Statistics for Business and Economics

Chapter 1

Describing Data: Graphical

1.1

Dealing with Uncertainty

Everyday decisions are based on incomplete information

Consider:

- Will the job market be strong when I graduate?
- Will the price of Yahoo stock be higher in six months than it is now?
- Will interest rates remain low for the rest of the year if the federal budget deficit is as high as predicted?

Dealing with Uncertainty

(continued)

Numbers and data are used to assist decision making

- Statistics is a tool to help process, summarize, analyze, and interpret data

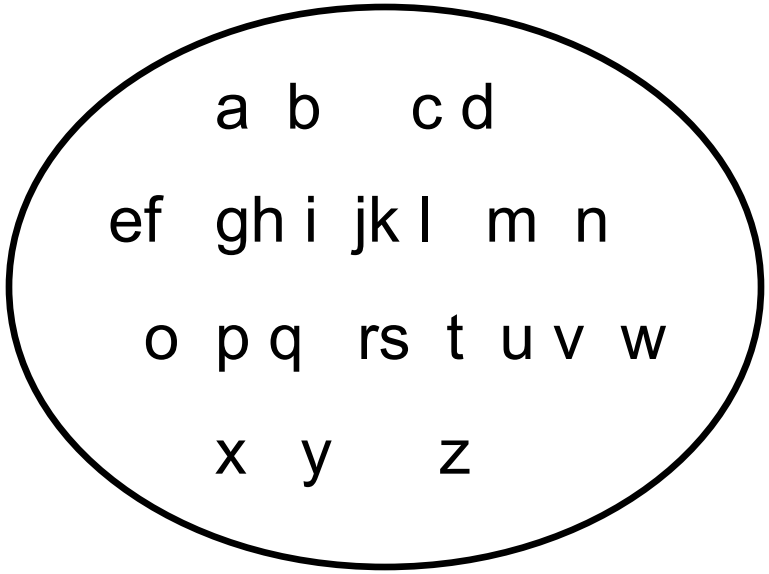
1.2

Key Definitions

- A population is the collection of all items of interest or under investigation
  - N represents the population size (N ≈ Infinity)
- A sample is an observed subset of the population
  - n represents the sample size
- A parameter is a specific characteristic of a population
- A statistic is a specific characteristic of a sample

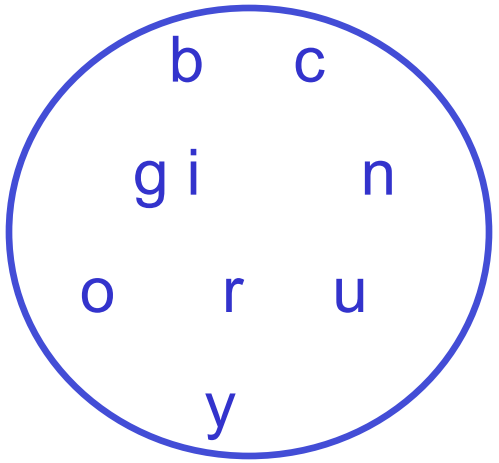
Population vs. Sample

Population



Values calculated using population data are called parameters

Sample



Values computed from sample data are called statistics

Examples of Populations

- Names of all registered voters in Canada
- Incomes of all families living in Vancouver
- Annual returns of all stocks traded on the Toronto Stock Exchange
- Grade point averages of all the students in UBC

Random Sampling

Simple random sampling is a procedure in which

- each member of the population is chosen strictly by chance,
- each member of the population is equally likely to be chosen,
- every possible sample of n objects is equally likely to be chosen

The resulting sample is called a random sample

Descriptive and Inferential Statistics

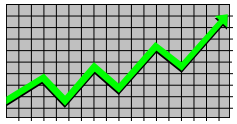
Two branches of statistics:

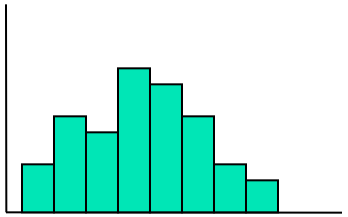
- Descriptive statistics
  - Graphical and numerical procedures to summarize and process data
- Inferential statistics
  - Using data to make predictions, forecasts, and estimates to assist decision making

Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs





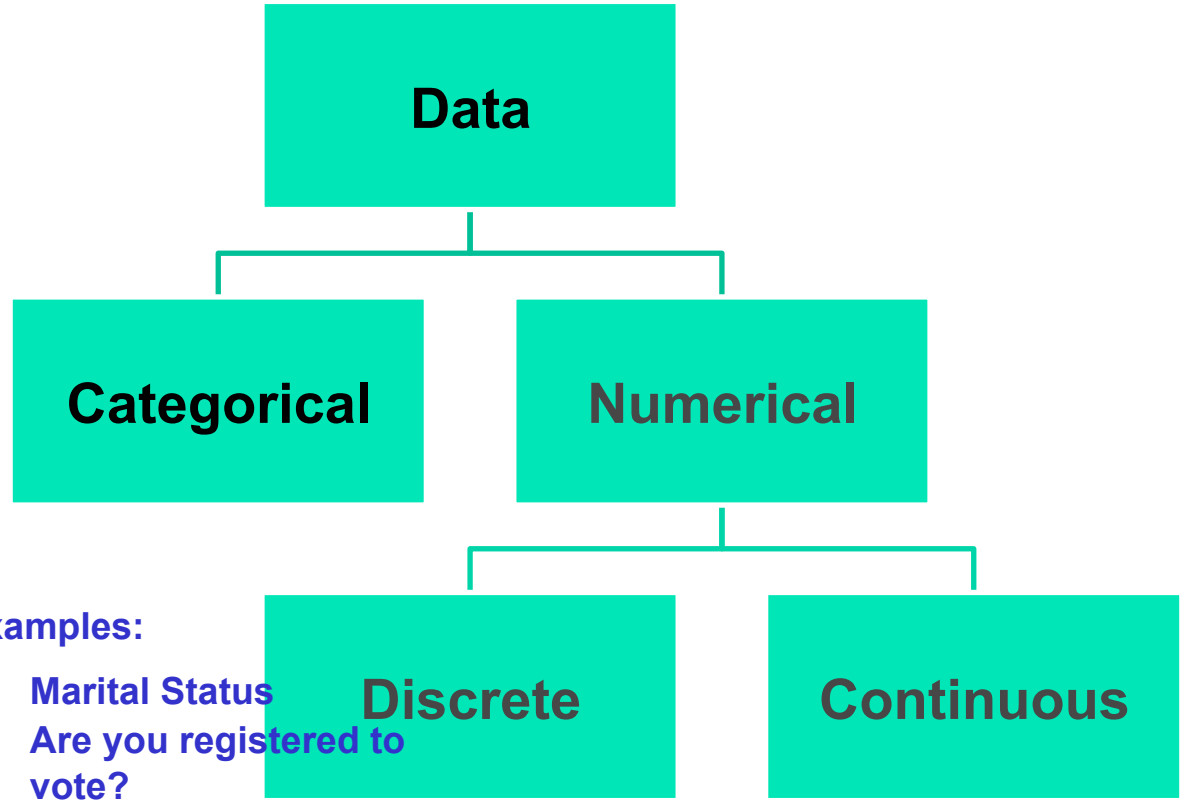
- Summarize data
  - e.g., Sample mean = $\frac{\sum X_i}{n}$

Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 140 pounds

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Types of Data



Examples:

- **Marital Status**
- **Are you registered to vote?**
- **Eye Color**

(Defined categories or groups)

Examples:

- **Number of Children**
- **Defects per hour**

(Counted items)

Examples:

- **Weight**
- **Voltage**

(Measured characteristics)

1.3

Graphical Presentation of Data

- Data in raw form are usually not easy to use for decision making
- Some type of organization is needed
  - Table
  - Graph
- The type of graph to use depends on the variable being summarized

Graphical Presentation of Data

(continued)

- Techniques reviewed in this chapter:

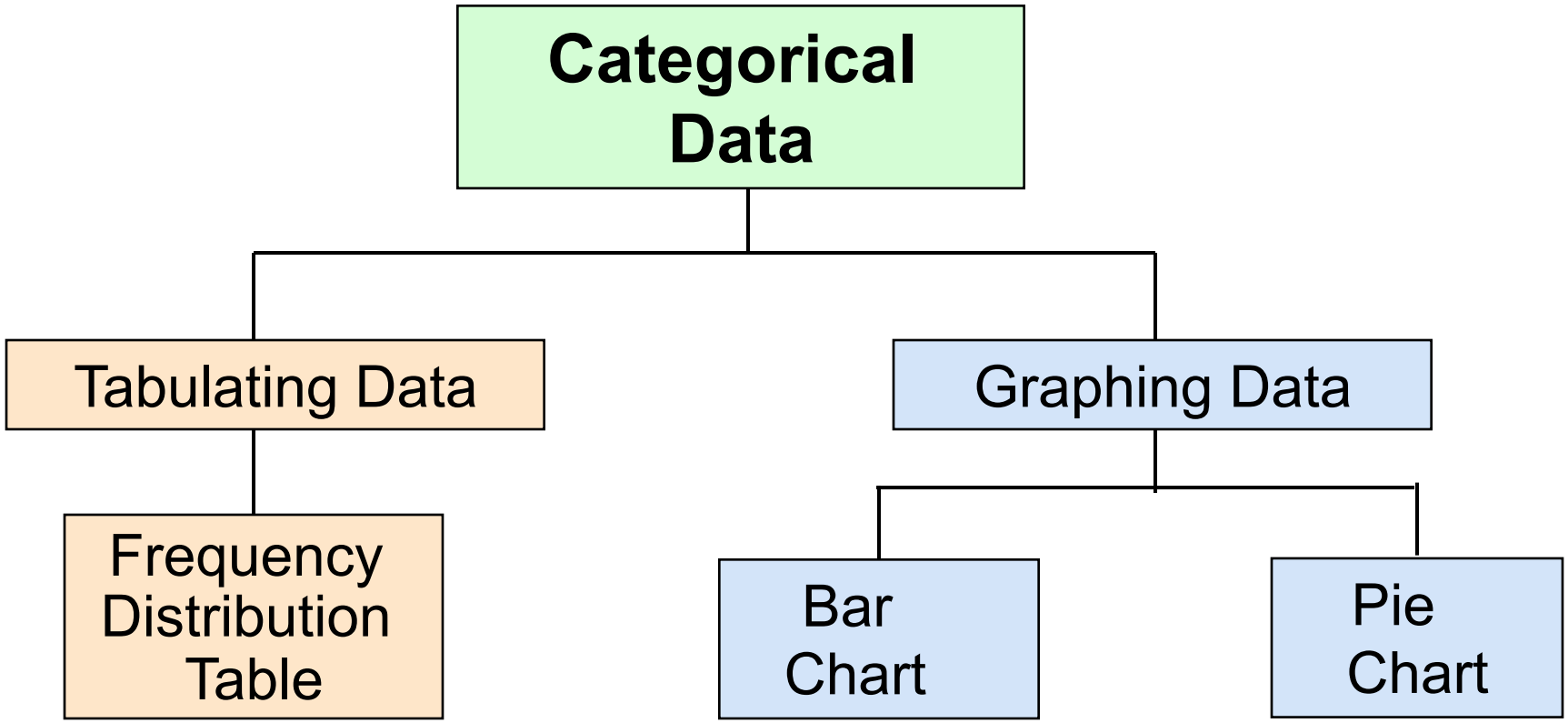
Categorical Variables

- Frequency distribution
- Bar chart
- Pie chart

Numerical Variables

- Line chart
- Frequency distribution
- Histogram
- Scatter plot

Tables and Graphs for Categorical Variables



The Frequency Distribution Table

Summarize data by category

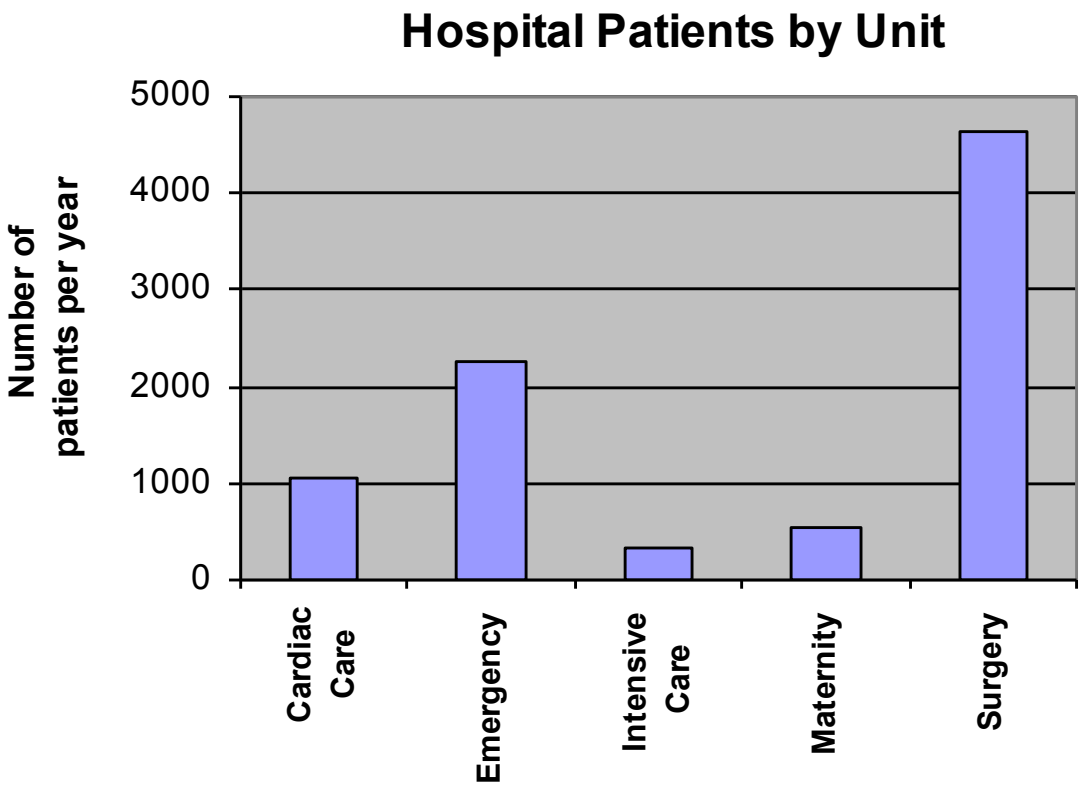
Example: Hospital Patients by Unit

Hospital Unit	Number of Patients
Cardiac Care	1,052
Emergency	2,245
Intensive Care	340
Maternity	552
Surgery	4,630

(Variables are categorical)

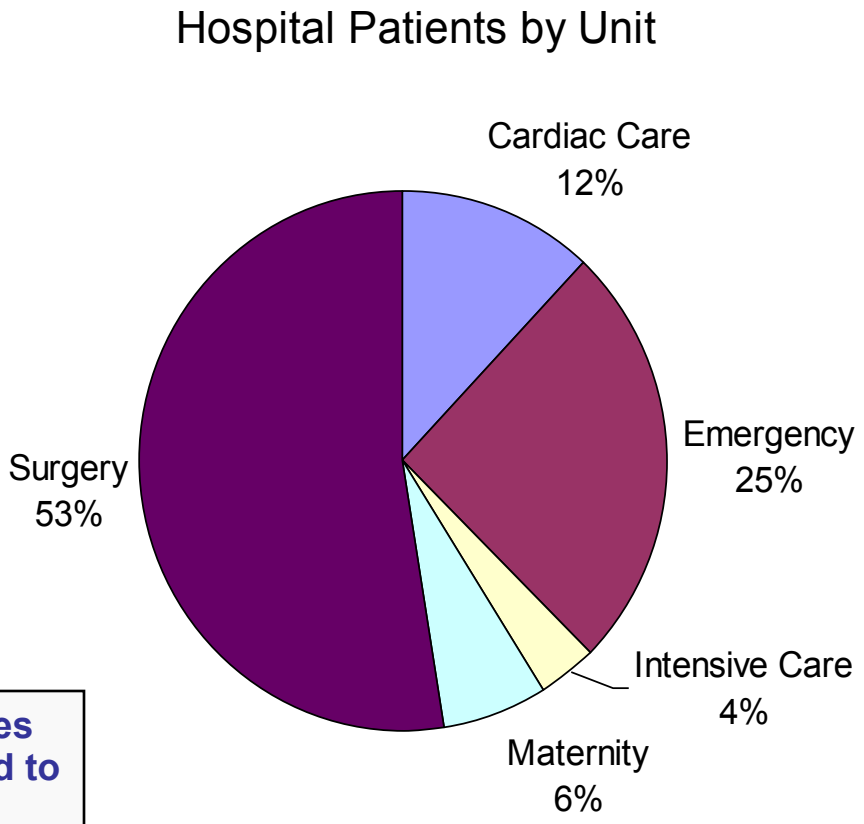
Bar Chart Example

Hospital Unit	Number of Patients
Cardiac Care	1,052
Emergency	2,245
Intensive Care	340
Maternity	552
Surgery	4,630



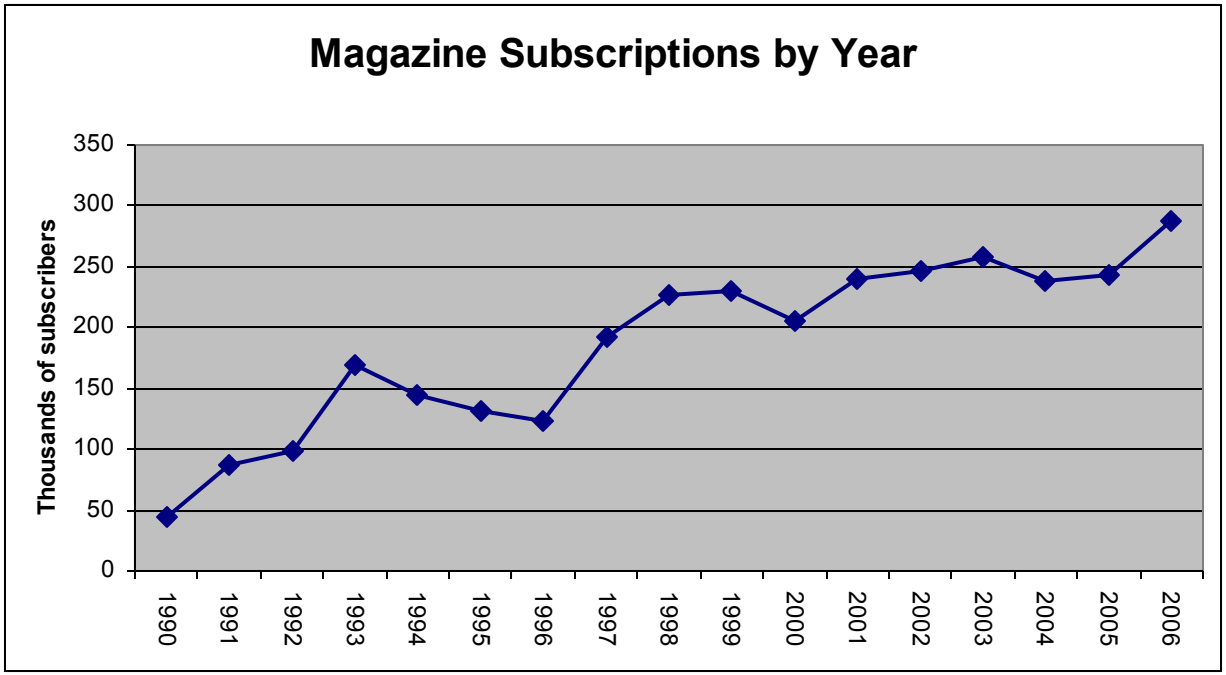
Pie Chart Example

Hospital Unit	Number of Patients	% of Total
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50



(Percentages are rounded to the nearest percent)

Line Chart Example



Frequency Distribution Example

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Frequency Distribution Example

(continued)

- Sort raw data in ascending order:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- Find range: **58 - 12 = 46**
- Select number of classes: **5**
- Compute interval width: **10** **(46/5 then round up)**
- Determine interval boundaries: **10 but less than 20, 20 but less than 30, . . . , 60 but less than 70**
- Count observations & assign to classes

Frequency Distribution Example

(continued)

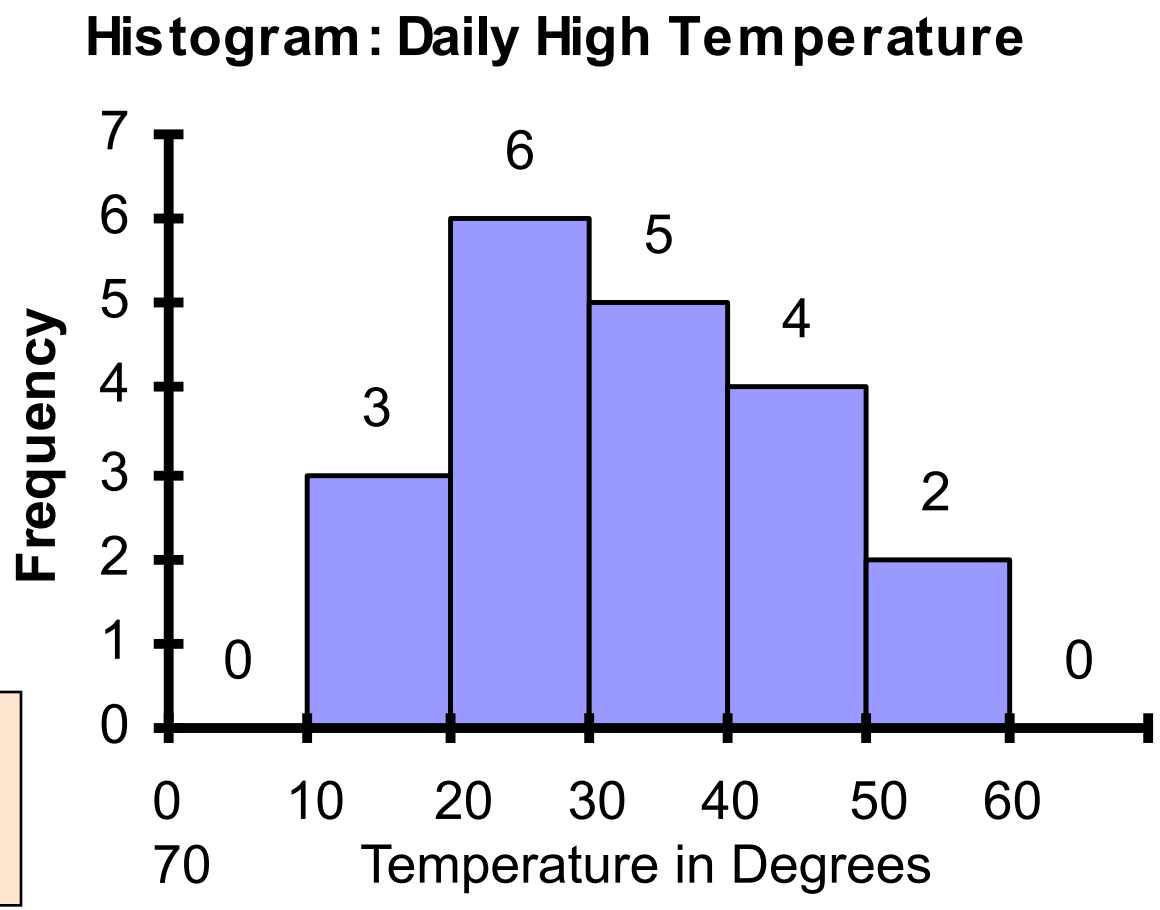
Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Interval	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Histogram Example

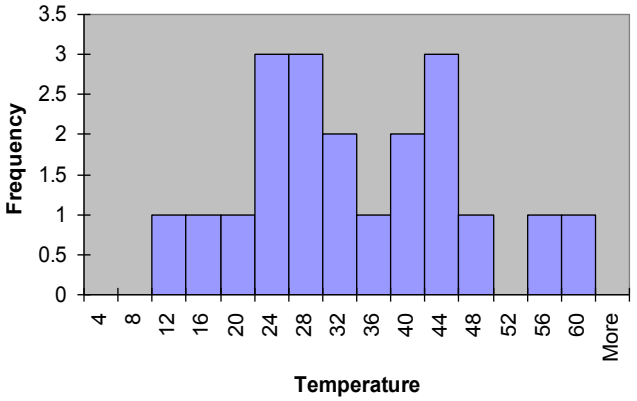
Interval	Frequency
10 but less than 20	3
20 but less than 30	6
30 but less than 40	5
40 but less than 50	4
50 but less than 60	2



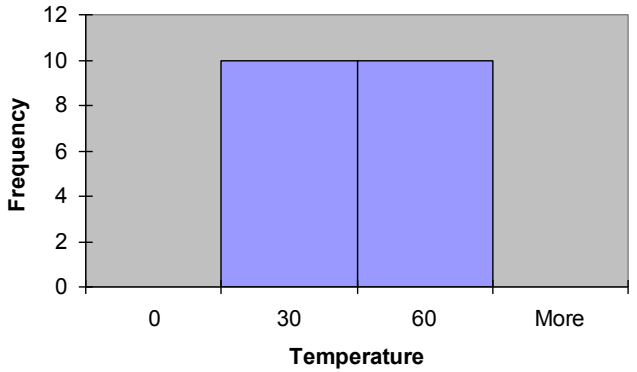
(No gaps between bars)

How Many Class Intervals?

- **Many (Narrow class intervals)**
  - may yield a very jagged distribution with gaps from empty classes
  - Can give a poor indication of how frequency varies across classes

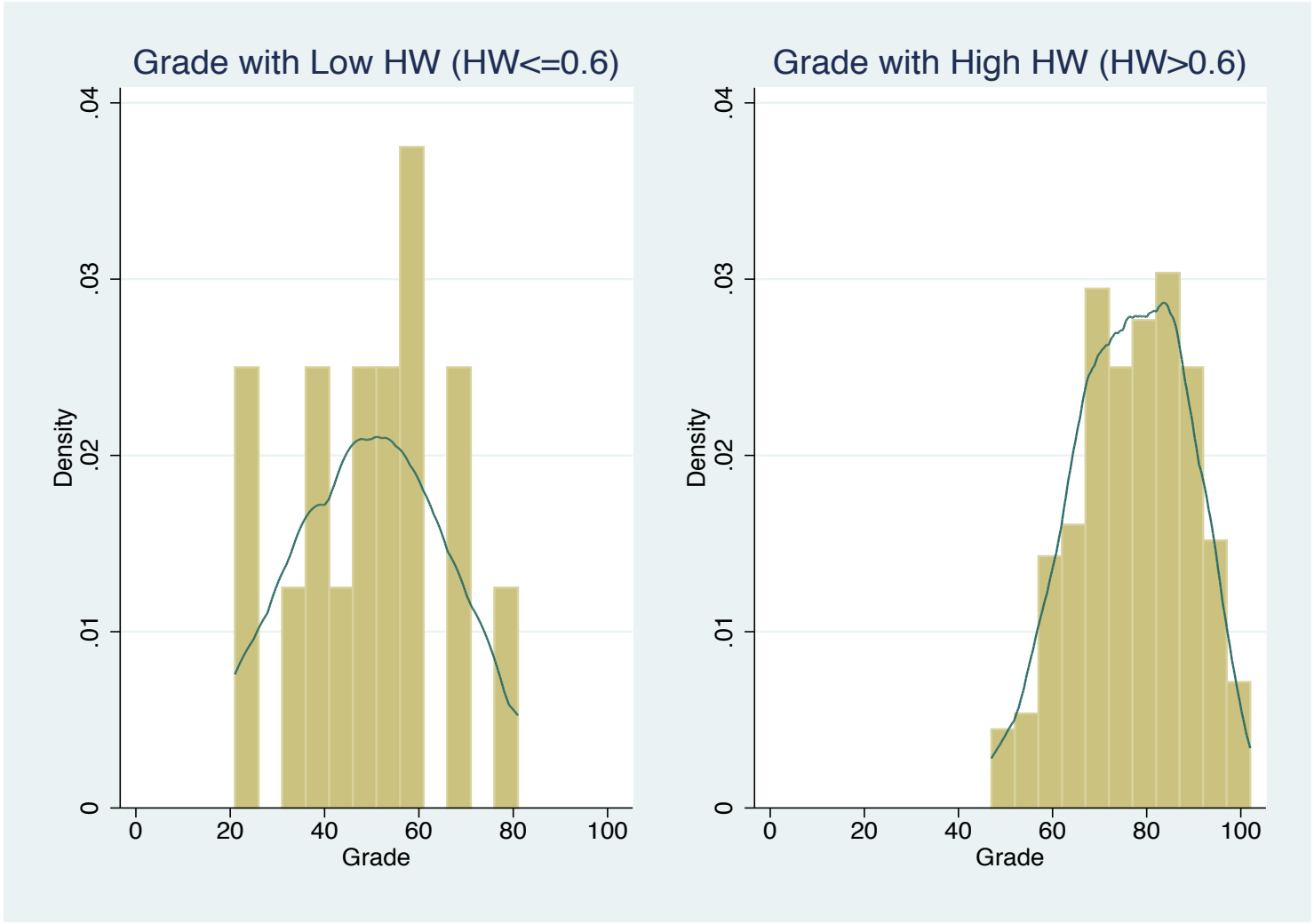


- **Few (Wide class intervals)**
  - may compress variation too much and yield a blocky distribution
  - can obscure important patterns of variation.



(X axis labels are upper class endpoints)

STATA Example

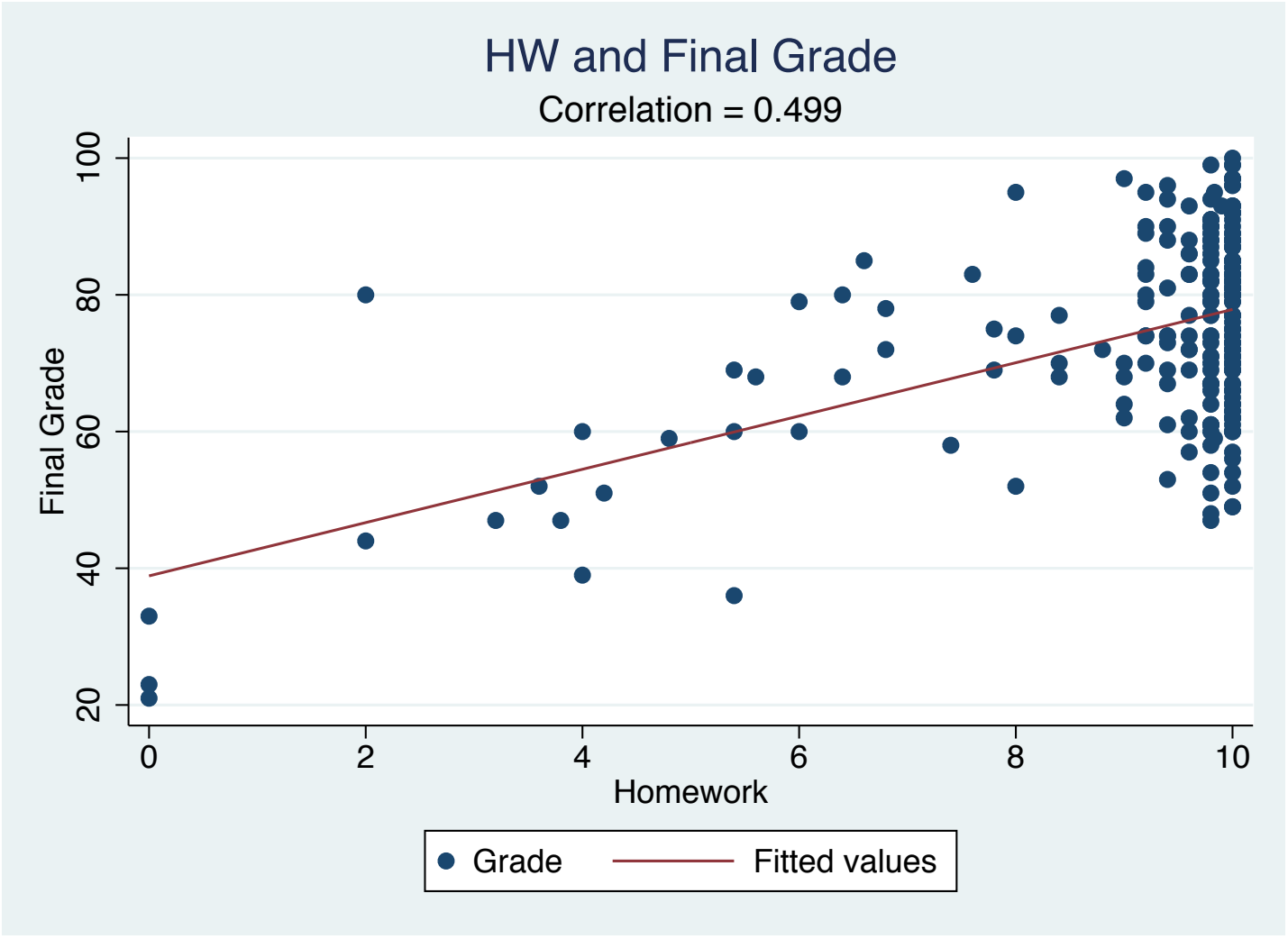


Scatter Diagrams

- Scatter Diagrams are used for paired observations taken from two numerical variables

- The Scatter Diagram:
  - one variable is measured on the vertical axis and the other variable is measured on the horizontal axis

STATA Example



Cross Tables

- Cross Tables (or contingency tables) list the number of observations for every combination of values for two categorical or ordinal variables

- If there are *r* categories for the first variable (rows) and *c* categories for the second variable (columns), the table is called an *r* x *c* cross table

Cross Table Example

- 2 x 4 Cross Table for type of patients and the daily average of smoking

Disease Group	Non-Smokers	1-14 Cigs.	15-24 Cigs.	25+ Cigs.	Total
lung-cancer	7	55	964	331	1357
Other dis.	61	129	1001	166	1357
Total	**68**	**184**	**1965**	**497**	**2714**