

Lecture 1: Describing Data with Graphs

Dr Md Rifat Ahmmad Rashid
Assistant Professor
East West University Bangladesh

What is Data?

Data - Consist of information coming from observations, counts, measurements, or responses.

- “People who eat three daily servings of cereal have been shown to reduce their risk of...stroke by 37%.”
- “Seventy percent of the 1500 IT students playing Fortnite and PUBG.”

What is Statistics?

Statistics

Applied mathematics that deals with collection, organization, presentation, analysis, and interpretation of numerical data in order to make decisions.



What is Data Science?

Data Science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.



Data Sets

Population

The collection of *all* outcomes, responses, measurements, or counts that are of interest.



Sample

A subset of the population.

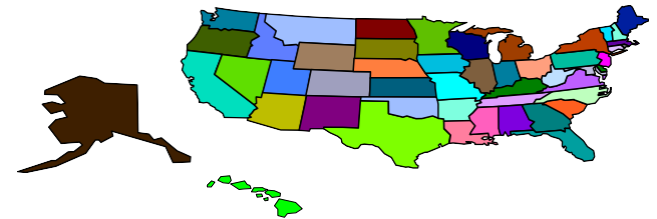


Parameter and Statistic

Parameter

A number that describes a population characteristic.

Average age of all people in the United States



Statistic

A number that describes a sample characteristic.

Average age of people from a sample of three states



Example: Identifying Data Sets

In a recent survey, 1526 adults in the Philippines were asked if they think global warming is a problem that requires immediate government action. 824 of the adults said yes and 702 of the adults said no.

Identify the population and the sample.

Describe the data set. *(Adapted from: Pew Research Center)*



Solution: Identifying Data Sets

- The population consists of the responses of all adults in the Philippines.
- The population consists of the responses of the 1526 adults in the Philippines in the survey.
- The sample is a subset of the responses of all adults in the Philippines.
- The data set consists of 824 yes's and 702 no's.

Responses of adults in the Philippines (population)

Responses of adults in survey (sample)

In a recent survey, 1526 adults in the Philippines were asked if they think global warming is a problem that requires immediate government action. 824 of the adults said yes and 702 of the adults said no.

Example: Identifying Data Sets

In June 2017, A survey asked to school students what they thought about Interactive Video Games.

There were 147 respondents, 97% of who are boys, and most are between 10 and 14 years old.

Middle child and only child came out about even with 20 saying they are the middle child and 19 saying they are an only child.

Most agreed or strongly agreed (60%) that they like interactive videogames and only 18% disagreed or strongly disagreed. The rest weren't sure (22%).

Example: Identifying Data Sets

In June 2017, A survey asked to school students what they thought about Interactive Video Games.

Population

There were **147 respondents**, **97% of who are boys**, and **most are between 10 and 14 years old**.

Sample

Middle child and only child came out about even with **20 students saying they are the middle child** and **19 students saying they are an only child**.

Sample

Most agreed or strongly agreed (60%) that they like interactive videogames and only 18% disagreed or strongly disagreed. The rest weren't sure (22%).

Example: Distinguish Parameter and Statistic

Decide whether the numerical value describes a population parameter or a sample statistic.

1. A recent survey of a sample of NBAs reported that the average salary for an NBA is more than 82,000\$. (*Source: The Wall Street Journal*)

Solution:



Example: Distinguish Parameter and Statistic

Decide whether the numerical value describes a population parameter or a sample statistic.

1. A recent survey of a sample of NBAs reported that the average salary for an NBA is more than 82,000\$. (*Source: The Wall Street Journal*)

Solution:

Sample statistic (the average of 82,000\$ is based on a subset of the population)



Example: Distinguish Parameter and Statistic

Decide whether the numerical value describes a **population parameter** or a **sample statistic**.

2. Starting salaries for the 667 IT graduates from Holy Angel University increased 8.5% from the previous year.

Solution:



Example: Distinguish Parameter and Statistic

Decide whether the numerical value describes a **population parameter** or a **sample statistic**.

2. Starting salaries for the 667 IT graduates from Holy Angel University increased 8.5% from the previous year.

Solution:

Population parameter (the percent increase of 8.5% is based on all 667 graduates' starting salaries)



Variables and Data

Variables

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

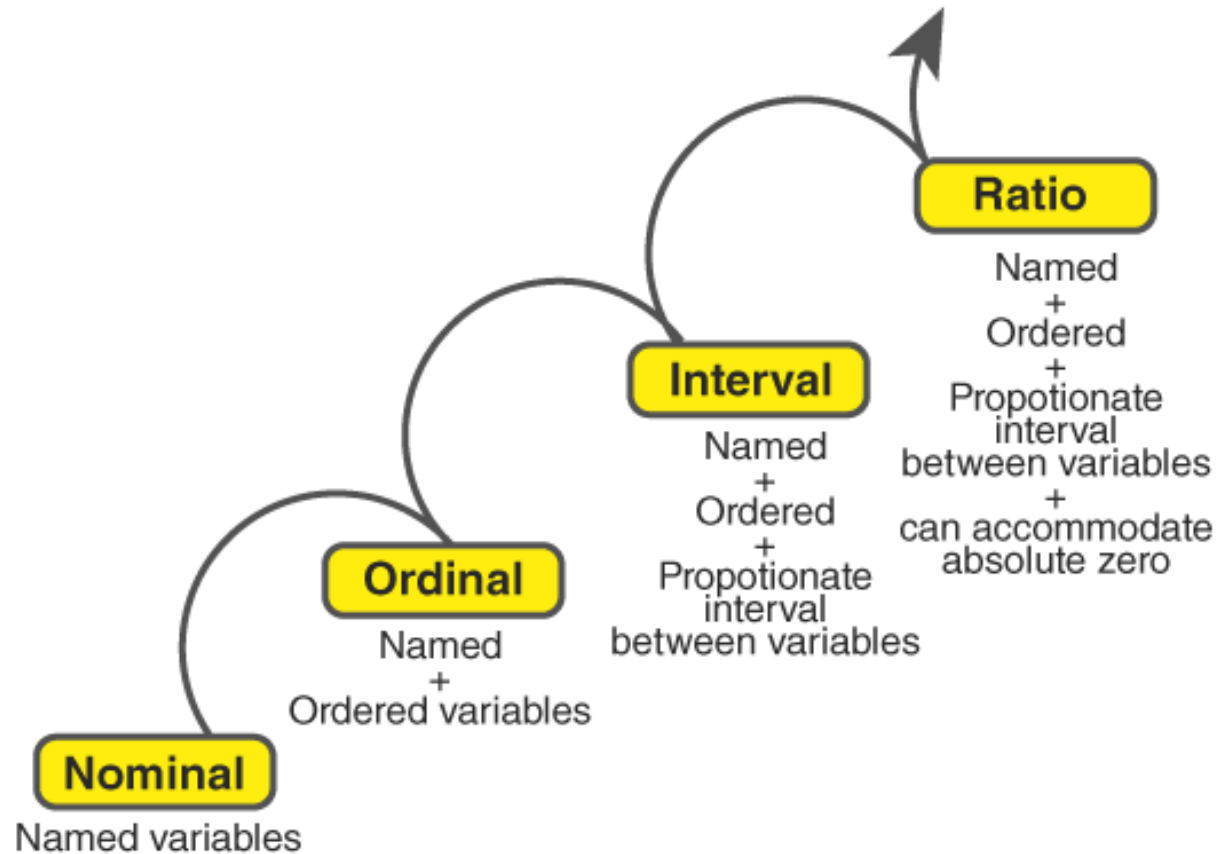


- **Variable**
 - Time until a light bulb burns out
- **Experimental unit**
 - Light bulb
- **Typical Measurements**
 - 1500 hours, 1535.5 hours, etc.

LEVELS OF MEASUREMENT

A nominal category or a nominal group is a group of objects or ideas that can be collectively grouped on the basis of a particular characteristic —a qualitative property.

A variable that codes whether each one in a set of observations is in a particular nominal category is called a categorical variable.



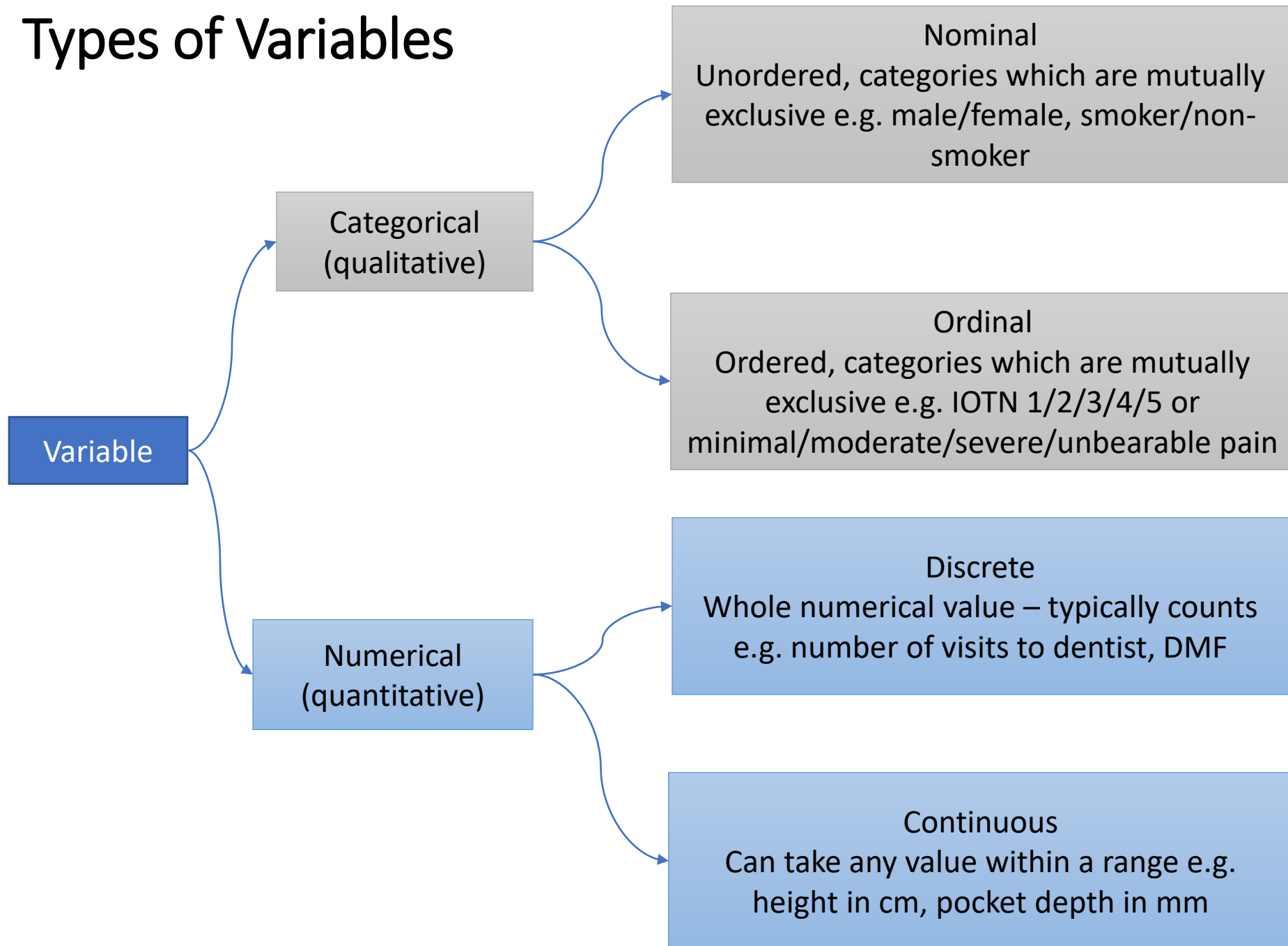
How many variables have you measured?

- **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.

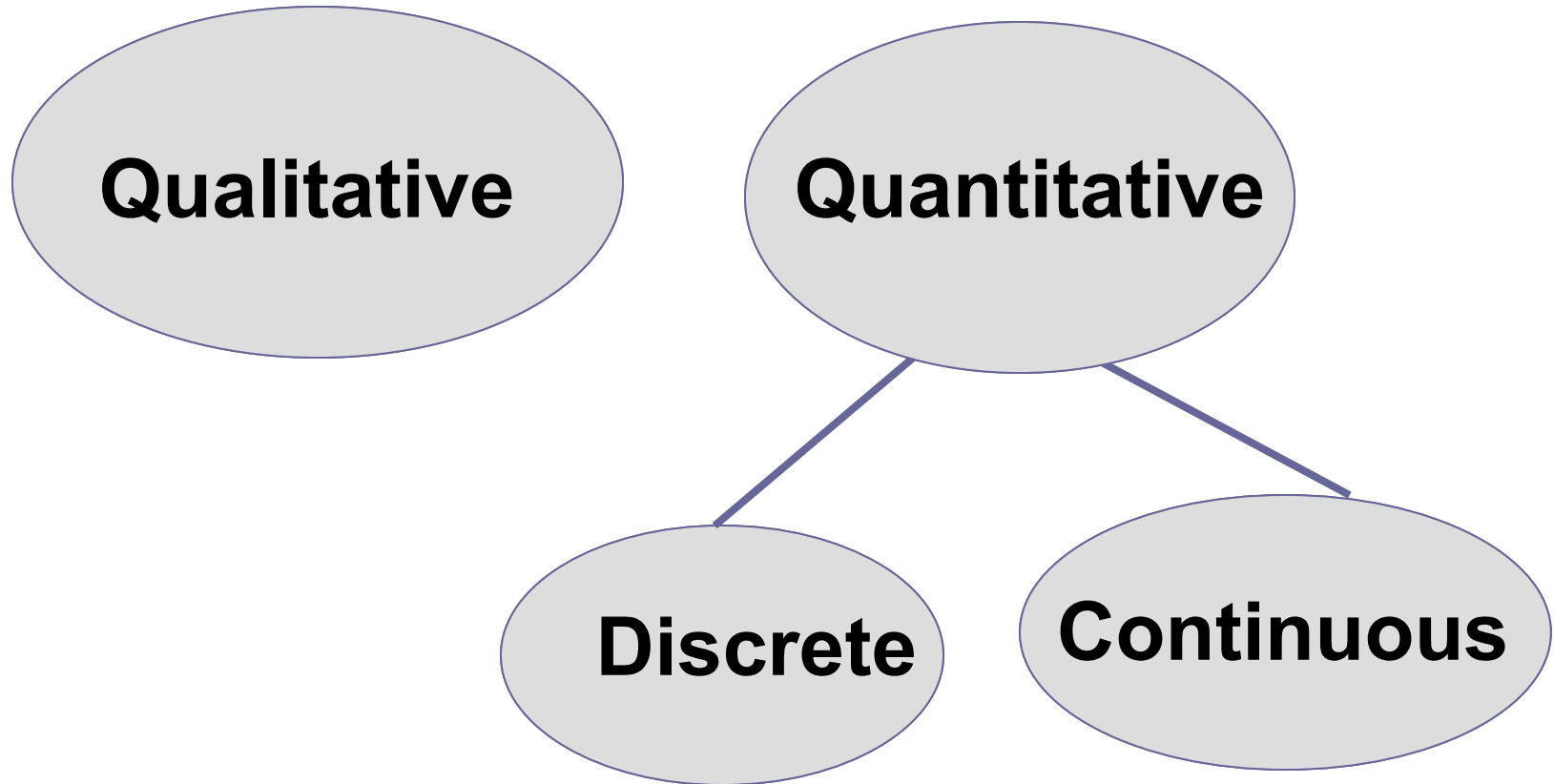
How many variables have you measured?

- **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



Types of Variables

- **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)
- State of birth (California, Arizona,....)

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 a college campus is measured.
 - **Quantitative discrete**
- Time until a light bulb burns out
 - **Quantitative continuous**

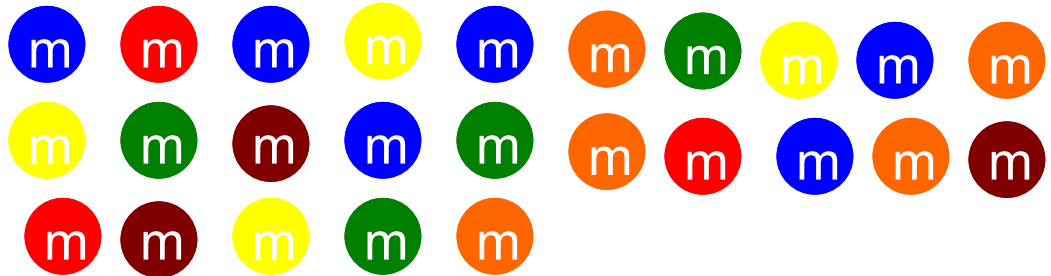
Graphing Qualitative Variables

- 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 = $\text{Frequency} / n$
 - Percent = $100 \times \text{Relative frequency}$







Example

- A bag of M&Ms contains 25 candies:

- Raw Data:

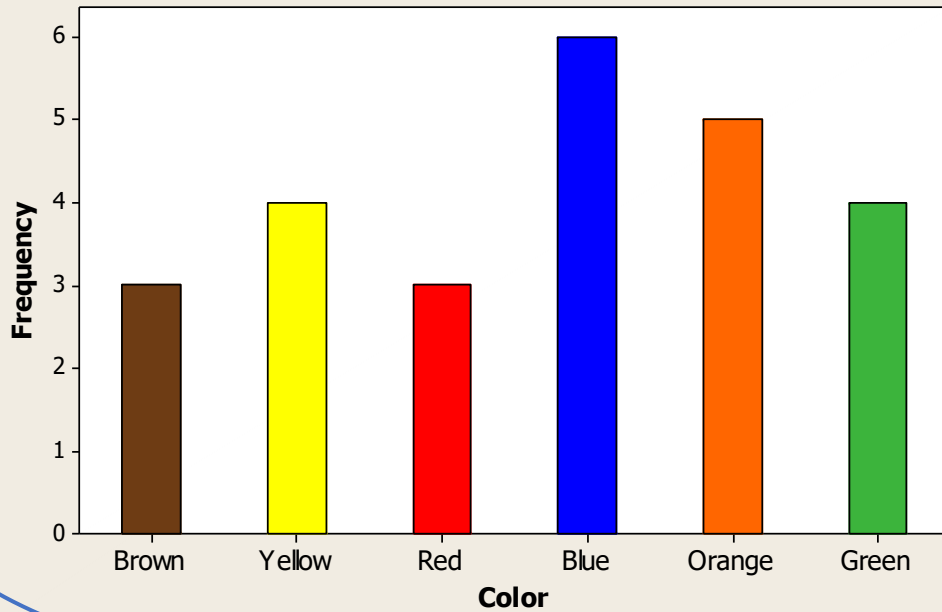


- Statistical Table:

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

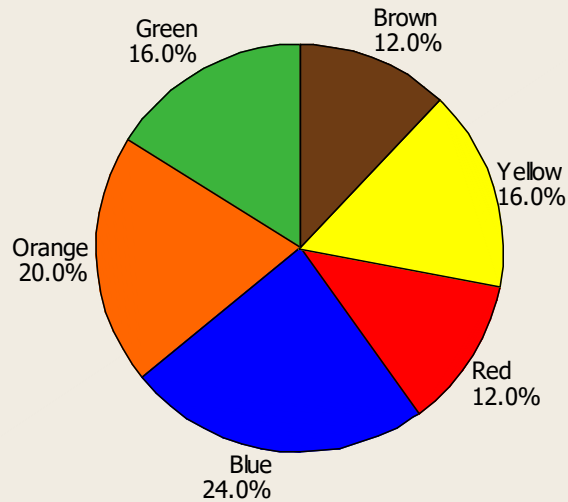
Bar Chart



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

Pie Chart

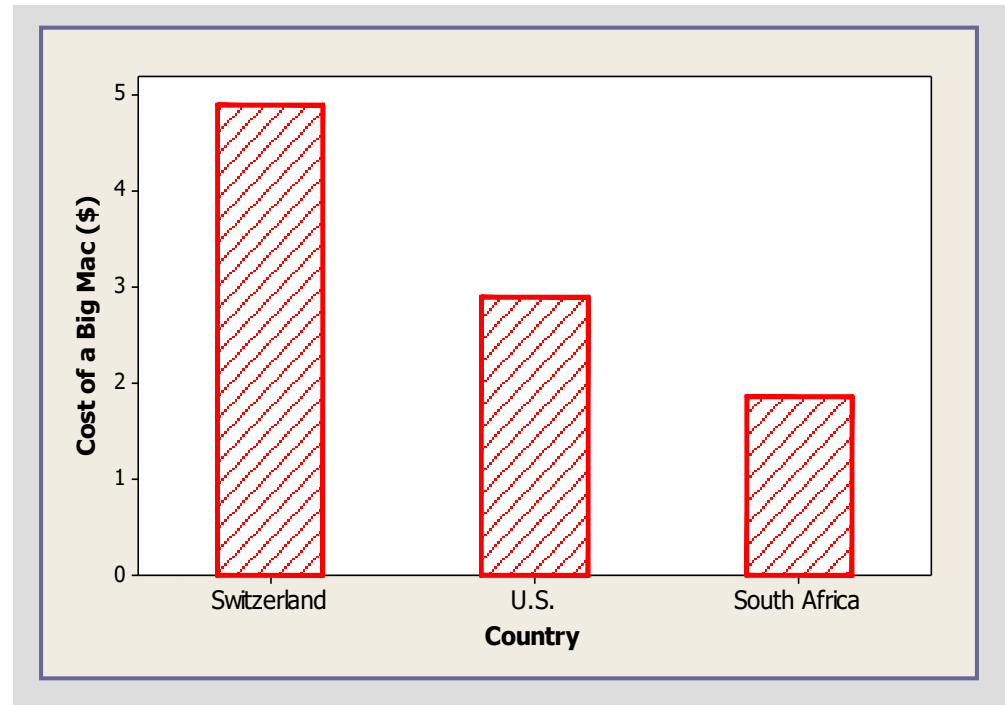


Color	Tally	Frequency	Relative Frequency	Percent
Red		3	$3/25 = .12$	12%
Blue		6	$6/25 = .24$	24%
Green		4	$4/25 = .16$	16%
Orange		5	$5/25 = .20$	20%
Brown		3	$3/25 = .12$	12%
Yellow		4	$4/25 = .16$	16%

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**.

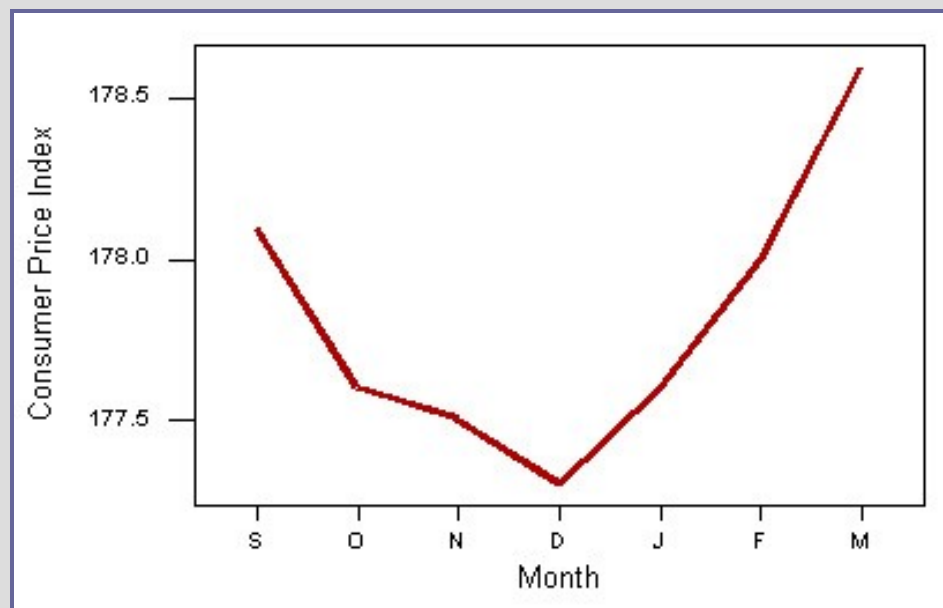
A Big Mac hamburger costs \$4.90 in Switzerland, \$2.90 in the U.S. and \$1.86 in South Africa.



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

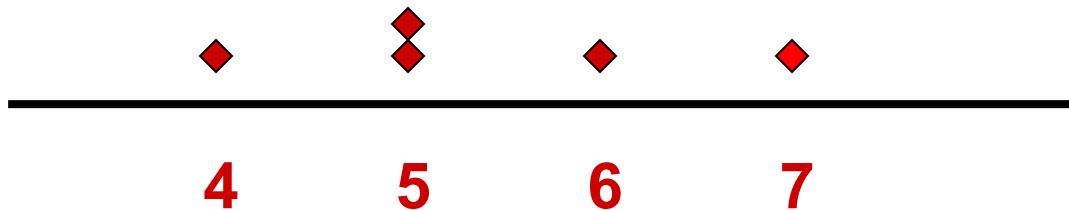
CPI: All Urban Consumers-Seasonally Adjusted

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

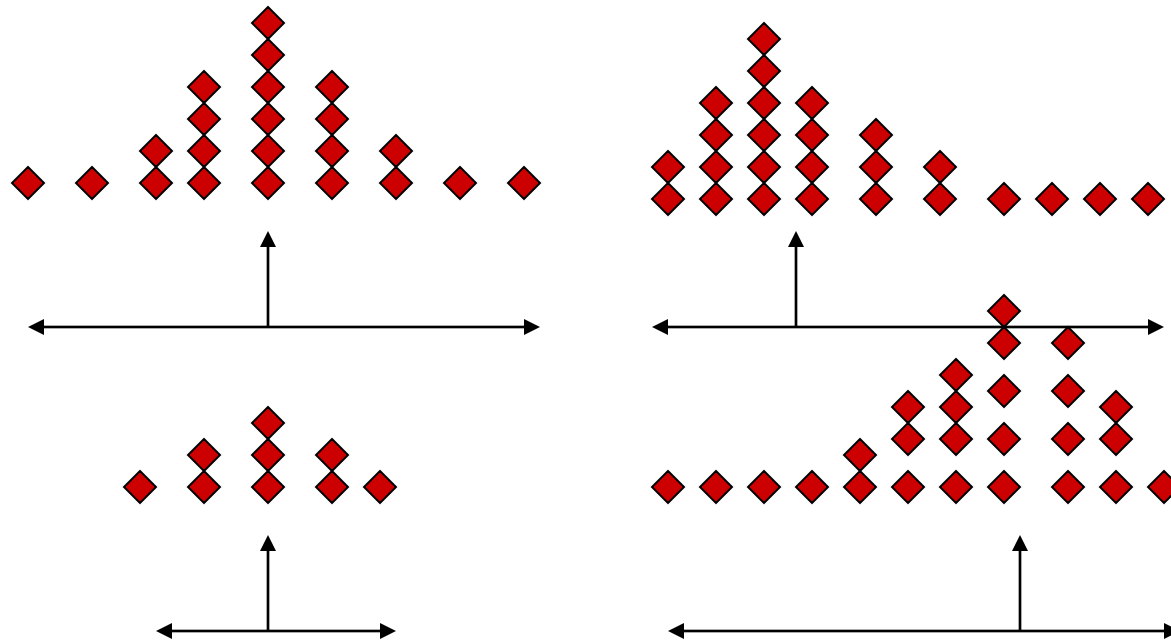


Dotplots

- 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

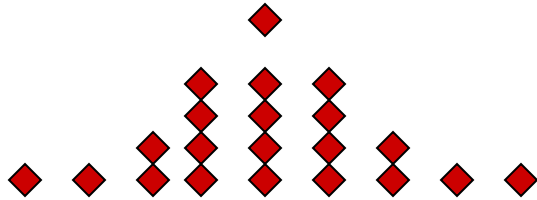


Interpreting Graphs: Location and Spread

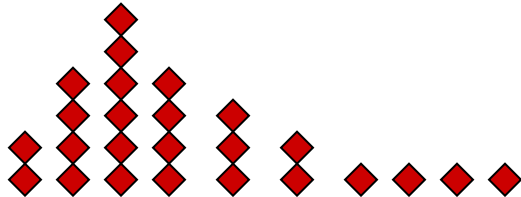


- Where is the data centered on the horizontal axis, and how does it spread out from the center?

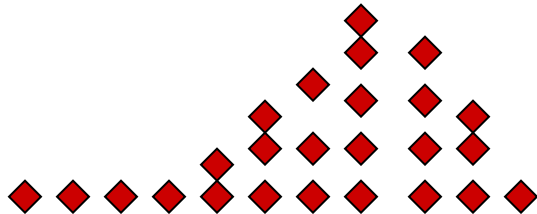
Interpreting Graphs: Shapes



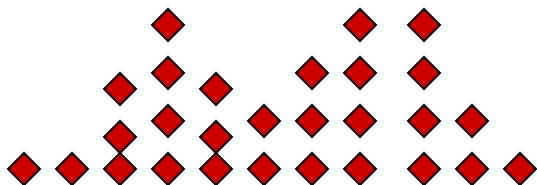
Mound shaped and symmetric (mirror images)



Skewed right: a few unusually large measurements

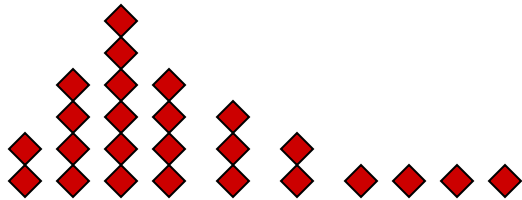


Skewed left: a few unusually small measurements

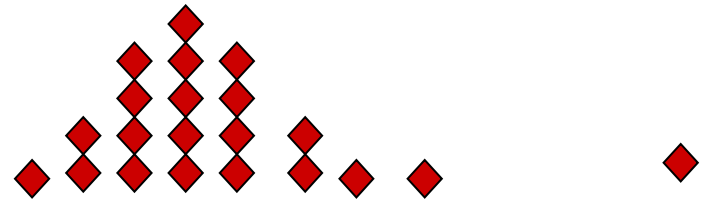


Bimodal: two local peaks

Interpreting Graphs: Outliers

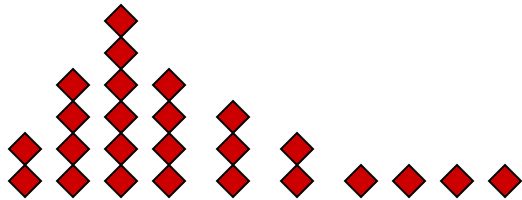


No Outliers

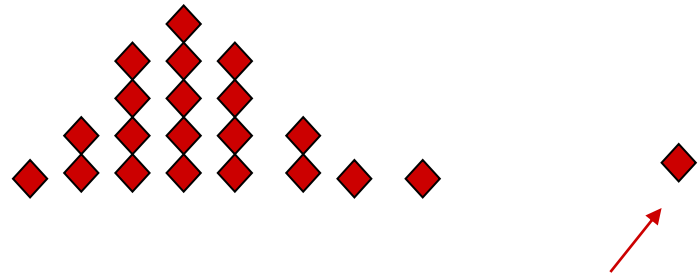


- Are there any strange or unusual measurements that stand out in the data set?

Interpreting Graphs: Outliers



No Outliers



Outlier

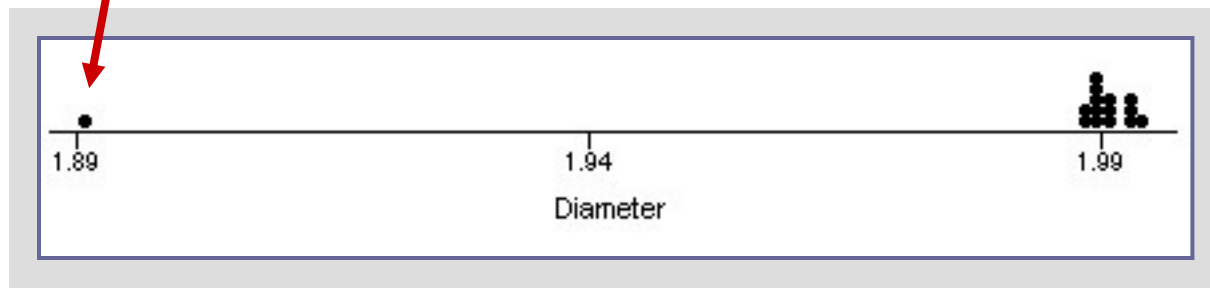
- Are there any strange or unusual measurements that stand out in the data set?

Example



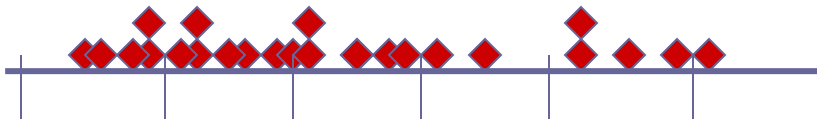
- A quality control process measures the diameter of a gear being made by a machine (cm). The technician records 15 diameters, but inadvertently makes a typing mistake on the second entry.

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

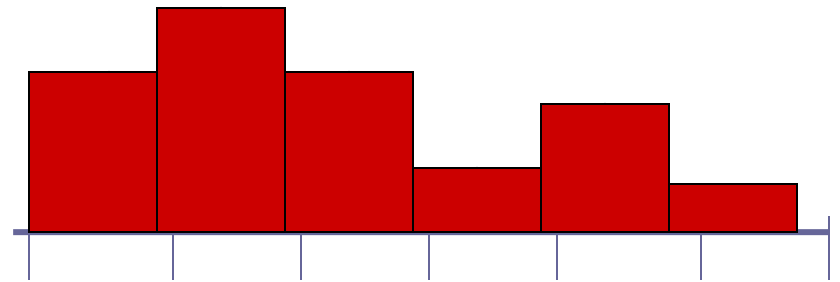


Relative Frequency Histograms

- 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 $\text{Range}/\text{number of subintervals}$.
- 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 faculty where minimum age is 25

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						

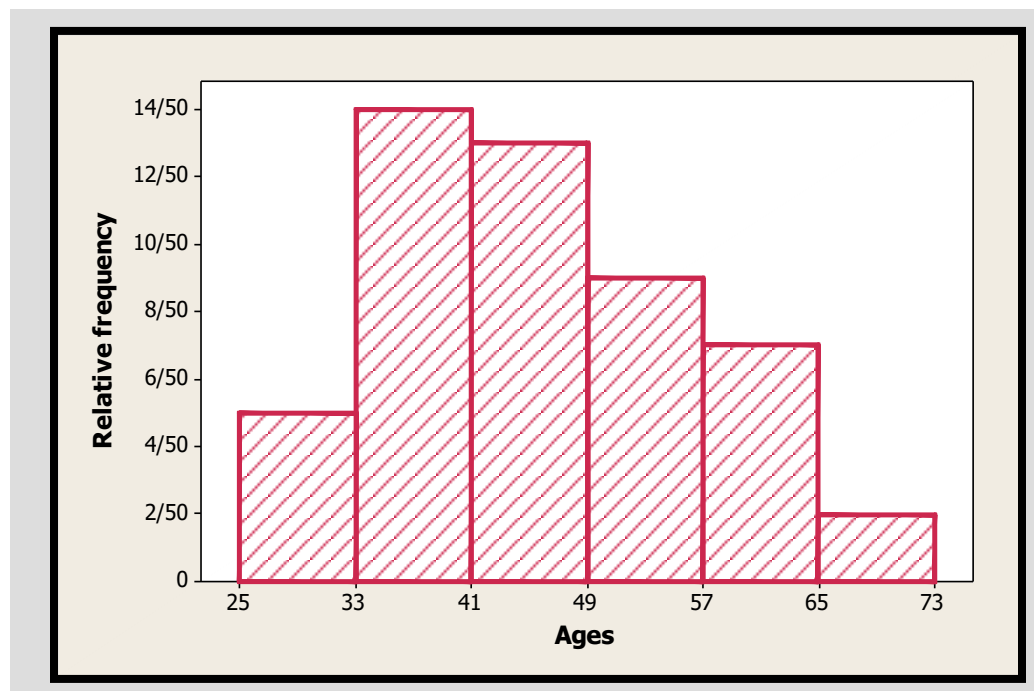
- We choose to use **6** intervals
- Minimum class width = $(70 - 26)/6 = 7.33$
- Convenient class width = **8**
- Use **6** classes of length **8**, starting at **25**.

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%

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

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?

Outliers?

What proportion of the tenured faculty are younger than 41?

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

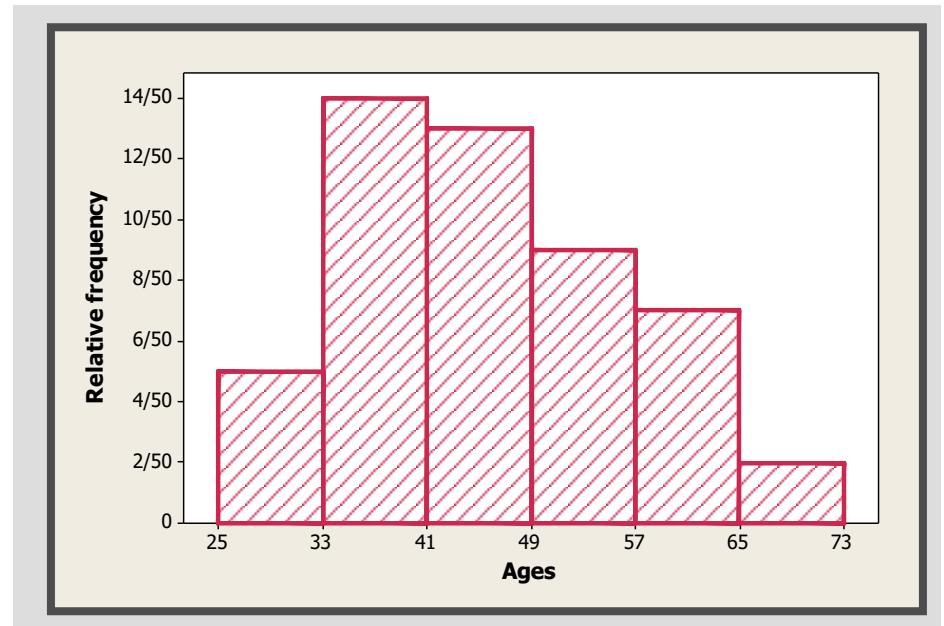
Describing the Distribution

Shape? **Skewed right**

Outliers? **No.**

What proportion of the tenured faculty are younger than 41?

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



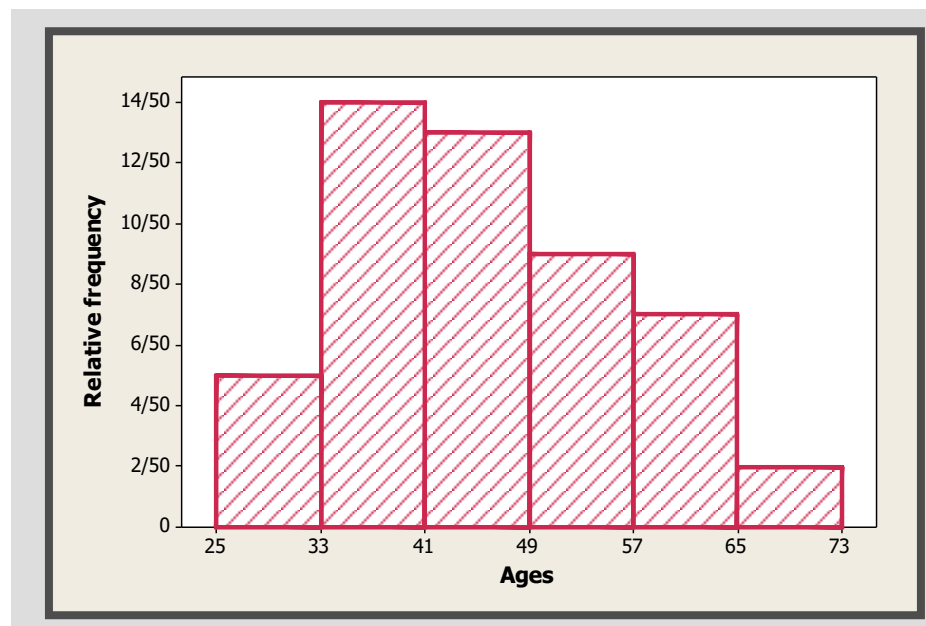
Describing the Distribution

Shape? **Skewed right**

Outliers? **No.**

What proportion of the tenured faculty are younger than 41?

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



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

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

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
 - a. Discrete
 - b. Continuous
2. Quantitative

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. 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