Introduction to Probability and Statistics Twelfth Edition

Chapter 1
Describing Data with Graphs

Some graphic screen captures from *Seeing Statistics* ®
Some images © 2001-(current year) www.arttoday.com

Variables and Data

- A **variable** is a characteristic that **changes** or **varies** over time and/or for different individuals or objects under consideration.
- **Examples:** Hair color, white blood cell count, time to failure of a computer component.
- **Symbolic variable:** e.g. x stands for hair color, y stands for white blood cell count

Definitions (1)

- An **experimental unit** is the individual or object on which a variable is measured.
- A **measurement** results when a variable is actually measured on an experimental unit.
- A set of measurements, called **data,** can be either a **sample** or a **population.**

Definitions (2)

- A **population** is the **set of all measurements** of interest to an investigator.
  - If a measurement is generated for every experimental unit in the entire collection, the resulting data constitute the population of interest.
- A **sample** is a **subset** of measurements selected from the population of interest.

Example 1.1

- **Variable**
  - –Hair color
- **Experimental unit**
  - –Person
- **Typical Measurements**
  - –Brown, black, blonde, etc.

Example 1.2

- **Variable**
  - Computer cost
- **Experimental unit**
  - Computer
- **Typical Measurements**
  - 20000 baht, 25000 baht, 50000.25 baht, etc.

Exercise

- **Variable**
  - Age
- **Experimental unit**

. . . . . .

. . . . . .

- **Typical Measurements**

. . . . . .

. . . . . .

Exercise

- **Variable**
  - Ice cream flavor
- **Experimental unit**

Ice cream

- **Typical Measurements**

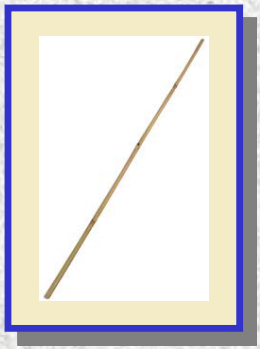
Vanilla, chocolate

How many variables have you measured?

- **Univariate data: One** variable is measured on a single experimental unit.
- **Bivariate data: Two** variables are measured on a single experimental unit.
- **Multivariate data: More than two** variables are measured on a single experimental unit.

Example 2

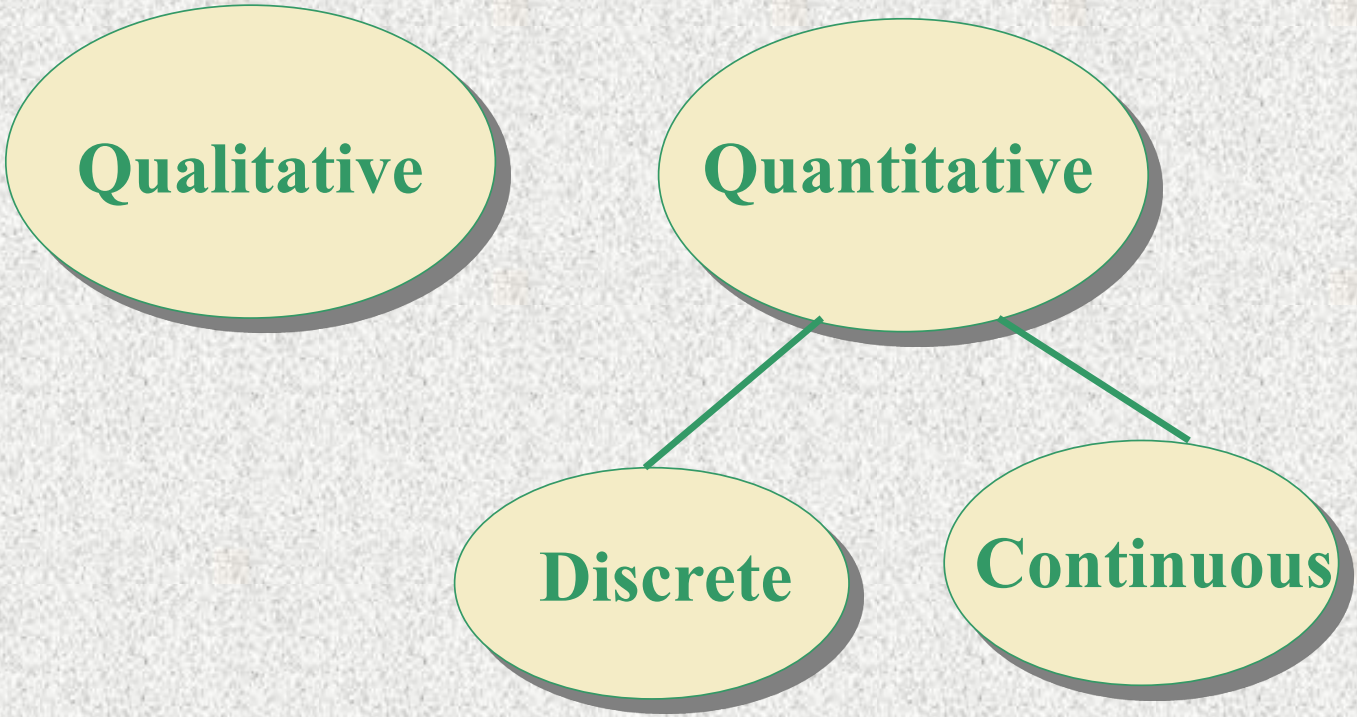
- **Univariate data:**
  - E.g. when a **length** of a **stick** is measured



- **Bivariate data:**
  - E.g. when a **dimension** (width + height) of a **rectangle** is measured

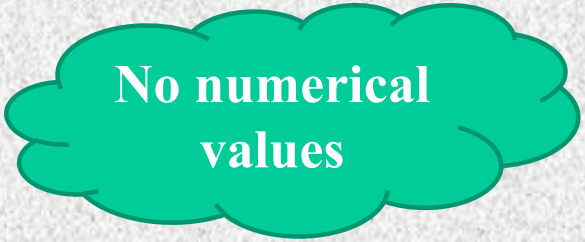
- **Multivariate data:**
  - E.g. when a **volume** (width + height + depth) of a **box** is measured

Types of Variables (1)



Types of Variables (2)

- **Qualitative variables** measure a quality or characteristic on each experimental unit.



- **Examples:**
  - Hair color (black, brown, blonde…)
  - Make of car (Dodge, Honda, Ford…)
  - Gender (male, female)
  - City of birth (Bangkok, Ubonratchatani,….)

Types of Variables (3)

- **Quantitative variables measure a numerical quantity on each experimental unit.**

  - ✓**Discrete** if it can assume only a **finite** or **countable** number of values.

    E.g. a set of {0, 1, 2, 3}

  - ✓**Continuous** if it can assume the infinitely many values corresponding to the points on a line interval.

    E.g. an interval of [0,3]

Exercise

Quantitative discrete OR Quantitative continuous ?

1. For each orange tree in a grove, the number of oranges is measured.

Discrete

2. For a particular day, the number of cars entering a college campus is measured.

Discrete

3. Time until a light bulb burns out.

Cont.

Class Activity

1. Qualitative or Quantitative?
   a) Amount of time it takes to assemble a simple puzzle - Quanti
   b) Number of students in a first-grade classroom - Quanti
   c) Province in which a person lives - Quali

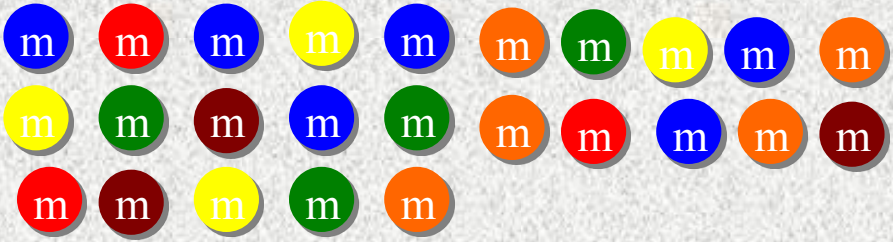
2. Discrete or Continuous?
   a) Number of consumers in a poll of 1000 who consider nutritional labelling on food products to be important - D
   b) Weight of ground beef in a package - C
   c) Number of brothers and sisters you have - D

Graphing Qualitative Variables

- Use a **data distribution** to describe:
  - **What values** of the variable have been measured
  - **How often** each value has occurred
- "How often" can be measured 3 ways:
  - Frequency
  - Relative frequency = Frequency/n
  - Percent = 100 x Relative frequency

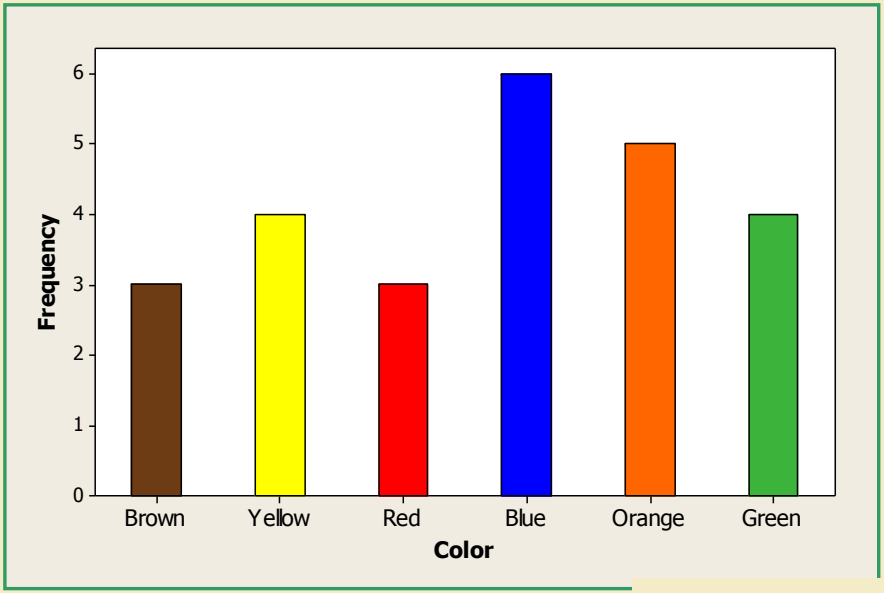
Example 3

- A bag of M&Ms contains 25 candies:
- Raw Data:



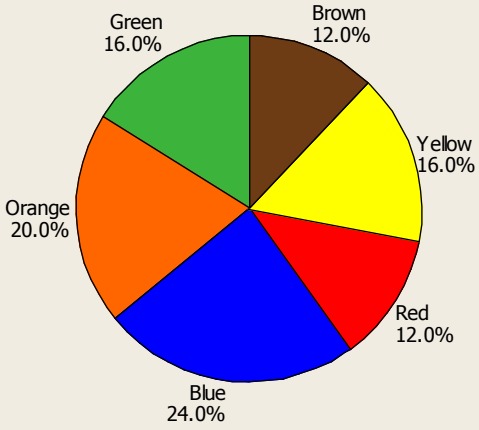
- Statistical Table:

Color	Tally	Frequency	Relative Frequency	Percent
Red	m m m	3	3/25 = .12	12%
Blue	m m m m m m	6	6/25 = .24	24%
Green	m m m m	4		
Orange	m m m m m	5		
Brown	m m m	3		
Yellow	m m m m	4		



Bar Chart

Pie Chart

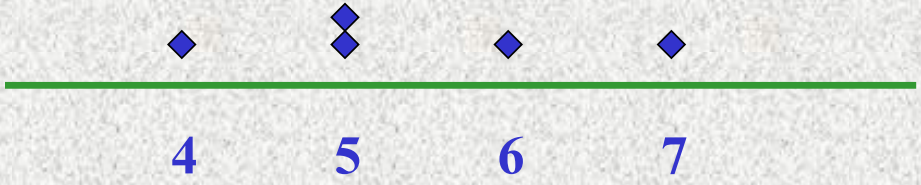


Describing Data with Graphs

- **Qualitative variable** (or categorical variable)
  - Bar chart
  - Pie chart
- **Quantitative variable**
  - Dot plot
  - Stem and leaf plot
  - Histogram

Dotplots

- The simplest graph for quantitative data
- Plots the measurements as points on a horizontal axis, stacking the points that duplicate existing points.
- **Example 4:** A set of data 4, 5, 5, 7, 6



Stem and Leaf Plots

- A simple graph for quantitative data
- Uses the actual numerical values of each data point.

–Divide each measurement into two parts: the **stem** and the **leaf.**

–List the stems in a column, with a **vertical line** to their right.

–For each measurement, record the leaf portion in the **same row** as its matching stem.

–**Order** the leaves from lowest to highest in each stem.

–Provide a **key** to you're your stem and leaf coding if necessary.

Example 5

The prices ($) of 18 brands of walking shoes:

90	70	70	70	75	70	65	68	60
74	70	95	75	70	68	65	40	65

4	0
5	
6	5 8 0 8 5 5
7	0 0 0 5 0 4 0 5 0
8	
9	0 5

Reorder

Leaf Unit = 1

4	0
5	
6	0 5 5 5 8 8
7	0 0 0 0 0 0 4 5 5
8	
9	0 5

Example 6

Test Score: 5.6 5.8 6.1 6.1 6.3 6.5
8.4 8.5 8.5 8.7 8.8 8.8 9.3

Descriptive statistics	
	Test Score
count	13
mean	7.492
sample variance	1.977
sample standard deviation	1.406
minimum	5.6
maximum	9.3
range	3.7

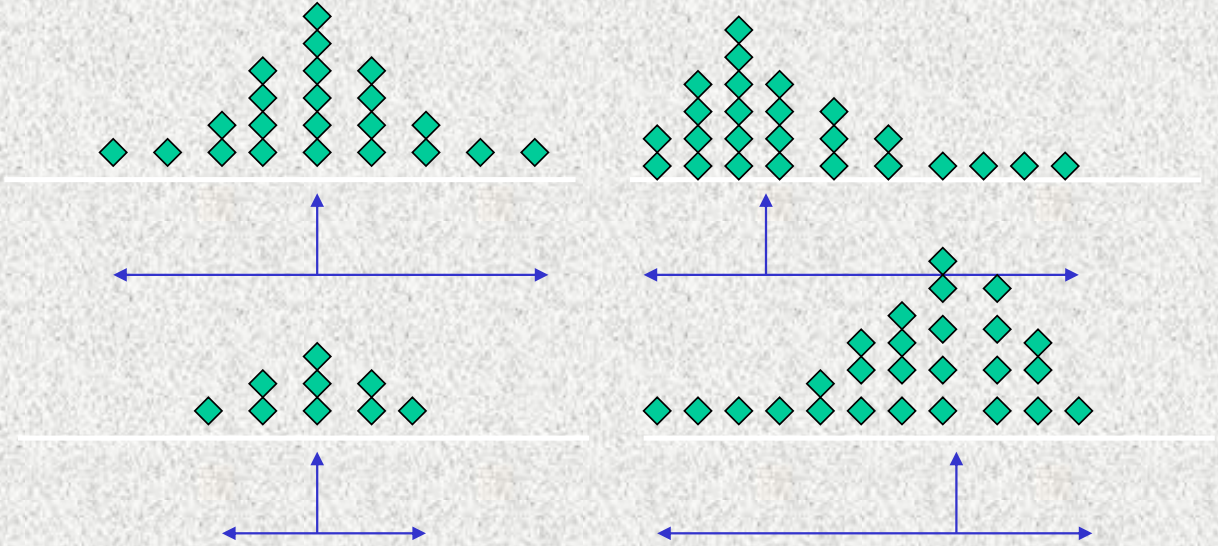
Stem and Leaf plot for Test Score

stem unit = 1 (*optional)

leaf unit = 0.1

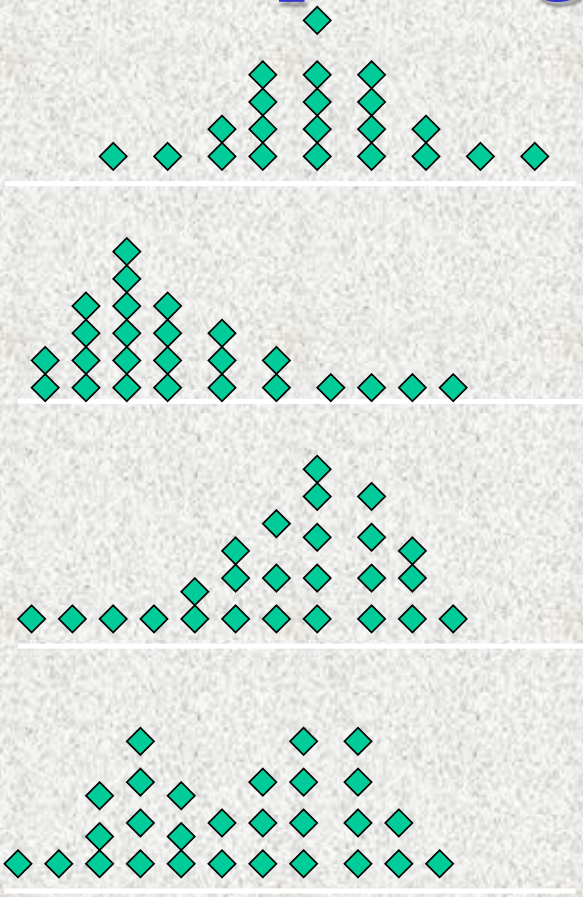
Frequency	Stem	Leaf
2	5	6 8
4	6	1 1 3 5
0	7	
6	8	4 5 5 7 8 8
1	9	3
13		

Interpreting Graphs: Location and Spread



- Where is the data centered on the horizontal axis, and how does it spread out from the center?

Interpreting Graphs: Shapes



Mound shaped and symmetric (mirror images)

Skewed right: a few unusually large measurements

Skewed left: a few unusually small measurements

Bimodal: two local peaks

Interpreting Graphs: Outliers

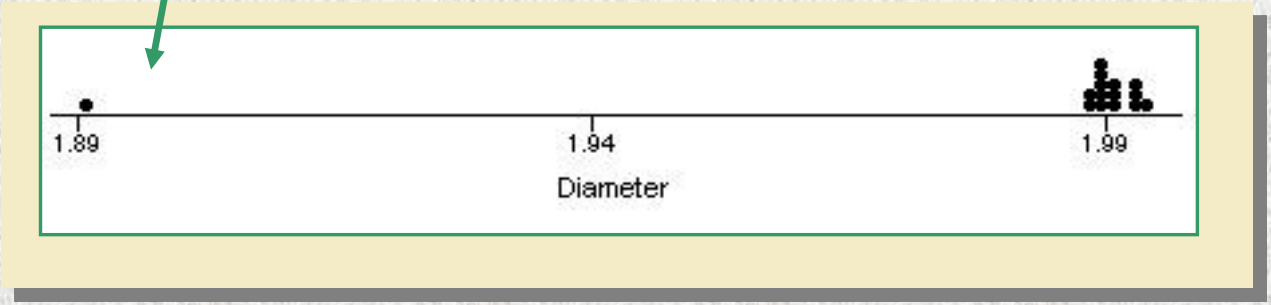
- Are there any strange or unusual measurements that stand out in the data set?

Example 7

- A quality control process measures the diameter of a gear being made by a machine (cm). The technician records 15 diameters, but inadvertently makes a typing mistake on the second entry.

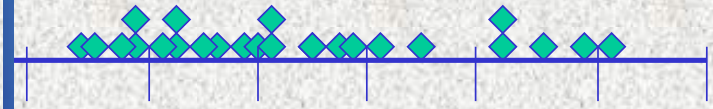
1.991 1.891 1.991 1.988 1.993 1.989 1.990 1.988

1.988 1.993 1.991 1.989 1.989 1.993 1.990 1.994

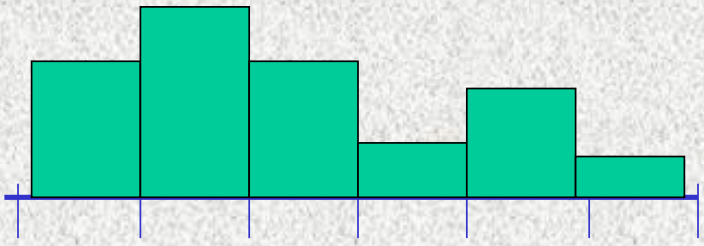


Relative Frequency Histograms

- A **relative frequency histogram** for a quantitative data set is a bar graph in which the height of the bar shows "how often" (measured as a proportion or relative frequency) measurements fall in a particular class or subinterval.



Create intervals



Stack and draw bars

Relative Frequency Histograms

- Divide the range of the data into **5-12 subintervals** of equal length.
- Calculate the **approximate width** of the subinterval as Range/number of subintervals.
- Round the approximate width up to a convenient value.
- Use the method of **left inclusion** including the left endpoint, but not the right in your tally.
- Create a **statistical table** including the subintervals, their frequencies and relative frequencies.

Relative Frequency Histograms

- Draw the **relative frequency histogram** plotting the subintervals on the horizontal axis and the relative frequencies on the vertical axis.
- The height of the bar represents
  - The **proportion** of measurements falling in that class or subinterval.
  - The **probability** that a single measurement, drawn at random from the set, will belong to that class or subinterval.

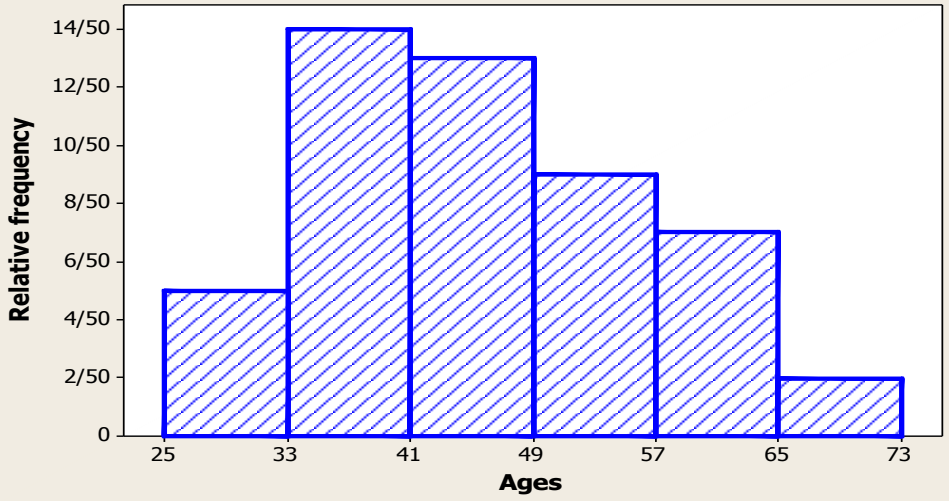
Example 8

The ages of 50 tenured faculty at a state university.

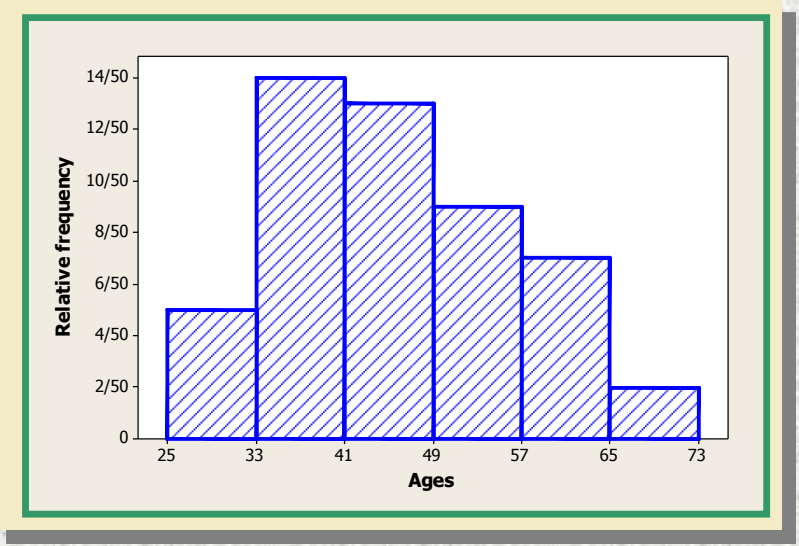
34	48	**70**	63	52	52	35	50	37	43	53	43	52	44
42	31	36	48	43	**26**	58	62	49	34	48	53	39	45
34	59	34	66	40	59	36	41	35	36	62	34	38	28
43	50	30	43	32	44	58	53						

- We choose to use **6** intervals.
- Minimum class width = **(70 – 26)/6 = 7.33** (Max, Min)
- Convenient class width = **8**
- Use **6** classes of length **8**, starting at **25.**

Age	Tally	Frequency	Relative Frequency	Percent
25 to < 33	1111	5	5/50 = .10	10%
33 to < 41	1111 1111 1111	14	14/50 = .28	28%
41 to < 49	1111 1111 111	13	13/50 = .26	26%
49 to < 57	1111 1111	9	9/50 = .18	18%
57 to < 65	1111 11	7	7/50 = .14	14%
65 to < 73	11	2	2/50 = .04	4%



Describing the Distribution



Shape? Skewed right

Outliers? No.

What proportion of the tenured faculty are younger than 41?

$(14 + 5)/50 = 19/50 = 0.38$

What is the probability that a randomly selected faculty member is 49 or older?

$(9 + 7 + 2)/50 = 18/50 = 0.36$

Key Concepts

I. How Data Are Generated

1. Experimental units, variables, measurements
2. Samples and populations
3. Univariate, bivariate, and multivariate data

II. Types of Variables

1. Qualitative or categorical
2. Quantitative
    a. Discrete
    b. Continuous

III. Graphs for Univariate Data Distributions

1. Qualitative or categorical data
    a. Pie charts
    b. Bar charts

Key Concepts

2. Quantitative data
    a. Pie and bar charts
    b. Line charts
    c. Dotplots
    d. Stem and leaf plots
    e. Relative frequency histograms
3. Describing data distributions
    a. Shapes—symmetric, skewed left, skewed right, unimodal, bimodal
    b. Proportion of measurements in certain intervals
    c. Outliers

Class Activity

4. For the following data sets, find a) the range, b) the minimum class width, c) a convenient class width, d) a convenient starting point, and e) the first two classes.

Number of measurements	Smallest and largest values	Number of classes	a)	b)	c)	d)	e)
40	0 to 100	6	100	100/6 = 16.66	17	0	0 to < 17 17 to < 34
130	1200 to 1500	7					

Class Activity

5. Consider this set of data:

4.5	3.2	3.5	3.9	3.5	3.9	4.3	4.8	3.6	3.3	4.3	4.2
3.9	3.7	4.3	4.4	3.4	4.2	4.4	4.0	3.6	3.5	3.9	4.0

a) Construct a stem and leaf plot by using the leading digit as the stem.

b) Construct a stem and leaf plot by using each leading digit twice. Does this technique improve the presentation of the data? Explain.

c) Construct a relative frequency histogram for the data.

Class Activity

6. A discrete variable can take on only the values 0, 1, or 2. A set of 20 measurements on this variable is shown here:

1	2	1	0	2	2	1	1	0	0
2	2	1	1	0	0	1	2	1	1

a) Draw a dotplot to describe the data.

b) What proportion of the measurements are greater than 1?

c) What proportion of the measurements are less than 2?

d) If a measurement is selected at random from the 20 measurements shown, what is the probability that it is a 2?

e) Describe the shape of the distribution. Do you see any outliers?