Descriptive Statistics

Math 105

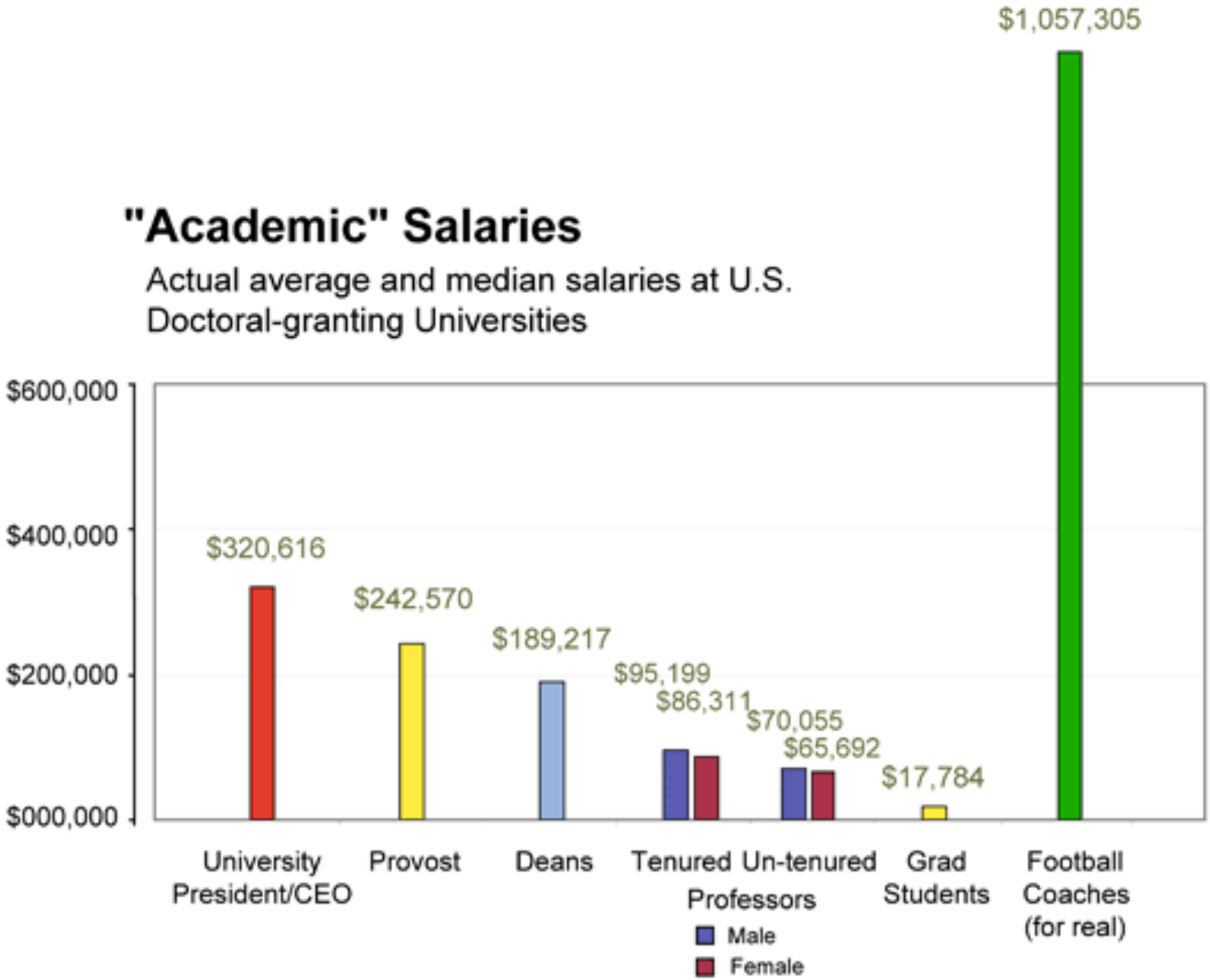
Section 003 - Prasad

October 5, 2010

Ways to organize and display data

Job Description	Average Salary
Grad Student	17784
Male un-tenured faculty	70055
Female un-tenured faculty	65692
Male tenured faculty	95199
Female tenured faculty	86311
Dean	189217
Provost	242570
University President/CEO	320616
Football Coach	1057305

Ways to organize and display data



Data sets

A **data set** is a collection of **data points**. Consider the following list of reading quiz scores:

10.0	7.0	0.0	10.0	10.0
8.0	10.0	6.0	9.0	10.0
8.0	8.0	7.0	10.0	9.0
0.0	10.0	9.0	9.0	8.0
8.0	9.0	10.0	10.0	5.0
0.0	8.0	10.0	0.0	10.0
10.0	0.0	6.0	6.0	

The collection of scores is the **data set**, but each individual score is **data point**.

Frequency tables

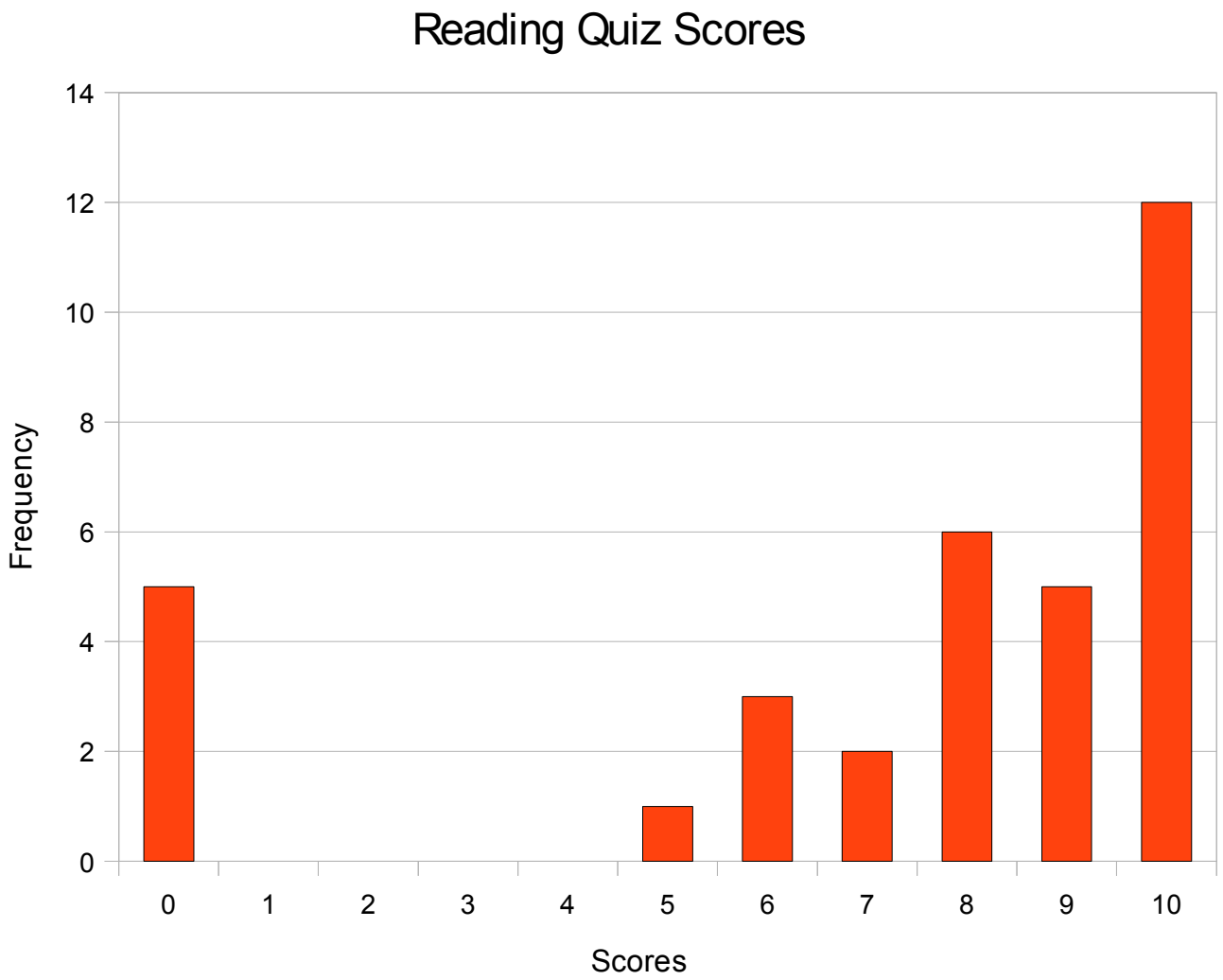
To make it easier to understand data sets at a glance, we create frequency tables:

Score	10.0	9.0	8.0	7.0	6.0	5.0	0.0
Frequency	12	5	6	2	3	1	5

The second row tells us how many students earned each score, that is, the **frequency** of each score.

Bar Graphs

We can take that information and put it in graph form.



Definitions

- **mean**: the average of the data points. The **mean**, A, of a set of N numbers: $d_1, d_2, d_3, \ldots, d_N$ is

$$A = \frac{d_1 + d_2 + d_3 + \ldots + d_N}{N}$$

- **percentile**: The data point in the data set that represents the p**th percentile** is the data point where p percent of the data points lie at or below this data point.

Using medians and percentiles

Using the frequency table given, find the **mean** of the data set.

Score	10.0	9.0	8.0	7.0	6.0	5.0	0.0
Frequency	12	5	6	2	3	1	5

Which score represents the 50th percentile of the data? Is this the same as the mean? Does it make sense if it is or isn't?
What percentile does the mean represent (think of how many data points are below the mean that you calculated)?
Given N data points, how would you calculate which data point represents the pth percentile? It may help to use the example above ($N = 34$) and a given percentile.

Definitions

- **median**: The **median**, M, of a data set is the 50th percentile of the data set.
- **quartiles**: The **first quartile**, Q_1, is the 25th percentile. The **third quartile**, Q_3, is the 75th percentile.
- **five-number summary**: A five number summary gives the following data:

$$\{Min, Q_1, M, Q_2, Max\}$$

Using medians and percentiles

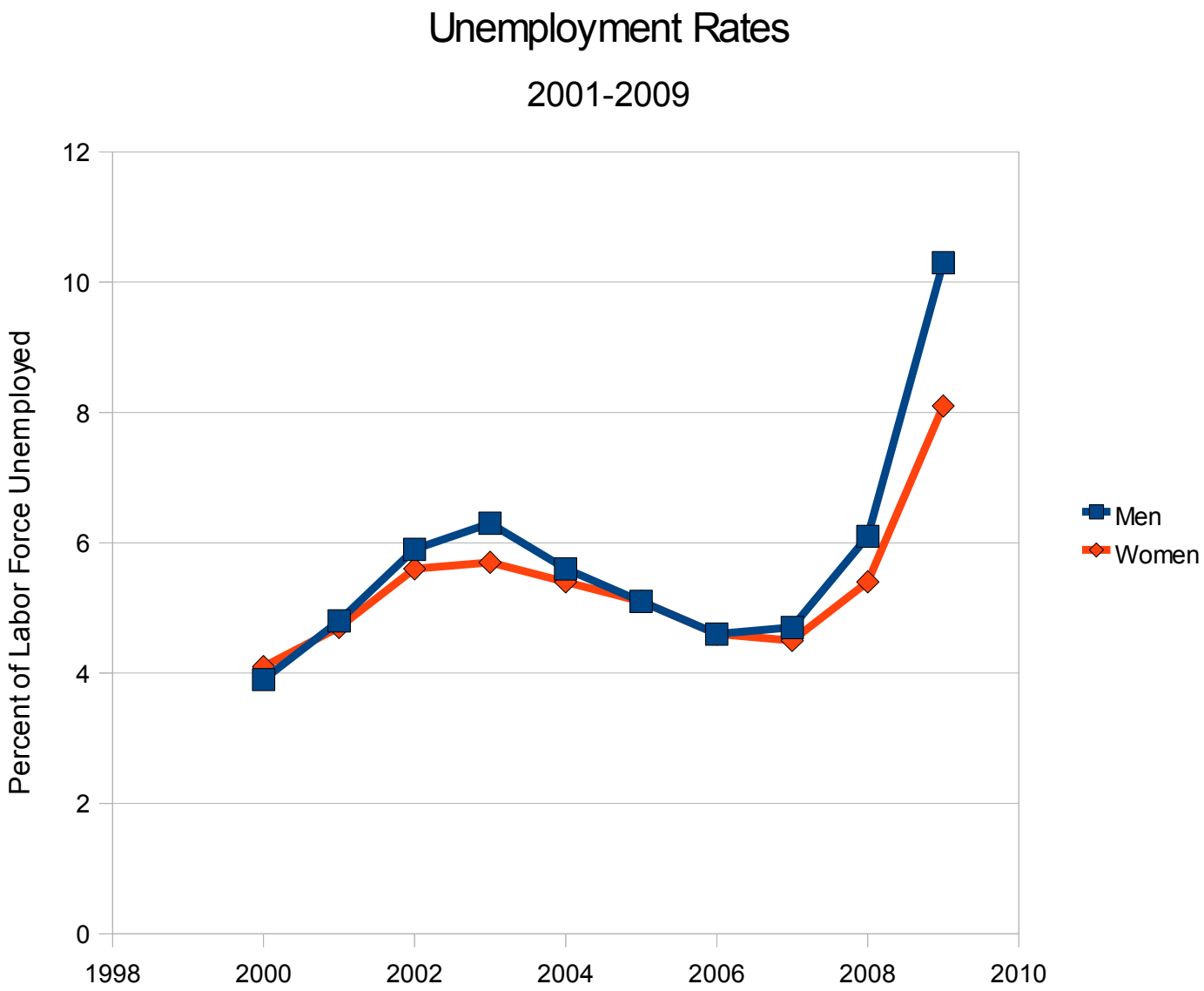
Using the frequency table given, find the **median** and **first and third quartiles** of the data set.

Score	10.0	9.0	8.0	7.0	6.0	5.0	0.0
Frequency	12	5	6	2	3	1	5

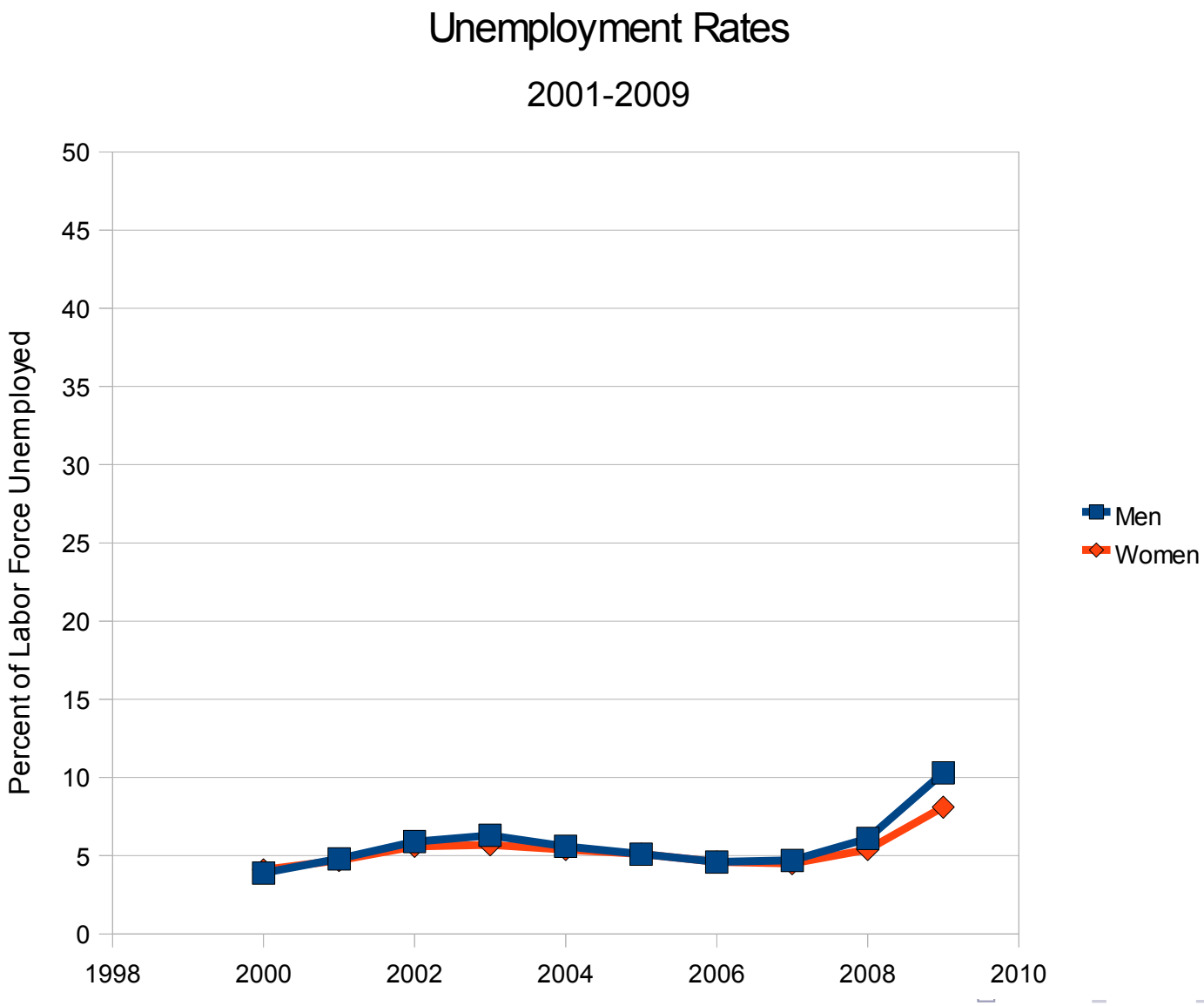
Organize this information into a five number summary.

For practice, compute the data point that is the 32nd percentile.

Umemployment rates by gender (BLS)



Unemployment rates by gender (BLS)



Bar graphs vs. histograms

Definitions:

- **variable**: any characteristic that varies among individuals in a population (i.e. scores on a test, number of spots a leopard has, hair color, etc)

Bar graphs vs. histograms

Definitions:

- **variable**: any characteristic that varies among individuals in a population (i.e. scores on a test, number of spots a leopard has, hair color, etc)
- **quantitative (numerical) variable**: variable that can be measured, like height, test scores, tail length

Bar graphs vs. histograms

Definitions:

- **variable**: any characteristic that varies among individuals in a population (i.e. scores on a test, number of spots a leopard has, hair color, etc)
- **quantitative (numerical) variable**: variable that can be measured, like height, test scores, tail length
  - continuous: differences between data points can be really small

Bar graphs vs. histograms

Definitions:

- **variable**: any characteristic that varies among individuals in a population (i.e. scores on a test, number of spots a leopard has, hair color, etc)
- **quantitative (numerical) variable**: variable that can be measured, like height, test scores, tail length
  - continuous: differences between data points can be really small
  - discrete: differences between data points have to be at least a certain size (like one)

Bar graphs vs. histograms

Definitions:

- **variable**: any characteristic that varies among individuals in a population (i.e. scores on a test, number of spots a leopard has, hair color, etc)
- **quantitative (numerical) variable**: variable that can be measured, like height, test scores, tail length
  - continuous: differences between data points can be really small
  - discrete: differences between data points have to be at least a certain size (like one)
- **qualitative (categorical) variable**: variable that produces categories, like hair color, or state of birth

Bar graphs vs. histograms

- We use bar graphs for categorical variables or discrete numerical variables.

Bar graphs vs. histograms

- We use bar graphs for categorical variables or discrete numerical variables.
- We use histograms for continuous numerical variables.

Bar graphs vs. histograms

- We use bar graphs for categorical variables or discrete numerical variables.
- We use histograms for continuous numerical variables.
- We can think of histograms as bar graphs where the categories are ranges of numbers, instead of a single number (and the bars will be pushed together).

Pie Charts

We use pie charts for categorical variables. Create a pie chart based on the hair colors of students in the class. Be precise with the angles of the slices in the pie chart.

Class Intervals

We can **aggregate** numerical data into categories that we call **class intervals**. Basically, it is a way of organizing quantitative data so that we can represent it in ways we would normally represent categorical data.

For example, if we have the following set of test scores:

93.0	82.0	40.0	79.0	95.0
93.0	65.0	61.0	66.0	48.0
82.0	66.0	70.0	76.0	66.0
92.0	65.0	95.0	82.0	75.0

How could we make a pie chart to represent the data? We could categorize the scores by letter grade. This would require defining class intervals.

Class Intervals

We can define the class intervals in the standard way:

$$A : 100 - 90,\ B : 89 - 80,\ C : 79 - 70,\ D : 69 - 60,\ F : 59 - 0$$

and build a frequency table:

Grade	A	B	C	D	F
Class interval	$100 - 90$	$89 - 80$	$79 - 70$	$69 - 60$	$59 - 0$
Frequency	5	3	4	6	2

Numerical Summaries 2: Data Spread

Definitions:

- **range**: The range of a data set is simply the difference between the maximum data point and the minimum data point:

$$\text{Range} = \text{Max} - \text{Min}.$$

- **interquartile range**: The interquartile range is the difference between the 75th percentile and the 25th percentile:

$$\text{Interquartile range} = Q_3 - Q_1.$$

- **standard deviation**: this measures how far each data point is from the mean and collects a sort of average of these deviations

Calculating Standard Deviation

Suppose we have a data set:

$$\{d_1, d_2, d_3, \ldots, d_N\}$$

which has a mean A. The **variance**, V, of this data set is

$$V = \frac{(d_1 - A)^2 + (d_2 - A)^2 + \ldots + (d_N - A)^2}{N}.$$

The **standard deviation**, σ, of the data set is

$$\sigma = \sqrt{V}.$$

Calculate the standard deviation of the data set

$$\{10, 5, 8, 2, 1, 1, 3\}.$$

Calculating Standard Deviation

You should get (something very close to):

$$V = \frac{75.43}{7} = 10.78$$

and

$$\sigma = \sqrt{V} = \sqrt{10.78} = 3.28.$$