Chapter 2: Summarizing and Graphing Data

Basic Terms

Raw data --- numbers and category labels that are collected, but not yet processed

Variable --- a characteristic that differs from one individual to the next

Observational unit (observation) --- single individual who participates in a study

1

Basic Terms (continued)

Statistic --- a summary measure computed from sample data

Parameter --- a summary measure computed for an entire population

Descriptive Statistics --- summary numbers for either a population or a sample

2

Types of Data

***Qualitative variables* (*categorical variables*)** --- cannot be measured on a natural numerical scale; data classified into categories

Quantitative variables --- recorded numerical values; the data are either measurements or counts taken on each **individual**

3

Explanatory and Response Variables

Many questions are about the **relationship** between ***two variables***.

It is useful to identify one variable as the **independent variable (explanatory variable, predictor, covariate)** and the other variable as the **dependent variable (response variable).**

Generally, the *value of the independent variable* for an individual is thought to **partially explain** the *value of the dependent variable* for that individual.

4

Explanatory and Response Variables

Example:
Age (continuous) + smoking (yes/no) ➔ cancer (yes/no)

Age and smoking are explanatory or independent variables; and cancer is the response

NOTE: unless data are from a randomized experiment, an observed relationship between exploratory and response variables ***does not*** imply a causal relationship.

5

Describing Qualitative Data

Class---a category into which qualitative data can be classified

Class frequency---number of observations in the data set falling in a given class

Class relative frequency---class frequency divided by the total number of observations in the data set

$$\text{class relative frequency} = \frac{\text{class frequency}}{n}$$

6

Describing Qualitative Data

Numerical summaries for one or two categorical variables

Count how many fall into each category.

Calculate the percent in each category.

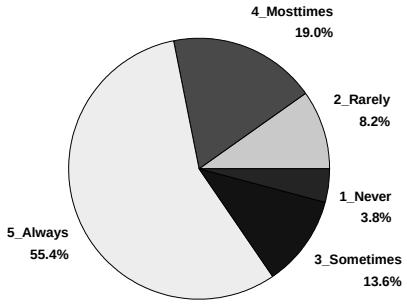
7

Table 2.1 Seatbelt Use by Twelfth-Graders When Driving

Response	Count	Percent
Always	1686	55.4%
Most times	578	19.0%
Sometimes	414	13.6%
Rarely	249	8.2%
Never	115	3.8%
Total	3042	100%

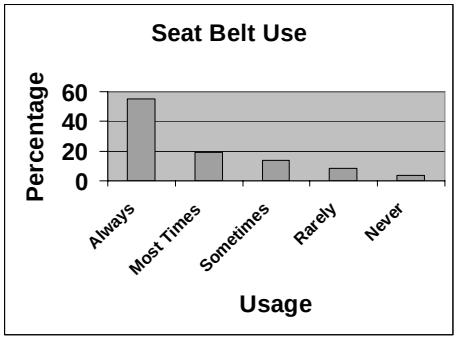
8

Describing qualitative data



9

Describing qualitative data—Bar Graph



10

If working with two variables, have the categories of the explanatory variable define the rows and compute row percentages.

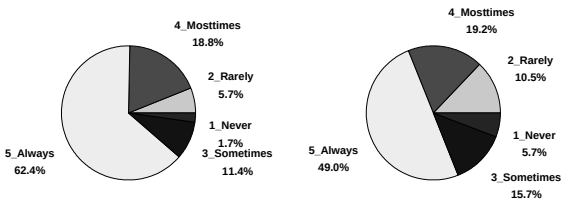
Table 2.2 Gender and Seatbelt Use by Twelfth-Graders When Driving

	Always	Most Times	Sometimes	Rarely	Never	Total
Female	915	276	167	84	25	1467
	(62.4%)	(18.8%)	(11.4%)	(5.7%)	(1.7%)	(100%)
Male	771	302	247	165	90	1575
	(49.0%)	(19.2%)	(15.7%)	(10.5%)	(5.7%)	(100%)

11

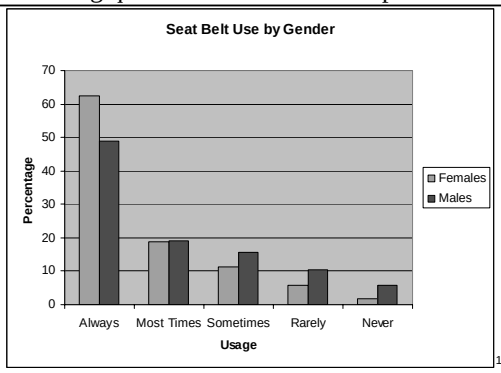
Describing qualitative data

Teen Seatbelt Usage: Females (left), Males (right)



12

Describing qualitative data—Bar Graph



13

Describing Quantitative Data

Example: 111 air temperature readings

67	72	74	62	65	59	61	69	66	68	58
64	66	57	68	62	59	73	61	61	67	81
79	76	82	90	87	82	77	72	65	73	76
84	85	81	83	83	88	92	92	89	73	81
80	81	82	84	87	85	74	86	85	82	86
88	86	83	81	81	81	82	89	90	90	68
86	82	80	77	79	76	78	78	77	72	79
81	86	97	94	96	94	91	92	93	93	87
84	80	78	75	73	81	76	77	71	71	78
67	76	68	82	64	71	81	69	63	70	75
76										

14

Describing Quantitative Data

Stem-and-leaf plots

5 : 7899
6 : 11122344
6 : 5566777888899
7 : 0111222333344
7 : 5566666777778888999
8 : 000111111111122222223333444
8 : 555666667778899
9 : 00012223344
9 : 67

In this example the stems correspond to the values of the 10's digits; the leafs are the values of the 1's digits

15

Describing Quantitative Data

5 : 7
5 : 899
6 : 111
6 : 223
6 : 4455
6 : 66777
6 : 888899
7 : 0111
7 : 2223333
7 : 4455
7 : 6666667777
7 : 8888999
8 : 0001111111111
8 : 2222222333
8 : 444555
8 : 66666777
8 : 8899
9 : 0001
9 : 22233
9 : 44
9 : 67

- Data arranged in ascending order
- Easy to identify individual measurements

16

Describing Quantitative Data

Histograms

- *x*-axis divided into intervals (best to use equal class/interval sizes); between 6 and 15 intervals is a good number
- *y*-axis gives the frequency (count) or relative frequency of the measurements that fall into each interval
  - Draw a bar with corresponding height
  - Decide rule to use for values that fall on the border between two intervals

17

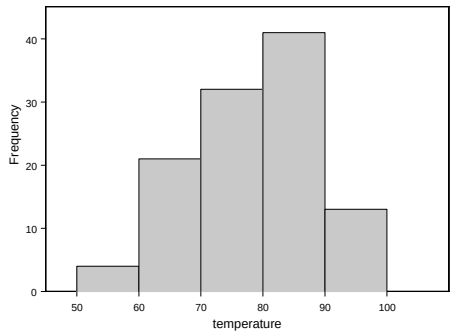
Describing Quantitative Data

Histograms (continued)

- The proportion of total area under the histogram that falls above a particular interval on the *x*-axis equals the relative frequency of measurements contained in the interval
- Cannot identify individual measurements

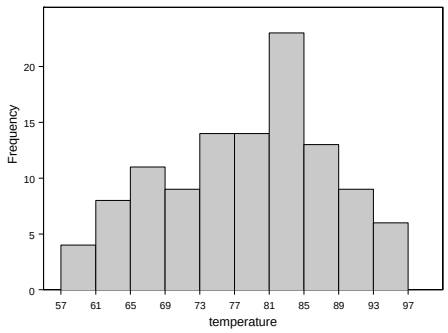
18

Describing Quantitative Data--Histogram

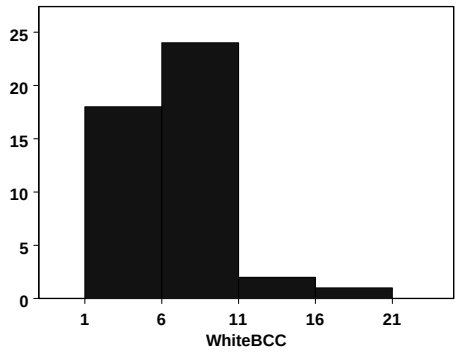


19

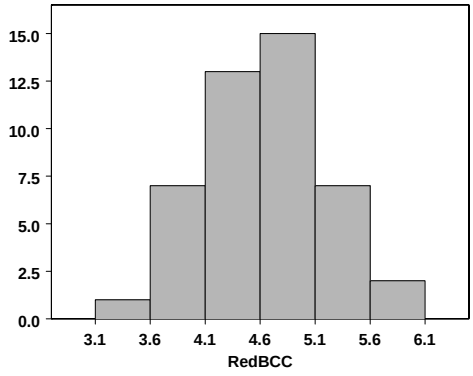
Describing Quantitative Data--Histogram



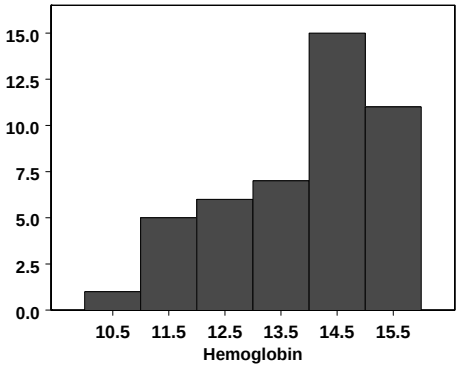
20



21

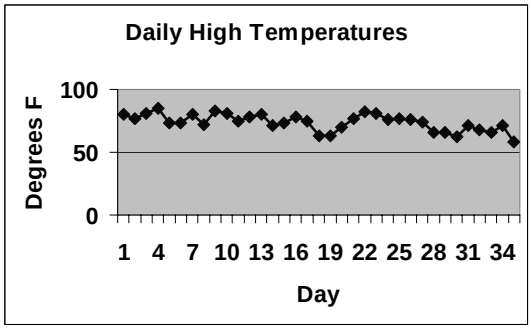


22



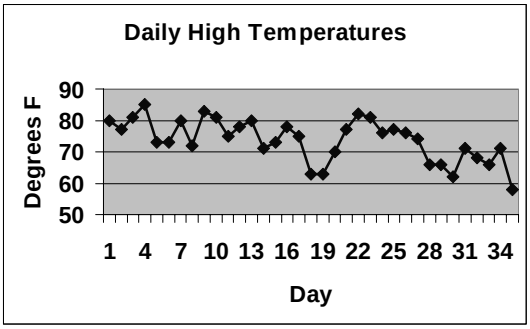
23

Describing Quantitative Data—Time Series



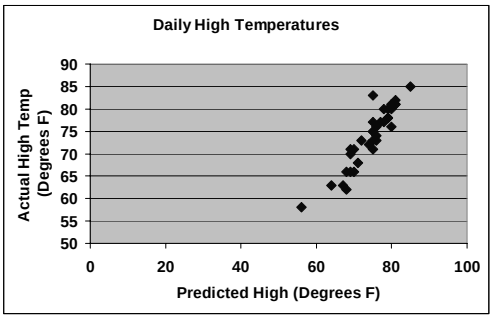
24

Describing Quantitative Data—Time Series



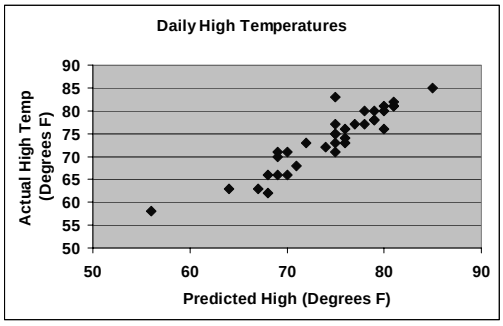
25

Describing Quantitative Data—Scatterplot



26

Describing Quantitative Data—Scatterplot



27