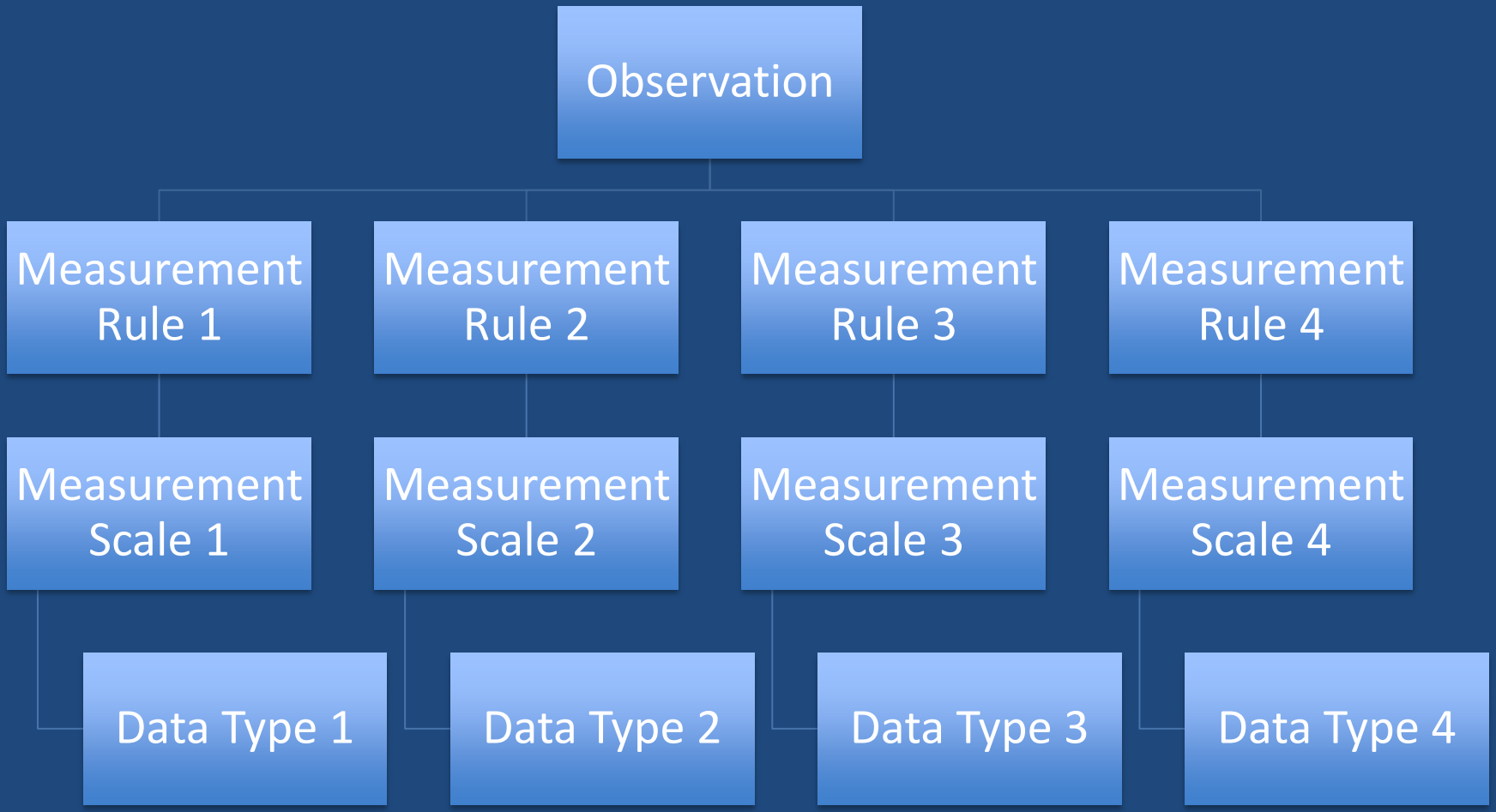
# Lecture 1: Basic Descriptive Statistics

1. Types of Biological Data
2. Summary Descriptive Statistics
   - Measures of Central Tendency
   - Measures of Dispersion
3. Assignments

# Scales of Measurement: General Comments

- Any observation or experiment in biology involves the collection of information (observe plants)
- Empirical observations become *statistical data* once they are cast as some type of *measurement* (plant height)
- Measurement is the assignment of numbers to objects or events according to rules (measure plant 1, plant 2, … )
- Different rules lead to different kinds of *scales of measurement*
- A dataset can thus be classified according to the type of scale by which it is measured
  - Different scales admit different permissible statistics (see table summary)

# Scales of Measurement: General Comments



# Scales of Measurement: NOIR

- Data on a Nominal Scale
  - A nominal scale assigns numbers as mere labels or types- words or letters would work just as well
  - Example: numbers on jerseys that serve to identify athletic participants
  - Example: rocks can be classified as igneous, sedimentary, and metamorphic
- Data on an Ordinal Scale
  - An ordinal scale assigns numbers according to some rank ordering
  - Example: order in which participant's finish a race ($1^{st}$, $2^{nd}$, … )

# Scales of Measurement: NOIR

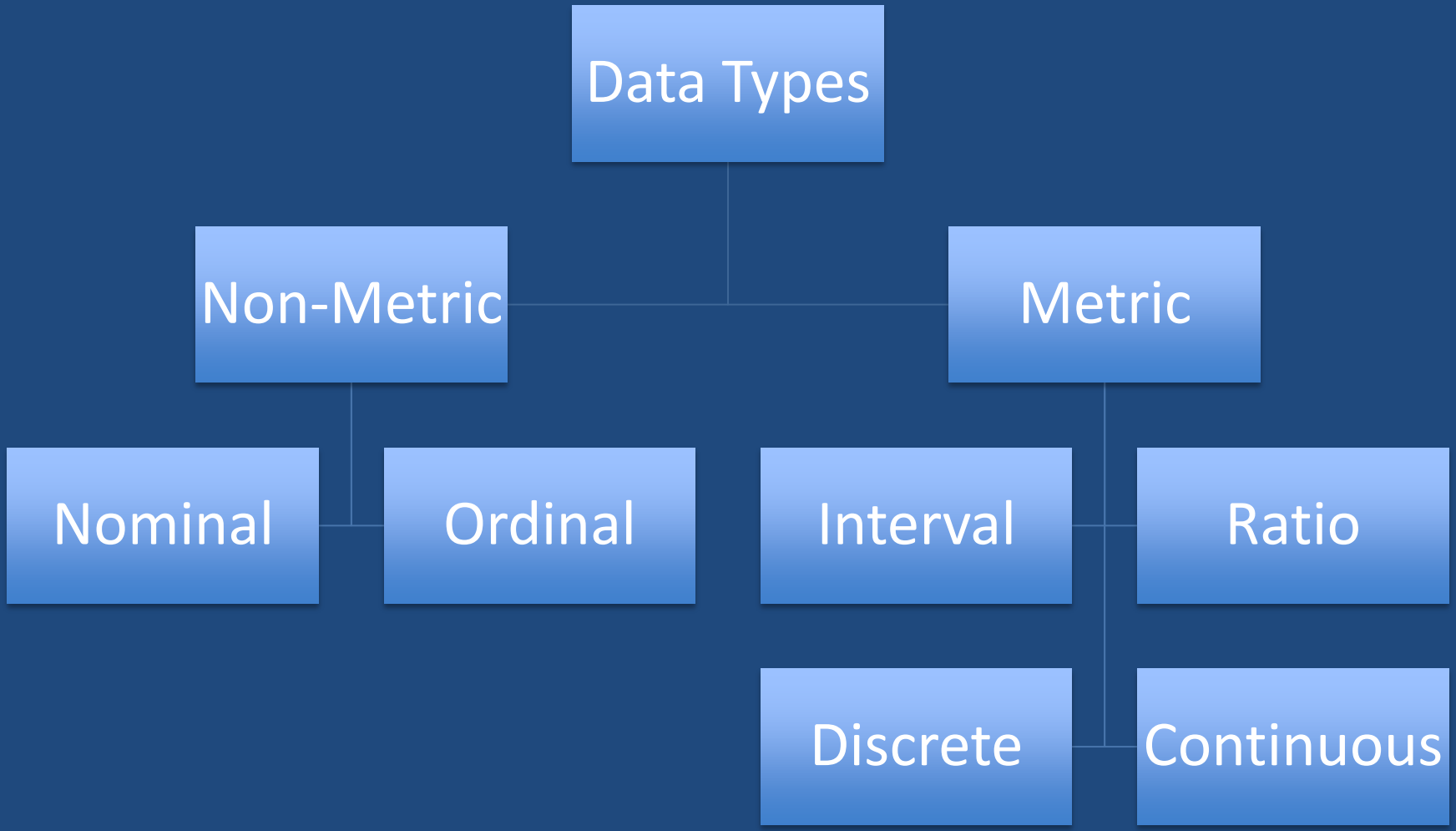
- Data on an Interval Scale
  - An interval scale assigns numbers according to some rank ordering **and** assigns the size of intervals in between data (**but** has no true zero point)
  - Example: The temperature scales of degrees Celsius and degrees Fahrenheit are interval scales
    - The amount of temperature change from 27° C to 32° C is the **same** as the temperature change from 104° C to 109° C
    - The choices for 0° C and 0° F are arbitrary; that is, it makes no sense to say that 98° F is twice as hot as 49° F
- Data on a Ratio Scale
  - A ratio scale is an interval scale with a true zero point.
  - Example: A participant's finish time for a race
    - A finish time of 25 seconds is better than 50 seconds (order) and is, indeed, twice as fast (true zero)
  - Example: The Kelvin temperature scale (has an absolute zero)

# Discrete vs. Continuous

Measurements may take on discrete or continuous values:

- A set of values is **discrete** if it is countable
  - Set of possible number of arms on a starfish
  - Set of possible number of leaves on a plant
  - Set of possible number of granules of sand on a beach
- A set of values is **continuous** if it is uncountable
  - Set of possible weights of starfish
  - Set of possible surface areas for leaves
  - Set of possible amounts of time spent counting sand granules

# Summary: Organizational Chart



# Overview

- When a dataset is summarized by its statistical information, there is a loss of information. That is, given the summary statistics, there is no way to recover the original data.

- Basic summary statistics may be grouped as:

  1. measures of **central tendency** (giving in some sense the central value of a data set)
  2. measures of **dispersion** (giving a measure of how spread out that data set is)

# Measures of Central Tendency

- Arithmetic Mean

  Dataset: $\{x_1, x_2, \mathrm{K}, x_n\}$

  Average: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(x_1 + x_2 + \mathrm{L} + x_n)}{n}$

  Example: $\{2, 12, 3\} \Rightarrow \bar{x} = \frac{2 + 12 + 3}{3} = \frac{17}{3} \approx 5.7$

  This statistic doesn't make sense for data on nominal or ordinal scales: jersey numbers, top ten list

# Measures of Central Tendency

- Median: half the dataset fall below this value; half above

Dataset: $\{ 30 , 40 , 50 , 1000000 , 40 , 45 \}$

Median: $\frac{40 + 45}{2} = 42.5$

This statistic doesn't make sense for data on nominal scales: jersey numbers

The median is less effected by outliers than the mean; in this case the mean is approximately 167,000

# Measures of Central Tendency

- Mode: The mode is the most frequently occurring value (or values - there may be more than one) in a data set

  Dataset: { 30 , 40 , 50 , 1000000 , 40 , 45 }

  Mode: 40

  This statistic is meaningful for all scales

# Measures of Central Tendency

- Midrange: The midrange is the value halfway between the largest and smallest values in the data set

  Dataset: $\{ x_1, x_2, \text{K} \; , x_{\max} \; , \text{K} \; , x_{\min} \; , \text{K} \; \}$

  Midrange: $\overline{x}_{mid} = \dfrac{x_{\min} + x_{\max}}{2}$

  This statistic doesn't make sense for data on nominal or ordinal scales: jersey numbers, top ten list

# Measures of Central Tendency

- Geometric Mean: The geometric mean of a set of n data is the $n^{th}$ root of the product of the n data values,

  Dataset: $\{x_1, x_2, \mathrm{K}\ , x_n\}$

  Geometric Mean: $\bar{x}_{geom} = \left( \prod_{i=1}^{n} x_i \right)^{1/n} = \sqrt[n]{x_1 \cdot x_1 \mathrm{L}\ x_n}$

  The geometric mean arises as an appropriate estimate of growth rates of a population when the growth rates vary through time or space

  It is always less than or equal to the arithmetic mean

# Measures of Dispersion

- Range

  Dataset: $\{ x_1, x_2, \mathrm{K}\ , x_{\max}\ , \mathrm{K}\ , x_{\min}\ , \mathrm{K}\ \}$

  Range: $x_{\max} - x_{\min}$

- Variance: the mean sum of the squares of the deviations of the data from the arithmetic mean
  - The “best” estimate of this (take a good statistics class to find out how “best” is defined) is the sample variance:

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

- Standard Deviation: $s = \sqrt{\mathrm{var}}$

# Table Summary

| Permissible Statistic/Operation | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| Mode | ✓ | ✓ | ✓ | ✓ |
| Median | ✗ | ✓ | ✓ | ✓ |
| Addition, Mean, Variance | ✗ | ✗ | ✓ | ✓ |
| Multiplication, Ratio | ✗ | ✗ | ✗ | ✓ |

# Homework, MATLAB

1. Homework: Chapter 1 Exercises 1.2 - 1.5.
2. Download MATLAB as soon as possible. We will begin working with MATLAB in class next Thursday.

Homework

# 1.1

Exercise capacity (in seconds) was determined for each of 11 patients who were being treated for chronic heart failure:
906, 1320, 711, 1170, 684, 1200, 837, 1056, 897, 882, 1008

(a) Determine the mean and the median of the data.

Solution:

$$mean = \frac{906 + 1320 + \text{L} + 1008}{11} = 970.09$$

To find the median, we first order the data:

684, 711, 837, 882, 897, 906, 1008, 1056, 1170, 1200, 1320

Since there are eleven (an odd number) data points, the median will be the $6^{th}$ data point. That is, the median is 906.

Homework

# 1.2

Daily crude oil output (in millions of barrels) is shown below for the years 1971 to 1990.

9.45 9.40 9.25 8.75 8.30 8.10 8.25 8.70 8.55 8.60

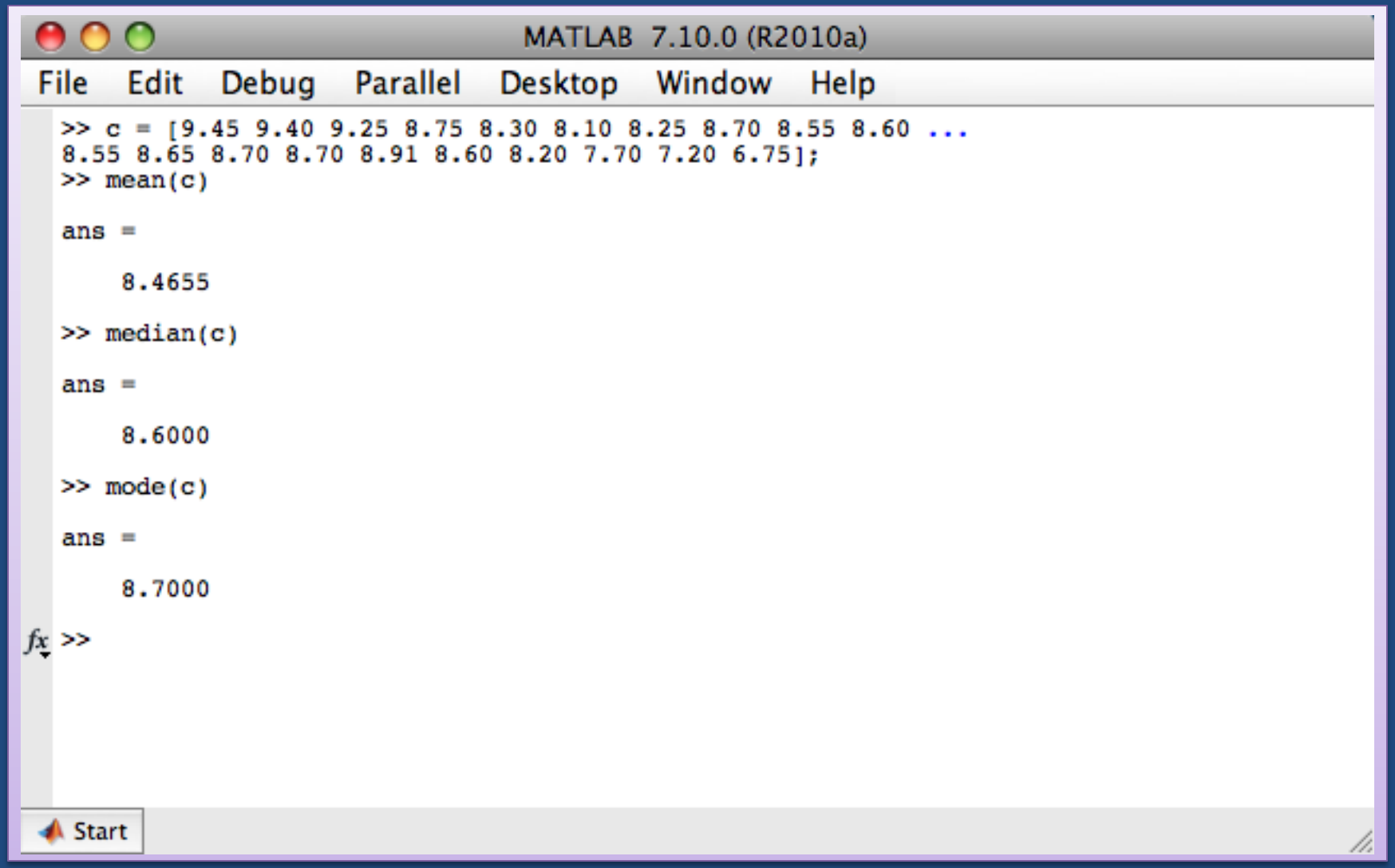
8.55 8.65 8.70 8.70 8.91 8.60 8.20 7.70 7.20 6.75

Compute the mean, median, and mode for the data.

Solution: Let's use MATLAB to solve this problem.

Homework

# 1.2


MATLAB 7.10.0 (R2010a)
File Edit Debug Parallel Desktop Window Help
>> c = [9.45 9.40 9.25 8.75 8.30 8.10 8.25 8.70 8.55 8.60 ...
8.55 8.65 8.70 8.70 8.91 8.60 8.20 7.70 7.20 6.75];
>> mean(c)
ans =
8.4655
>> median(c)
ans =
8.6000
>> mode(c)
ans =
8.7000
>>
Start

Homework

# 1.4

Ten hospital employees on a standard American diet agreed to adopt a vegetarian diet for one month. Below is the change in the serum cholesterol level (before - after).

49, −10, 27, 13, 36, 19, 48, 21, 8, 16

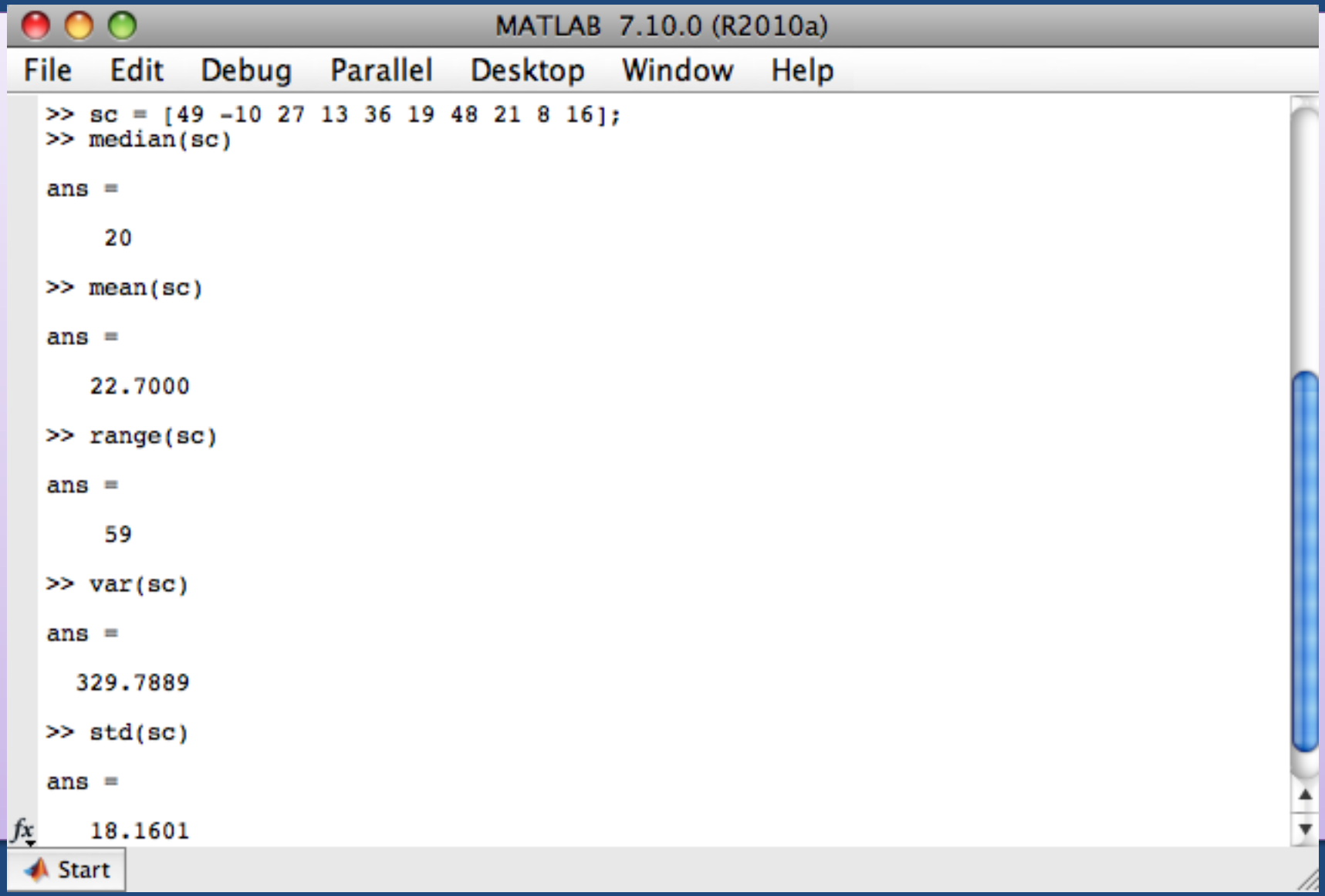
(a) Compute the median and mean change in cholesterol.

(b) Compute the range, variance and standard deviation of the data. Is the data fairly spread out or close together?

Solution: Again we use MATLAB.

Homework

# 1.4


MATLAB 7.10.0 (R2010a)
File Edit Debug Parallel Desktop Window Help
>> sc = [49 -10 27 13 36 19 48 21 8 16];
>> median(sc)
ans =
20
>> mean(sc)
ans =
22.7000
>> range(sc)
ans =
59
>> var(sc)
ans =
329.7889
>> std(sc)
ans =
18.1601
Start

Homework

# 1.4

In order to study for the quiz, we now do these by hand. First we rewrite the dataset in numerical order:

-10, 8, 13, 16, 19, 21, 27, 36, 48, 49

Since there are ten (an even number) data points, the median will be halfway between 19 and 21. That is, the median is 20. Finding the variance is more work:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10-1}\sum_{i=1}^{n}(x_i - 22.7)^2$$

$$= \frac{1}{9}\left((-10 - 22.7)^2 + (8 - 22.7)^2 + \mathrm{L} + (49 - 22.7)^2\right)$$

$$= 329.7889 \quad \Rightarrow \quad std = s = \sqrt{329.7889} = 18.1601$$

Homework

# 1.5

Twelve sheep were fed pingue as a part of an experiment and died as a result. The time of death in hours after the administering of pingue for each sheep follows:

44 27 24 24 36 36 44 120 29 36 36 36

Compute the range, variance and standard deviation of the sample.

Answer:

range: 96

variance: 663.8182

std: 25.7647