Probability and Statistics
Lecture 1: Introduction to Data Analysis

to accompany

Probability and Statistics for Engineers and Scientists

Fatih Cavdur

Chapter 1: Introduction to Data Analysis

- Statistical Inference
- Samples, Populations
- The Role of Probability

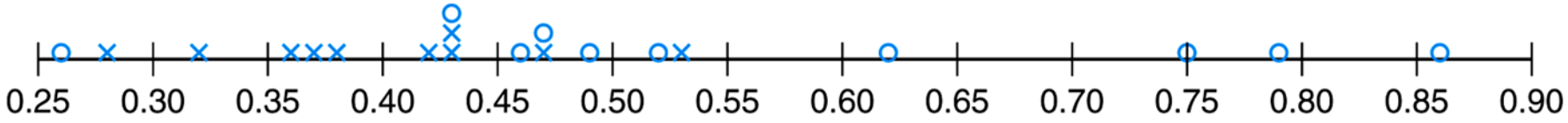
Example 1.2

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

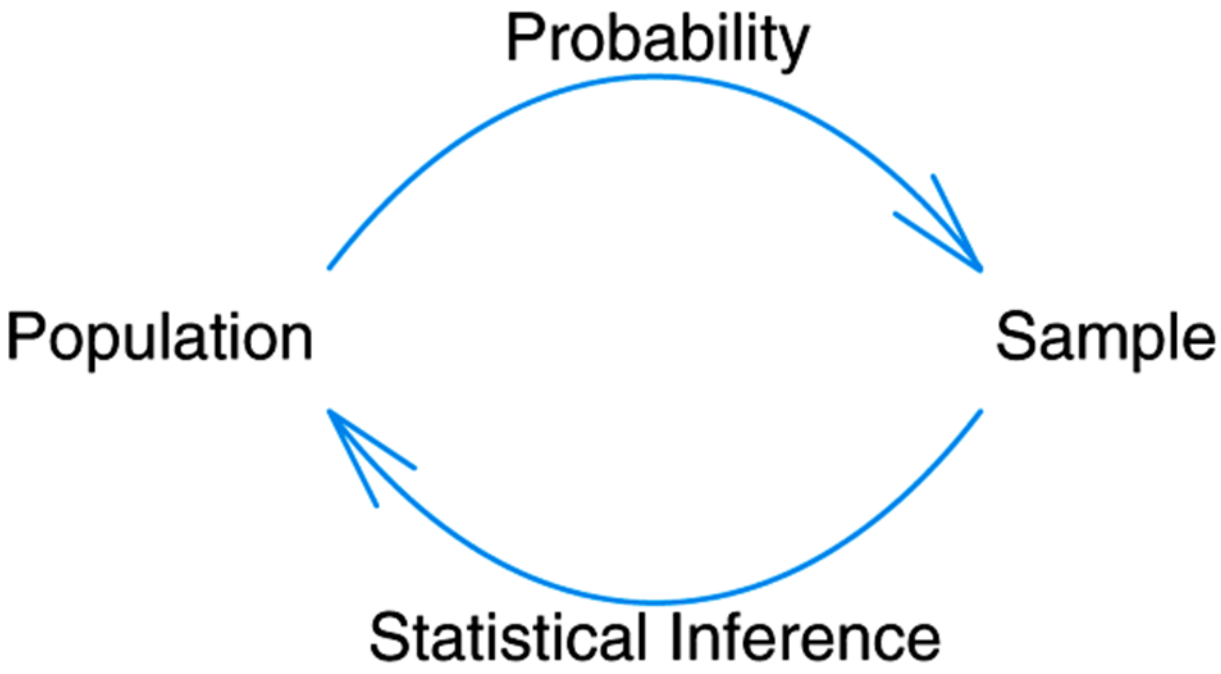
- Study to develop a relationship between the roots of trees and the action of a fungus.
- Minerals are transferred from the fungus to the trees and sugar from the trees to the fungus.
- 2 samples of 10 seedlings planted treated with nitrogen and no-nitrogen where all other conditions were held constant.
- Stem weights are recorded in grams after the end of 140 days as shown in the table.

Example 1.2 (cont.)

- A dot plot of stem weight data is shown below where “o” and “x” correspond to nitrogen and no nitrogen, respectively.
- How do you interpret the plot?



Probability and Statistical Inference



Sampling Procedures: Data Collection

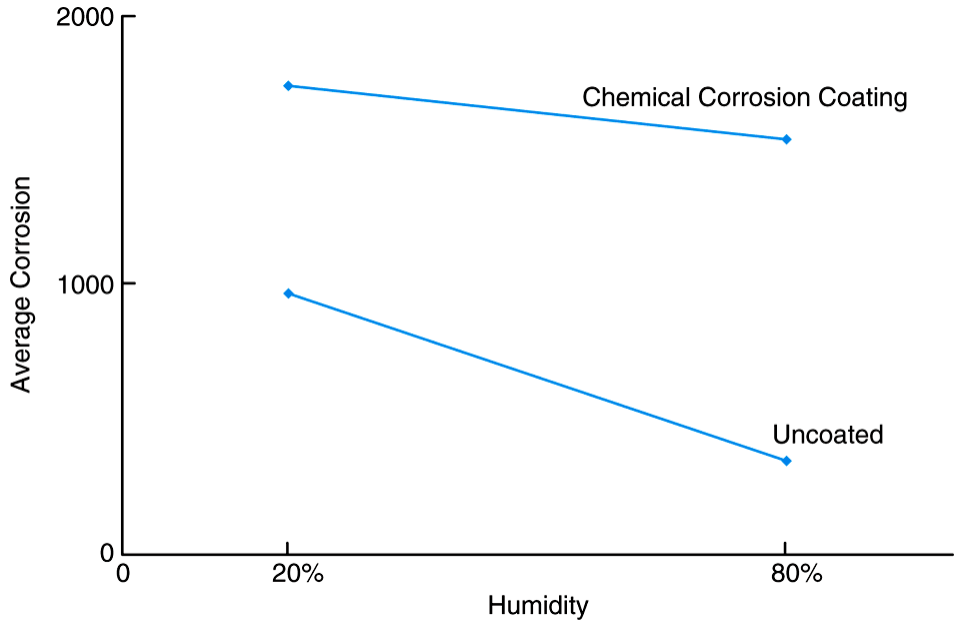
- *Simple Random Sampling* implies that any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size.
- If it is not true, we can mention a *biased sample*.
- The concept of randomness or random assignment plays a huge role in the area of *experimental design*.
- In Example 1.2, we can mention two different of variability in the experiment, within a group and between the groups. What do they mean?

Example 1.3

- A corrosion study was made in order to determine whether coating an aluminum metal with a corrosion retardation substance reduced the amount of corrosion.
- Also the influence of humidity on the amount of corrosion is of interest.
- A corrosion measurement can be expressed in thousands of cycles to failure.
- 2 levels of 2 factors considered:
  - Factor 1: Coating; Levels: Coating and No Coating
  - Factor 2: Humidity; Levels: 20% Humidity and 80% Humidity

Example 1.3 (cont.)

Coating	Humidity	Average Corrosion in Thousands of Cycles to Failure
Uncoated	20%	975
	80%	350
Chemical Corrosion	20%	1750
	80%	1550



Measures of Location: Mean and Median

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

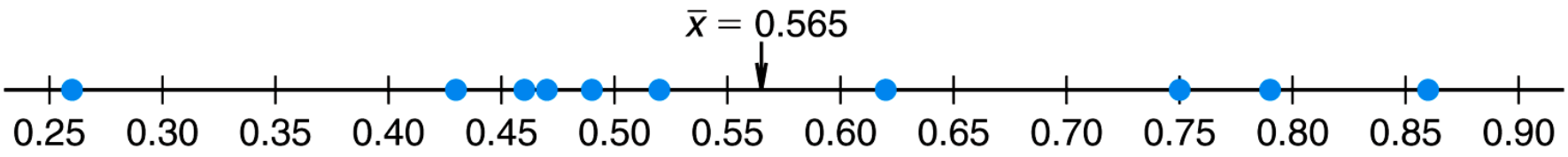
$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

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

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

Measures of Location: Mean and Median (cont.)

- In Example 1.2, the sample mean is computed as $\bar{x} = 0.565$ whereas the median is $\tilde{x} = \frac{0.49+0.52}{2} = 0.505$ within with nitrogen sample.
- Note that the mean is influenced considerably by the extremes whereas the median places emphasis on the true center of the data set.



Measures of Variability: Variance & Std. Dev.

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,

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

Measures of Variability: Variance & Std. Dev.

Assume that we are given the following data:

7.07; 7.00; 7.10; 6.97; 7.00; 7.03; 7.01; 7.01; 6.98; 7.08

The mean is

$$\bar{x} = \frac{7.07 + 7.00 + \cdots + 7.08}{10} = 7.025$$

and the variance of the above data is

$$s^2 = \frac{(7.07 - 7.025)^2 + \cdots + (7.08 - 7.025)^2}{9} = 0.001939$$

So the sample standard deviation is $s = \sqrt{0.001939} = 0.044$ with $n - 1 = 9$ degrees of freedom.

Graphical Diagnostics

- Scatter Plots
- Stem-and-Leaf Plot
- Histogram
- Box-and-Whisker or Box Plot

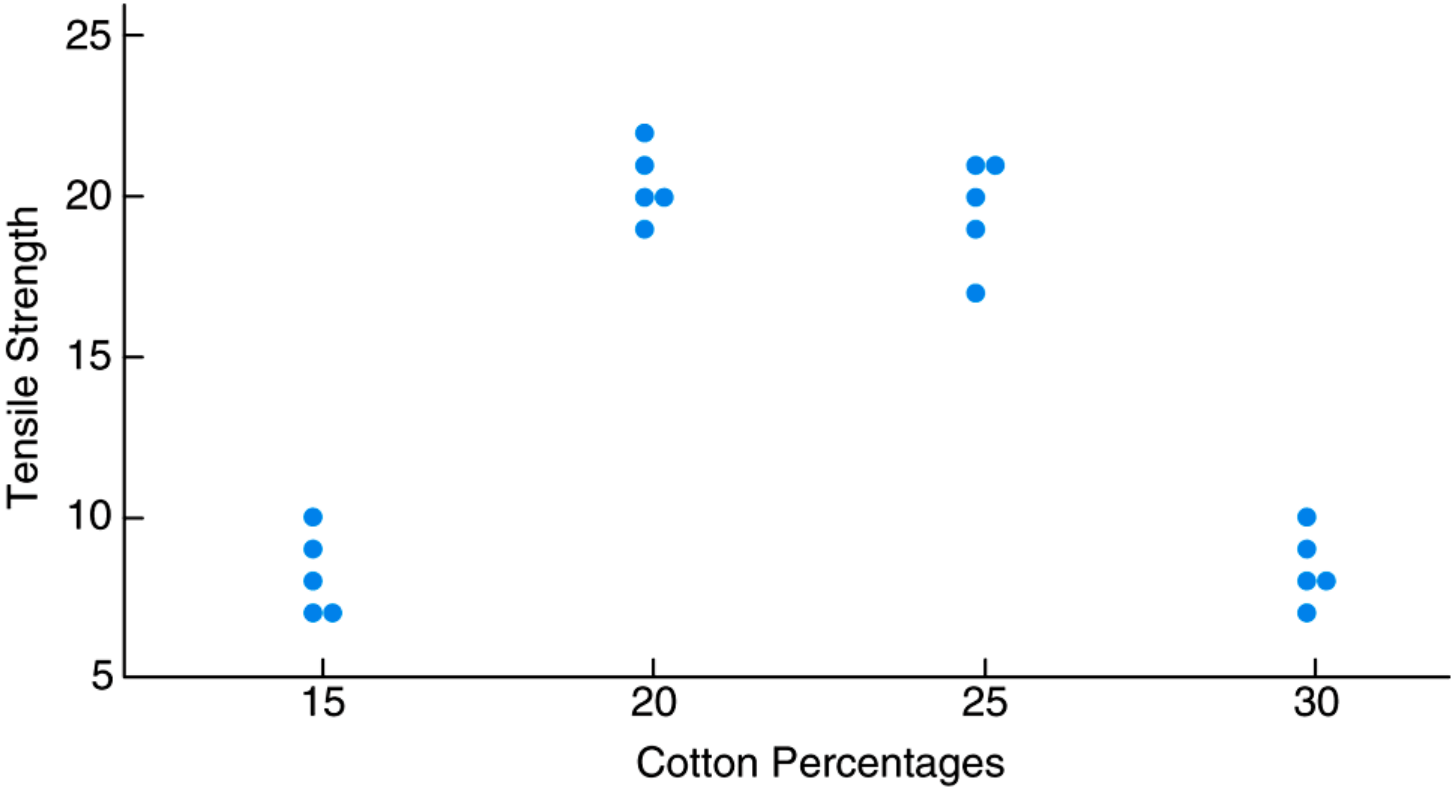
Scatter Plot

- Consider the data in Table 1.3 where 5 cloth specimens are manufactured for each of the 4 cotton percentages.
- What do you want to visualize here, and how can you do it?

Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Scatter Plot (cont.)

- Scatter Plot of Tensile Strength and Cotton Percentages



Stem-and-Leaf Plot

- Consider the data in Table 1.4 where the lives of 40 similar car batteries recorded to the nearest tenth of a year.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Stem-and-Leaf Plot (cont.)

- Stem-and-Leaf Plot of Car Battery Life

Stem	Leaf	Frequency
1	69	2
2	25669	5
3	00111122233344455677889 9	25
4	11234577	8

Stem-and-Leaf Plot (cont.)

- Double Stem-and-Leaf Plot of Car Battery Life

Stem	Leaf	Frequency
1·	69	2
2⋆	2	1
2·	5669	4
3⋆	001111222333444	15
3·	5567778899	10
4⋆	11234	5
4·	577	3

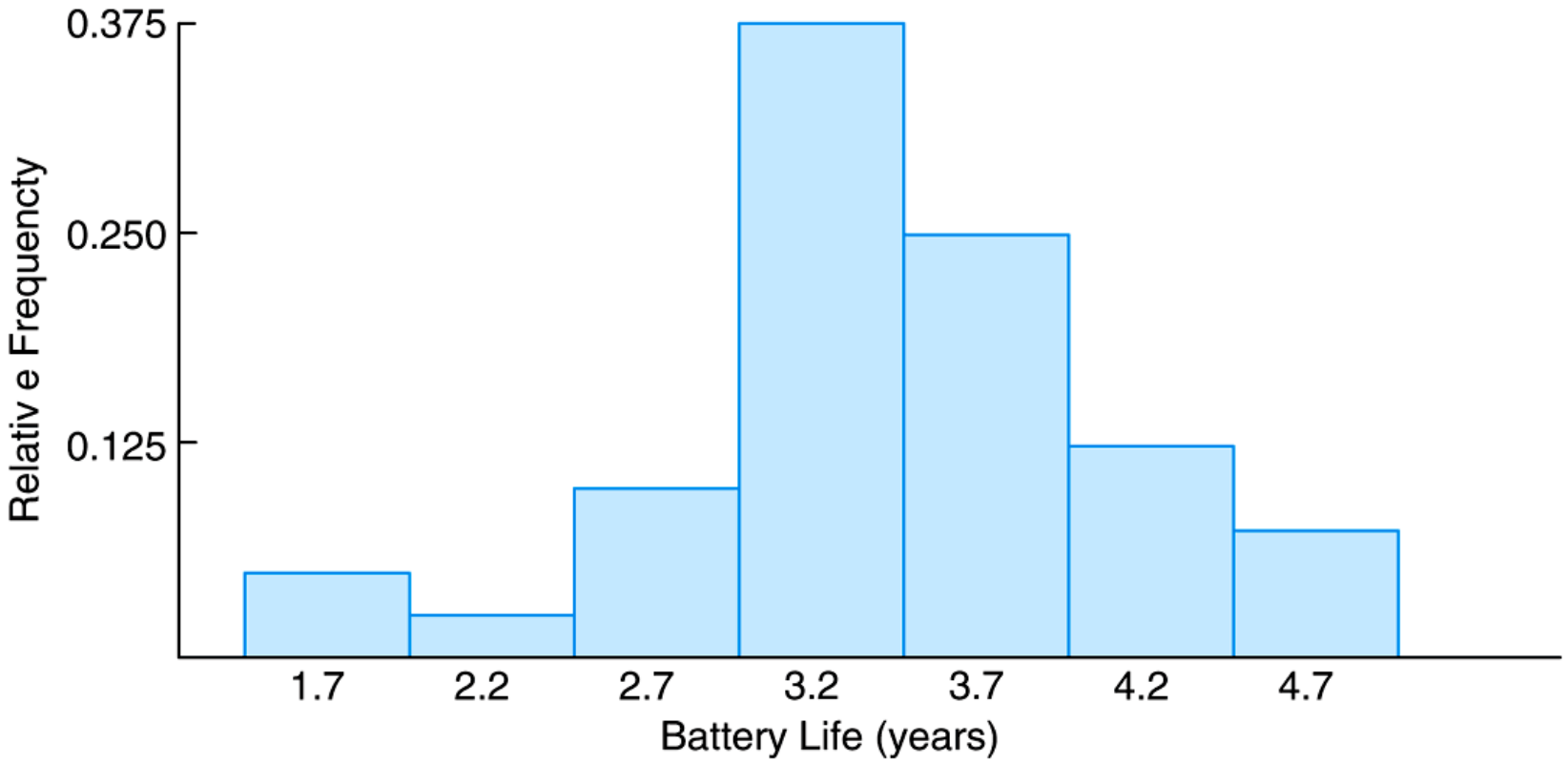
Histogram

- A table listing the frequencies is called a relative frequency distribution as shown below for the car battery life data:

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

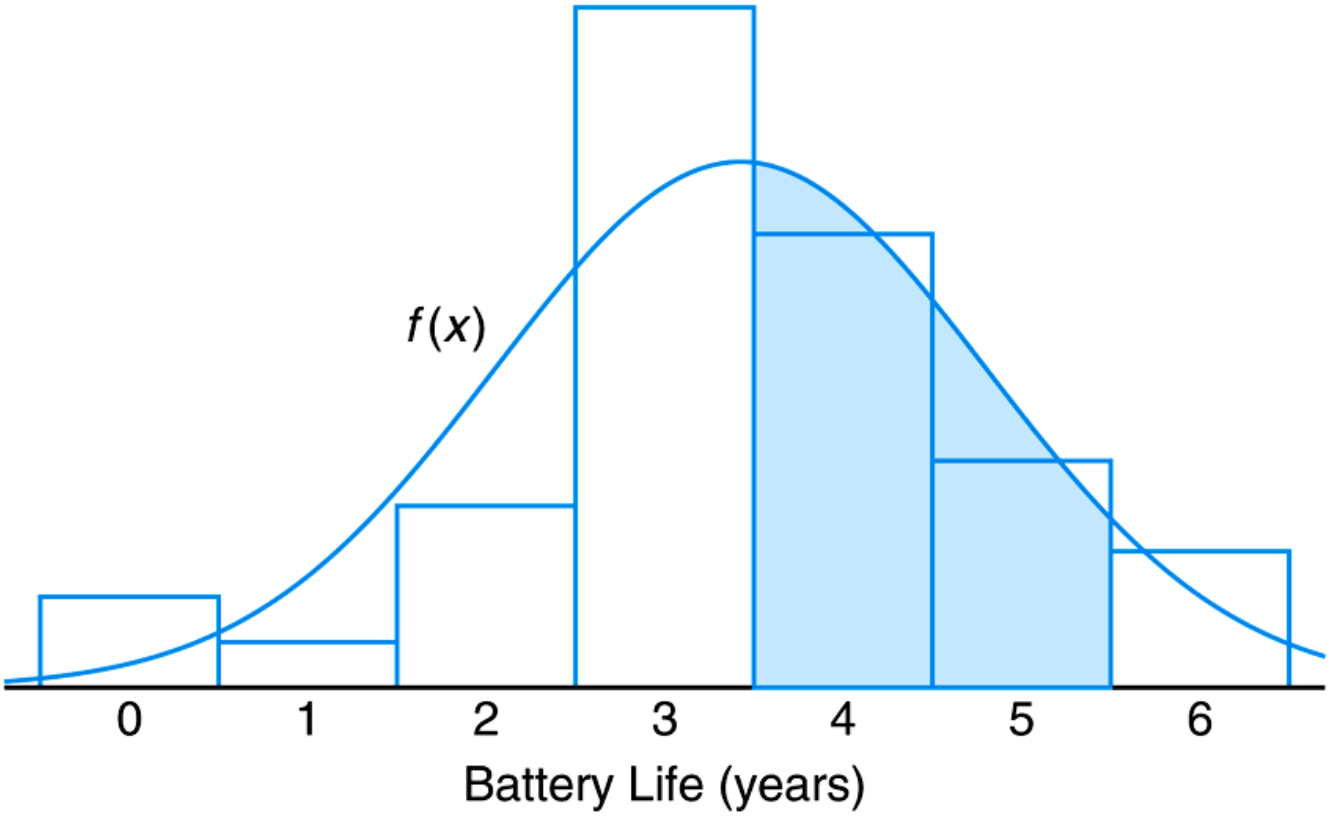
Histogram (cont.)

- Relative Frequency Histogram of Car Battery Life Data



Histogram (cont.)

- Can you estimate the underlying distribution from the data?



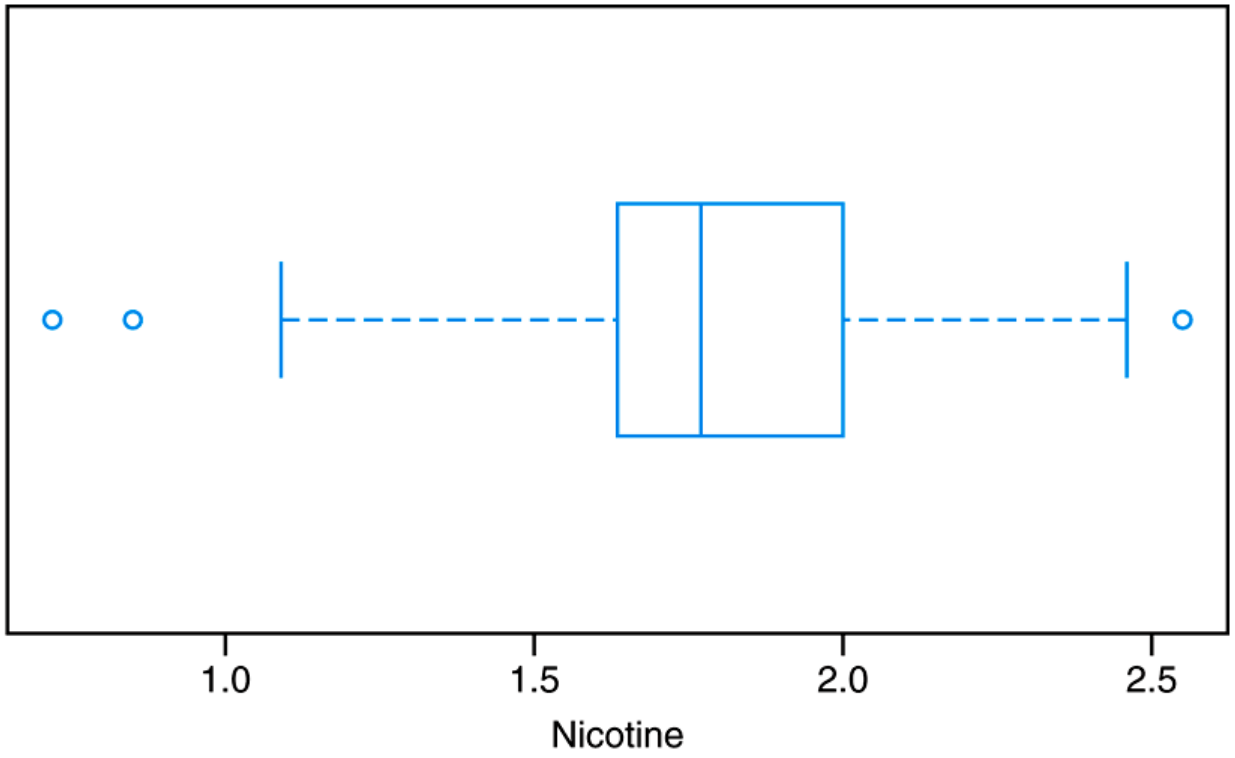
Box-and-Whisker Plot or Box Plot

- Nicotine content was measured in a random sample of 40 cigarettes as follows:

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

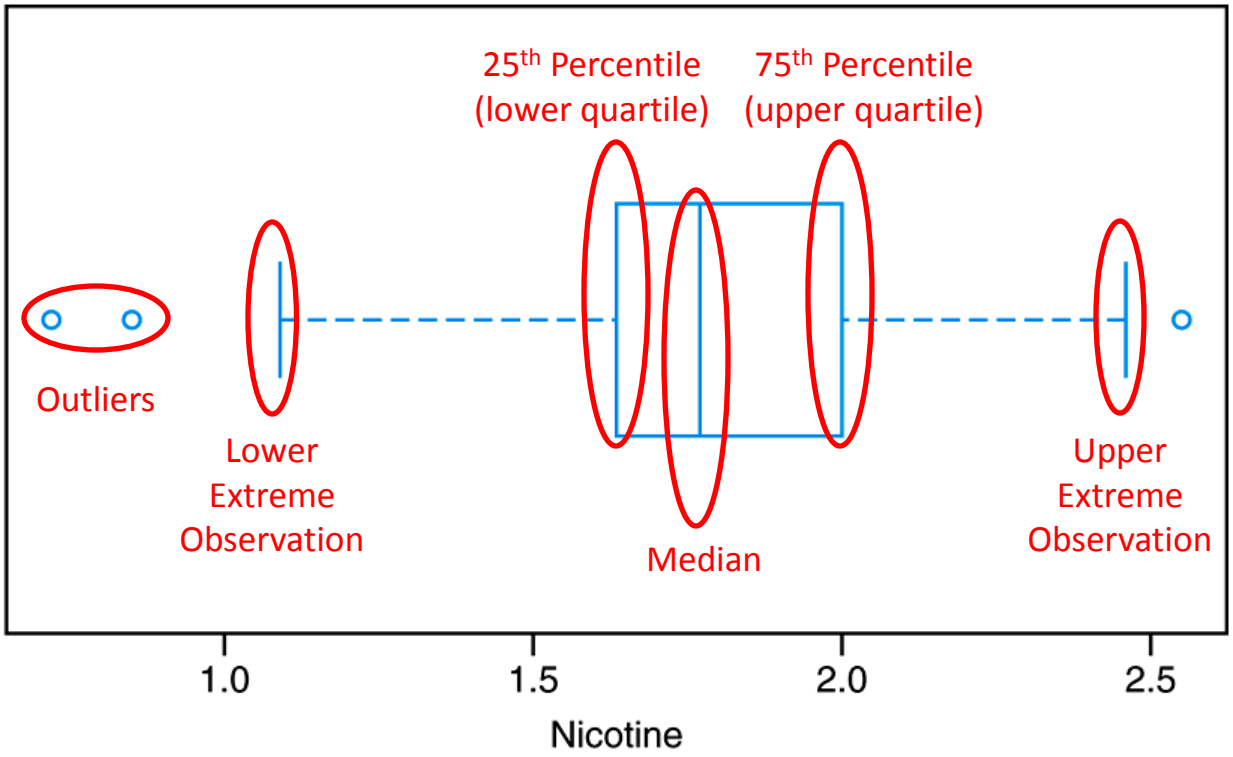
Box-and-Whisker Plot or Box Plot (cont.)

- Box-and-Whisker Plot for Nicotine Data



Box-and-Whisker Plot or Box Plot (cont.)

- Box-and-Whisker Plot for Nicotine Data



End of Lecture

Thank you! Questions?