Chapter 1

Introduction to Statistics and Data Analysis

Chapter Outline

1.1 Overview: Statistical Inference, Samples, Populations, and the Role of Probability

1.2 Sampling Procedures; Collection of Data

1.3 Measures of Location: The Sample Mean and Median

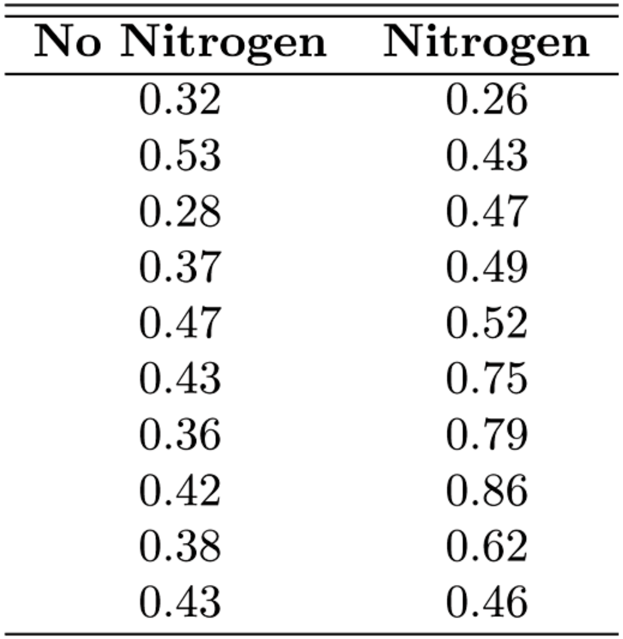
1.4 Measures of Variability

1.5 Discrete and Continuous Data

1.6 Statistical Modeling, Scientific, Inspection, and Graphical Diagnostics

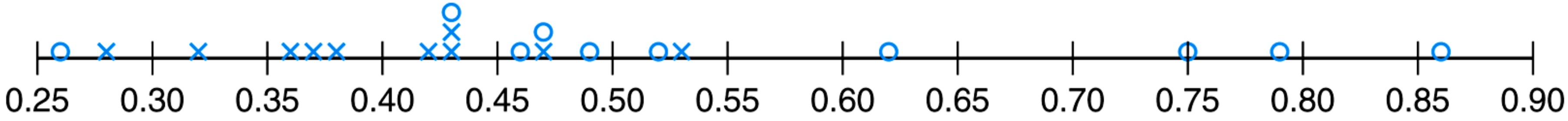
Example

Two samples of 10 northern red oak seedlings were planted in a greenhouse, one containing seedlings treated with nitrogen and the other containing seedlings with no nitrogen. The stem weights in grams were recorded after the end of 140 days. The data are given as follows:

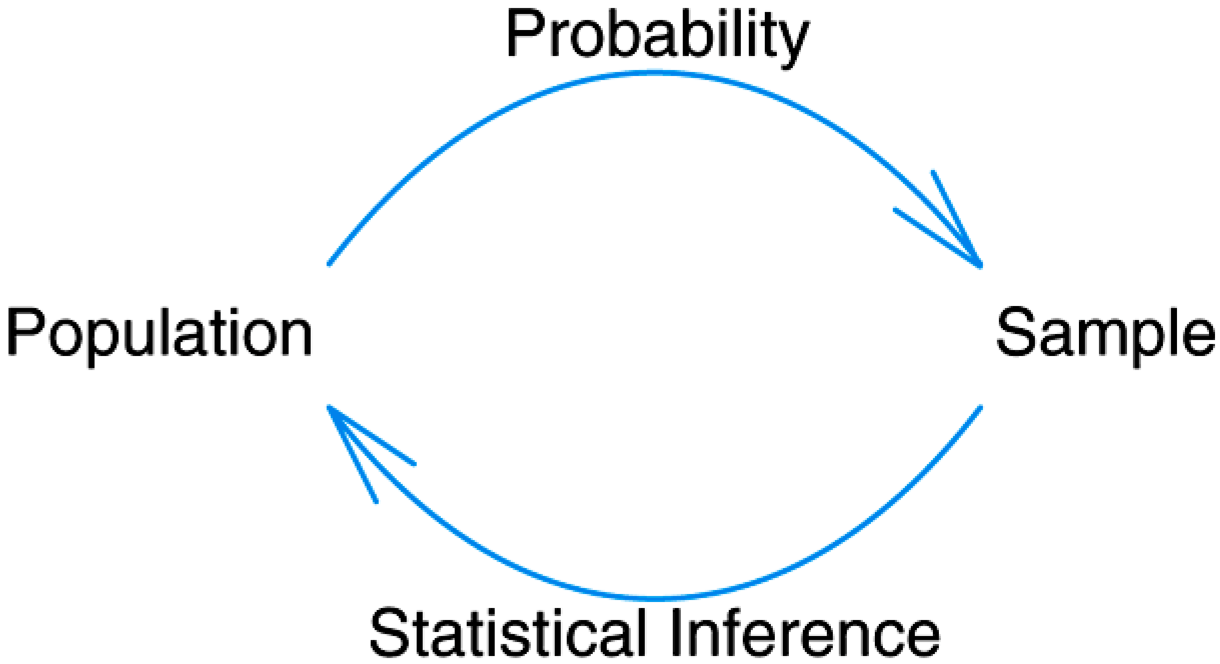


No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

The Dot Plot



Fundamental Relationship between Probability and Inferential Statistics



Probability

Population

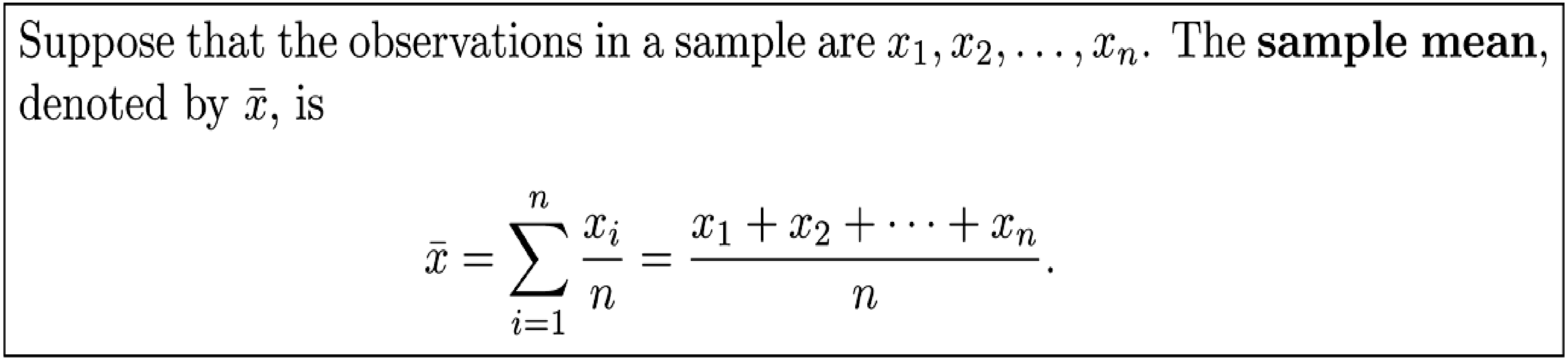
Sample

Statistical Inference

Measures of Location (Central Tendency)

- The data (observations) often tend to be concentrated around the center of the data.
- Some measures of location are: the mean, mode, and median.
- These measures are considered as representatives (or typical values) of the data. They are designed to give some quantitative measures of where the center of the data is in the sample.

Sample Mean



Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Example

Suppose that the following sample represents the ages (in year) of a sample of 3 men:

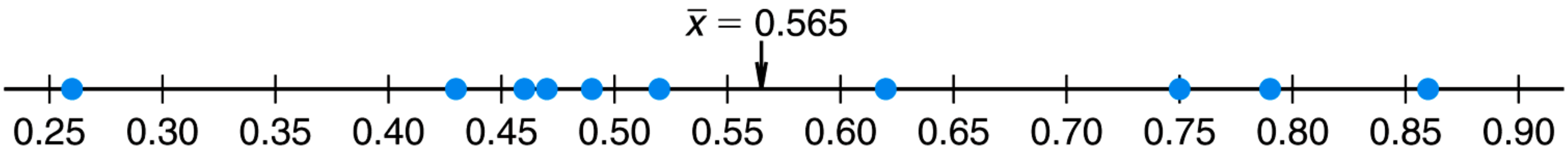
$$x_1 = 30, x_2 = 35, x_3 = 27.$$

Then, the sample mean is:

$$\bar{x} = \frac{30+35+27}{3} = \frac{92}{3} = 30.67$$

Note: $\sum_{i=1}^{3} (x_i - \bar{x}) = (30-30.67)+(35-30.67)+(27-30.67)=0$

Sample Mean as a Centroid of the with-nitrogen stem weight



Median

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

- e.g. 4, 2, 1, 4, 5, 2, 1

  1, 1, 2, 2, 4, 4, 5

  Therefore the median is 2

- e.g. 4, 2, 1, 4, 5, 2

  1, 2, 2, 4, 4, 5

  Therefore the median is (2 + 4)/2 = 3

Mode

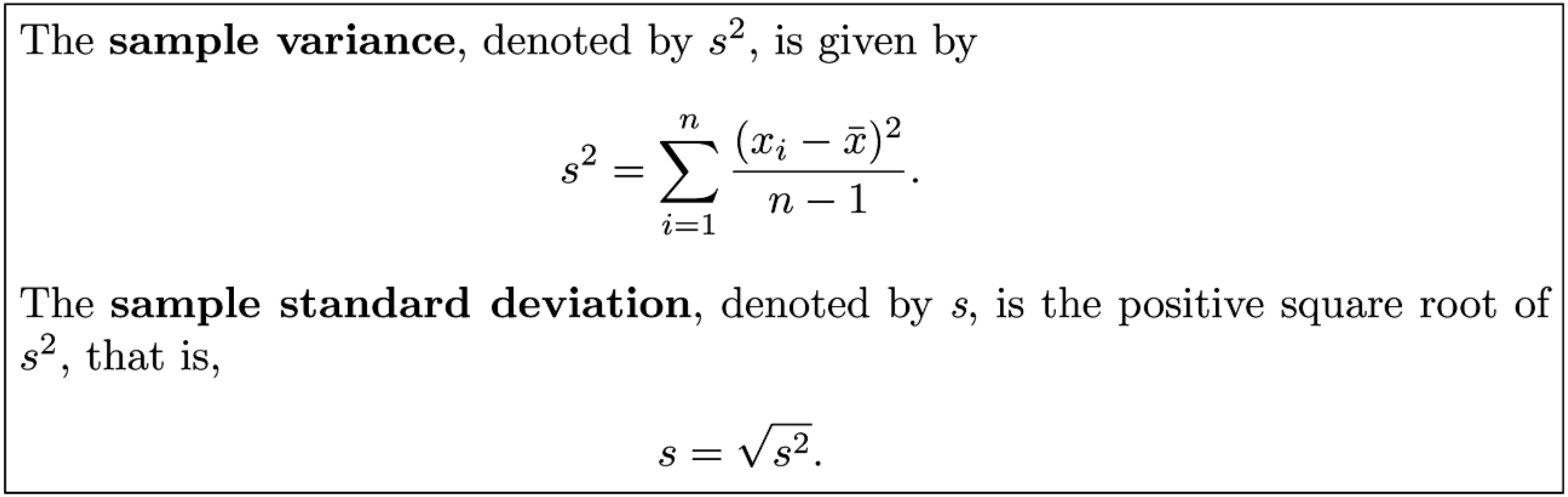
- The **mode** of a set of quantitative data is the most frequently occurring measurement in a data set.
- If no measurements occurring more than once, then there is no mode.
- There may be several modes if there are more than one data with the same most frequently occurring.

e.g. 2, 4, 5, 1, 7, 9, 0 : No mode

2, 4, 2, 5, 4, 2 : Mode is 2

2, 4, 2, 5, 4, 2, 4, 7 : Modes are 2 and 4

Sample Variance



The **sample variance**, denoted by s^2, is given by

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

The **sample standard deviation**, denoted by s, is the positive square root of s^2, that is,

$$s = \sqrt{s^2}.$$

Example 1

Compute the sample variance and standard deviation of the following observations (ages in year): 10, 21, 33, 53, 54.

Solution:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{10+21+33+53+54}{5} = \frac{171}{5} = 34.2 \text{ year}$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{5} (x_i - 34.2)^2}{5-1}$$

$$= \frac{(10-34.2)^2 + (21-34.2)^2 + (33-34.2)^2 + (53-34.2)^2 + (54-34.2)^2}{4}$$

$$= \frac{1506.8}{4} = 376.7 \quad (\text{year})^2$$

$$s = \sqrt{s^2} = \sqrt{376.7} = 19.41 \text{ year}$$

Example 2

A sample of 10 students scored the following grades: 40, 42, 35, 54, 57, 54, 46, 42, 54, 57.

(i) Find the sample mean, mode and median.

(ii) Compute the standard deviation.

Solution:

(i) Listing the score in order: 35, 40, 42, 42, 46, 54, 54, 54, 57, 57

$$\text{Mean} = \bar{x} = \frac{35+40+42+42+46+54+54+54+57+57}{10} = 48.1$$

$$\text{Mode} = 54 \qquad \text{Median} = \tilde{x} = \frac{46+54}{2} = 50$$

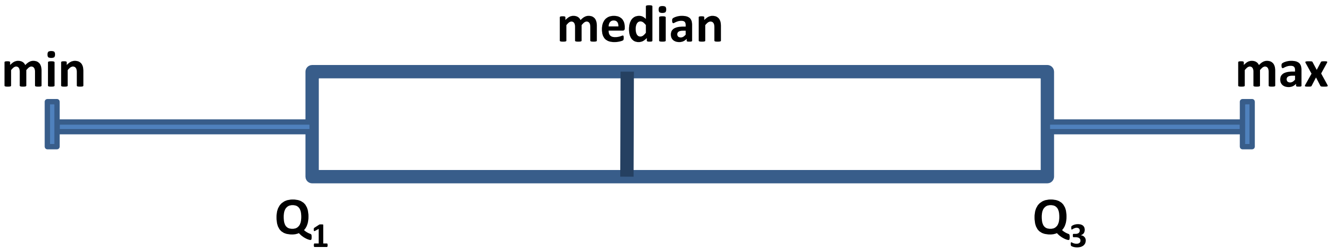
$$\text{(ii)}\ s = \sqrt{\frac{1}{9}[(35-48.1)^2 + (40-48.1)^2 + \cdots + (57-38.1)^2]} = 8.1$$

Comparing

- The range is the numerical difference between the largest and the smallest value of a set of a batch of data:

$$\text{range} = \max - \min$$

- The lower quartile, denoted by Q_1, is the median of the lower half of the batch of data.
- The upper quartile, denoted by Q_3, is the median of the upper half of the batch of data.
- The inter-quartile range, is defined by $Q_3 - Q_1$.
- A Box-plot is a diagram consisting of box and whiskers displays the median, the quartiles and maximum and minimum values in a batch of data.



Example 1

For the batch of data

4, 5, 6, 6, 7, 11, 12, 14, 16, 20, 22, 29

Min = 4

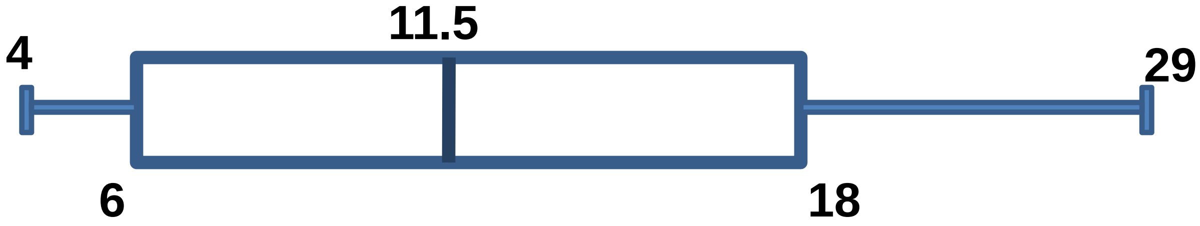
Max = 29

$Q_1 = (6 + 6)/2 = 6$

$Q_3 = (16 + 20)/2 = 18$

Inter-quartile range = 18 – 6 = 12

Median = (11 + 12)/2 = 11.5



Example 2

The table below gives the gross weekly earning including overtime in pounds of 20 actors working in a theatre (9 women and 11 men):

Women	221	272	334	361	372	399	415	456	510		
Men	258	315	333	353	398	420	435	462	495	523	587

(a) Draw an accurate diagram of the box-plots.

(b) What do box-plots tell you about the relative earnings of male and female actors.

Example 2

For women

Min = 221

Max = 510

$Q_1 = (272 + 334)/2 = 303$

$Q_3 = (415 + 456)/2 = 435.5$

Median = 372

For men

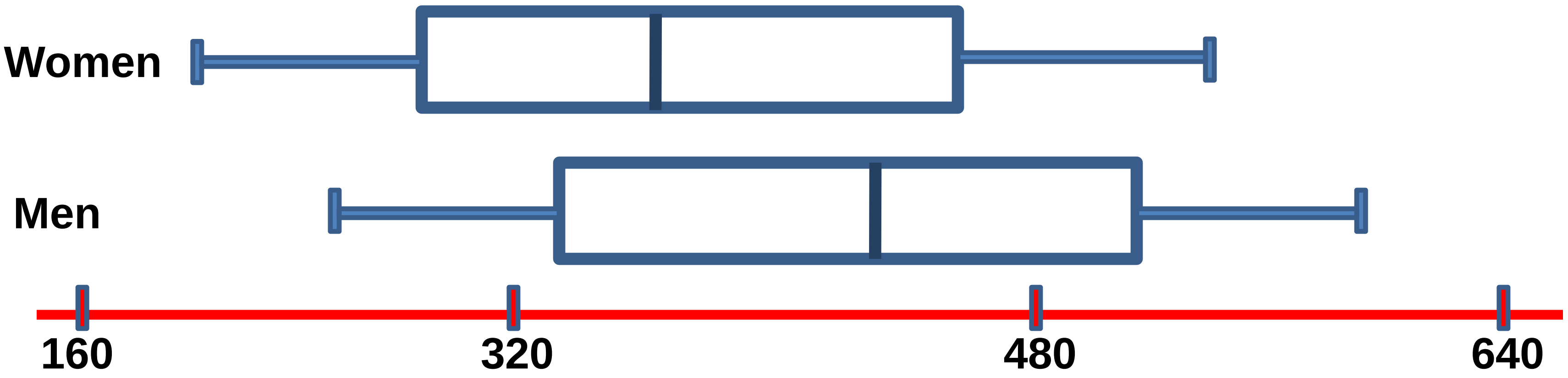
Min = 258

Max = 587

$Q_1 = 333$

$Q_3 = 495$

Median = 420



Example 2

CONTINUED

From the box-plots it is clear that the men’s earnings are higher than the women’s: all the five values marked on the box-plots are higher for men than for the women.