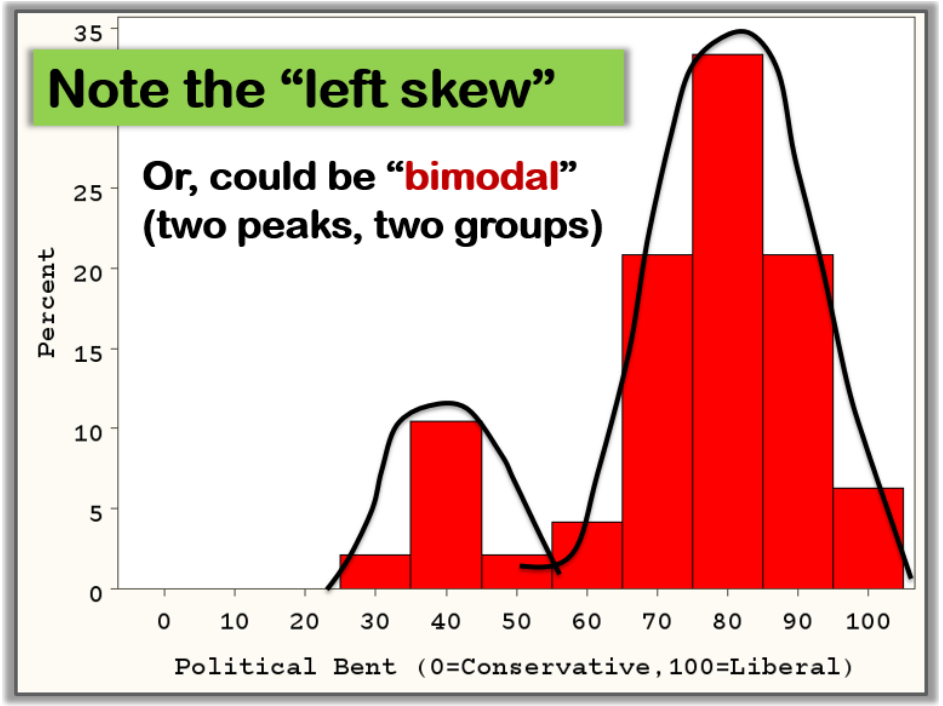
Univariate EDA: Distributions

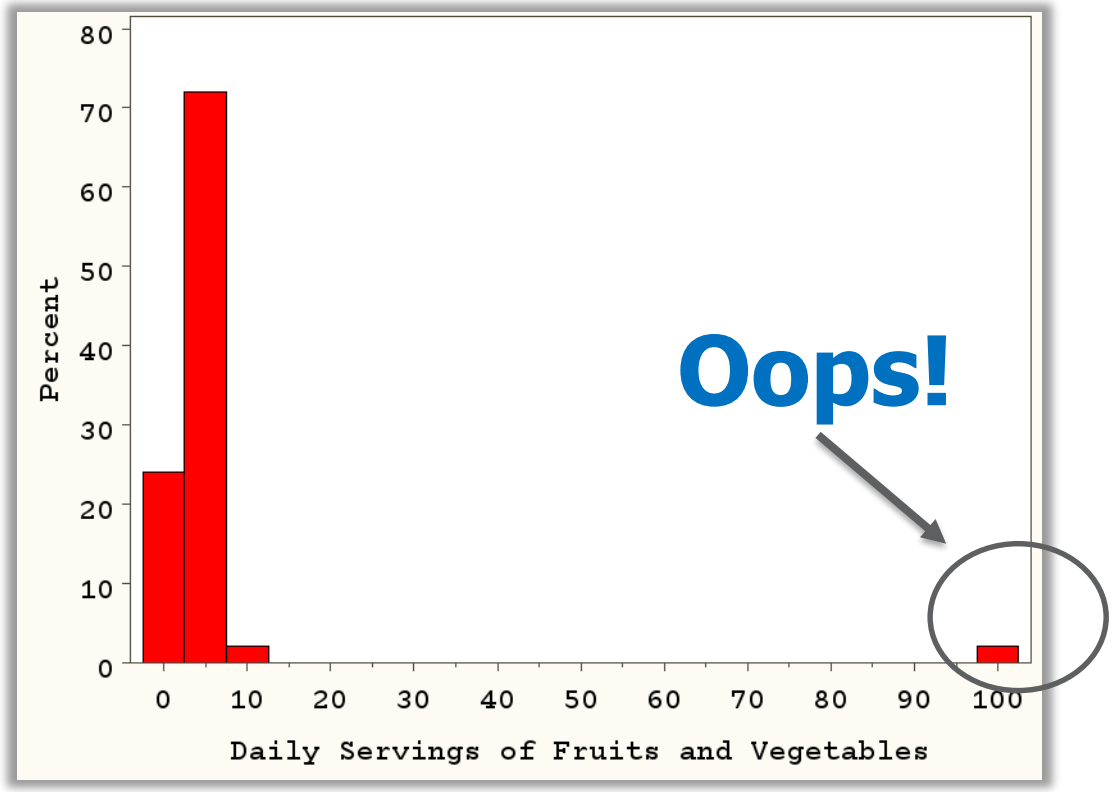


- Frequency plots: bar charts, box plots, histograms, density plots
- Normal distribution
- Bi-modal distribution
- Outliers

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

Always Plot Your Data!

- Are there "outliers"?
- Are there data points that don't make sense?
- How are the data distributed?



Are there points that do not make sense?

How are the data distributed?

- **Categorical data:**
  - What are the N's and percentiles in each category?
- **Quantitative data:**
  - What's the shape of the distribution
    - is it normally distributed or skewed?
  - Where is the center of the data?
  - What is the spread/variability of the data?

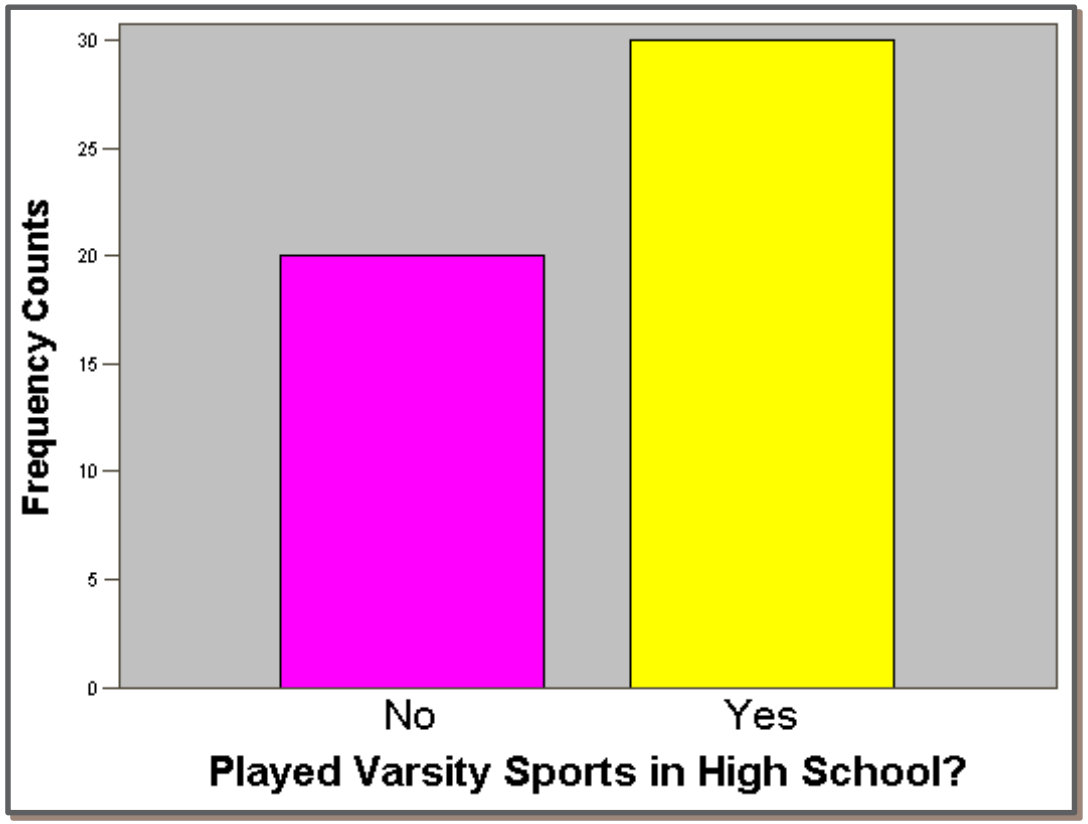
Frequency Plots (univariate)

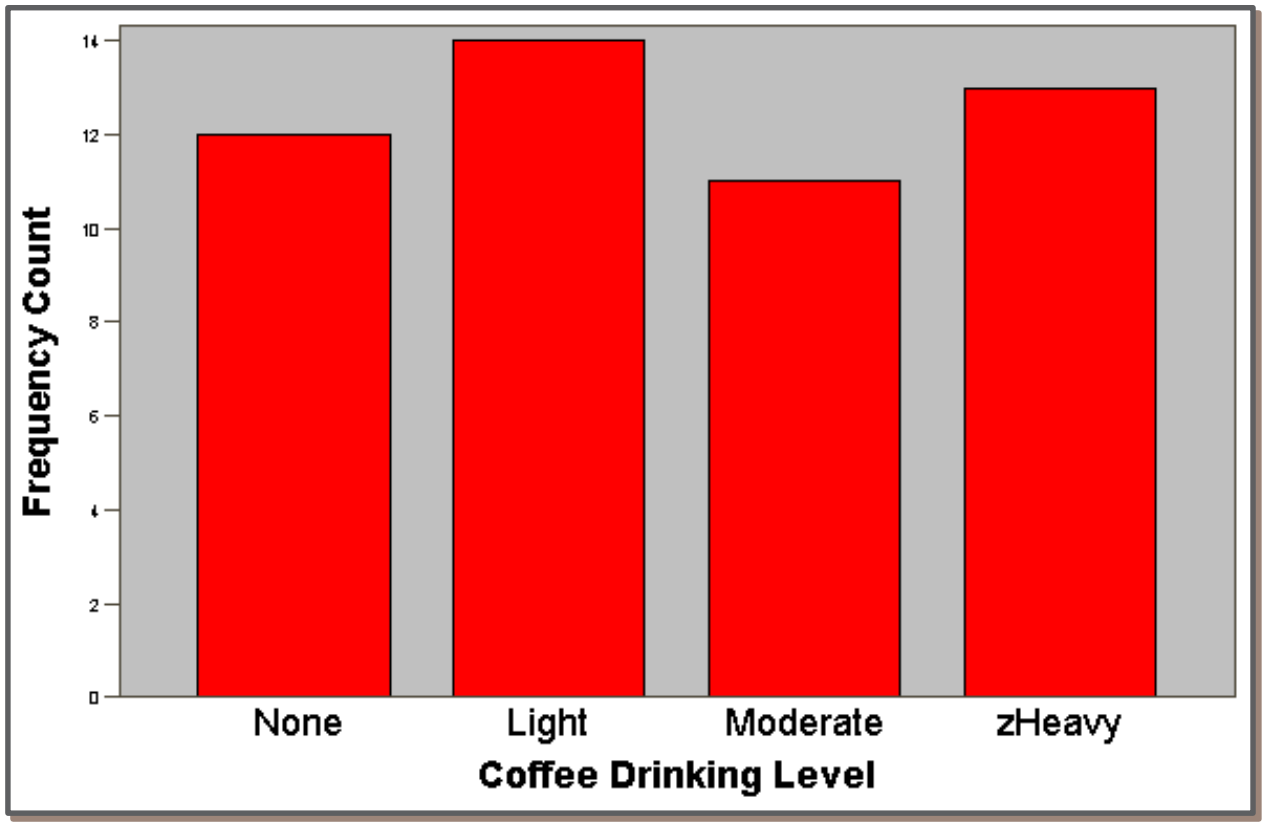
- **Categorical variables**
  - Bar Chart

- **Quantitative / continuous variables**
  - Box Plot
  - Histogram

Bar Chart: Categorical Variables

Show frequency or proportion in each category

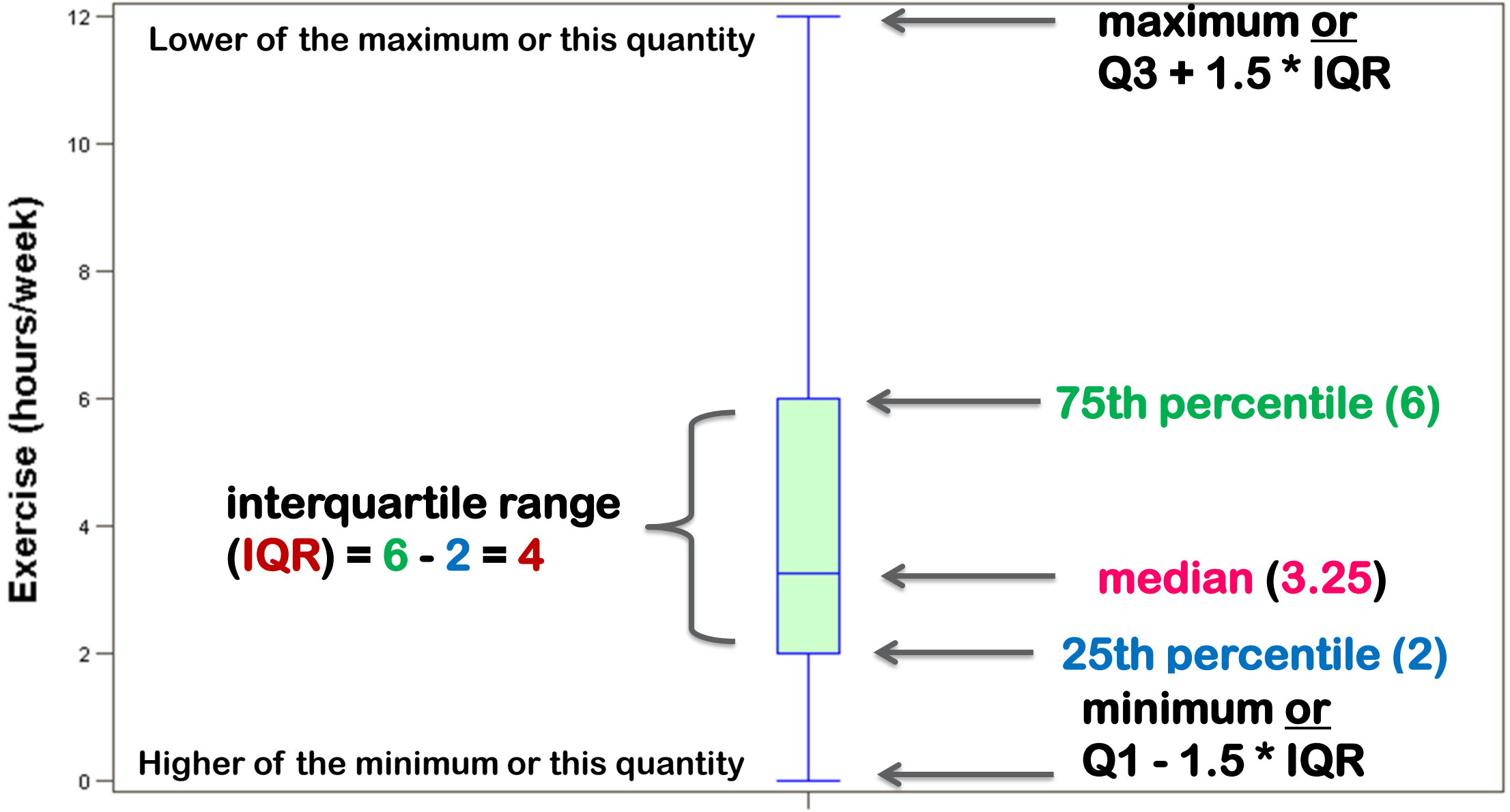




Box Plot and Histograms: Quantitative Variables

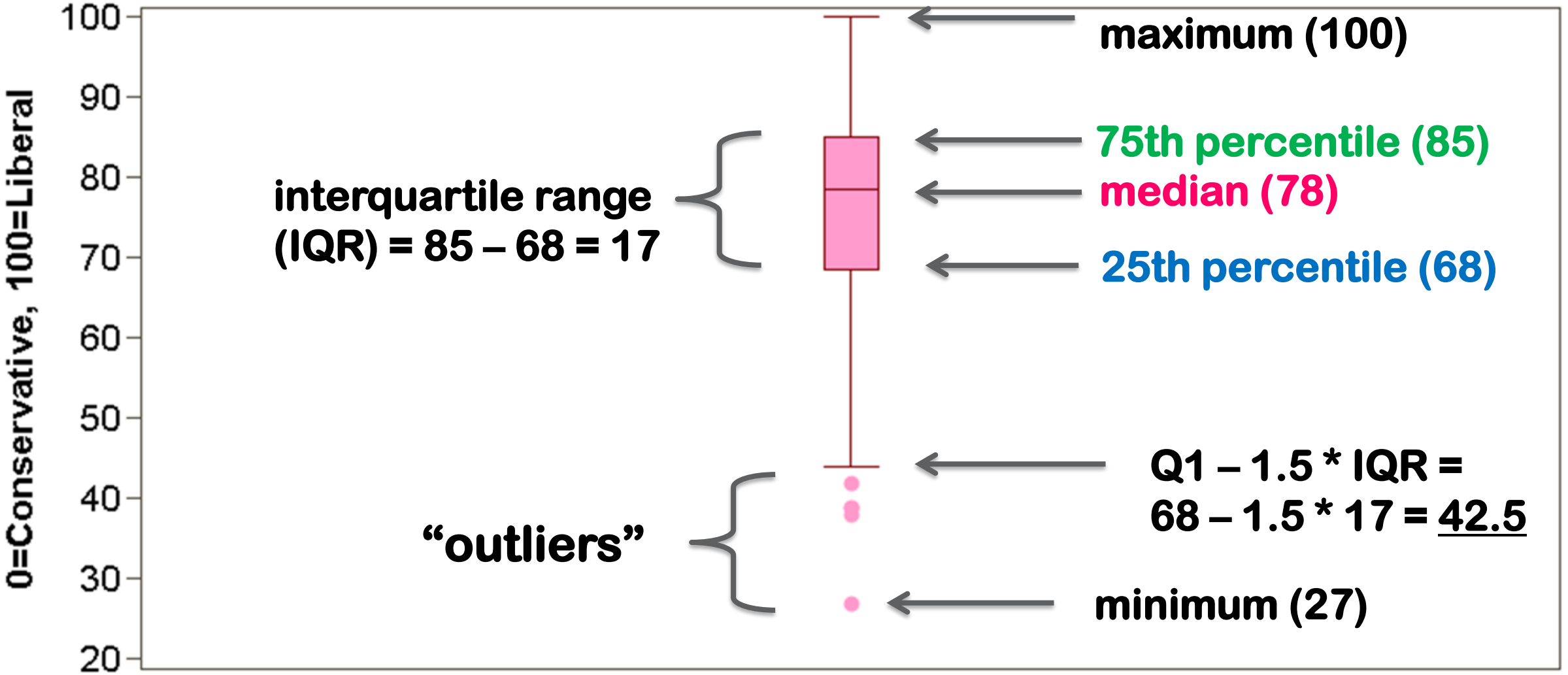
- To show the distribution of quantitative variables
  - shape
  - center
  - range
  - variation

Boxplot of Exercise



Boxplot of Political Bent

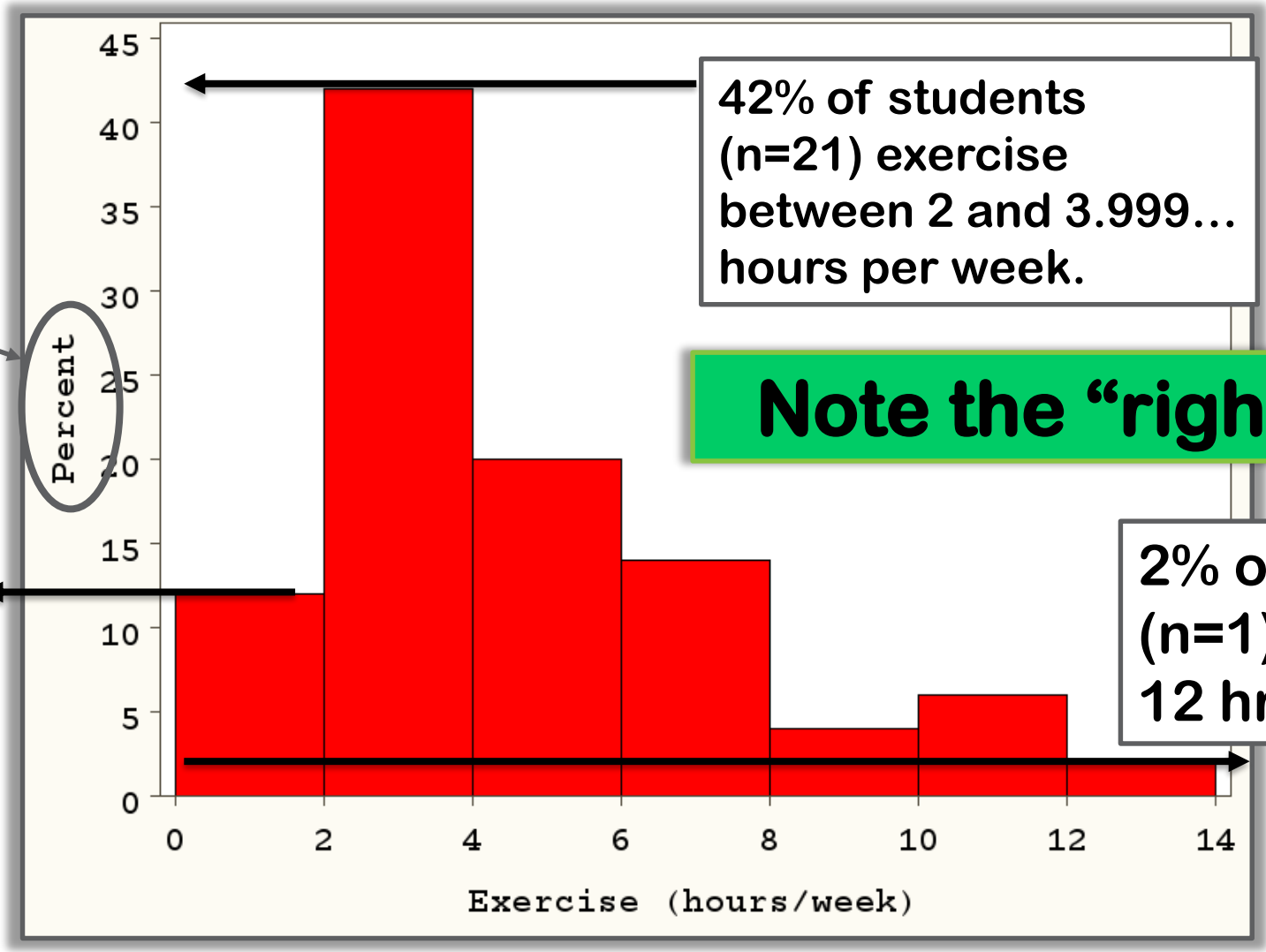
(0=Most Conservative, 100=Most Liberal)



Histogram of Exercise

Bins of size = 2 hours/week

Y-axis: The percent of observations that fall within each bin.



42% of students (n=21) exercise between 2 and 3.999… hours per week.

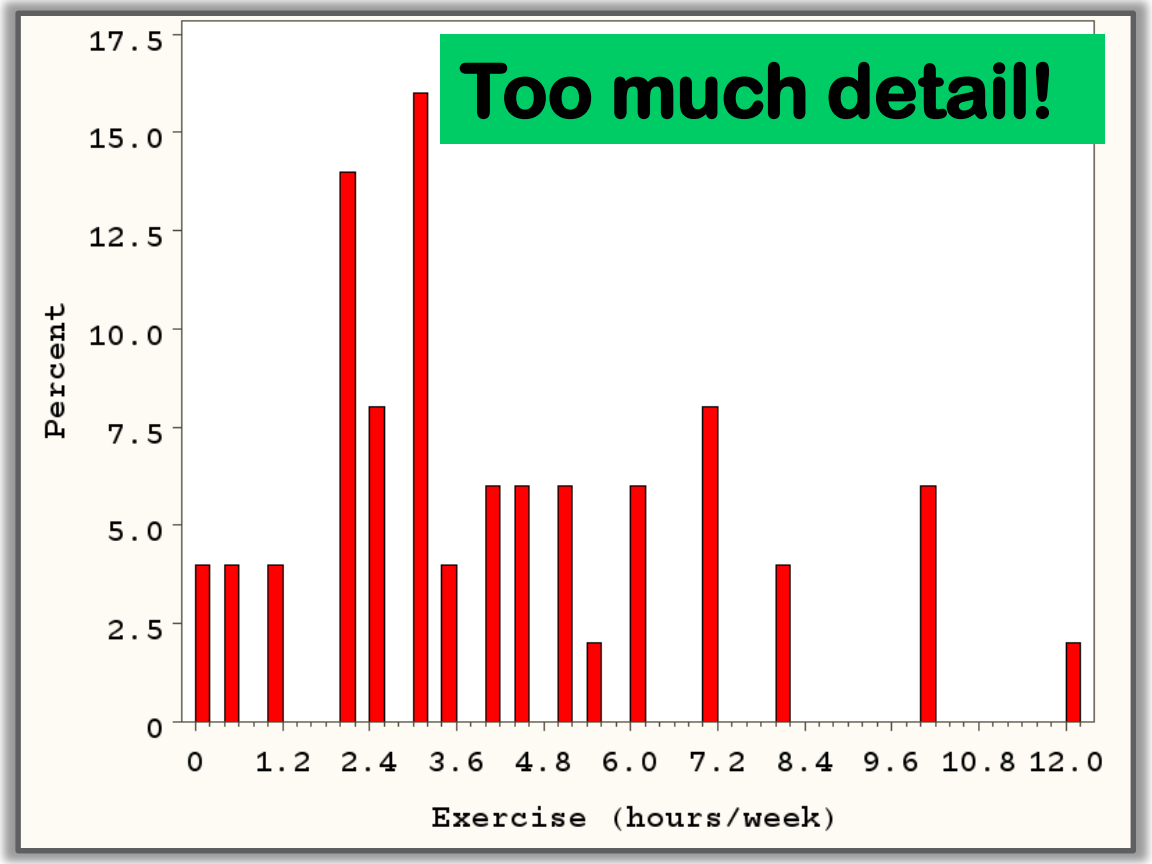
Note the "right skew"

2% of students (n=1) exercise ≥ 12 hr/wk

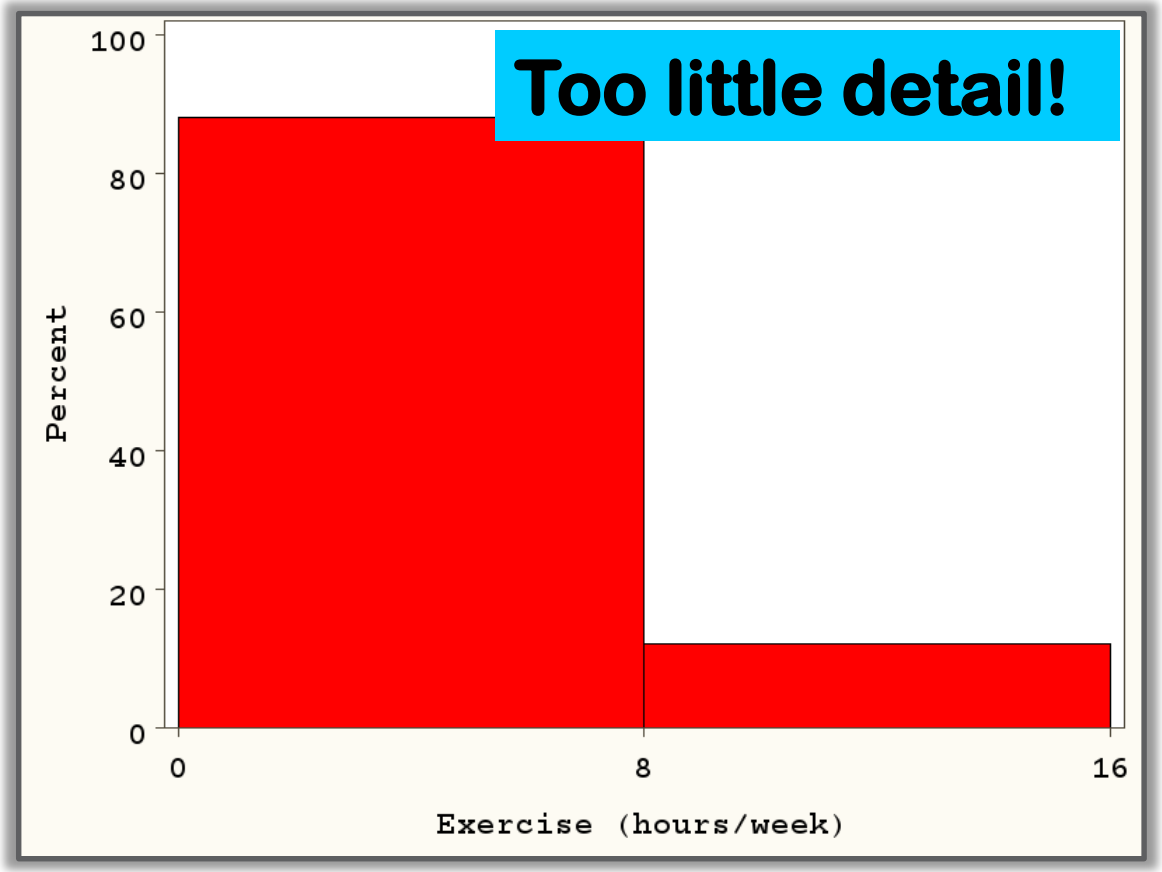
12% of students (n=6) exercise between 0 and 1.999… hours per week.

Histogram of Exercise

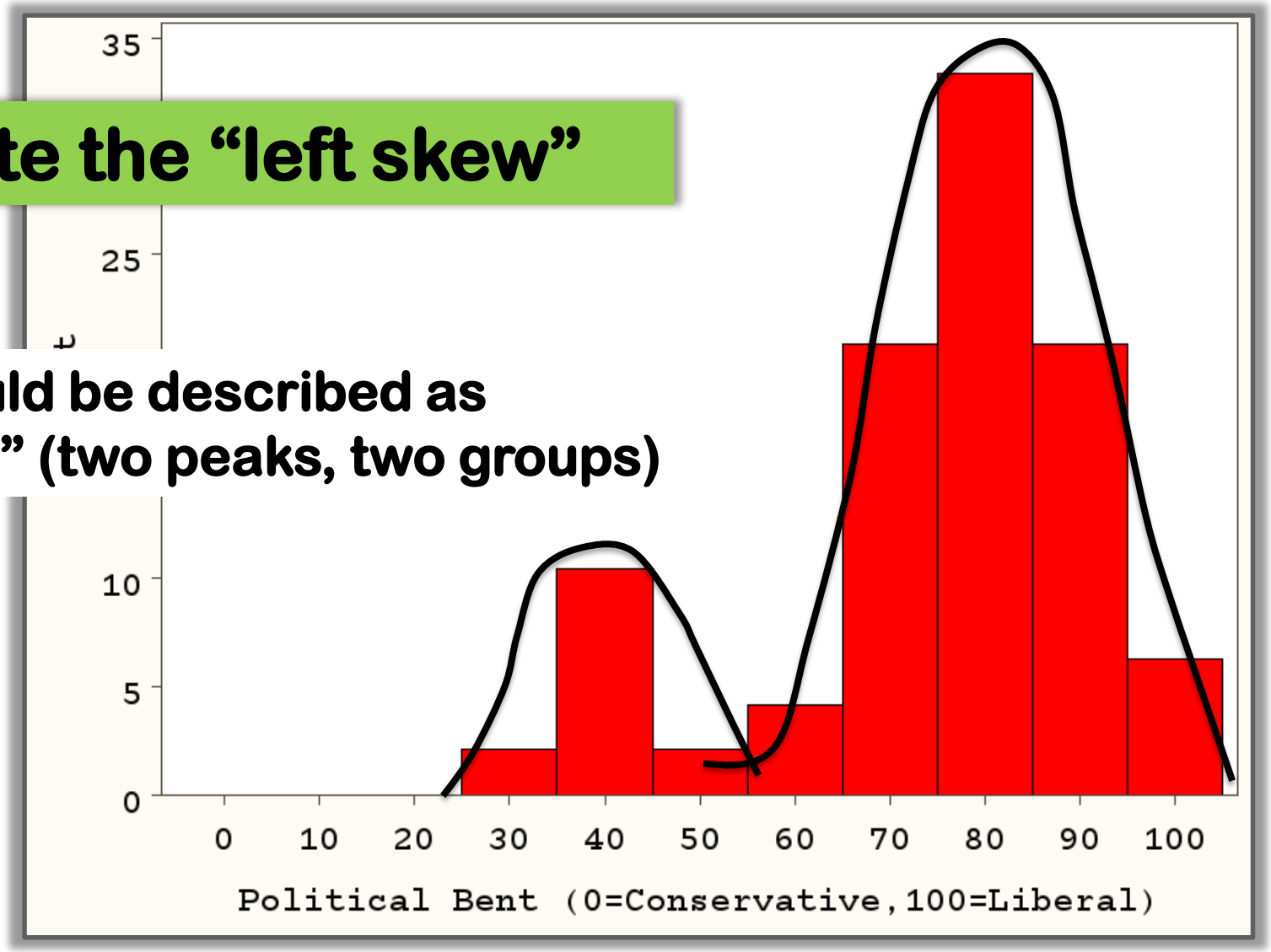
Bins of size = 0.2 hours/week



Bins of size = 8 hours/week



Histogram of Political Bent



Note the "left skew"

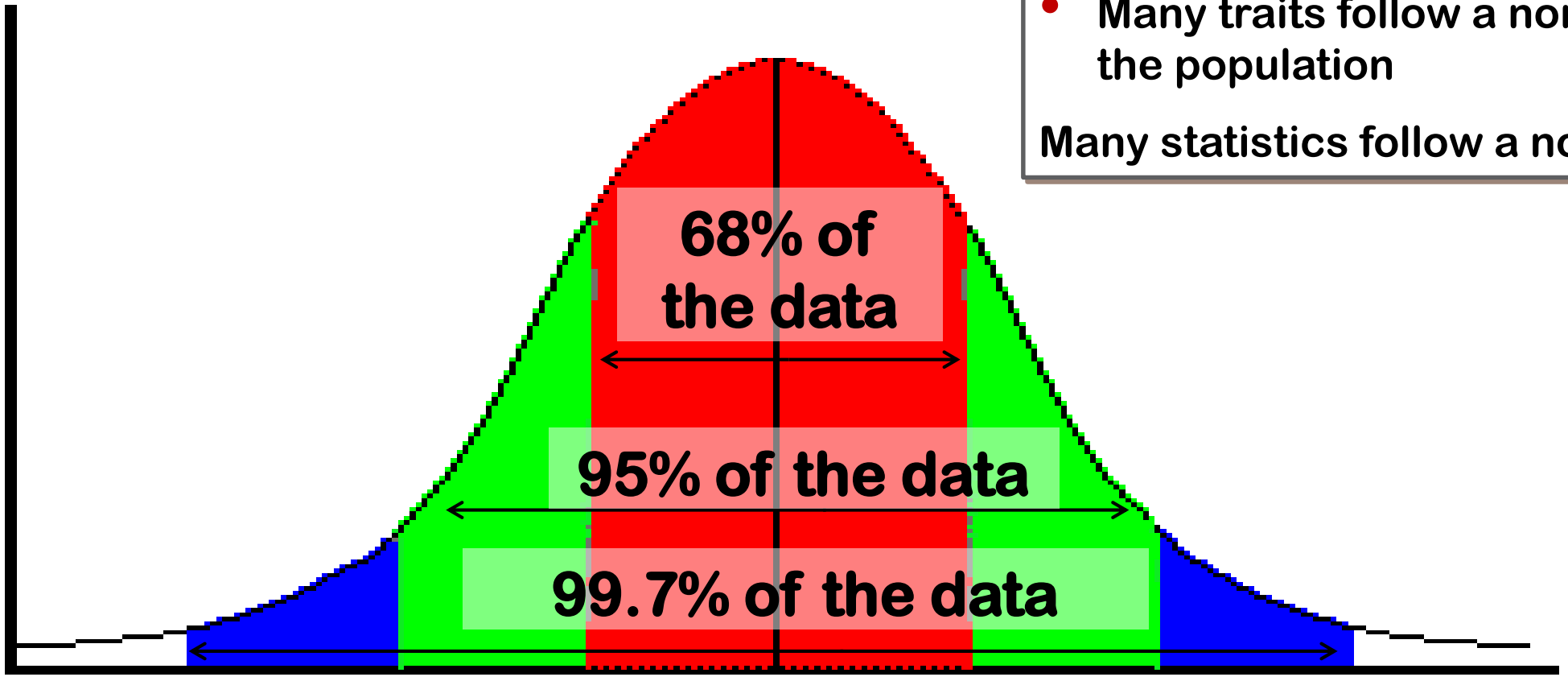
Also, could be described as "bimodal" (two peaks, two groups)

Normal distribution (bell curve)

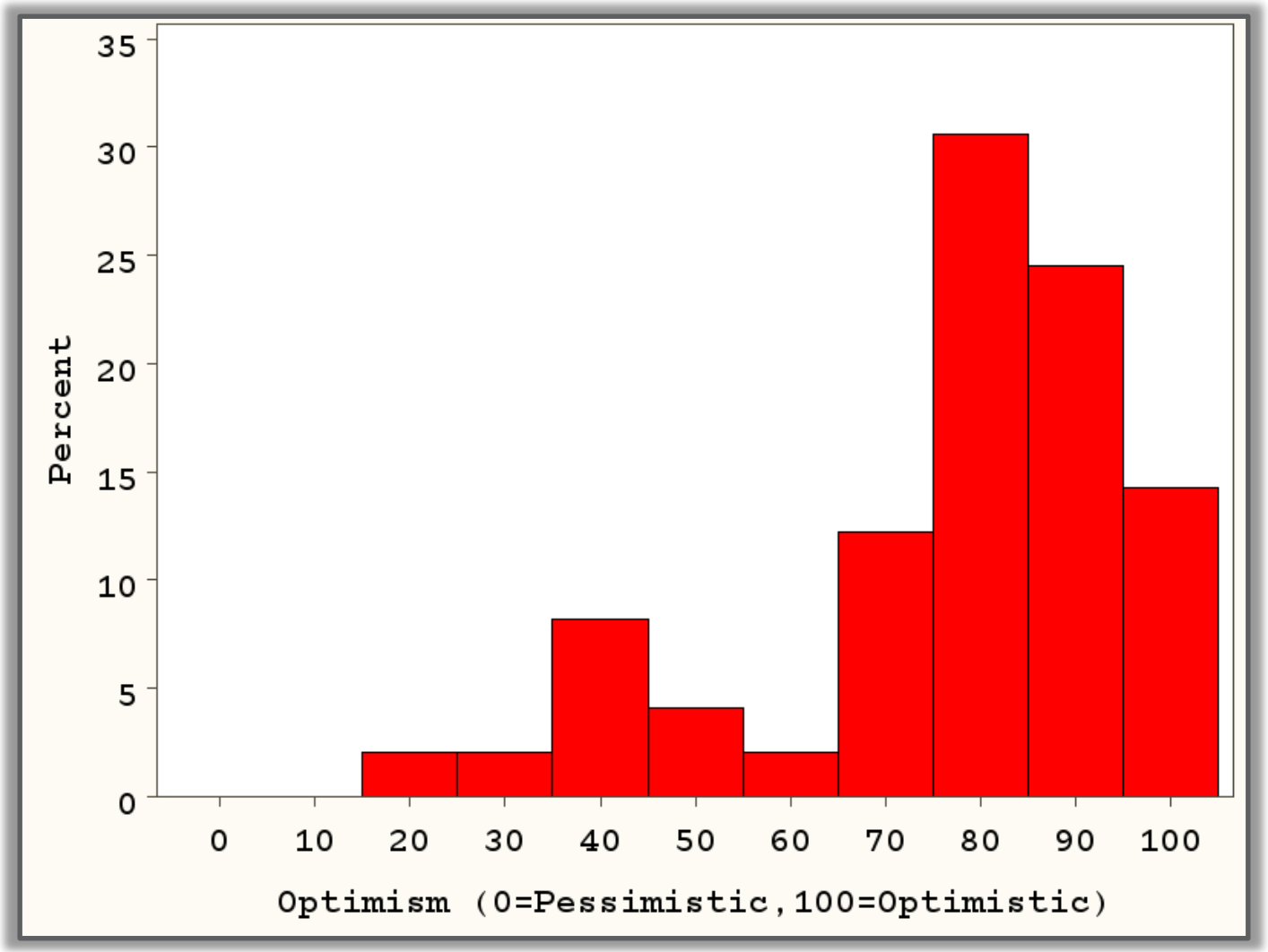
Useful for many reasons:

- Has predictable behavior
- Many traits follow a normal distribution in the population

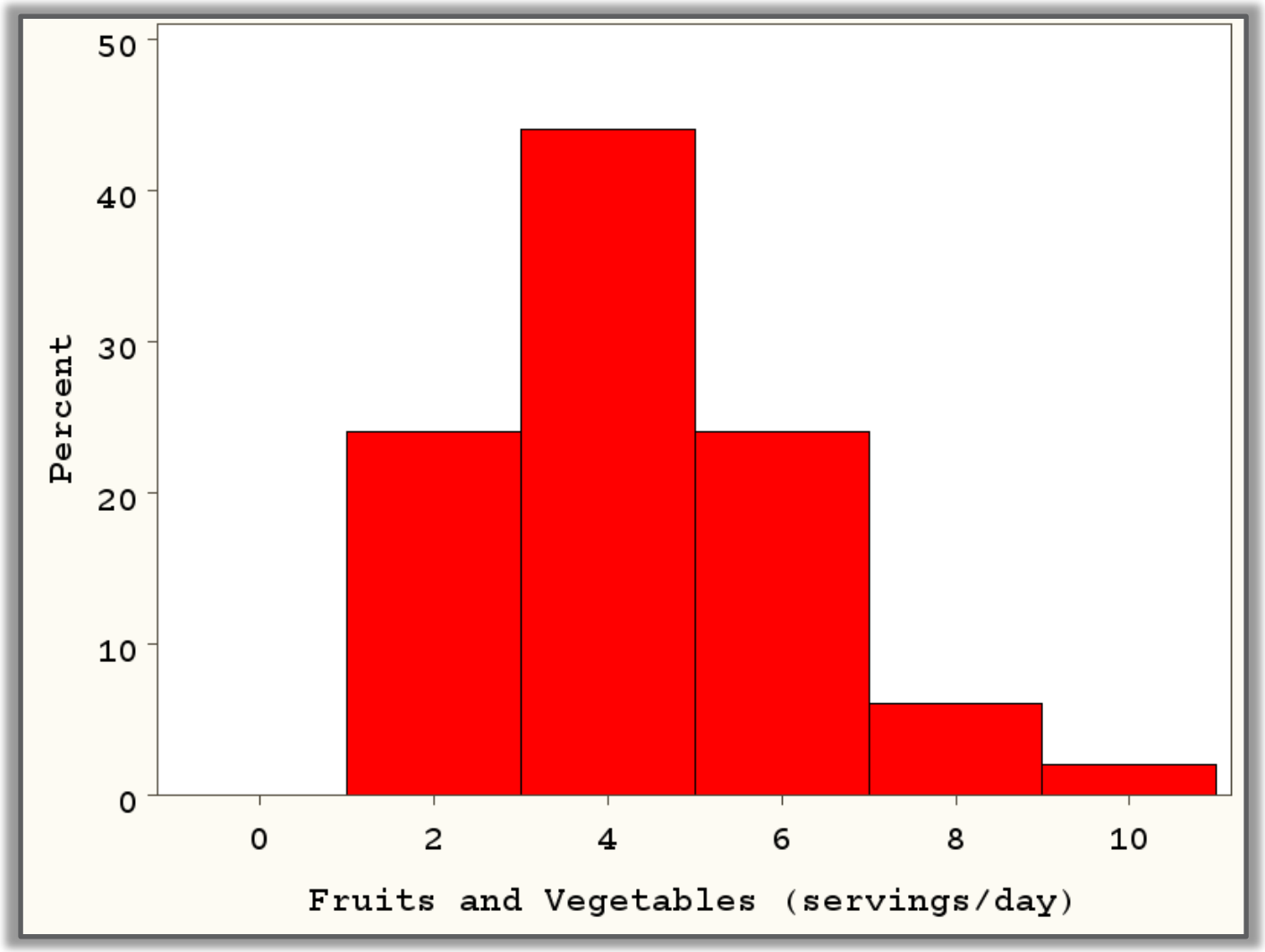
Many statistics follow a normal distribution



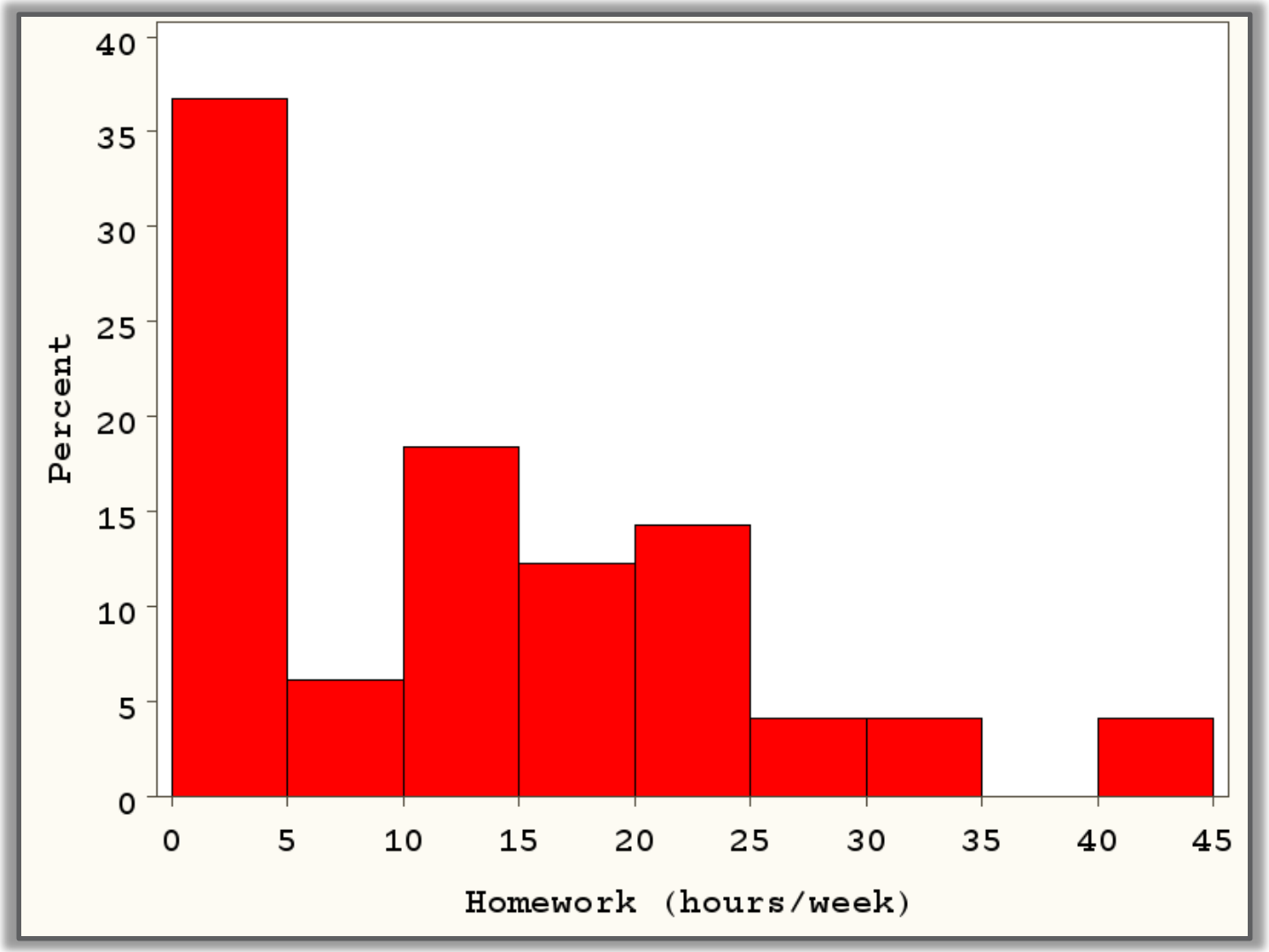
Example data: Optimism…



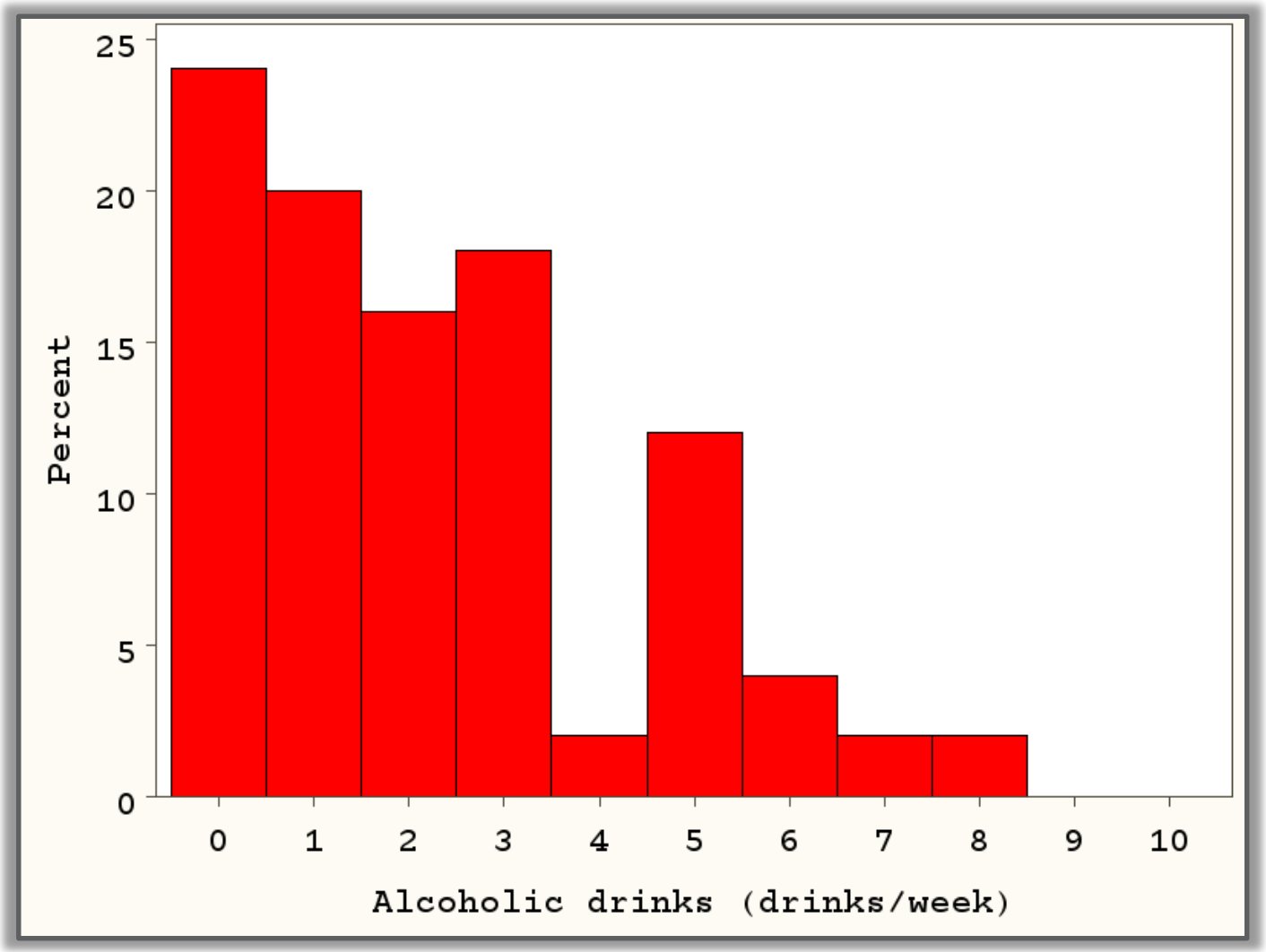
Fruit and vegetable consumption (servings/day)…



Homework (hours/week)…



Alcohol (drinks/week)



Feelings about math (0=lowest, 100=highest)

Closest to a normal distribution!