Recap

Measures of spread: SD vs. IQR

- SD: Appropriate when the dist. is symmetric looking.
- IQR: Appropriate when the distn. is skewed or the data has outliers.

Rule of thumb for outlier detection: 1.5(IQR) rule.

"Outlier" if it falls outside $[\hat{Q}_1 - 1.5\,\hat{IQR},\ \hat{Q}_3 + 1.5\,\hat{IQR}]$.

Graphical Statistics

"Plot the data before you do anything with it."

Boxplot: Displays the 5-number summary of the data, i.e., $(\min, \hat{Q}_1, \hat{Q}_2, \hat{Q}_3, \max)$. It shows

- the data distribution (e.g., symmetric, right-skewed or left-skewed)
- outliers

Alternative form: The bottom whisker extends from $\hat{Q}_1$ to $\max\{\min, \hat{Q}_1 - 1.5 \times I\hat{Q}R\}$ and the top whisker extends from $\hat{Q}_3$ to $\min\{\max, \hat{Q}_3 + 1.5 \times I\hat{Q}R\}$

Side-by-side boxplots: Draw side-by-side boxplots on the same scale to compare distributions of more than one data set — see Figure 8.10 in the textbook.

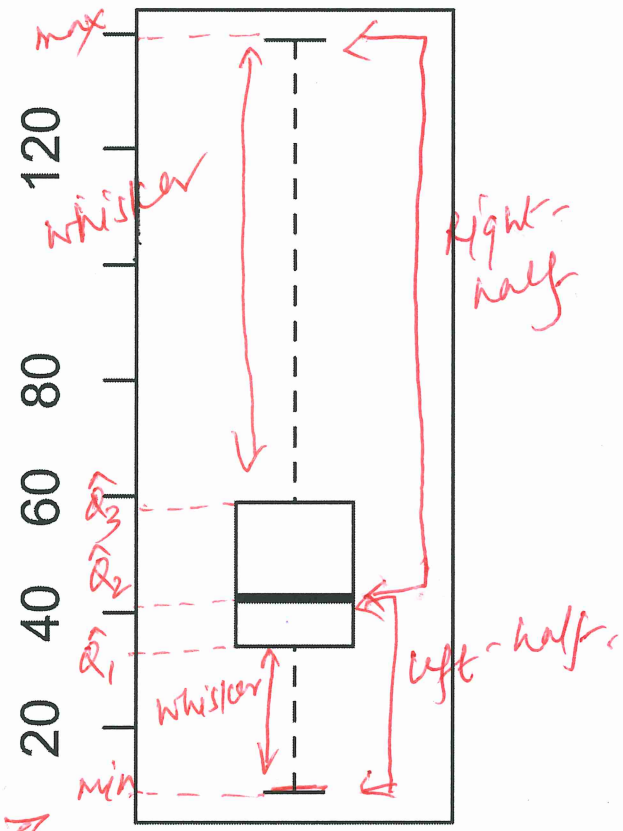
Ex: CPU data

```
?boxplot # see help
par(mfrow=c(1,2)) # 2 plots in 1 row
# plot of 5-number summary
boxplot(cpu, range=0)
# uses 1.5 (IQR) rule (also default), i.e.,
# same as boxplot(cpu)
boxplot(cpu, range=1.5)
par(mfrow=c(1,1)) # back to the default, 1 plot per row
```

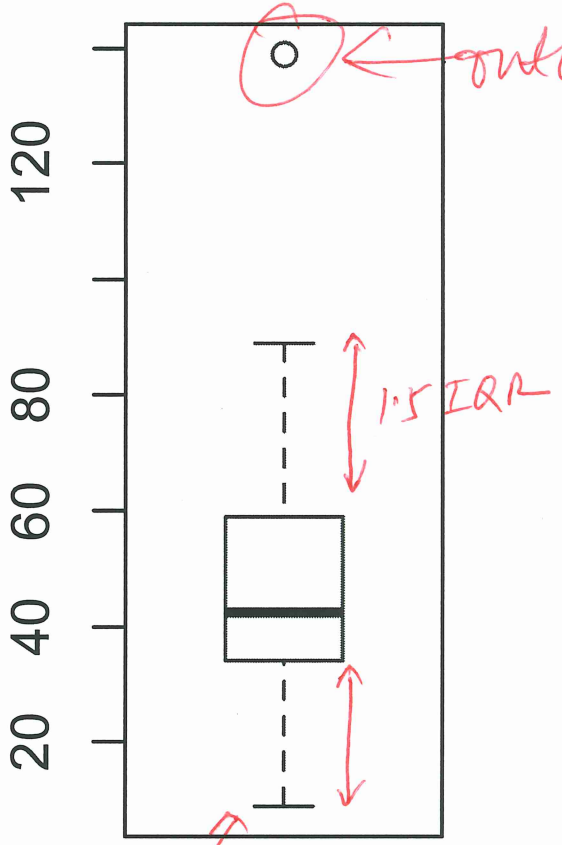
→ 1.5 IQR rule.

Boxplots for CPU data

Dist. is right-skewed.



Display of 5-# summary



Box plot with 1.5(IQR) rule

outlier.

1.5 IQR

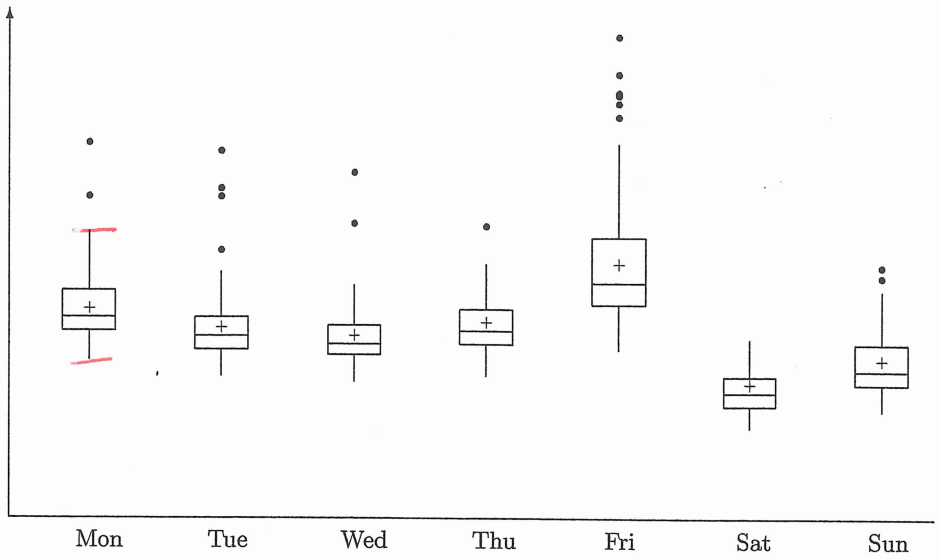


FIGURE 8.10: *Parallel boxplots of internet traffic.*

8.3.4 Scatter plots and time plots

Scatter plots are used to see and understand a relationship between two variables. These can be temperature and humidity, experience and salary, age of a network and its speed, number of servers and the expected response time, and so on.

To study the relationship, both variables are measured on each sampled item. For example, temperature and humidity during each of n days, age and speed of n networks, or experience and salary of n randomly chosen computer scientists are recorded. Then, a **scatter plot** consists of n points on an (x, y)-plane, with x- and y-coordinates representing the two recorded variables.

Example 8.20 (ANTIVIRUS MAINTENANCE). Protection of a personal computer largely depends on the frequency of running antivirus software on it. One can set to run it every day, once a week, once a month, etc.

During a scheduled maintenance of computer facilities, a computer manager records the number of times the antivirus software was launched on each computer during 1 month (variable X) and the number of detected worms (variable Y). The data for 30 computers are in the table.

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

Is there a connection between the frequency of running antivirus software and the number of worms in the system? A scatter plot of these data is given in Figure 8.11a. It clearly shows that the number of worms reduces, in general, when the antivirus is employed more frequently. This relationship, however, is not certain because no worm was detected on some "lucky" computers although the antivirus software was launched only once a week on them. ◊

Histogram

Shows the data distribution and suggests possible outliers. Its shape is similar to the population pdf/pmf, especially if the sample size is large.

Frequency histogram: Consists of bars, one over each bin, whose heights represent the *number* of observations in the bins.

Relative frequency histogram: Consists of bars, one over each bin, whose heights represent the *proportion* of observations in the bins.

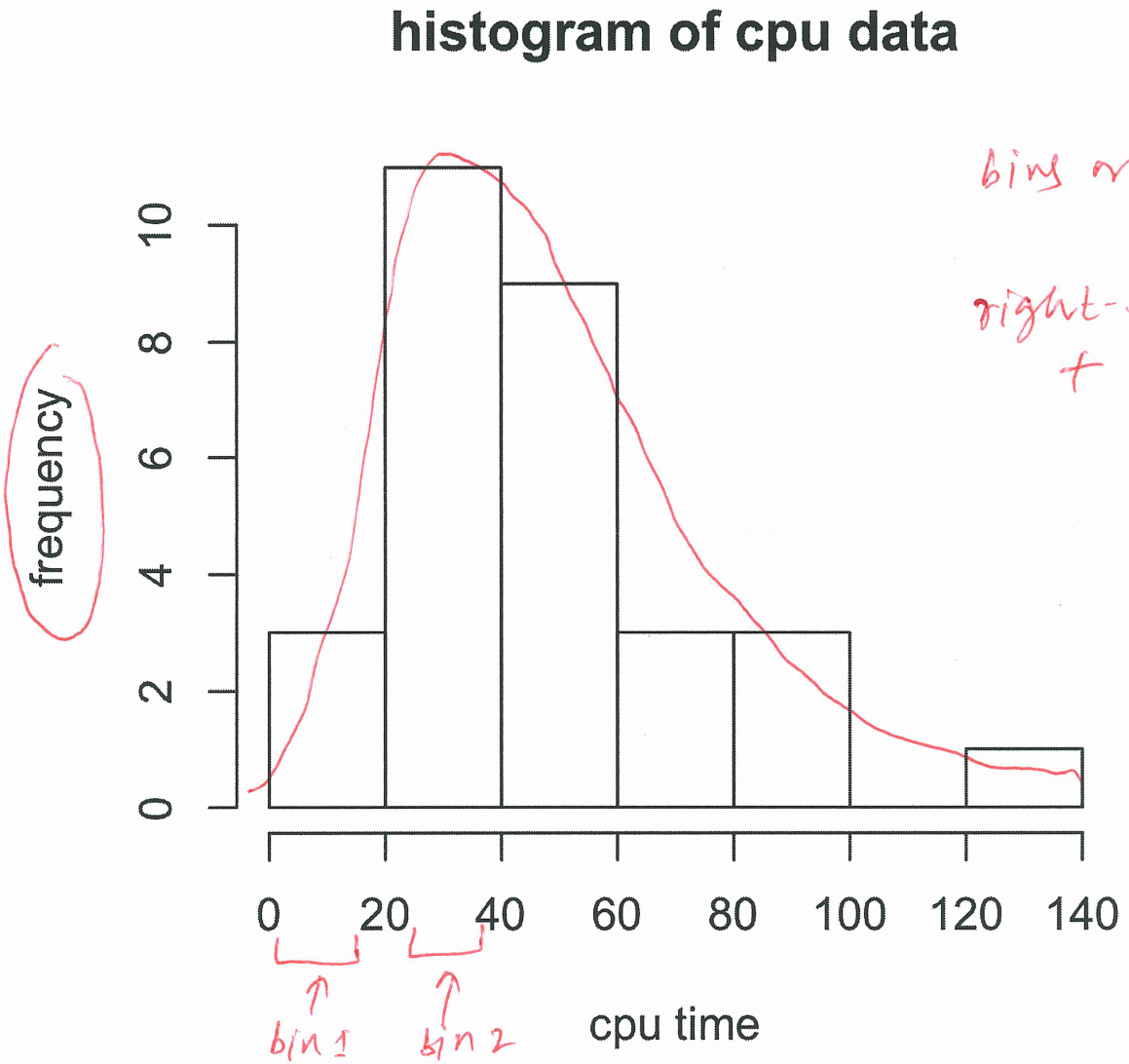
How to construct a histogram?

- effect of number of bins (too many or too few)
- bins of unequal sizes

avoid.

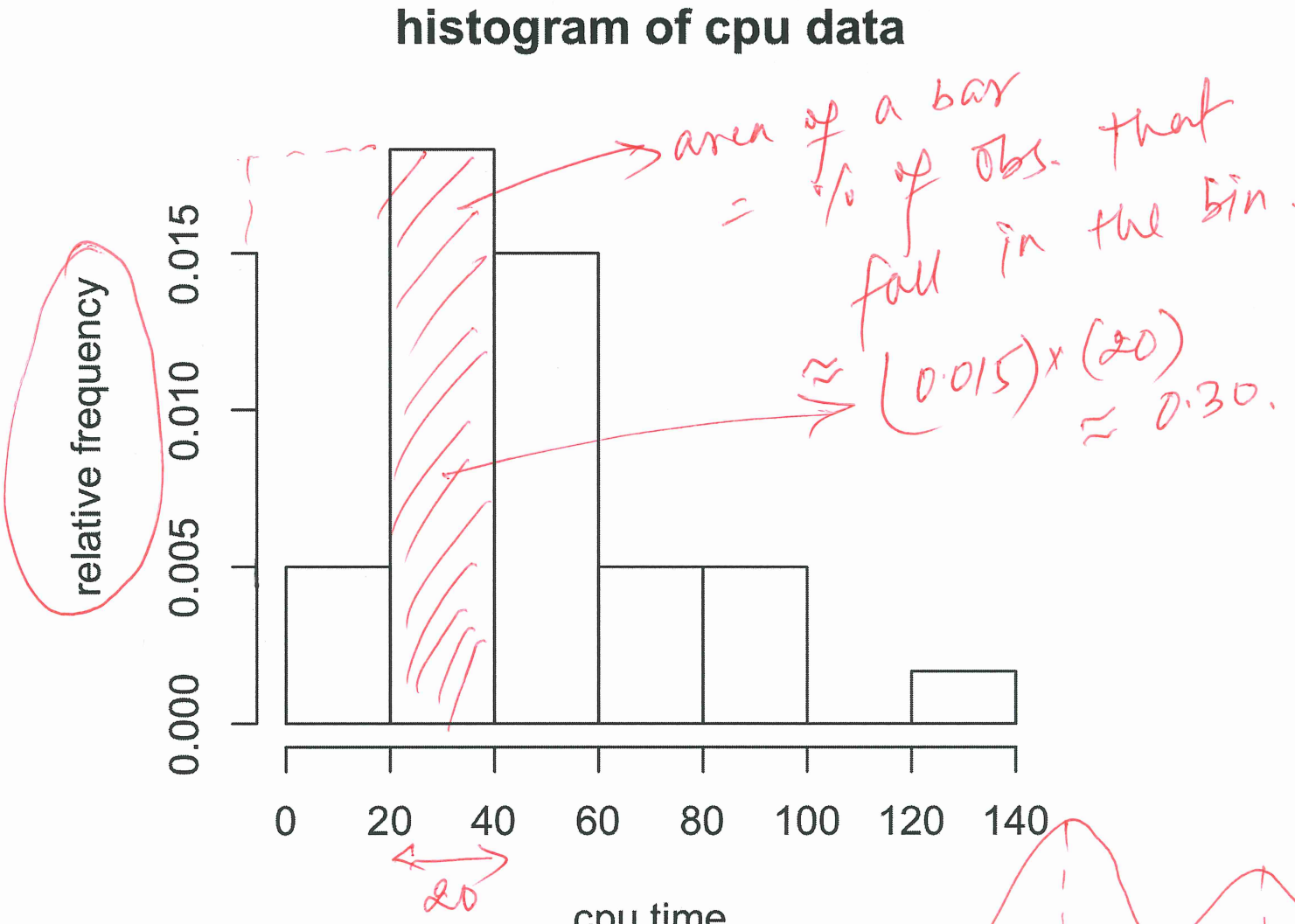
avoid this.

```
# frequency histogram by default
hist(cpu, xlab="cpu time", ylab="frequency",
            main="histogram of cpu data")

# relative frequency (density) histogram
hist(cpu, freq=FALSE, xlab="cpu time",
ylab="relative frequency", main="histogram of cpu data")
```


histogram of cpu data
frequency
0
2
4
6
8
10
0
20
40
60
80
100
120
140
cpu time
bin 1
bin 2
bins or class intervals.
right-skewed dist.
+ evidence of
outlier.

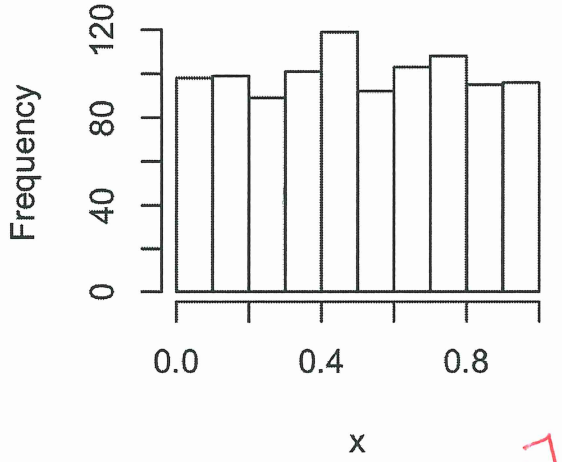
Sample mode: Most frequent obs. in the data

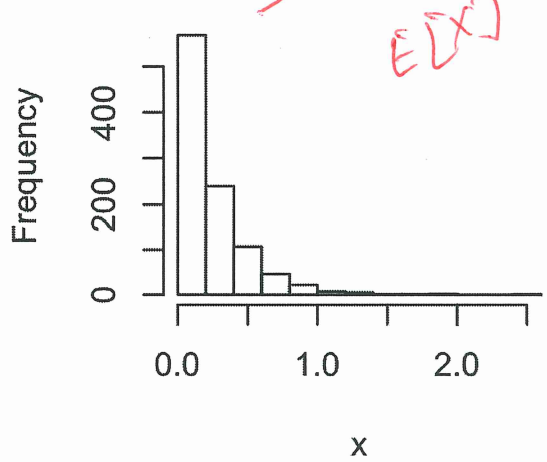
Population mode: The point with the highest pmf/pdf

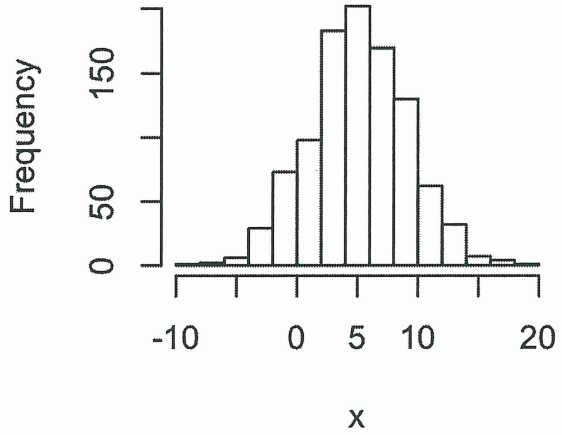


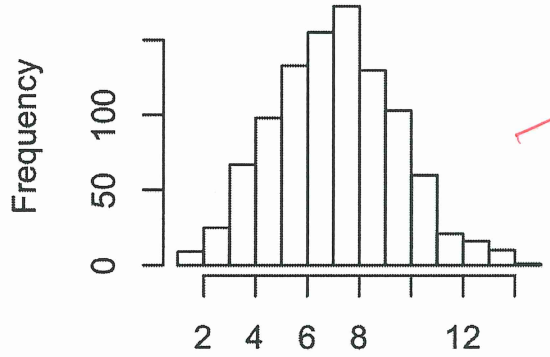
Histograms of some simulated data:

```
nsim <- 1000
# uniform (0,1) distribution
par(mfrow=c(2,2))
hist(runif(nsim), xlab="x", main="")

# exponential (lambda = 4) distribution
hist(rexp(nsim, rate=4), xlab="x", main="")

# normal (mu=5, sigma^2=16) distribution
hist(rnorm(nsim, mean=5, sd=4), xlab="x", main="")

# binomial (n=50, p=0.25)
hist(rbinom(nsim, size=30, prob=0.25), main="")
par(mfrow=c(1,1))
```


Frequency
0
40
80
120
0.0
0.4
0.8
x


Frequency
0
200
400
0.0
1.0
2.0
x


Frequency
0
50
150
-10
0
5
10
20
x


Frequency
0
50
100
2
4
6
8
12
rbinom(nsim, size = 30, prob = 0.25)

Why does the last histogram have a "normal shape?"

Recall: $X \sim \text{Bin}(n, p)$

X = # successes in n BT's.

NOTE: $X = \sum_{i=1}^{n} X_i$, $X_i \sim \text{Bernoulli}(p)$, (indep.)

X = sum of n IID BRVs.

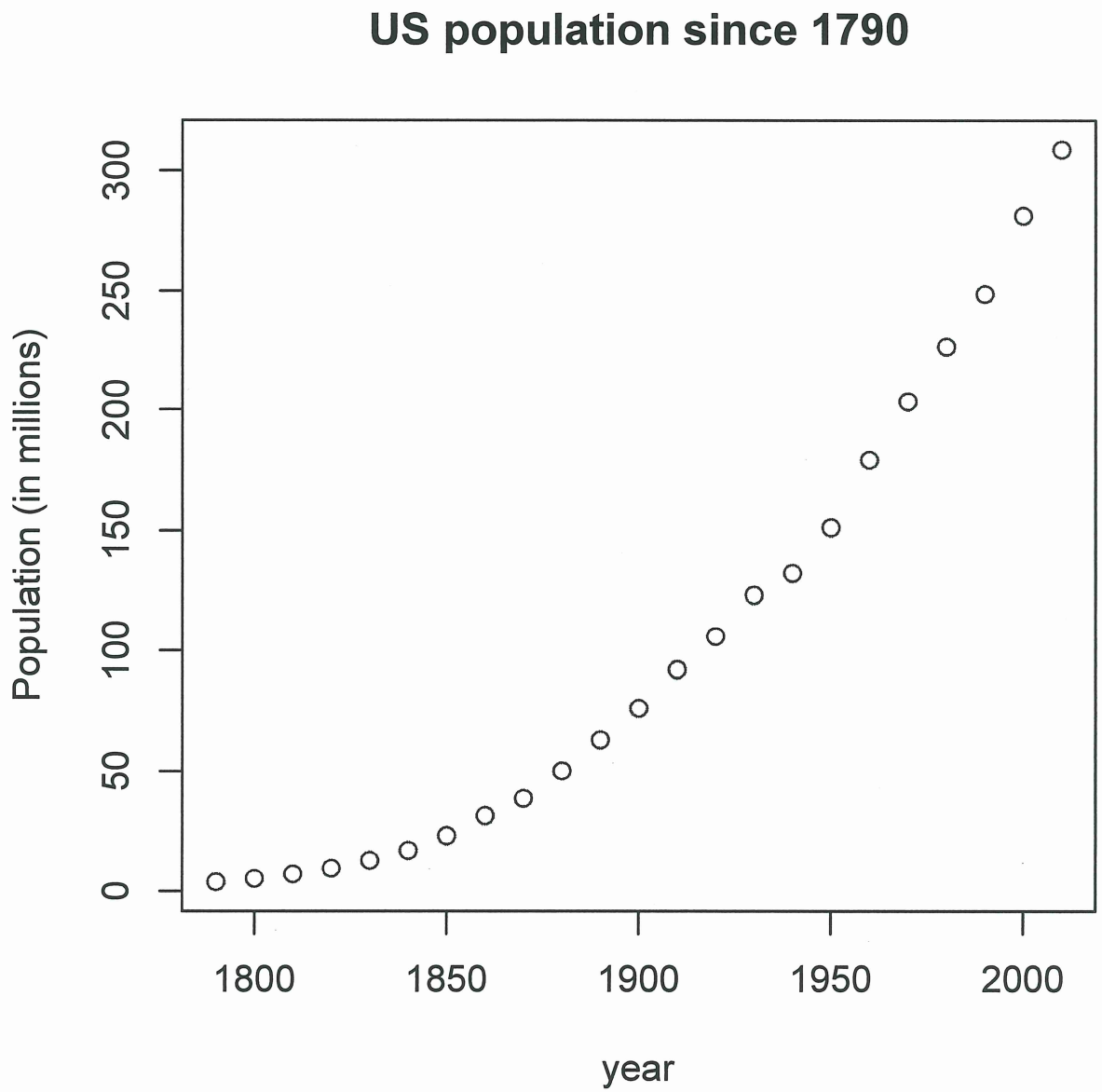
CLT: If n is large, $X \approx \text{Normal}\left[\mu = np, \sigma^2 = np(1-p)\right]$.

works if p is not too small.

Time series plot: Plot of a data on a variable against time — shows how the variable changes over time.

```
# Data from Exercise 8.5
year <- seq(from=1790, to=2010, by=10)
# > year
 # [1] 1790 1800 1810 ...  2010
# >
uspop <- c(3.9, 5.3, 7.2, 9.6, ..., 281.4, 308.7)

plot(year, uspop, ylab="Population (in millions)",
  main="US population since 1790")
```

Scatterplot: Plot of one variable (X) against another variable Y — shows the relationship between the two variables. See Figure 8.11 of the textbook.



Later in:
Fitting an equation that describes the 'trend' here.

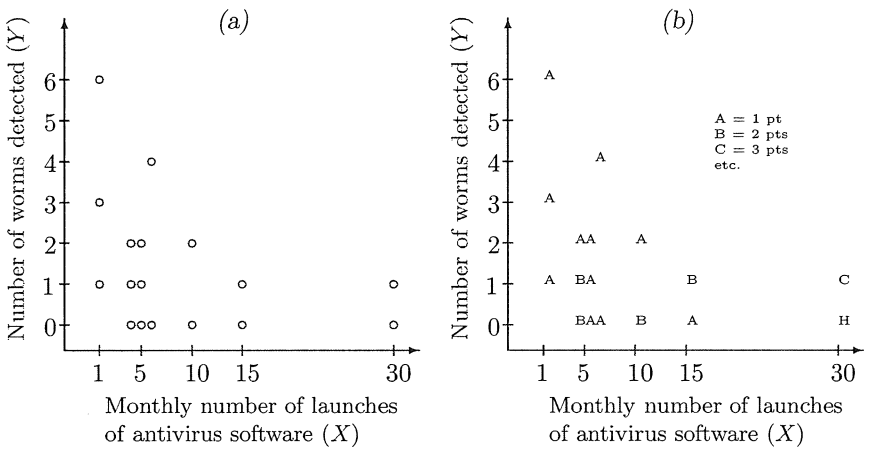


Figure 8.10 *Scatter plots for Examples 8.18 and 8.19.*

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0
X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

Is there a connection between the frequency of running antivirus software and the number of worms in the system? A scatter plot of these data is given in Figure 8.10a. It clearly shows that the number of reduces, in general, when the antivirus is employed more frequently. This relationship, however, is not certain because no worm was detected on some "lucky" computers although the antivirus software was launched only once a month on them. ◇

Example 8.19 (PLOTTING IDENTICAL POINTS). Looking at the scatter plot in Figure 8.10a, the manager in Example 8.18 realized that a portion of data is hidden there because there are identical observations. For example, no worms were detected on 8 computers where the antivirus software is used daily (30 times a month). Then, Figure 8.10a may be misleading.

When the data contain identical pairs of observations, the points on a scatter plot are often depicted with either numbers or letters ("A" for 1 point, "B" for two identical points, "C" for three, etc.). You can see the result in Figure 8.10b. ◇

When we study time trends and development of variables over time, we use **time plots**. These are scatter plots with x-variable representing time.