and preferable to use instead of the ordinary mean
  - ex.: to eliminate the influence of a single judge to manipulate the score at the competition

scores:	4	1	4	9	5	4	**mean = 4.5**
sorted:	1	4	4	4	5	9	
trimmed:		4	4	4	5		**trimmed mean = 4.2**

Weighted Mean

$$\bar{x}_{\mathrm{w}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Reason #1:**
  - Some values are more variable than others
    - Highly variable observations are given a lower weight
  - Ex.: the average from multiple sensors and one sensor is less accurate → down-weight the data from that sensor
- **Reason #2:**
  - The data collected does not equally represent the different groups
  - To correct for group representation bias, a higher weight might be given to the values from the underrepresented groups

Median and Robust Estimate

- **Median**
  - the middle number on a sorted list of the data
  - if the sample size is even, then the middle value is NOT in the data set but the average of the two values that divide the sorted data into upper and lower halves
  - If there are an odd number of observations, find the middle value

scores: 4 1 5 9 5 4

mean = 4.7

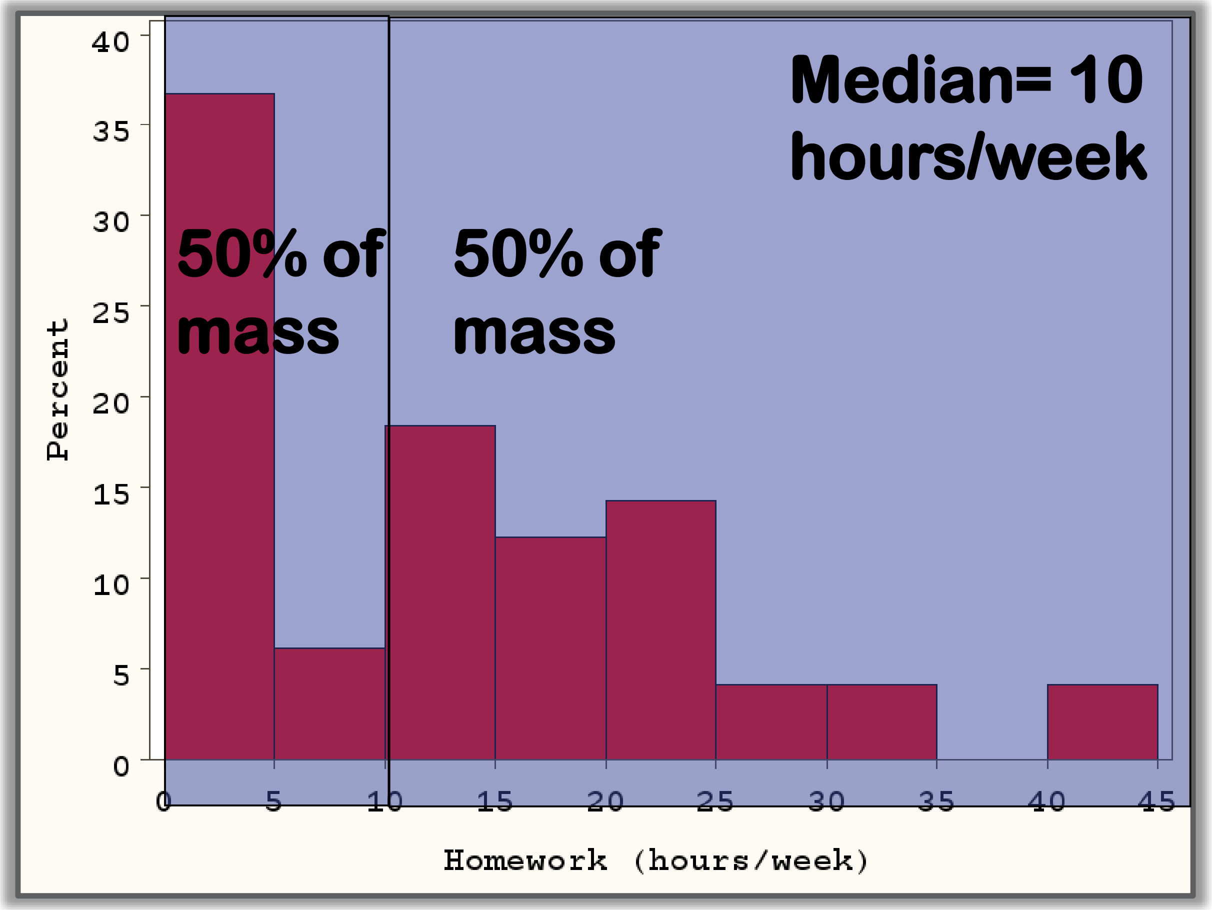
sorted: 1 4 4 | 5 5 9

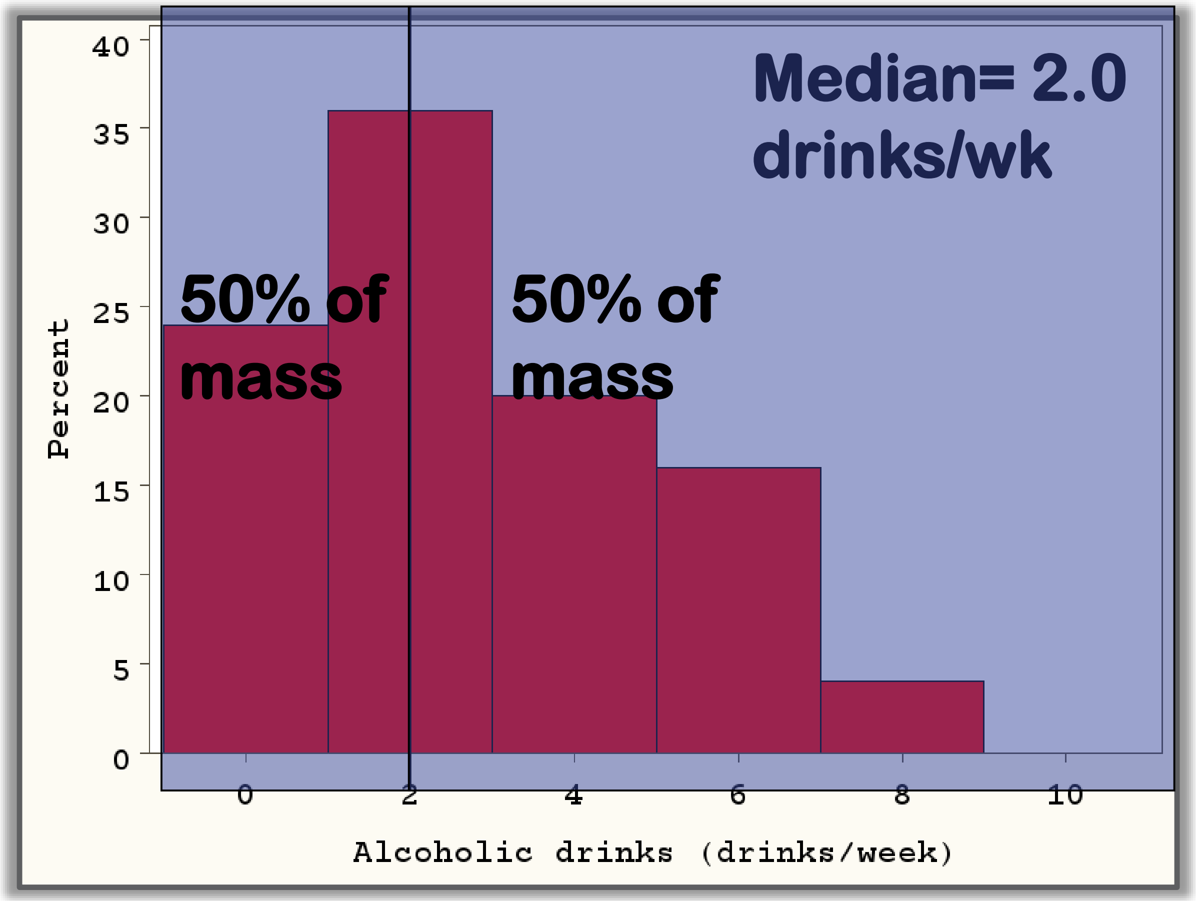
lower half | upper half

median = 4.5

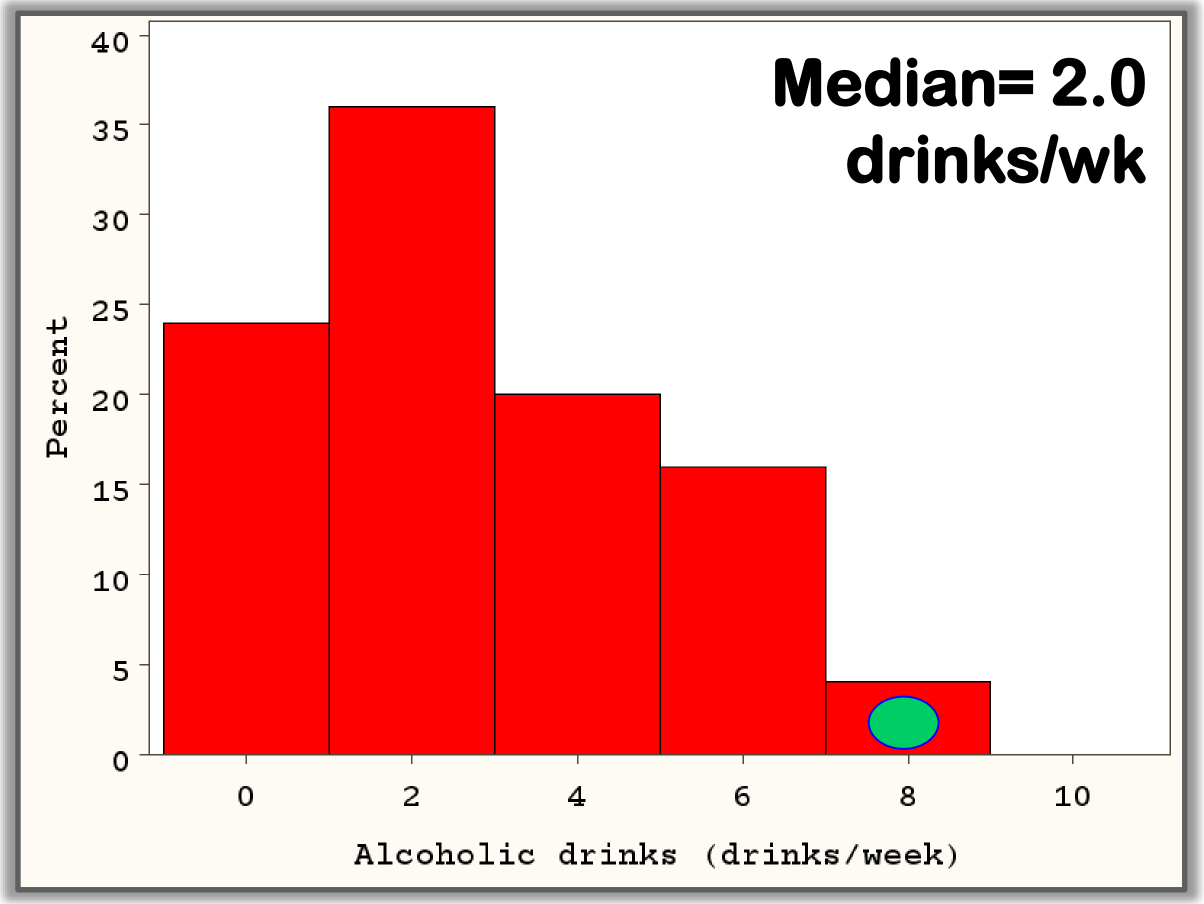
$$\frac{4+5}{2} = 4.5$$

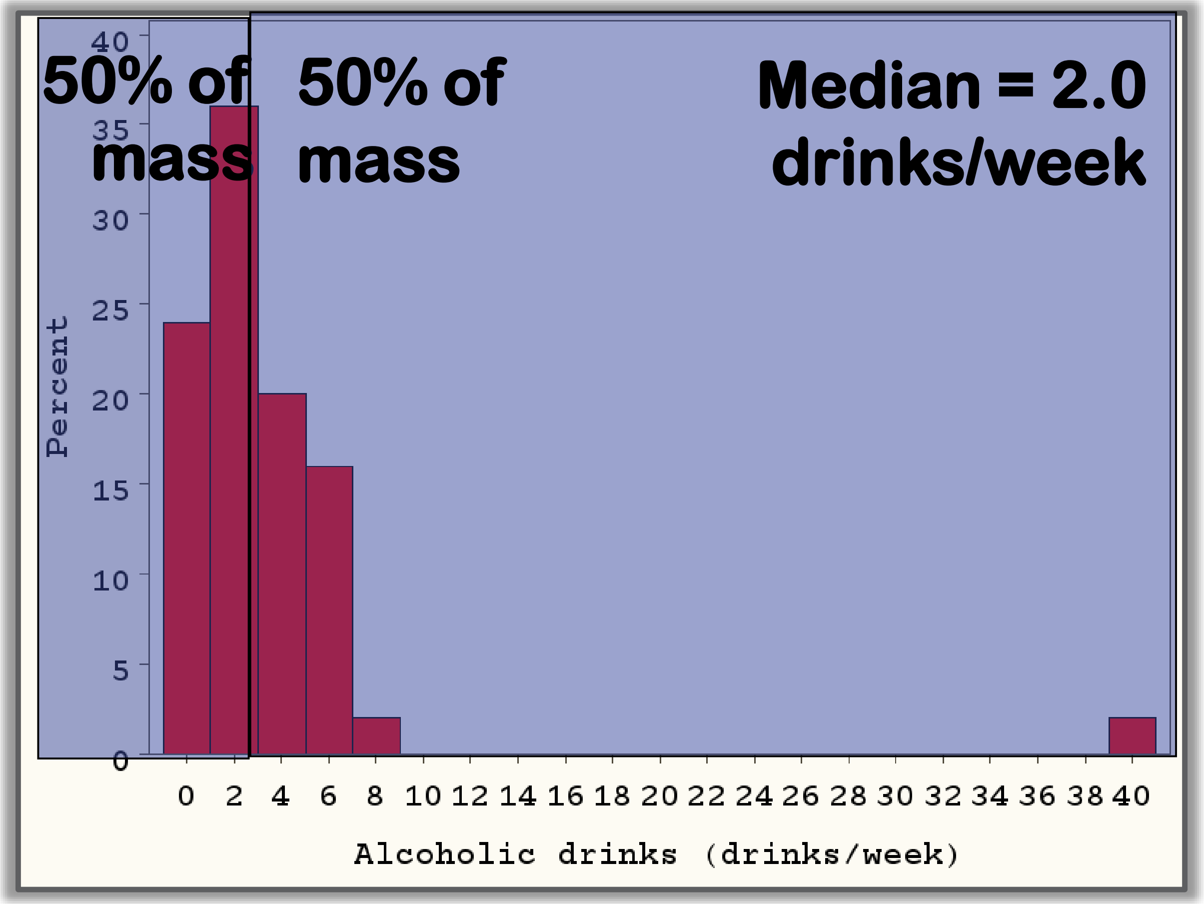
Examples: Median





Median: NOT Affected by Extreme Values





Outliers

- **Outlier (or extreme case)**
  - any value that is very different from the other values in the data
  - subjective definition
- **Conventions on what constitutes an outlier**
  - tails (the outer range) of the distribution
  - e.g., one-percenters: people in the top 99th percentile of the wealth
- **Reasons for outliers**
  - The result of data errors
    - mixing data if different units (kilometers vs. meters)
    - bad reading from a sensor
- **How to handle outliers**
  - Examine and identify the true reason
  - Trim (10% top and bottom values) when estimating the statistics: trimmed mean

Mean vs. Median vs. Trimmed Mean

- **Mean**
  - uses all the observations
  - much more sensitive to data
  - very sensitive to outliers or extreme values
- **Median**
  - depends only on the values in the center of the sorted data
  - robust estimate because it is NOT influenced by outliers (extreme cases)
  - Example: Typical household income
    - Mean of the neighborhood where Bill Gates lives will be influenced by his income
    - Yet, median will not depend on how rich Bill Gates is: the position of the median will remain the same
- **Weighted Median**
  - Robust to outliers, like the median but still depends only on the values in the center
- **Trimmed Mean**
  - Compromise between the mean and the median:
    - it is robust to extreme values in the data
    - but uses most of the data to calculate the estimate for location

Example: Location Estimates of Population & Murder Rates

```
state = pd.read_csv("../data_raw/eda_state.csv")
state.head()
```

	State	Population	Murder.Rate	Abbreviation
0	Alabama	4779736	5.7	AL
1	Alaska	710231	5.6	AK
2	Arizona	6392017	4.7	AZ
3	Arkansas	2915918	5.6	AR
4	California	37253956	4.4	CA

```
state.Population.mean()
```

```
6162876.3
```

```
stats.trim_mean(state.Population,
                proportiontocut = 0.1)
```

```
4783697.125
```

```
state.Population.median()
```

```
4436369.5
```

Population:
mean vs. median

```
np.average(state['Murder.Rate'],
           weights=state.Population)
```

```
4.445833981123393
```

```
ws.weighted_median(state['Murder.Rate'],
                   weights=state.Population)
```

```
4.4
```

Murder Rate:
weighted mean vs. weighted median

Bivariate Exploration

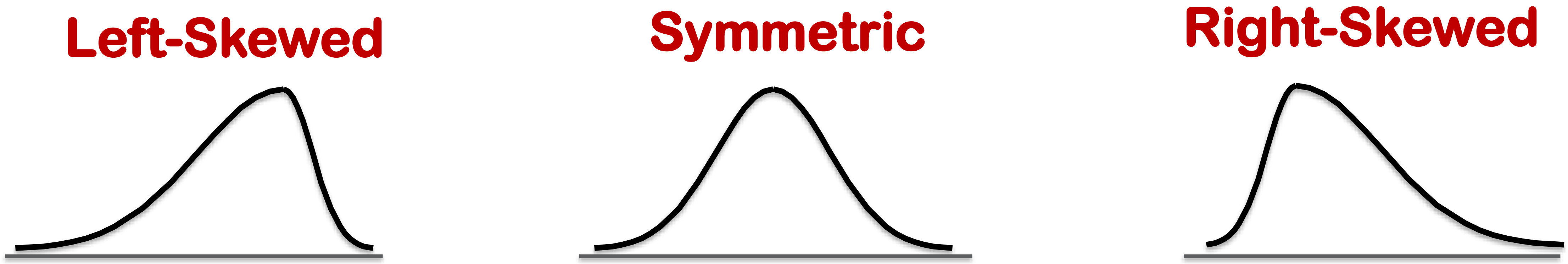
MEAN VS. MEDIAN

Tests: Central Tendency and Variability

Comparison	Groups	Normal or Almost Normal	Not Normal	Binomial (Proportions)	Variances
Compare data within one group to a standard or target value	1	One sample t-test	Wilcoxon Rank-Sum test	One proportion z-test (or exact Binomial test)	Chi-square for one variance
Compare data within two unpaired groups	2	Two sample t-test	Mann Whitney Wilcoxon Rank-Sum test (or U-test)	Two proportions z-test, Chi-square test of independence (or Fisher's exact test if counts in cells <5)	F-test for homogeneity of variances
Compare two paired groups	2	Paired t-test	Wilcoxon Rank-Sum test	McNemar's test	Bonett's test
Compare data among many groups	>2	One-Way ANOVA	Kruskal-Wallis test	Chi-square test of independence (or Fisher's exact test)	Levene's test or Bartlett's test for normal data

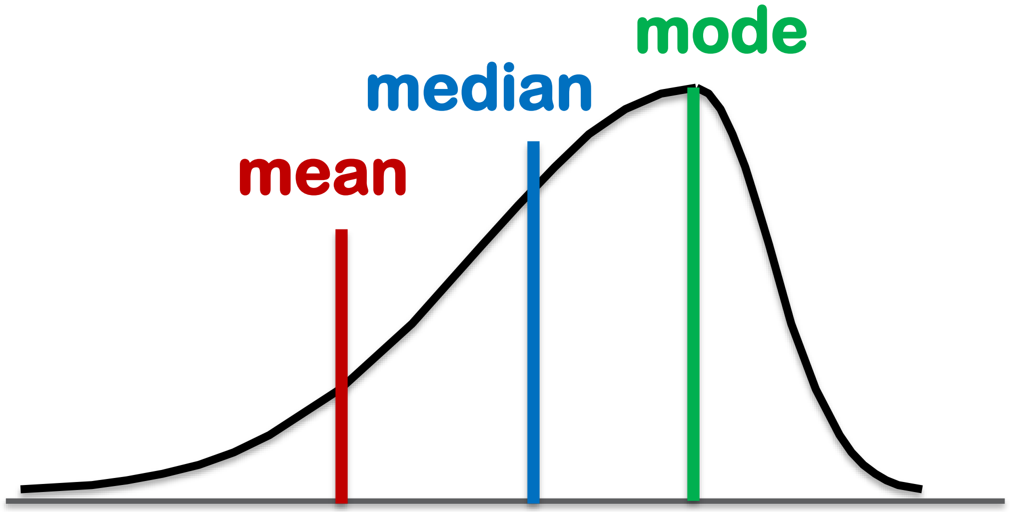
Compare Means or Medians?

- **For skewed data:**
  - the median is preferred because the mean can be highly misleading…
- **The shape of the distribution:**
  - left-skewed
  - symmetric:
    - Bell curve (“normal distribution”)
  - right-skewed



Left-Skewed Location Measures

mean is to the left of the median

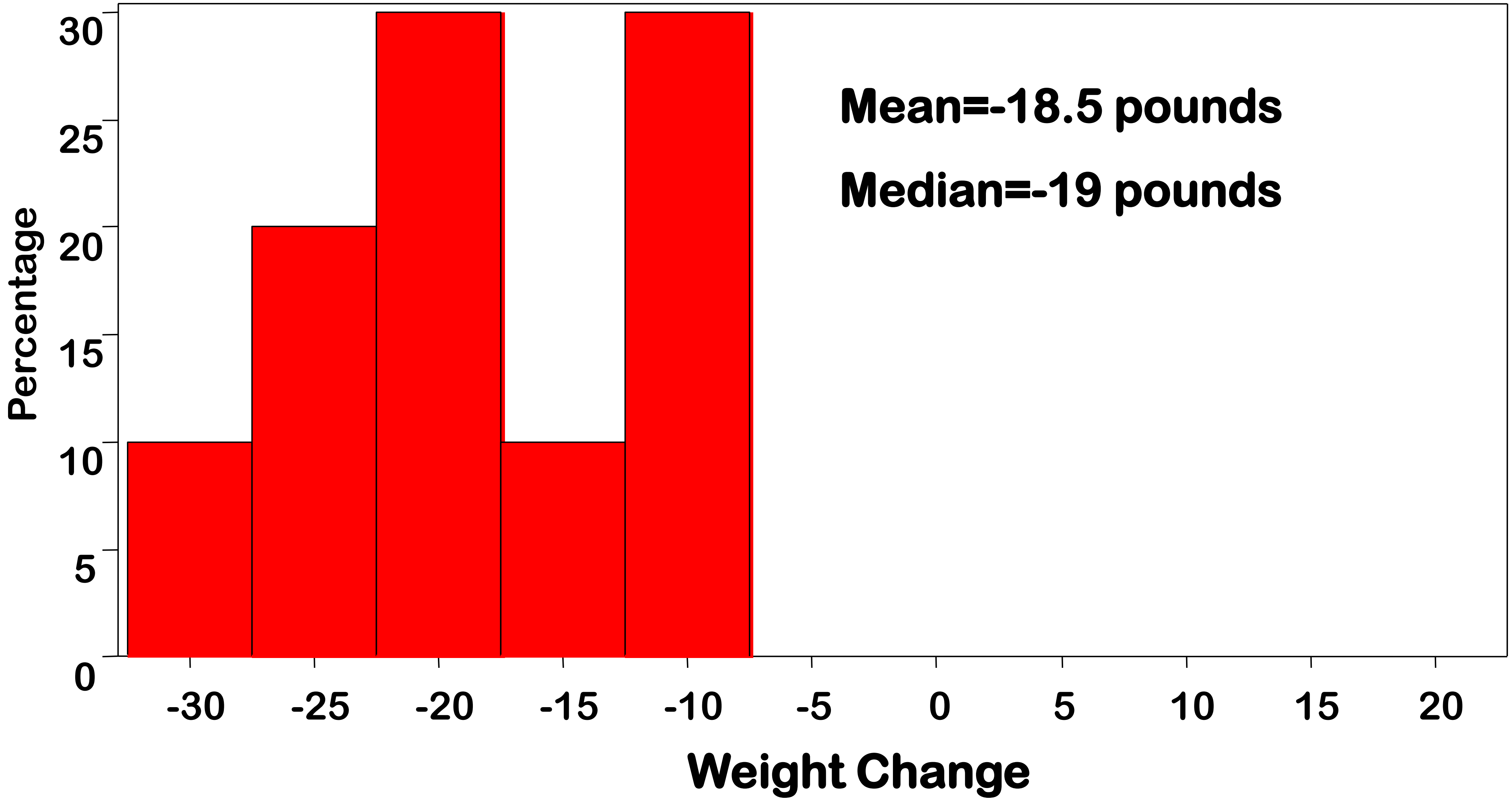


left-skewed distribution

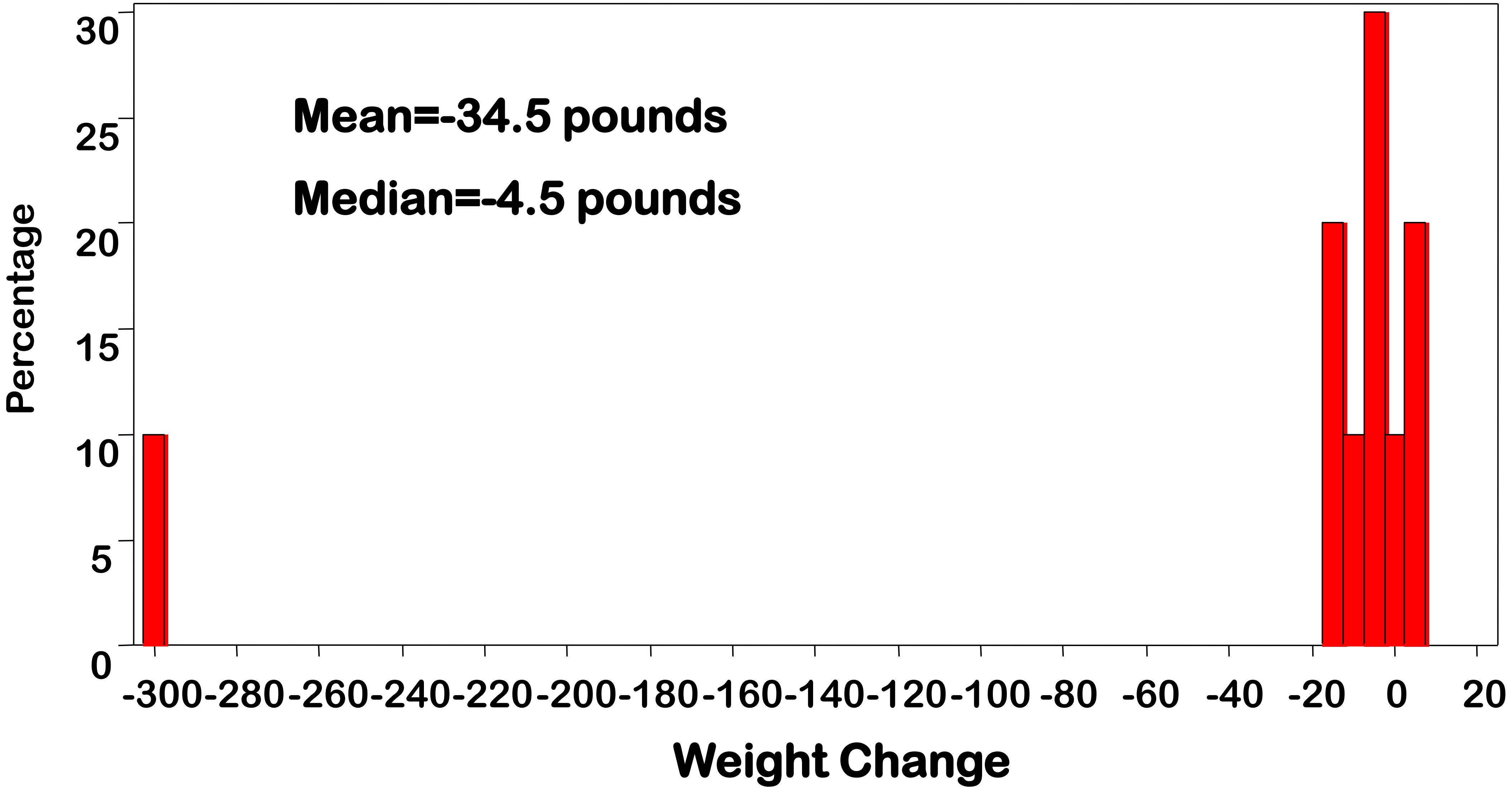
Hypothetical Example: Means vs. Medians…

- 10 dieters following diet 1 vs. 10 dieters following diet 2
- Group 1 (n=10) loses an average of 34.5 lbs.
- Group 2 (n=10) loses an average of 18.5 lbs.
- Conclusion: diet 1 is better?

Weight Change Histogram, Diet 2…



Weight Change Histogram, Diet 1…



Compare Medians via a “Non-parametric Test”

- We need to compare medians (ranked data) rather than means:
  - requires a “non-parametric test”
- Apply the Wilcoxon rank-sum test (wilcox.test())
  - also known as the Mann-Whitney U test
  - non-parametric test

Wilcoxon rank-sum test: Rank the data…Sum the ranks

- **Diet 1, change in weight (lbs):**
  - Weight change: +4, +3, 0, -3, -4, -5, -11, -14, -15, -300
  - Ranks: 1 2 3 4 5 6 9 11 12 20
  - Sum of ranks: 1+2+3+4 +5 +6+9+11+12 +20 = 73

```
diet_1_change  = (+4, +3, 0, -3, -4, -5, -11, -14, -15, -300)
print ("{} : median : diet_1_change".format(np.median (diet_1_change)))
print ("{} : mean : diet_1_change".format(np.mean (diet_1_change) ) )

-4.5 : median : diet_1_change
-34.5 : mean : diet_1_change
```

- **Diet 2, change in weight (lbs)**
  - Weight Change: -8, -10, -12, -16, -18, -20, -21, -24, -26, -30
  - Ranks: 7 8 10 13 14 15 16 17 18 19
  - Sum of ranks: 7+8+10+13 +14 +15+16+17+18 +19 = 137

```
diet_2_change  = (-8, -10, -12, -16, -18, -20, -21, -24, -26, -30)
print ("{} : median : diet_2_change".format(np.median (diet_2_change)))
print ("{} : mean : diet_2_change".format(np.mean (diet_2_change) ) )

-19.0 : median : diet_2_change
-18.5 : mean : diet_2_change
```

```
from scipy.stats import wilcoxon
wilcoxon(x = diet_1_change, y = diet_2_change)

WilcoxonResult(statistic=10.0, pvalue=0.07389705510759269)
```

Diet 2 is superior to Diet 1, p=.007

Summary: Central Tendency: Location Metrics

- The basic metric for location of central tendency is the arithmetic mean
  - can be sensitive to extreme values (outliers)
- Robust estimates of central tendency
  - Median
  - Trimmed Mean