Marmara University, 2015

# Probability and Statistics

## Chapter 1
## Describing Data with Graphs

Mujdat Soyturk, Ph.D.
Asst. Prof.

## Contents

- Variables and Data
- Types of Variables
- Graphs for Categorical Data
- Graphs for Quantitative Data
- Relative Frequency Histograms

- Most parts of the slides are derived from the textbook: "Mendenhall, Beaver, Beaver, *Introduction to Probability and Statistics*, 14th Ed., Brooks/Cole, Cengage Learning, 2013"

## 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.



## Definitions

- 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.**

## Example

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



## Example

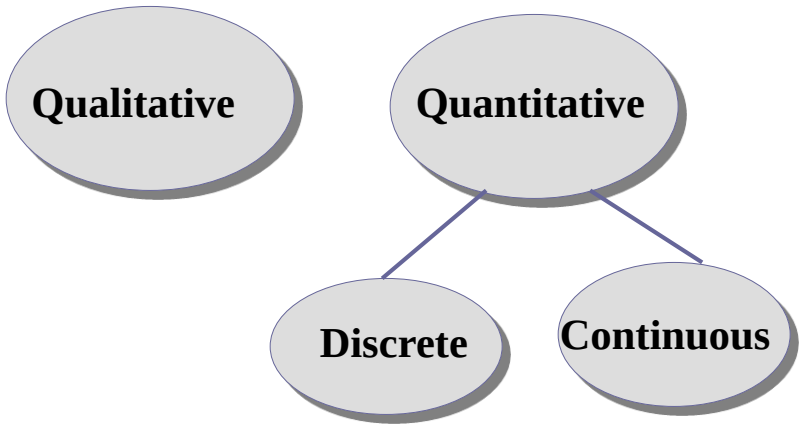
- **Variable**
  - –Time until a light bulb burns out
- **Experimental unit**
  - –Light bulb
- **Typical Measurements**
  - –1500 hours, 1535.5 hours, etc.

## 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.



## Types of Variables



# Types of Variables


- **Qualitative variables** measure a quality or characteristic on each experimental unit.
- Specifies the similarities or differences in kind → categorical data
- **Examples:**
  - Hair color (black, brown, blonde…)
  - Make of car (Fiat, Opel, Honda, Ford…)
  - Gender (male, female)
  - City of birth (Istanbul, Ankara,….)



# Types of Variables


- **Quantitative variables measure a numerical quantity on each experimental unit.**
  - **Discrete** if it can assume only a finite or countable number of values.
  - **Continuous** if it can assume the infinitely many values corresponding to the points on a line interval.

## Examples

- For each orange tree in a grove, the number of oranges is measured.
  - **Quantitative discrete**
- For a particular day, the number of cars entering the Goztepe campus is measured.
  - **Quantitative discrete**
- Time until a light bulb burns out
  - **Quantitative continuous**



## 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
- Three steps to a data distribution:
  1. Raw data ➔
  2. Statistical Table ➔
  3. Graph

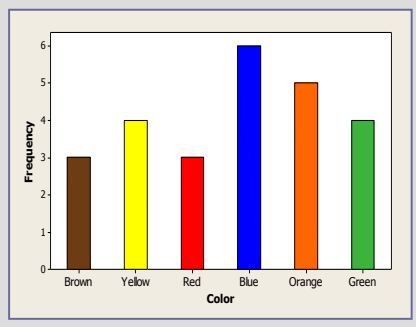
# Example

- **A bag of candies that contains 25 candies:**
- **Raw Data:**
- **Statistical Table:**

| Color | Tally | Frequency | Relative Frequency | Percent |
|---|---|---|---|---|
| Red | | 3 | 3/25 = .12 | 12% |
| Blue | | 6 | 6/25 = .24 | 24% |
| Green | | 4 | 4/25 = .16 | 16% |
| Orange | | 5 | 5/25 = .20 | 20% |
| Brown | | 3 | 3/25 = .12 | 12% |
| Yellow | | 4 | 4/25 = .16 | 16% |

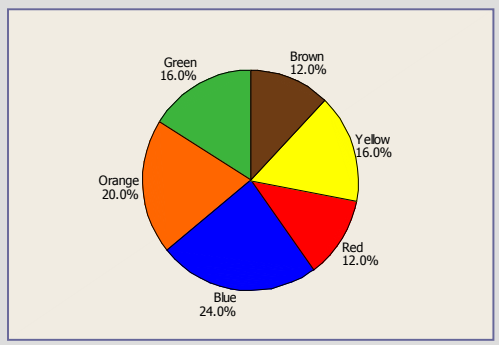


# Graphs




Bar Chart

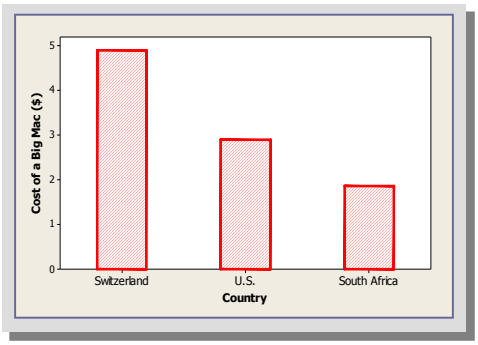
Pie Chart



# Graphing Quantitative Variables

A single quantitative variable measured for different population segments or for different categories of classification can be graphed using a **pie** or **bar chart**.

A Big Mac hamburger costs $4.90 in Switzerland, $2.90 in the U.S. and $1.86 in South Africa.




**Pareto chart**:→ bars ordered from largest to smallest



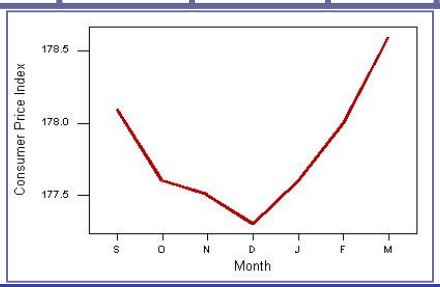
# Graphing Quantitative Variables

A single quantitative variable measured over time is called a **time series.** It can be graphed using a **line** or **bar chart.**

CPI: All Urban Consumers-Seasonally Adjusted

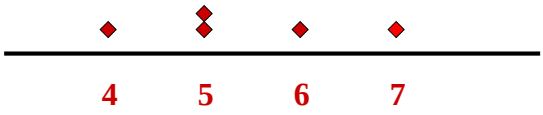
| Sept | Oct | Nov | Dec | Jan | Feb | Mar |
|---|---|---|---|---|---|---|
| 178.10 | 177.60 | 177.50 | 177.30 | 177.60 | 178.00 | 178.60 |

BUREAU OF LABOR STATISTICS



# Dotplots

- The simplest graph for quantitative data
- Plots the measurements as points on a horizontal axis, stacking the points that duplicate existing points.
- **Example:** The set 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 your coding.

# Example

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 |

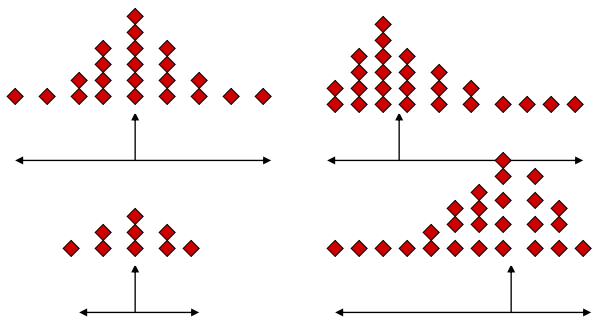
Reorder

| | |
|---|---|
| 4 | 0 |
| 5 | |
| 6 | 0 5 5 5 8 8 |
| 7 | 0 0 0 0 0 0 4 5 5 |
| 8 | |
| 9 | 0 5 |



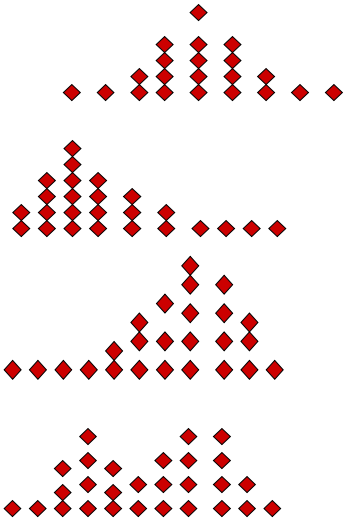
# Interpreting Graphs: Location and Spread

1. Scales
2. Location
3. Shape
4. Outliers



- 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**

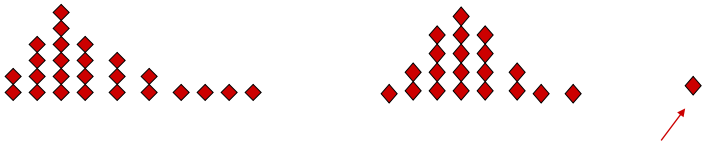
**Skewed left: a few unusually small measurements**

**Bimodal: two local peaks**



# Interpreting Graphs: Outliers

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



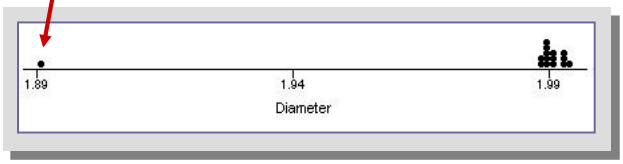
No Outliers

Outlier

# Example

- 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.

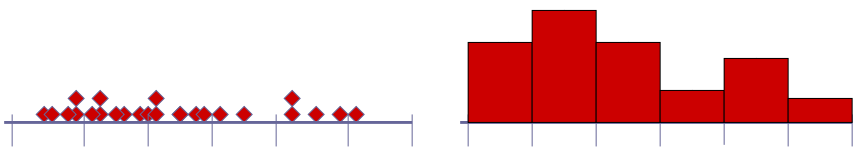
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.991 | 1.891 | 1.991 | 1.988 | 1.993 | 1.989 | 1.990 | 1.988 |
| 1.988 | 1.993 | 1.991 | 1.989 | 1.989 | 1.993 | 1.990 | 1.994 |






# 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

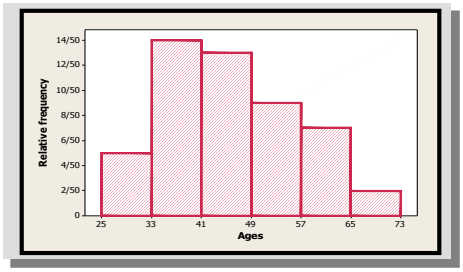
The ages of 50 tenured faculty at a state university.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | | | | | | |

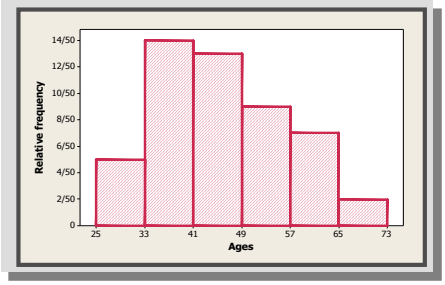
- We choose to use **6** intervals.
- Minimum class width **= (70 – 26)/6 = 7.33**
- Convenient class width **= 8**
- Use **6** classes of length **8**, starting at **25.**



| 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 = .38

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

(9 + 7 + 2)/50 = 18/50 = .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