Welcome to ENGR 21: Probability and Statistics for Engineers I

Introduction to Statistics and Data Analysis

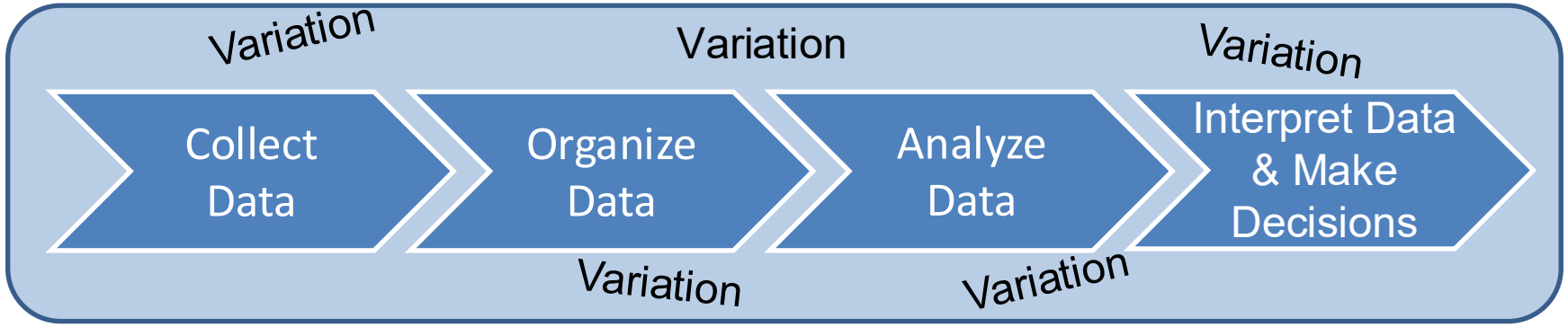
Chapter 1

Learning Objectives

- Statistical Inference, Samples, Populations and the Role of Probability
- Sampling Procedures; Collection of Data
- Measures of Location: The Sample Mean and Median
- Measures of Variability
- Statistical Modeling, Scientific Inspection and Graphical Diagnostics

Introduction

- Statistics helps us to make judgments and decisions when there is uncertainty
- How?
  - Methods for carrying out experiments
  - Methods for organizing and analyzing experiment results
  - Methods for interpreting data and making decisions



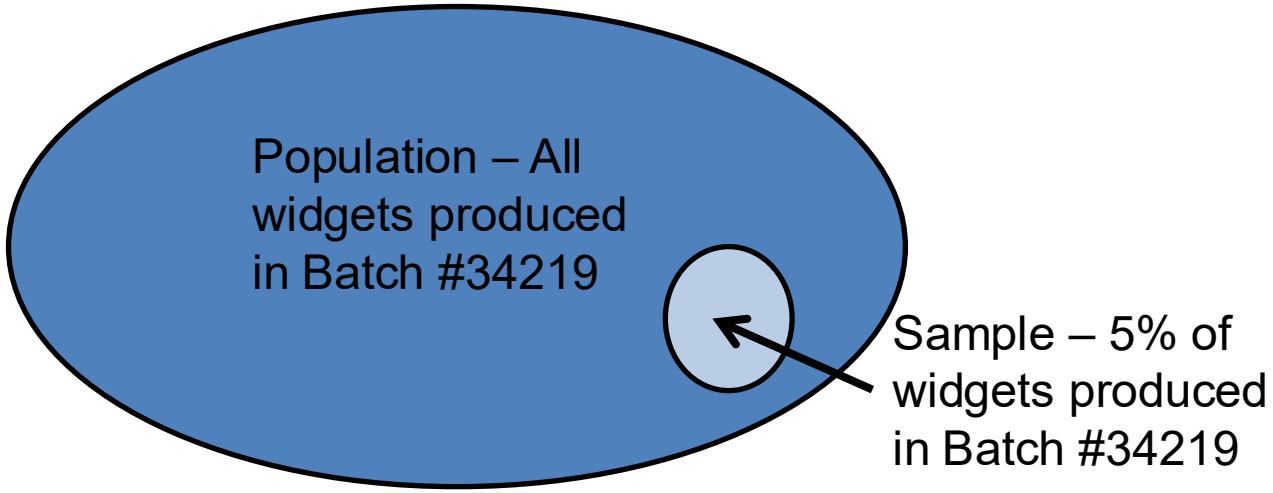
Variability in Scientific Data

- Example – Manufacturing
  - Differences in the material from product to product
  - Differences between batches
  - Differences within batches
- Another Example?

temperature
price of gas

Populations & Samples

- **Population** – All individuals or individual items of a particular type
- If collect data on entire population – **census**
- Normally, we only have a subset of the population – **sample**
- Example:



Populations & Samples

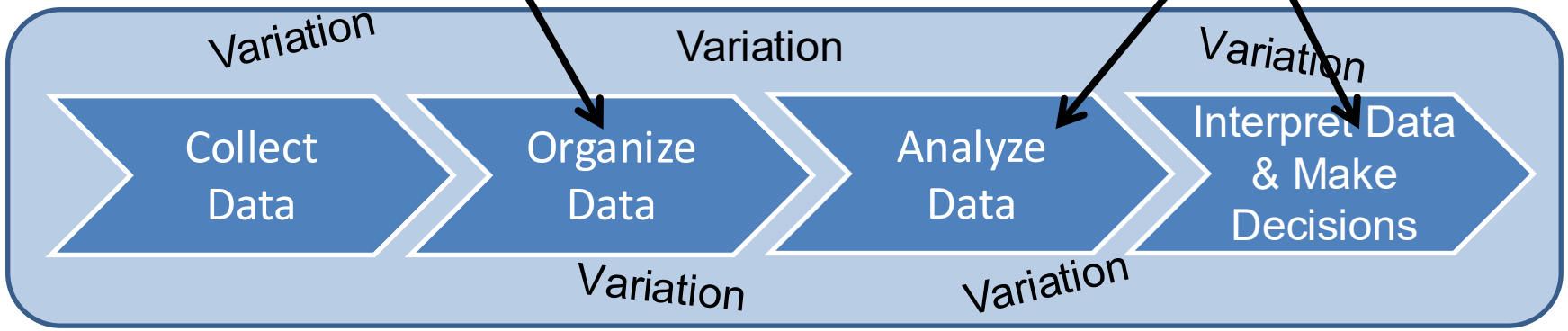
- Often, it's important to collect the sample and data in a systematic way (**experimental design**)

- Interested in the effect of certain characteristics (**factors**) on the widget
  - **Factor 1** – machine temperature (make observations at 3 different temps)
  - **Factor 2** – production shift (make observations on all shifts)

The Role of Probability

Descriptive Statistics – summarize and describe important features of the *sample* data

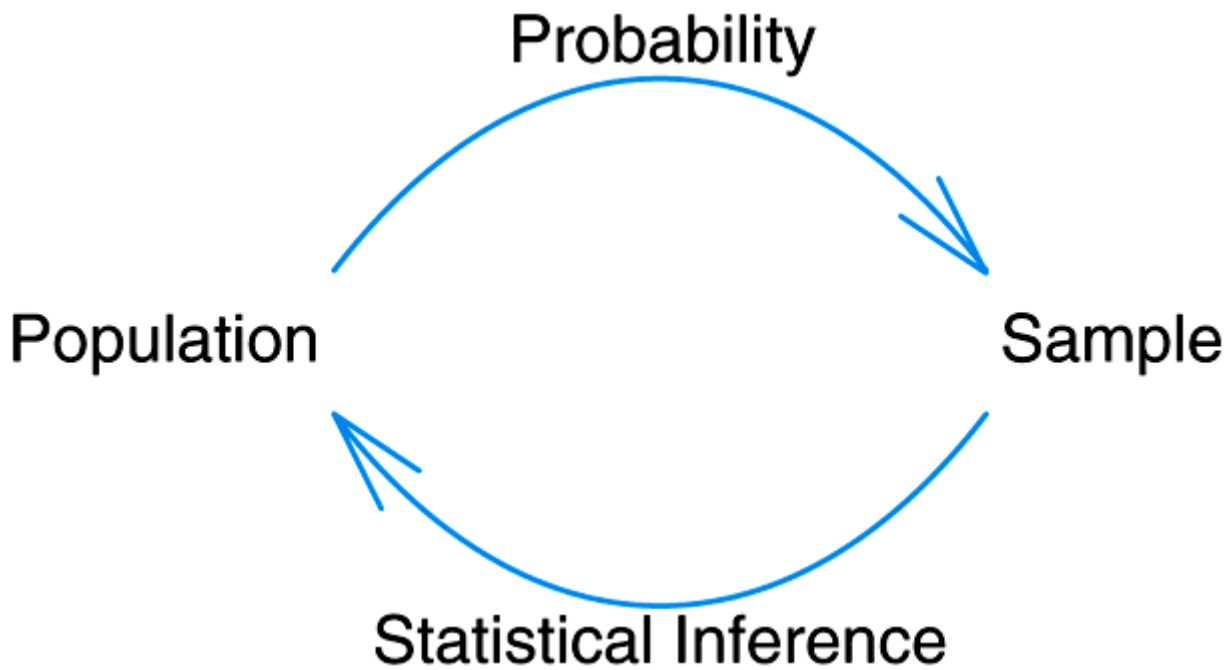
Inferential Statistics – techniques for drawing conclusions about the *population* based on information from the *sample*



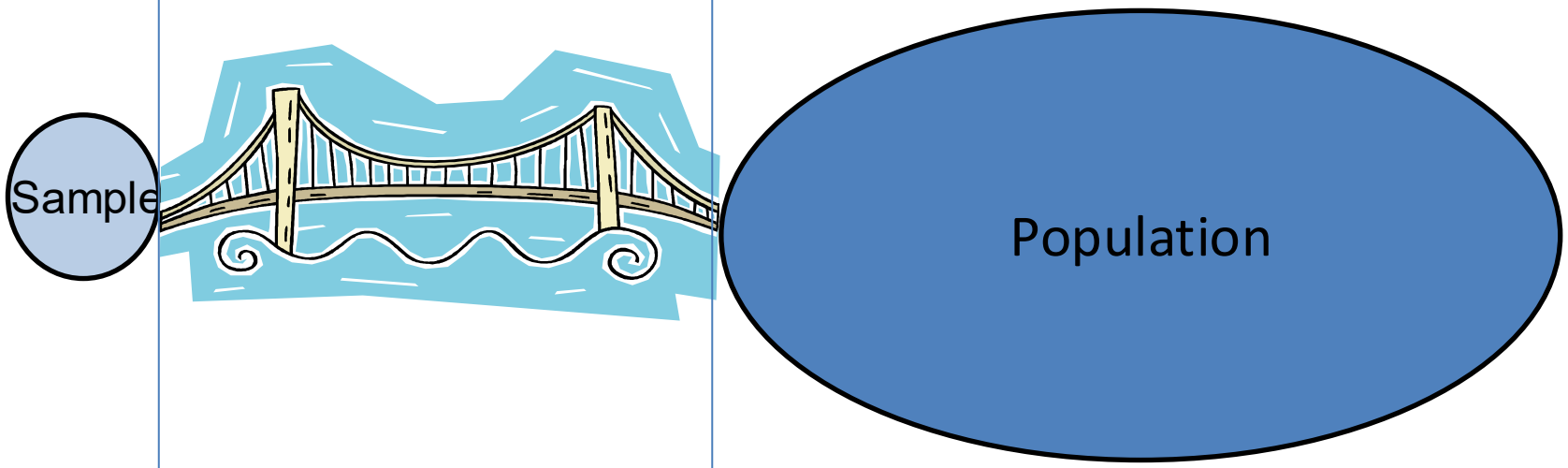
Probability – bridge between descriptive and inferential statistics. Explains the uncertainty associated with the *sample*.

NOTE: Before we can understand what a particular sample tells us about the population, we need to understand the uncertainty associated with taking a sample from the population.

How do Probability & Statistical Inference Work Together?



Course Roadmap



Descriptive Statistics	Probability	Inferential Statistics
Chapter 1	Chapters 2 – 6	Chapters 8 – 11, 13 – 14

Sampling / Data Collection

- If data is not properly collected, then analysis can not be done with certainty
  - Sample might not be representative of population
  - Use entire population, choose a simple random sample, stratified sample, or design an experiment
- Example – if entire sample of widgets is taken at one time, might not be representative of the entire batch

Samples

- A **sample** is a subset of measurements selected from the population of interest

- Different Sampling methods:
  - Simple random sample - *n* items and each is equally likely to be chosen
  - Systematic random sample - take every m^{th} item
  - Convenience sample - a sample population selected because it is readily available and convenient
  - Stratified sample (subpopulation (stratum)) - take proportionally equally from m different groups

Measures of Location

- Sample Mean
- Median
- Mode

- Other measures of location that we will not cover:
  - Trimmed Mean
  - Quartiles & Percentiles
  - Sample Proportions

Notation

- Notation
  - n – number of observations in sample (sample size)
  - N – number of elements in population (if finite)
  - $x_1, x_2, x_3, \ldots, x_n$ – individual observations

The Sample Mean

- Arithmetic average of the set
- Notation:
  - Sample mean = $\bar{x}$
  - Population mean = μ
  - $\bar{x}$ is an estimate of μ

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- $\bar{x}$ is the center of the sample as it relates to location

Example - Mean

- Grade point averages of 15 University of Pittsburgh Engineering Freshmen:

2.0	2.7	2.8
2.3	3.1	2.8
2.9	2.6	1.9
2.7	3.0	3.1
2.6	2.5	2.7

$\bar{x} \approx 2.65$

Measures of Location: The Sample Median

- **Median**, unlike the mean, represents the middle value in the data where one-half the data is above and one-half the data is below when the data is ordered from lowest to highest.
- It is denoted in by $\tilde{x}$.
- The population median is denoted with $\tilde{\mu}$**.**

To compute the median:

  - If n is odd, order the data, the data point in the middle is the median.
  - If n is even, order the data, sum the two middle data points and divide by two.

Example - Median

- Grade point averages of 15 University of Pittsburgh Engineering Freshmen:

2.0	2.7	2.8
2.3	3.1	2.8
2.9	2.6	1.9
2.7	3.0	3.1
2.6	2.5	2.7

- First we order the numbers:

1.9	2.0	2.3	2.5	2.6	2.6	2.7	2.7	2.7	2.8	2.8	2.9	3.0	3.1	3.1

- The median is
- What is the median of the following sample? (n=14)

10	14	15	16	16	16	17	24	27	29	32	34	39	47

- The median is 20.5

Measures of Central Tendency or Location

- **Mode**, a third measure of central tendency, is the single value that occurs most often in the data.
- **Example**: Determine the mode(s) in the GPA data.

2.0	2.7	2.8
2.3	3.1	2.8
2.9	2.6	1.9
2.7	3.0	3.1
2.6	2.5	2.7

2.7

Mean and Median – Sensitivity to Outliers

- Grade point averages of 15 University of Pittsburgh Engineering Freshmen:

2.0	2.7	2.8
2.3	3.1	2.8
2.9	2.6	1.9
2.7	3.0	3.1
2.6	2.5	2.7
Mean	2.65	

- First we order the numbers:

1.9	2.0	2.3	2.5	2.6	2.6	2.7	2.7	2.7	2.8	2.8	2.9	3.0	3.1	3.1

- Since *n* =15 is odd, the median is the $\{(n+1)/2\}^{th}$ ordered value.

2.7

Mean and Median – Sensitivity to Outliers

- Grade point averages of 15 University of Pittsburgh Engineering Freshmen:

2.0	2.7	2.8
2.3	3.1	2.8
2.9	2.6	1.9
2.7	3.0	33.1
2.6	2.5	2.7
Total	39.7 + (33.1 – 3.1) = 39.7 + 30 = 69.7	
Mean	4.65	

- First we order the numbers:

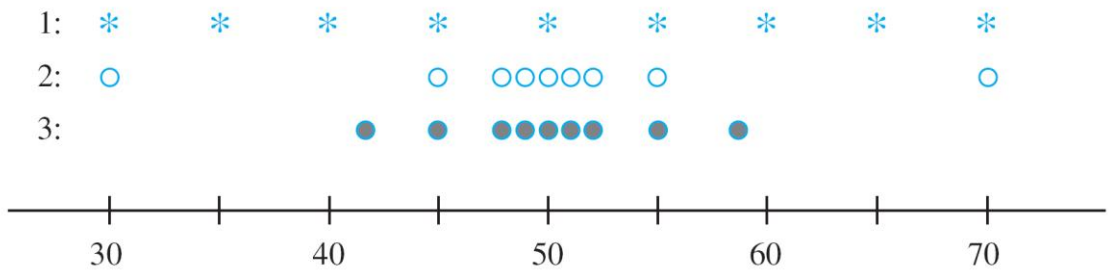
1.9	2.0	2.3	2.5	2.6	2.6	2.7	2.7	2.7	2.8	2.8	2.9	3.0	3.1	33.1

- Since *n* =15 is odd, the median is the {(n+1)/2}th ordered value.

2.7

Measures of Variability

- In addition to measures of location, we need measures to tell us how spread out the data is:
  - Range
  - Variance
  - Standard Deviation



Samples with identical measures of center but different amounts of variability

Measures of Variability

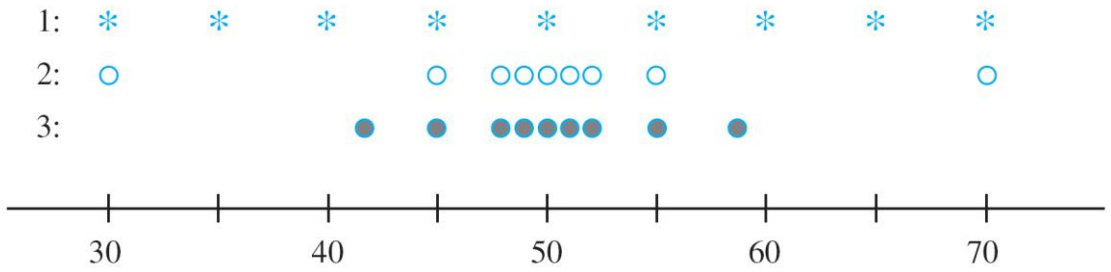
- In addition to a measure of location, we need a measure that tells us how spread out the data is.

- A simple measure of variability is **range**.

  Range = largest value - smallest value

Range

- Simplest measure of variability
- Range = largest value – smallest value
  - When range is large, variability is high
  - Limitation – only focuses on most extreme values



  - We are more interested in the deviations of individual observations from the mean…

Variance

- Deviations from the mean – subtract $\bar{x}$ from each of the n sample observations $(x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})$
  - Negative deviation – observation smaller than the mean
  - Positive deviation – observation larger than the mean
  - Small deviations – less variability
  - Large deviations – more variability
- Need to combine these deviations so we can easily interpret large sets of data. How?...
  - Thought 1: Average the deviations? No, this always results in a value of zero.
  - Thought 2: Average the absolute values of the deviations? OK, but algebra is kind of complicated
  - Thought 3: Square the deviations?

Variance & Standard Deviation

- Sample Variance $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
- Sample Standard Deviation $s = \sqrt{s^2}$
  - square root of the variance
- Population Variance $\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$
- Population Standard Deviation $\sigma = \sqrt{\sigma^2}$

Why Divide by n-1?

- Sample Variance: $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$
- Divide by $n-1$ rather than n?
  - The sample variance is based on n – 1 **degrees of freedom**.
  - If we know the sample mean and the values of all but one of the data points, the final data point is automatically determined.
  - Only $n-1$ observations are freely determined.

Example 1

- Observations: 4, 6, 9, 13, 18
- Estimate σ^2 , σ ?

Example 2

Observations: 4, 5, 9, 17, 18

- Estimate σ^2 , σ ?

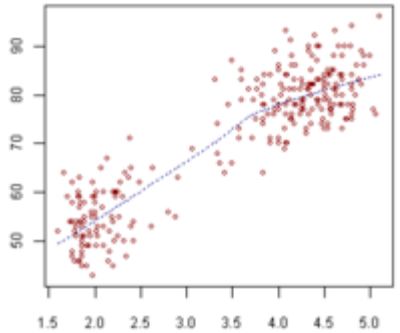
Compare Examples 1 & 2

Statistic	Example 1	Example 2
Sample Observations	4, 6, 9, 13, 18	4, 5, 9, 17, 18
Sample Mean, $\bar{x}$	10	10.6
Sample Range	14	14
Sample Variance, s^2	31.5	43.3
Sample Std Dev, s	5.61	6.58

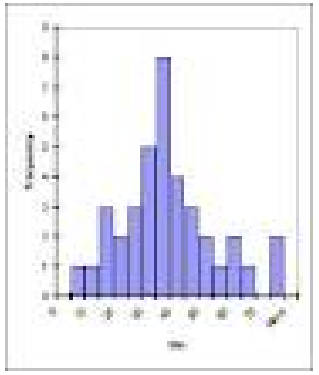
What can we say about these 2 samples?

Statistical Modeling, Scientific Inspection and Graphical Diagnostics

- Scatter Plot



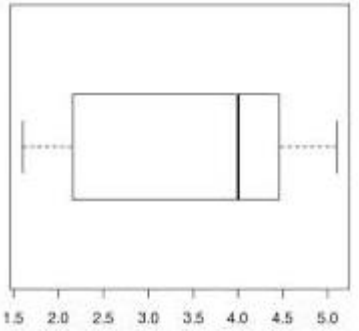
- Histograms



- Stem-and-Leaf Displays

Stem	Leaf
13	6, 9, 9
14	2, 3, 3, 3, 3, 4
14	6, 7, 7, 8, 9
15	1, 3, 4
15	6, 7
16	2, 4

- Box-and-W or Box Plot



Scatter Plot

- Example from p. 19
  - Textile manufacturer produces cloth with various percentages of cotton

Table 1.3: Tensile strength

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

  - How can graphics be used to determine distinction between samples?...

Example – Page 19

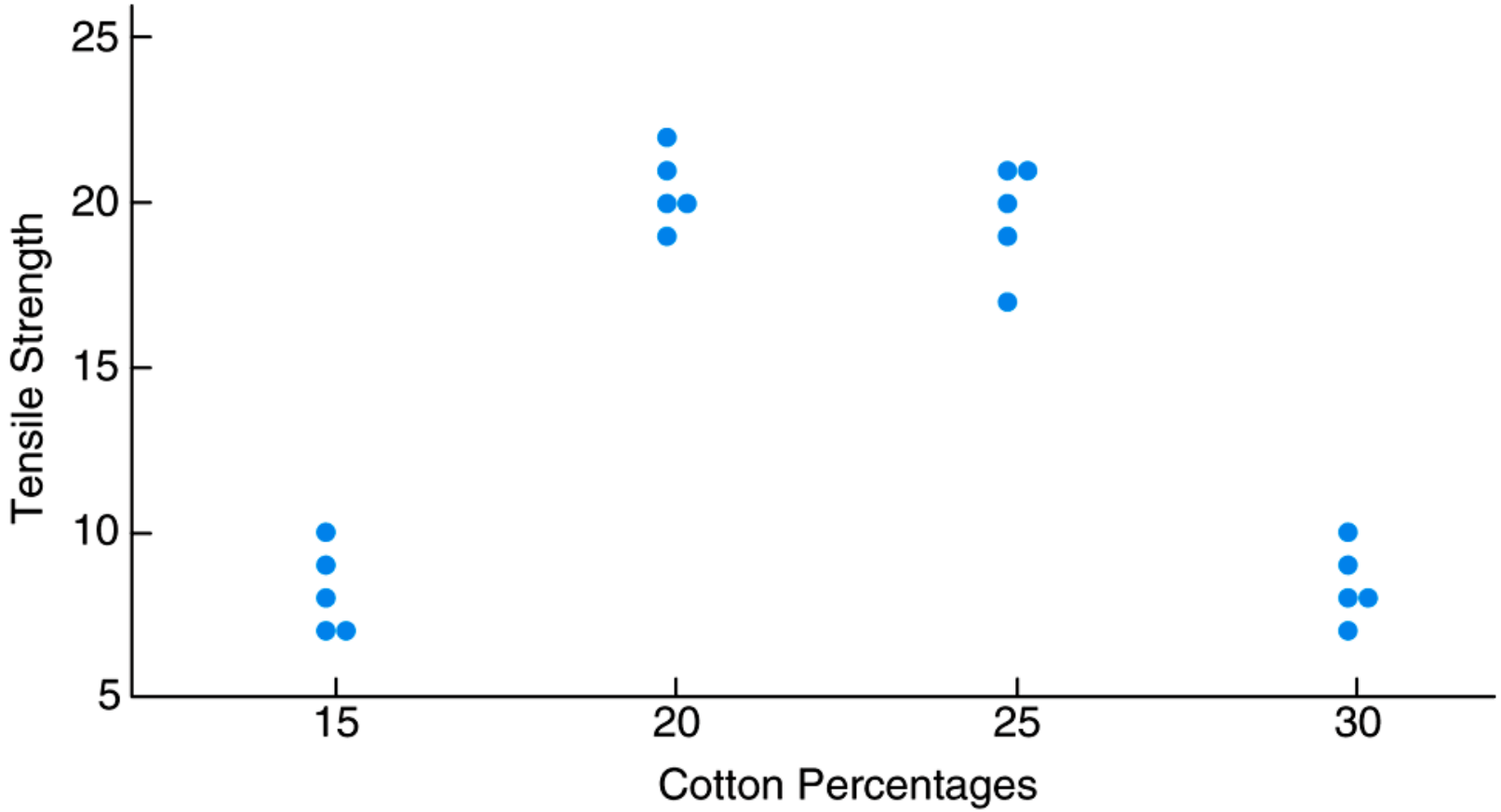


Figure 1.5 Scatter plot of tensile strength and cotton percentages

Stem-and-Leaf Displays

- Quick way to get visual representation of data
- How to make:
  1. Select one or more digits for stem values
  2. List stem values in vertical column
  3. Record leaf for each observation besides corresponding stem value
  4. Indicate units somewhere in display

Table 1.4 Car Battery Life

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

Table 1.5 Stem-and-Leaf Plot of Battery Life

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

Table 1.6 Double-Stem-and-Leaf Plot of 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

Example: Stem-and-Leaf Displays

Grade point averages of 15 University of Pittsburgh Engineering Freshmen:

2.0	2.7	2.8
2.3	3.1	2.8
2.9	2.6	1.9
2.7	3.0	3.1
2.6	2.5	2.7

Frequency Distribution and Histograms

- **Frequency** – # of times a particular value occurs in the data set
- **Relative Frequency** – proportion of times the value occurs

$$\text{Rel Freq} = \frac{\#\text{ times value occurs}}{\#\text{ observations in data set}}$$

- **Frequency Distribution** – a tabulation of the frequencies and/or relative frequency
- **Histogram** – vertical bar graph of the frequency distribution

Example – page 21

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

Table 1.7 Relative Frequency Distribution of Battery Life

Example – page 21(cont.)

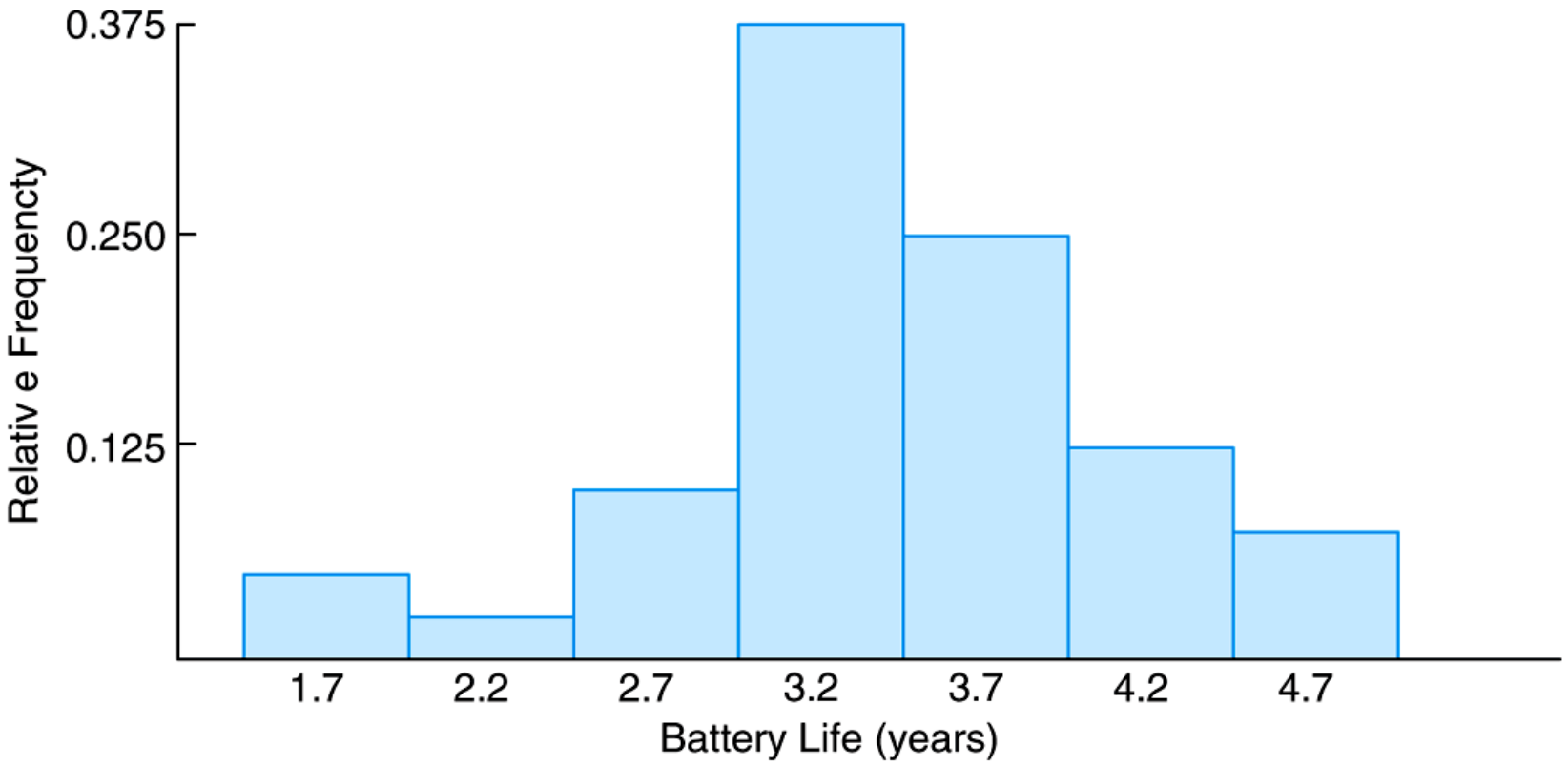


Figure 1.6 Relative frequency histogram

Example - Frequency Distribution and Histogram

- Grade point averages of 40 University of Pittsburgh Engineering Freshmen:

2.0	3.0	2.5	3.4	2.1
2.3	2.5	1.4	1.9	2.1
2.9	2.8	2.8	2.2	2.3
2.7	2.8	2.6	2.3	2.2
2.6	1.9	3.0	2.4	2.1
2.7	3.1	2.7	2.5	2.7
3.1	2.7	3.8	2.6	2.8
2.6	3.5	3.0	2.7	2.9

- We want a compact summary of data.

Example - Frequency Distribution and Histogram

2.0	3.0	2.5	3.4	2.1
2.3	2.5	1.4	1.9	2.1
2.9	2.8	2.8	2.2	2.3
2.7	2.8	2.6	2.3	2.2
2.6	1.9	3.0	2.4	2.1
2.7	3.1	2.7	2.5	2.7
3.1	2.7	3.8	2.6	2.8
2.6	3.5	3.0	2.7	2.9

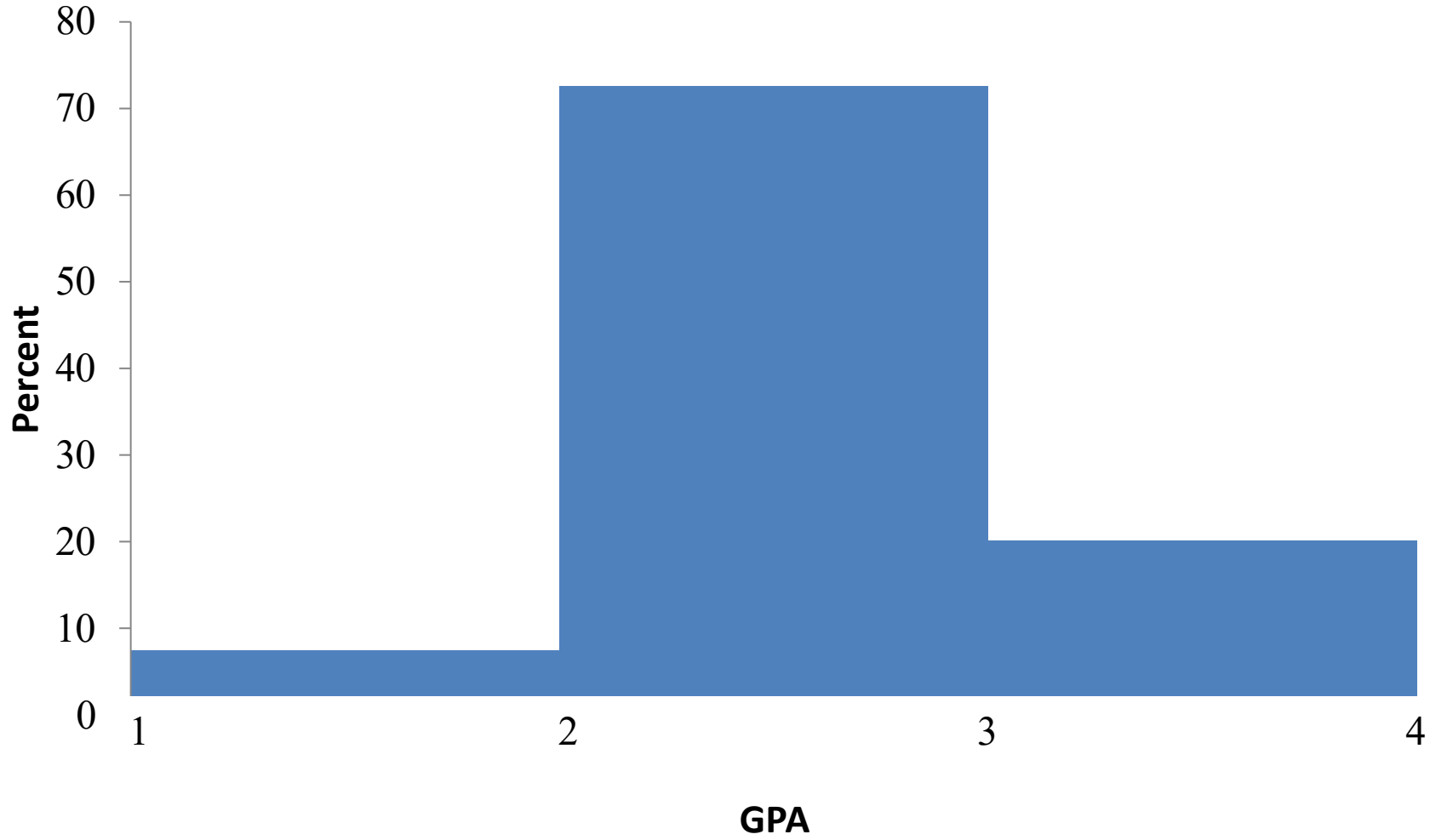
Class	1.0 - <2.0	2.0 - <3.0	3.0 - <4.0
Frequency			
Relative Frequency			

Example - Frequency Distribution and Histogram

- *Frequency distribution* - a grouped tabulation of data arranged according to size / a more compact summary of data than the original observations
- *Histogram* - a vertical bar graph used to make a picture of the frequency distribution.

Class	1.0 - <2.0	2.0 - <3.0	3.0 - <4.0
Frequency	3	29	8
Relative Frequency	(%7.5)	(%72.5)	(%20)

Example - Frequency Distribution and Histogram

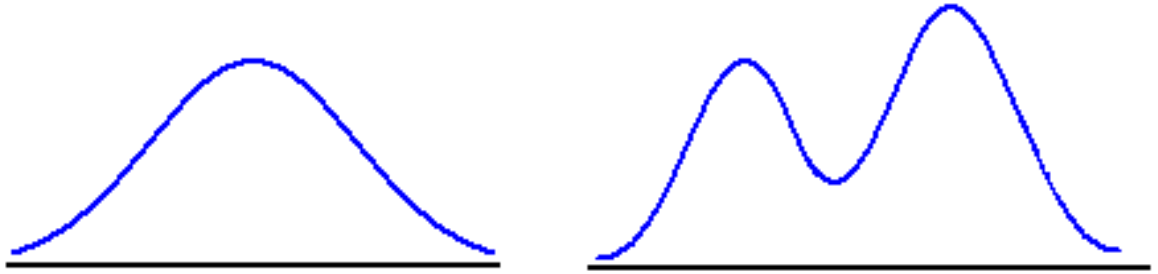


Frequency Distribution and Histogram

Steps for creating a frequency distribution and histogram:

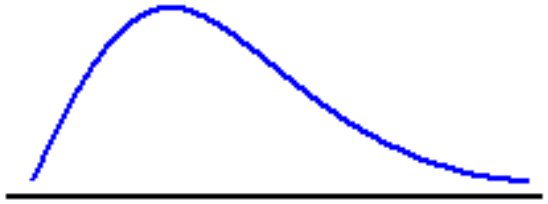
1. Divide the relevant measurement scale into a collection of disjoint (non-overlapping) equal size intervals such that each observation in the data can be placed into one of the intervals.
2. Once these class intervals (classes) have been chosen, list along the left margin.
3. Count the number of observations that fall into each class (frequencies).
4. Determine the relative frequency by dividing the frequencies by n.
5. Draw a horizontal line (x-axis) to represent the measurement scale and mark the boundaries of the adjacent class intervals.
6. Draw a vertical line (y-axis) to represent the relative frequencies.
7. Draw the bars to represent the relative frequency in each class.

Histogram Shapes

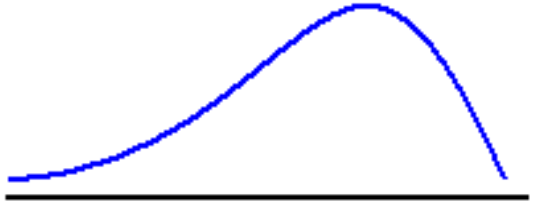


symmetric unimodal

bimodal



positively skewed



negatively skewed

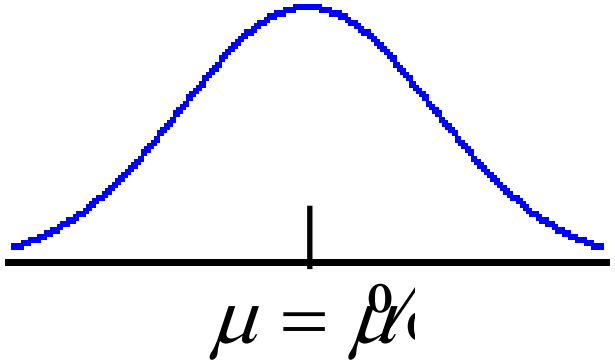
Peaks: A histogram may be unimodal, bimodal, or multimodal

Skewness: Looking at a histogram or a stem and leaf diagram helps us to determine if data are **symmetric**, **positively skewed** (skewed right), or **negatively skewed** (skewed left)

Outliers: These graphs also show **outliers**

Three Different Shapes for a Population Distribution

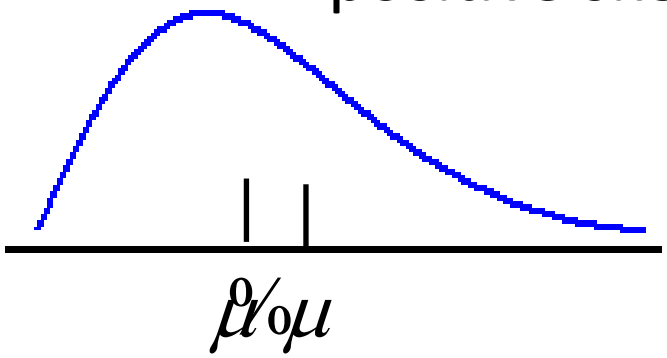
μ ; $\tilde{\mu}$



symmetric

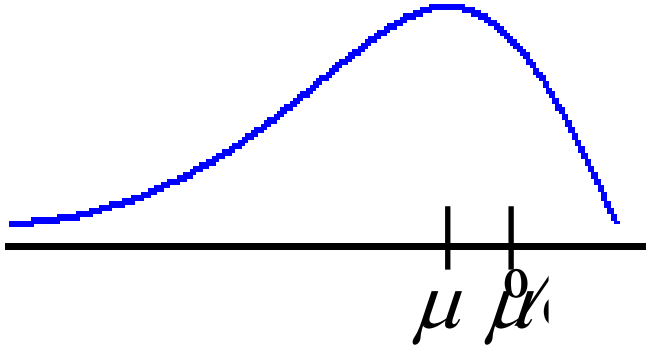
$\mu = \tilde{\mu}$

positive skew



$\tilde{\mu} < \mu$

negative skew



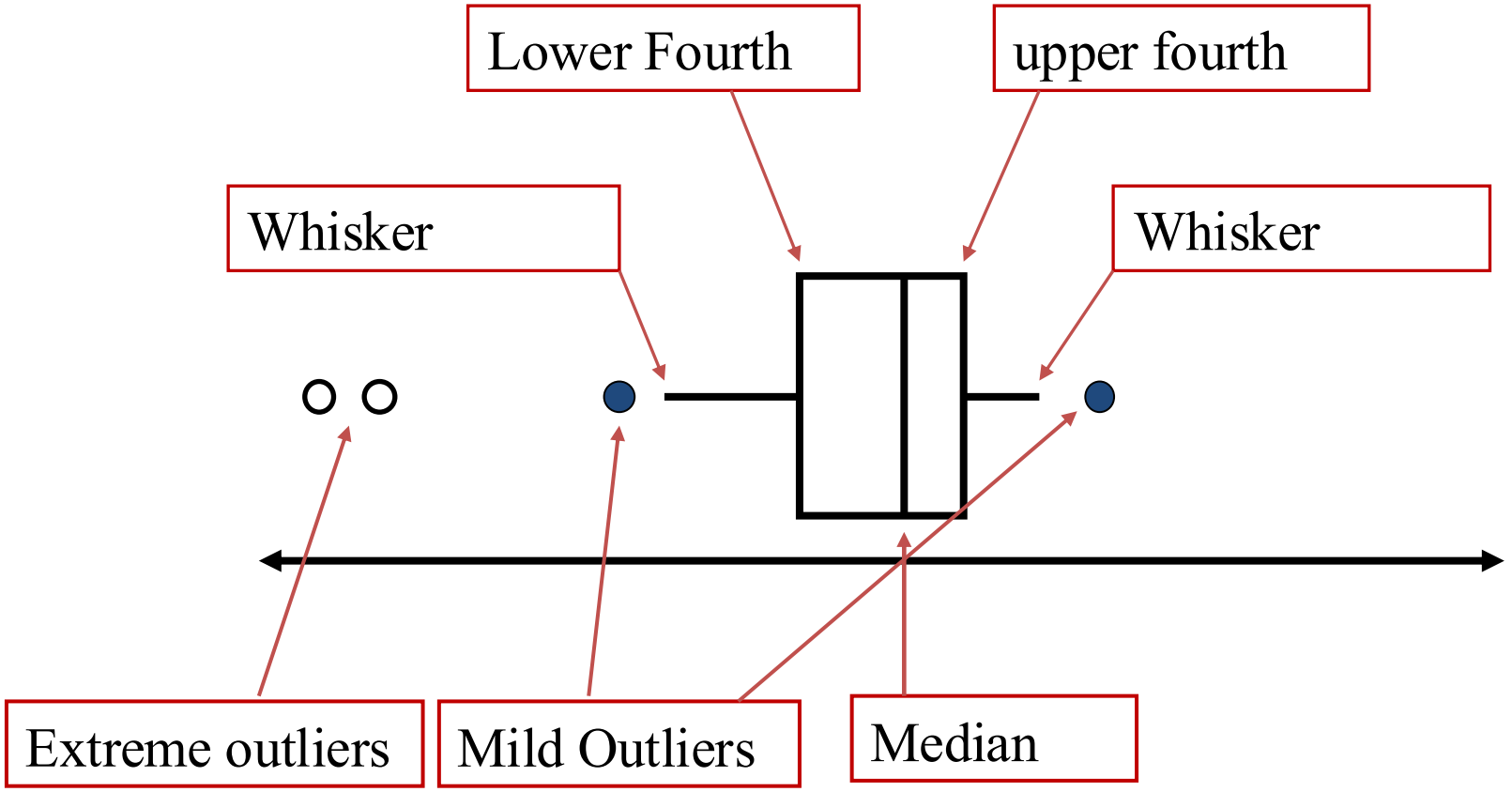
$\mu < \tilde{\mu}$

Boxplots

- We want a pictorial summary which describes the most important features of the data set, e.g., central tendency, variability, symmetry, outliers.
- Also, we want this summary to not be too sensitive to the presence of a few outliers.
- Reminder: median is not much sensitive to outliers!
- We want the pictorial summary to give us some information about the spread as well, but median only divides the data into two parts.

Boxplots

- Pictorial summaries of several important features

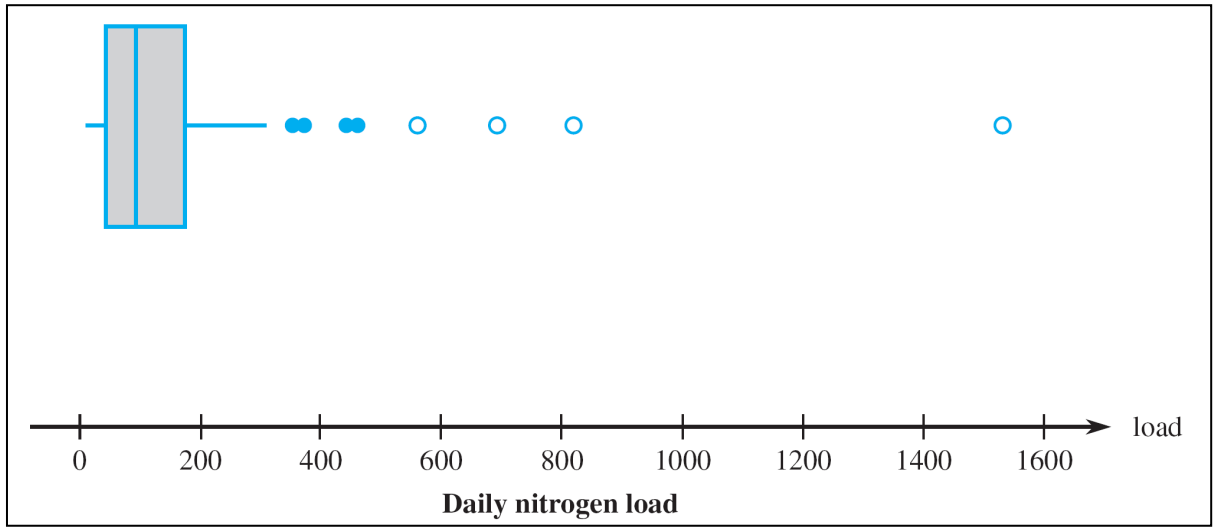


Constructing a Box Plot

1. Order the data from smallest to largest
2. Determine the median, $\tilde{x}$
3. Determine the **lower fourth** – median of smaller half of data
4. Determine the **upper fourth** – median of larger half of data
5. Calculate the **fourth spread** – $f_s = \text{upper fourth} - \text{lower fourth}$
6. Draw a horizontal measurement scale (x-axis)
7. Place rectangle above the axis
   - Left edge at lower fourth
   - Right edge at upper fourth
   - Vertical line inside rectangle at median
8. Draw “whiskers”
   - Lines coming out of either side of rectangle to smallest and largest observations

Boxplots with Outliers

- Values farther than $1.5f_s$ from either the upper or lower fourth are mild outliers
  - Indicated on boxplot with ●
- Values farther than $3.0f_s$ from either the upper or lower fourth are extreme outliers
  - Indicated on boxplot with ○



Example 1.5 – Page 25

Table 1.8 Nicotine Data for Example 1.5

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

Example 1.5 – Page 25

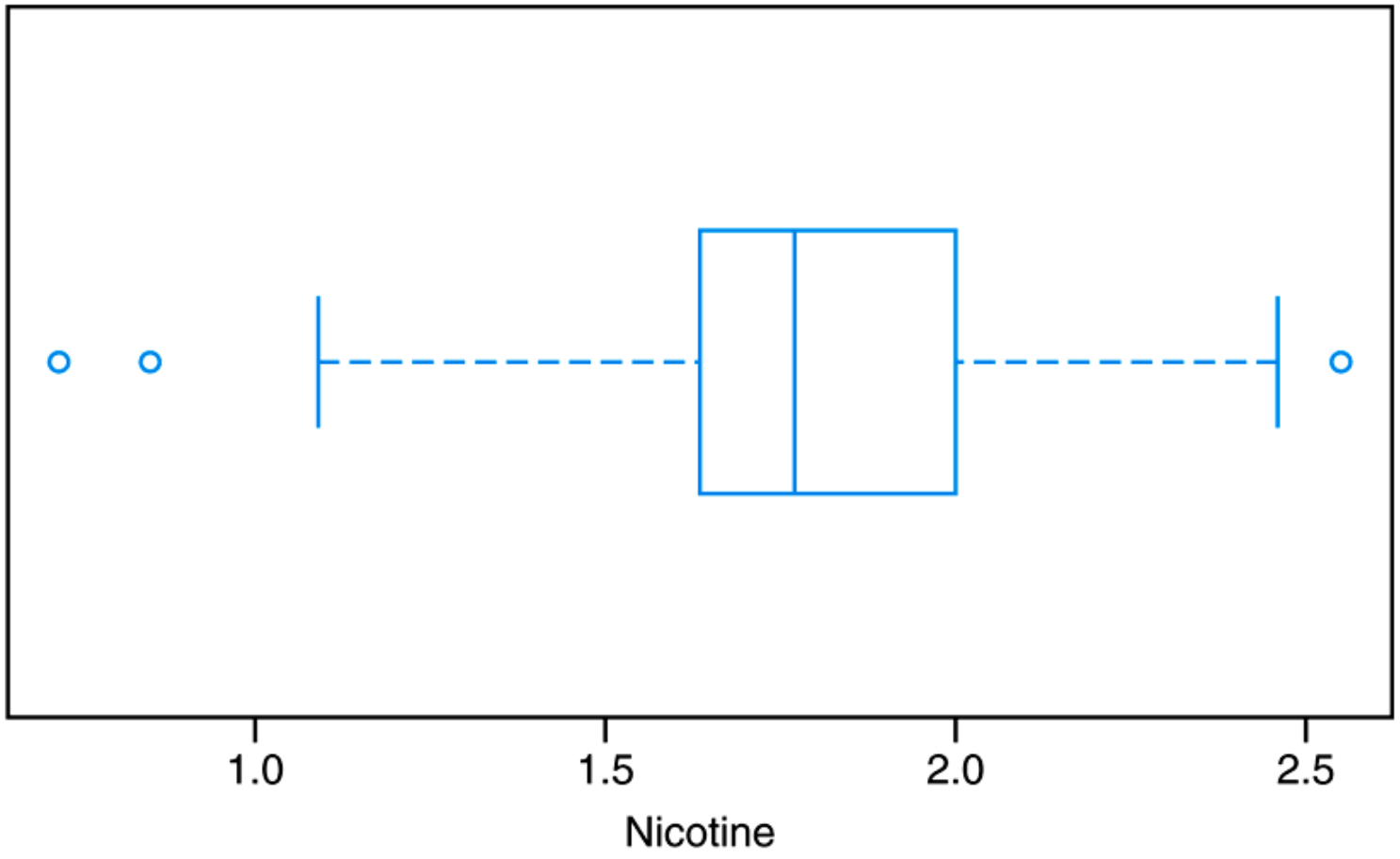


Figure 1.9 Box-and-whisker plot for Example 1.5

Example: Box Plot

Grade point averages of 15 University of Pittsburgh Engineering Freshmen:

2.0	2.7	2.8
2.3	3.1	2.8
2.9	2.6	1.9
2.7	3.0	3.1
2.6	2.5	2.7

Summary

- We have talked about several graphical display techniques.
  - Scatter plot
  - Stem-and-Leaf plot
  - Histogram
  - Box plot
- Much easier for human to see patterns, understand the data and identify problems.
- Need to know how to construct = How to interpret the plots
- Data Viz or Data visualization is an extremely important subject within the field of statistics and data analysis
  - Python package: matplotlib and seaborn
  - Interactive Data Viz Packages
  - Plotly