Chapter 1

Descriptive Statistics

"强国需知十三数" (商鞅, 390B.C.)

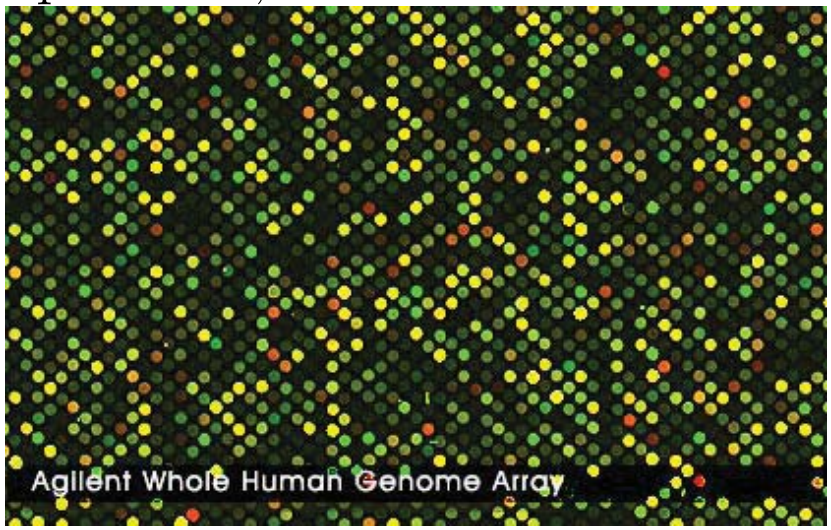
"*Every country to be powerful needs to know 13 numbers*"

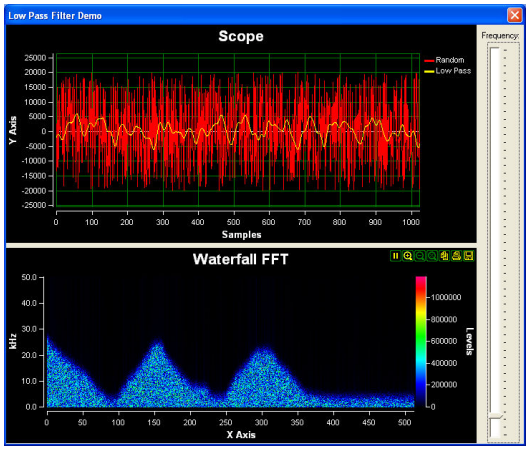
(Shang Yang, statesman and thinker)

"I keep saying that the sexy job in the next 10 years will be statisticians, and I'm not kidding" (Hal Varian, Google's chief economist in 2009 with The McKinsey Quarterly)

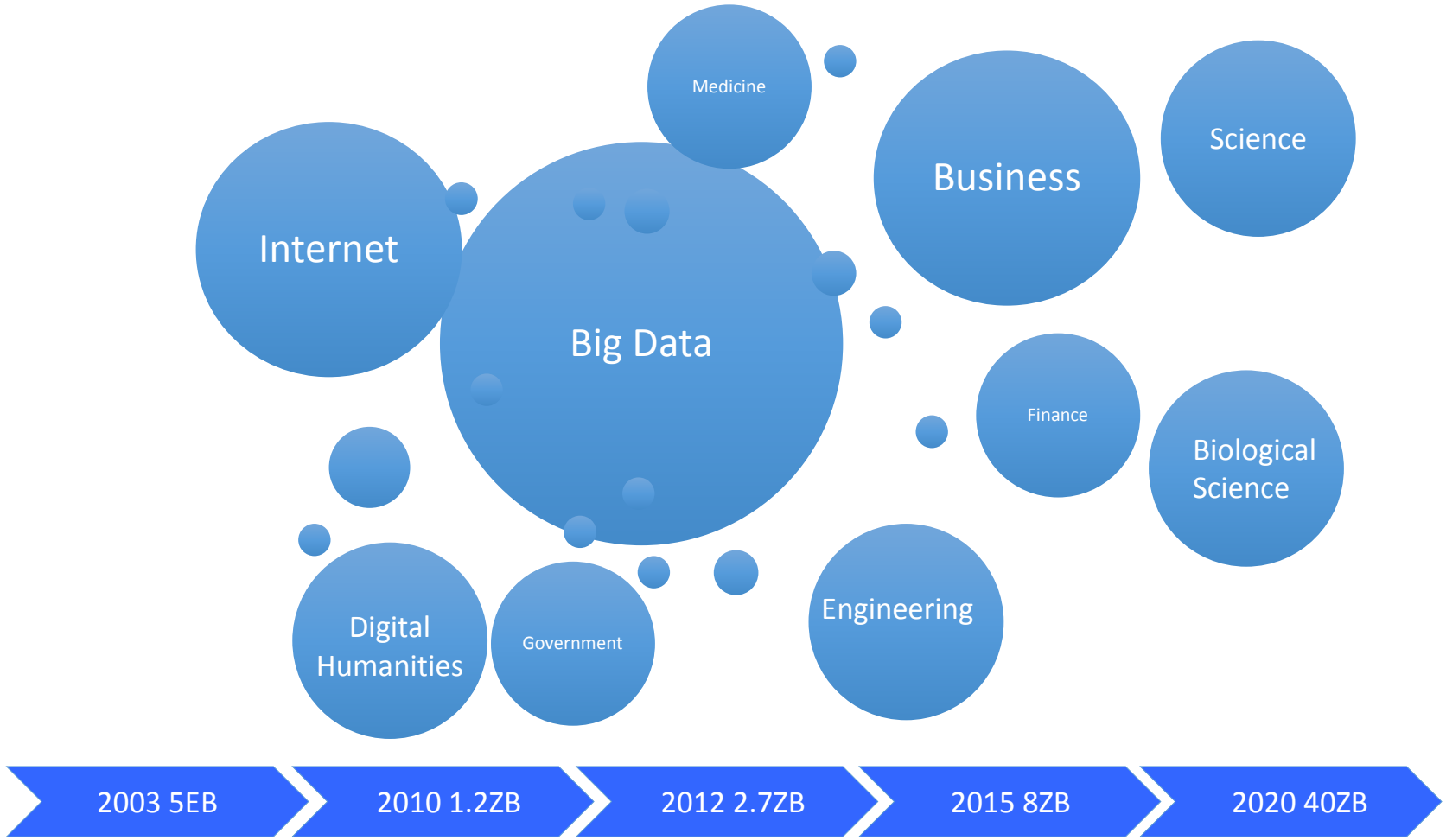
1.1 Introduction

- Evolution of Dimensions, Complexities, and Sizes





Big Data are ubiquitous



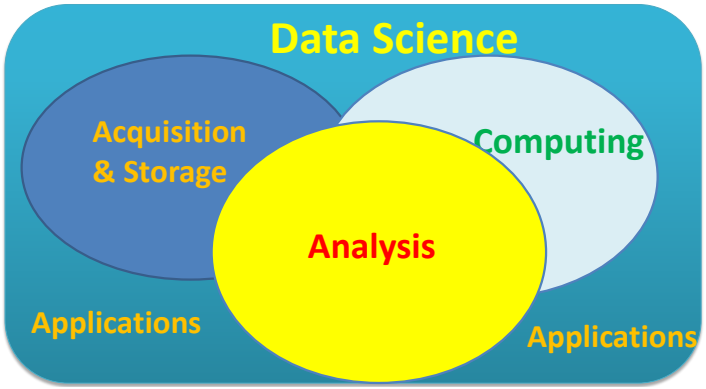
"*There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days*", said Eric Schmidt, CEO of Google, in 2010.

Volume **Velocity** **Variety**

They have huge impacts on

System: storage, communication, computation architectures

Analysis: statistics, computation, optimization



Big Data $\Longrightarrow$ **Smart Data**

What can big data do? Hold great promises for understanding

★ **Heterogeneity**: personalized medicine or services

★ **Commonality**: in presence of large variations (noises)

from large pools of variables, factors, genes, environments and their interactions as well as **latent factors**.

"Big data is not about the data" (Gary King, Harvard University)

- It is about **smart statistics**, **not size**
- It powers modern machine learning and AI

What is statistics?

★ infer about populations from samples via **prob. modeling**;

★ predict future outcomes

Probability: Describe how data were drawn from a population.

Statistics: Infer about population via probabilistic reasoning.

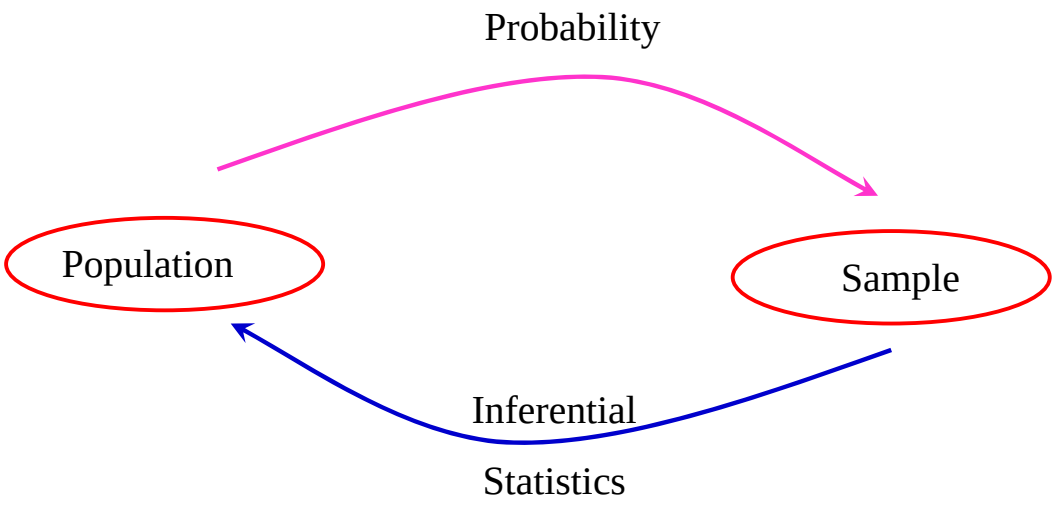


Figure 1.1: An illustration of probability versus statistics

1.2 Sampling

Population and sample:

- A **population** is a well-defined collection of objects.
- A **sample** is a subset of the population.
- A **simple random sample** is a sample chosen with equal chance.

Example 1.1 *Political poll.*

- to predict the outcome of an election; too expensive and slow to ask everyone; hence, ask some and hope they are representative
- random sample is used to reduce **selection** bias.

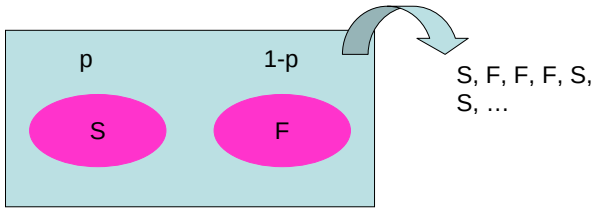


Figure 1.2: Schematic illustration of random sampling

```
> x=rbinom(100,1,0.5)   #draw 100 tickets from the box with P(S) = 0.5
> x                     #display the data
  [1] 0 0 0 1 1 1 1 0 1 0 0 0 0 1 1 1 1 0 1 0 0 1 1 1 1 1 1 0 1
 [30] 1 0 0 0 1 0 0 1 1 1 1 0 0 0 1 0 0 1 1 1 0 0 0 0 1 0 1 0 0
 [59] 1 0 0 1 0 0 0 0 0 0 1 1 0 0 1 0 1 1 1 0 1 1 1 1 0 1 0 1 1
 [88] 0 0 1 0 0 1 1 0 1 0 1 1 1
```

```
> mean(x)                  #compute the average = sample proportion
[1] 0.51
```

Statistical questions (on population aspects):

★ What is the population percentage of voters for the candidate?

(point estimation)

★ Can a candidate win an election? (Hypothesis testing)

$$H_0 : p \leq 0.5 \quad \longleftrightarrow \quad H_1 : p > 0.5.$$

★ In what interval does p lie with high confidence? (confidence interval).

Understand uncertainty of estimation (the size of poll errors).

Probabilistic question: (On the sample aspects)

■ If $p = 0.45$, what is the chance the sample proportion exceeds 51%?

■If there is 50% chance that daily stock returns are positive, what is the probability to get 5 consecutive negative daily returns?

Conceptual or Hypothetical population:

For example: the people who might benefit from a new drug to be introduced in the market.

1.3 Graphical Summaries

Purpose: Visualize data; extract salient features; examine overall data pattern.

Example 1.2 *Time series plot: used to examine serial dependence, time trend, and seasonal patterns. See Figure 1.3.*

Histogram: Histogram is the used tool to examine the distribution

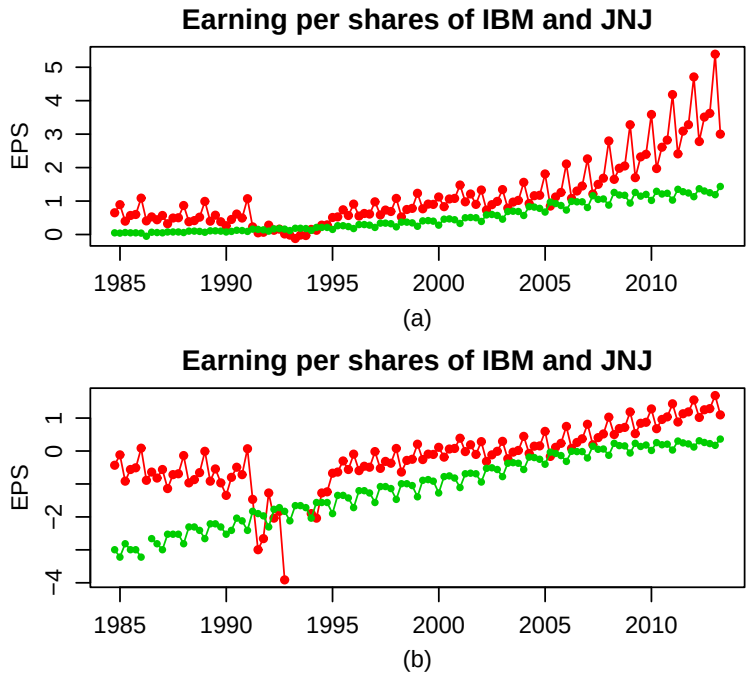


Figure 1.3: Quarterly earnings per share of IBM (red) and Johnson and Johnson (green) from 1984– 2013. Top panel: earning per share; bottom panel: logarithm of earings per share.

of data. To plot it, we need first to obtain a frequency table.

Example 1.3 *Frequency table and histogram*

The following are the salaries of 152 data scientists (in thousands per year) with master degree.

```
122 127 130 131 132 133 134 134 135 135 135 136 137 138 139 140 143 124
127 130 132 132 133 134 134 135 135 136 136 137 138 139 140 143 124 128
131 132 133 133 134 134 135 135 136 136 137 138 139 141 143 125 128 131
132 133 133 134 134 135 135 136 136 137 138 139 141 143 126 129 131 132
133 133 134 134 135 135 136 137 137 138 139 141 144 126 129 131 132 133
133 134 134 135 135 136 137 138 138 140 141 144 126 129 131 132 133 133
134 134 135 135 136 137 138 138 140 142 144 127 129 131 132 133 133 134
134 135 135 136 137 138 138 140 143 147 127 130 131 132 133 134 134 135
135 135 136 137 138 139 140 143
```

What are the distribution and features of the data?

To get a frequency table,

1. Choose non-overlapping intervals that cover the range of the data.

   (An extra decimal point is often used to define class boundaries to avoid ambiguities.)

2. Count the number of data, called **frequency**, in each interval.

3. Compute relative frequency = frequency / (sample size n).

For this example, $n = 152$. The minimum value is 122 and the maximum is 147. We will choose classes from 121.5 to 147.5.

Table 1.1: **Frequency distribution of heights**

Intervals	121.5 – 123.5	123.5 – 125.5	125.5 – 127.5	127.5 - 129.5	129.5 – 131.5	···	145.5-147.5
Frequency	1	3	7	6		···	1
Rel. Freq.	.0066	.020	.046			···	

Histograms are used to visualize the **distributions** of the data.

They repesent relative frequencies by **area**, **not** by height.

$$\text{Height of blocks} = \textbf{density} = \frac{\text{relative frequency of in an interval}}{\text{interval width}}.$$

For equal-width classes, frequencies can also be used as the heights.

This results in a **frequency histogram**.

```
> x = scan()                    #read data: hit return and cut and paste data
> hist(x);                      #plot of histogram with defaults
```

```
> hist(x,nclass=13)                    #histogram with No. class = 13
> hist(x,nclass=13,freq=F)             #histogram with relative frequency
```

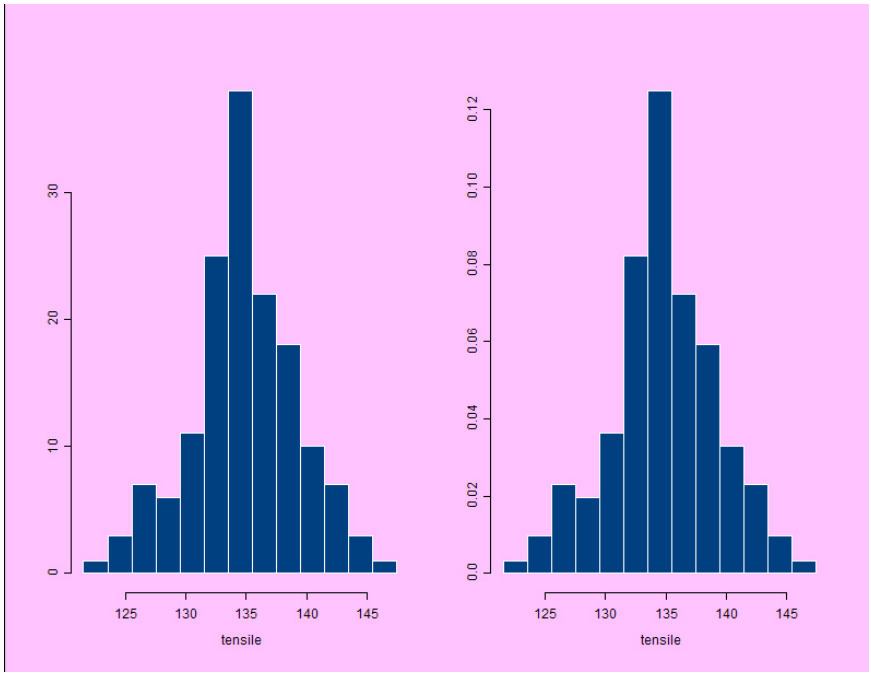


Figure 1.4: A frequency histogram and relative frequency histogram for the tensile data.

Remarks: Choosing class intervals involves trial-and-error.

♠ Too few intervals lose details of the data pattern.

♠ Too many may result in too chaotic histograms.

Example 1.4 *Histogram with different class widths*

The Adjust Gross Incomes (AGI) of US taxpayers in 2014 are summarized as follows (https://www.irs.gov/uac/soi-tax-stats-individual-income-tax-return-form-1040-statistics)

Table 1.2: **Distribution of Adjust Gross Income in 2014.**

Intervals	0–5	5–10	10–15	15–20	20–25	25–30	30–40	40–50	50–75	75–100	100–200	200– 500	500–1000	above
Rel. Freq	8.3	7.9	8.3	7.6	6.8	5.9	9.8	7.7	13.1	8.6	11.8	3.4	0.6	.2

```
> breaks = c(0,5,10,15,20,25,30,40,50,75,100, 200, 400)  #create break points
> AGI = c(rep(3,83), rep(8,79), rep(13,83), rep(18, 76), rep(23,68), rep(28,59),
  rep(35,98), rep(45,77), rep(60,131), rep(85,86), rep(150,118), rep(250, 34+6+2))
        #create data with correct frequency in each class interval
>  hist(AGI, breaks, col="blue")    #plot of histogram
> text(62.5,.0056, ".524%", col="red");              #add text to the plot
> text(87.5,.0037, ".344%", col="red");  text(150,.0015, ".118%", col="red");

>  dev.off()    #turn the device off -- close the file
```

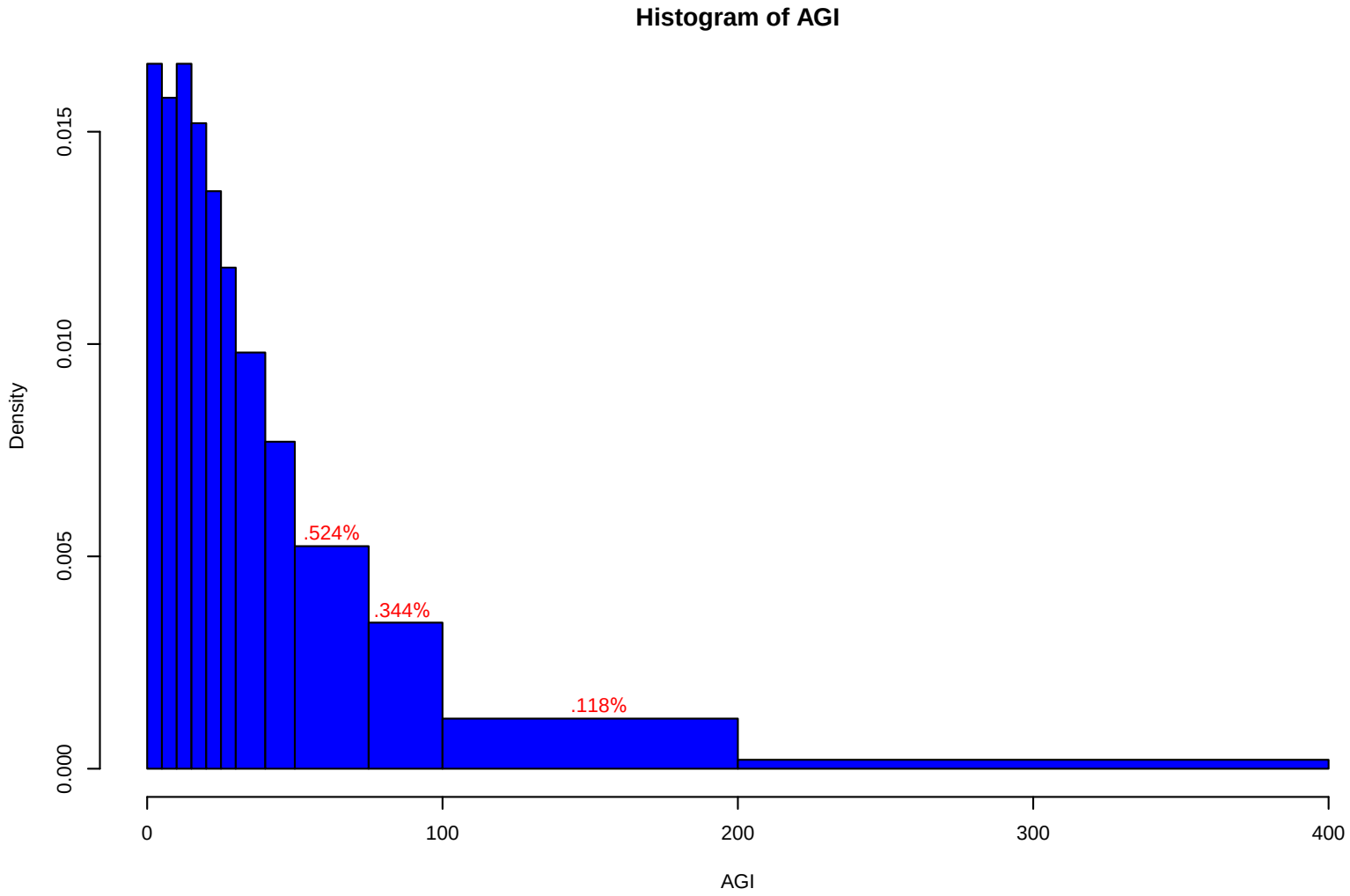


Figure 1.5: The distribution of AGI in 2014.

Based on the previous histogram. Let's answer some basic questions regarding the AGI in the US:

1. What is the shape of the income distribution?

2. Which interval is more crowded (dense)? (10, 15) or (50, 75) former

3. Which interval has more families, (10, 15) or (50, 75)? latter

4. Where is the mode (most dense one)? (0, 5) and (10, 15) with 8.3/5=1.66%

5. What is the density (height) of block (50, 75)? 13.1/25=0.524%

6. What percentage of families has income in (60, 120)?

$$(75 - 60) * .524\% + 25 * .344\% + (120 - 100) * .118\% = 18.82\%$$

Shapes of histograms. Here are a few commonly-seen shapes of histograms.

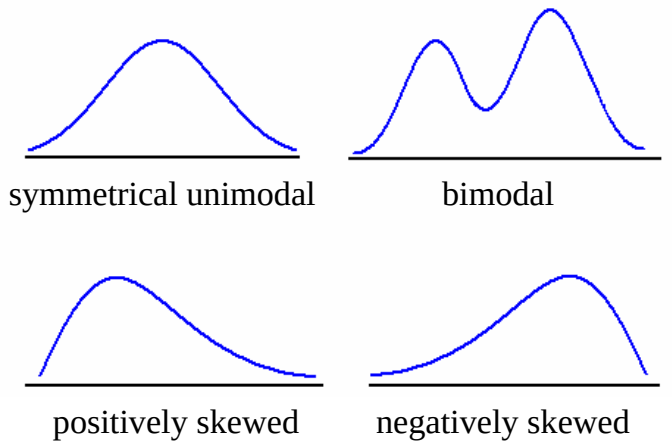


Figure 1.6: Commonly-seen shapes of histograms.

What is the shape of the income distribution? right/positive skewed

1.4 Summary Statistics: center of data

Observed data: $x_1, \cdots, x_n$, sample size $= n$.

Commonly-used measures of location (center):

1. **sample mean** (average): $\bar{x} = \sum_{i=1}^{n} x_i/n$.

2. **sample median**: $\tilde{x}$ = middle-value of the data.

   — the $\left(\frac{n+1}{2}\right)$-*th* largest value, when n is **odd**.

   — average of $\left(\frac{n}{2}\right)$-*th* and $\left(\frac{n}{2}+1\right)$-*th* largest value, when n is **even**.

Example 1.5 *Mean and median.*

The survival times (in days) of 6 patients after heart transplants in a hospital are 15, 3, 46, 623, 126, 64. Then, the average is

$$\bar{x} = \frac{15 + 3 + 46 + 623 + 126 + 64}{6} = 146.2\text{days}$$

and the median is

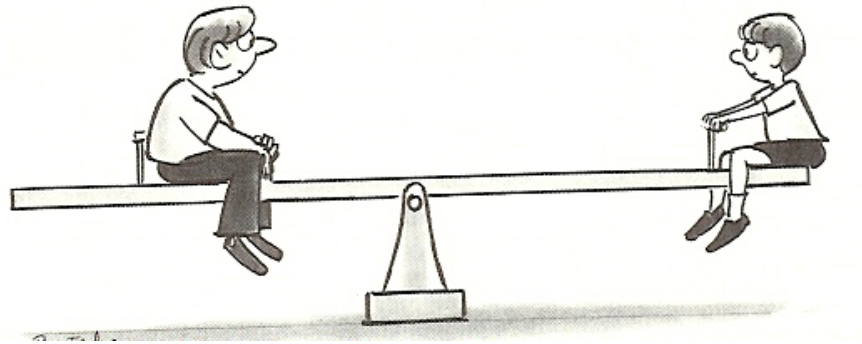
$$\tilde{x} = \frac{46 + 64}{2} = 55\text{days}.$$

Only 1 out of 6 patients survived longer than the average. Median is a better summary.

Fact 1: Median is robust against outliers (unusually large or small observations), while the average is not.

Relations with histogram:

- A histogram balances when supported at the average.



- The median divides the histogram so that half area is to its left and half to its right.

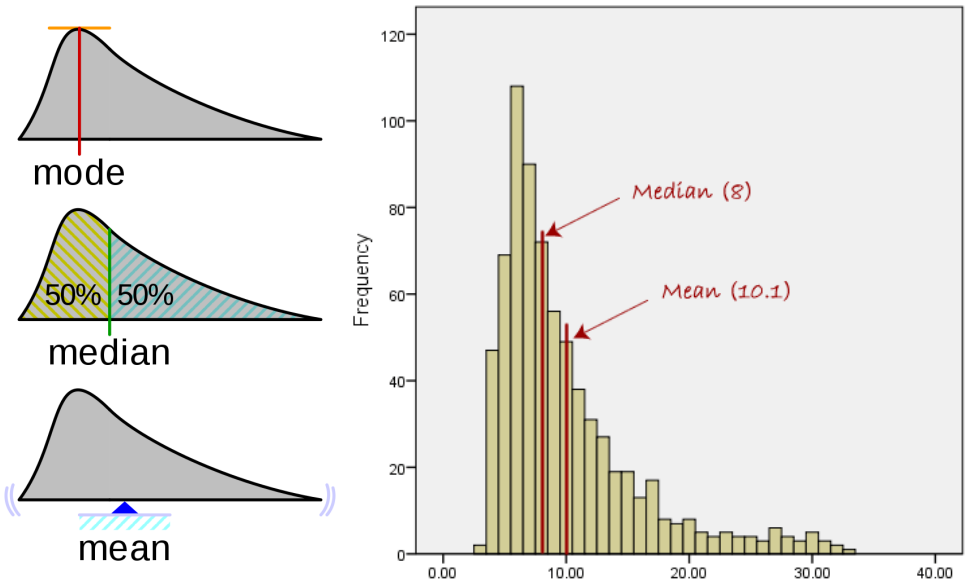


Figure 1.7: Reading average and median from a histogram.

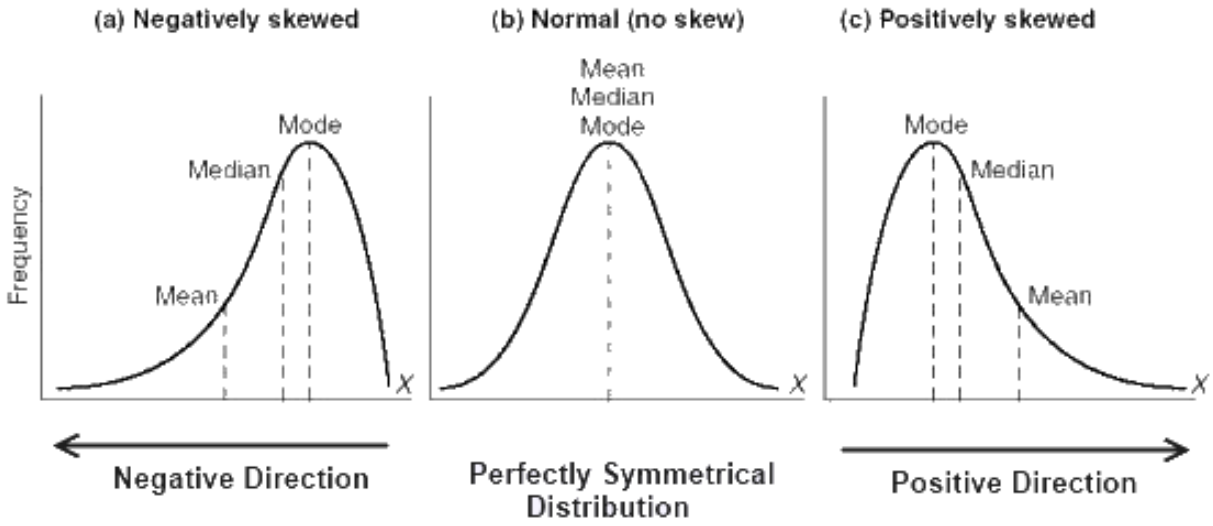


Figure 1.8: Relations between median and average for negatively skewed, positively skewed and symmetrical histograms.

Fact 2: For binary data, sample average = sample proportion.

Trim mean: 10% trimmed mean is the average after eliminating the smallest 10% and the largest 10% of the sample.

Percentile and quantiles: the 20th percentile = 0.2 quantile of the data = "20% largest value" in the sample

- For $p \in (0, 1)$, the p-quantile is a smallest value such that there are $\lceil np \rceil$ data points $\leq$ this value; if there is another value for which there are $\lceil n(1-p) \rceil$ data points $\geq$ that value, take their average.
- For any real number x, flooring $\lceil x \rceil$ means the smallest integer greater or equal than x. For instance $\lceil 6.1 \rceil = 6$ and $\lceil 3 \rceil = 3$.
- The median is just the 0.5 quantile (50th percentile) of the sample!

Example **1.6** *Percentiles and Quantiles*

The following data show the grades of a class of 30 students.

25 34 55 59 63 63 65 71 73 74 75 77 78 80 81 81 82 84 85 85 86 86 87 88 90 91 92 95 98 99

Find the 14th, 50th, 20th, 75th percentiles.

- 14th percentile = 63 ($.14 \times 30 = 4.2 \rightarrow 5$).
- 50th percentile = (81+81)/2 ($0.5 \times 30 = 15 \leftrightarrow 16$)
- 20th percentile = (63+65)/2 = 64 (two qualifications)
- 75th percentile = 87 (23^{rd} largest)

Terminology:

- **Lower (first) quartile** (lower fourth): $Q_1 = 25$th percentile.
- **Second quartile (median)**: $Q_2 = 50$th percentile.
- **Upper (third) quartile** (upper fourth): $Q_3 = 75$th percentile.

1.5 Summary Statistics: measures of variability

It is inadequate to summarize a histogram by using the average or median only, as we don't know the amount of variability.

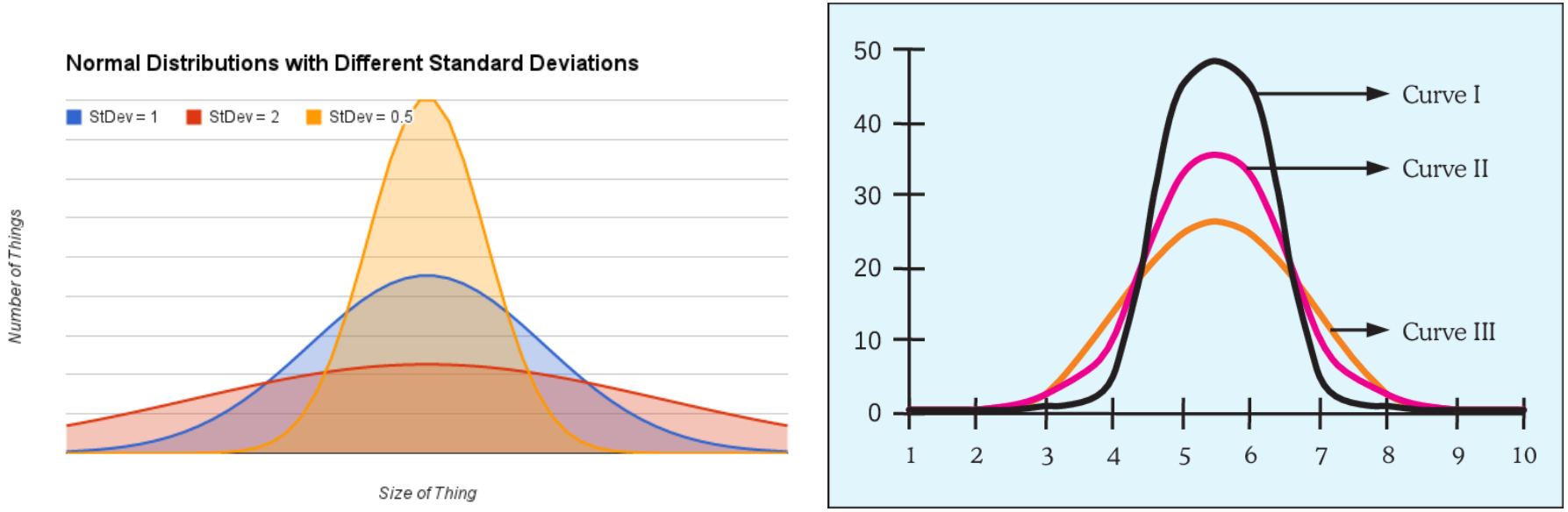


Figure 1.9: Histograms with the same center but different amount of spread.

Two measures of variability:

♣standard deviation, **SD**, associated with mean.

♣sample interquartile range, **IQR** $= Q_3 - Q_1$ (**fourth spread**), associated with median.

Deviation from $\bar{x}$: $(x_1 - \bar{x}), (x_2 - \bar{x}), \cdots, (x_n - \bar{x})$.

Square deviations: $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \cdots, (x_n - \bar{x})^2$.

Sample variance: $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$.

Sample SD s is its square-root.

Shortcut formula: $s = \sqrt{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)/(n-1)}$.

Box plot is a compact and powerful tool to summarize data distribution. It shows quartiles and outliers, as well as skewness.

Example 1.7 *Box plot*

Consider the following **25 pulse widths** from slow discharges in a cylindrical cavity made of polyethylene.

5.3, 8.2, 13.8, 74.1, 85.3, 88.0, 90.2, 91.5, 92.4, 92.9, 93.6, 94.3, 94.8, 94.9, 95.5, 95.8, 95.9, 96.6, 96.7, 98.1, 99.0, 101.4, 103.7, 106.0, 113.5

$$Q_1 = 90.2,\ Q_2 = 94.8,\ Q_3 = 96.7$$

$$\text{IQR} = Q_3 - Q_1 = 6.5$$

$$1.5 * \text{IQR} = 9.75$$

$$3 * \text{IQR} = 19.50$$

Outlier: Data further than **1.5 IQR** from the closest quartile. An **extreme outlier** is more than 3 IQR away from nearest quartiles.

Wiskers: To the last data that are not outliers from bothsides.

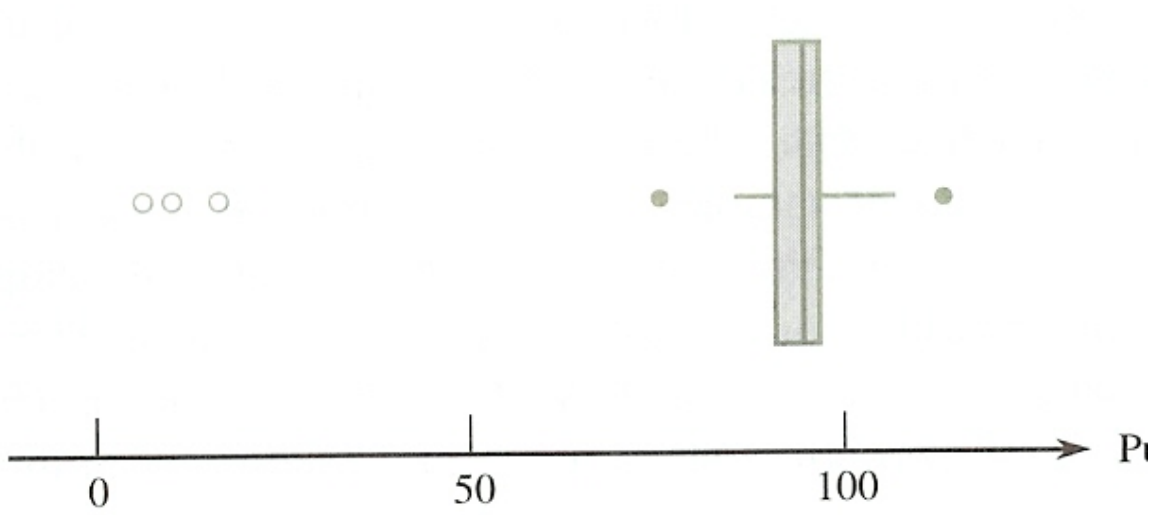


Figure 1.10: A box plot of the pulse width data showing mild (filled circles) and extreme outliers (blank circles).

Example 1.8 *Consider adjusted closing prices of SP500 index and IBM stock from Jan. 1, 2000 to September 8, 2016. We compare the distributions of their returns and the returns of SP500 before and after the 2008 financial crisis.*

■We first give a time series plot of the stock prices in Figure 1.11(a).

For a given price series $\{p_t\}$, its log-return r_t at time t is defined as $r_t = \log(p_t/p_{t-1})$.

```
>  getwd()                              #getting the working directory;  setwd(wd)
>  IBM = read.csv("IBM.csv",header=T) #read data
> IBM[1:3,]                                    #display firt 3 rows
        Date   Open   High    Low  Close  Volume Adj.Close
1 2016-09-08 160.55 161.21 158.76 159.00 3919300    159.00
2 2016-09-07 160.19 161.76 160.00 161.64 2867300    161.64
3 2016-09-06 159.88 160.86 159.11 160.35 2994100    160.35

>  SP500 = read.csv("SP500.csv",header=T)    #read data

  ############Getting Adjusted Closing Price and Returns ##################
>  pSP500 = SP500[,7]                          #take adj close price column
>  pSP500 = rev(pSP500)                        #reverse the time order
```

```
>  rSP500 = diff(log(pSP500))*100          #percentage of returns
>  pIBM = rev(IBM[,7])                     #Closing prices of IBM
>  rIBM    = diff(log(pIBM))*100           #percentage of log-returns
>  Dates = as.vector(IBM[,1])              #dates of Data
>  Dates = strptime(Dates, "%Y-%m-%d")     #convert to  POSIXlt (a date class)
>  Dates = rev(Dates)                      #time from past to future

>  pdf("Boxplot.pdf", width=6, height=3, pointsize=8)   #print to pdf file
>  par(mfrow = c(2,2), mar=c(4,3,1.5,1)+0.1, cex=0.8)   #2x2 subplots

>  plot (Dates,  pSP500,  ylim=c(500, 2500), col=4, type="l", xlab="(a)", ylab="")
>  lines(Dates,  8*pIBM,  col=2)
>  title("Prices of SP500 and 8*IBM")
```

■We then give the boxplots of returns of SP500 during different periods and that of IBM stock returns. Clearly, IBM stock returns are riskier (larger volatility) than the SP500 stock index. In addition, the volatility increases 3 times during the 2008 financial crisis than that before the crisis.

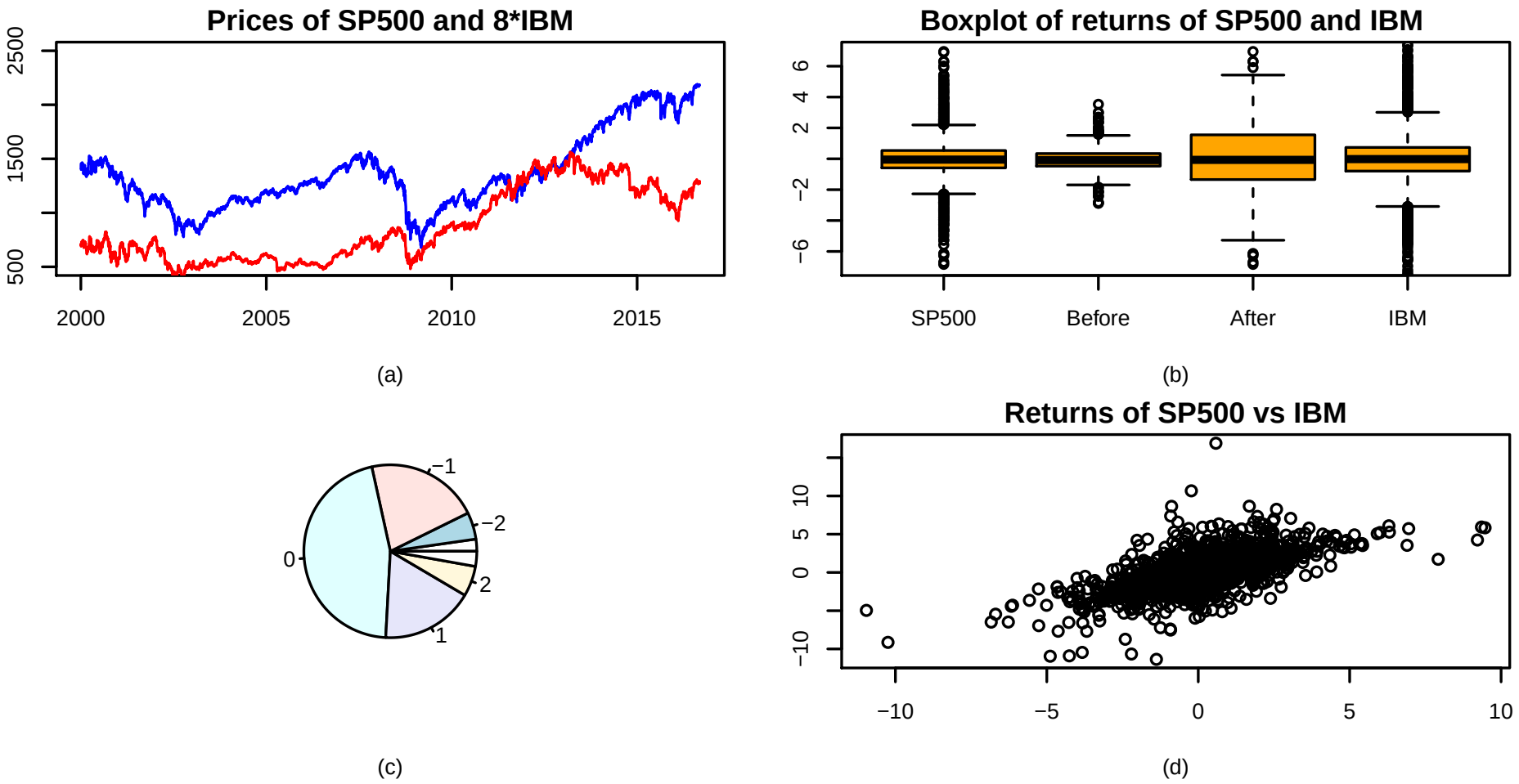


Figure 1.11: Distributions of the returns of SP500 and IBM stocks from January 1, 2000 to September 8, 2016. Included is also the returns of SP500 before the 2008 financial crisis (01/03/06 – 07/31/07) and after the 2008 financial crisis (07/01/08–6/30/09)

```
> rSP500b = rSP500[1059:2010]                    #returns in 01/03/06--12/31/07
> rSP500a = rSP500[2136:2387]                    #returns from 7/1/08-- 6/30/09
> boxplot(list(rSP500, rSP500b, rSP500a, rIBM), names=c("SP500", "Before",
        "After", "IBM"), col="Orange", ylim=c(-7,7), xlab="(b)")
```

```
>  title("Boxplot of returns of SP500 and IBM")
```

More quantitative calculation can be done as follows:

```
> summary(rSP500);summary(rSP500b);summary(rSP500a)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-9.469512 -0.536993  0.052386  0.009644  0.588574 10.957197
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.53427 -0.36632  0.07644  0.02925  0.45627  2.87896
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-9.46951 -1.53979  0.06895 -0.13113  1.34223 10.95720
```

■The distributions of SP500 returns can also be summarized by the pie-chart. For example, the category "-1" means the returns between -1.5% to -0.5%. This is also an effort to see if distribution is negative skew (market reacts more to negative news).

```
>  freq = table(round(rSP500))       #round to nearest integer
>  freq = c(sum(freq[1:7]), freq[8:12], sum(freq[13:19]))
       #consolidating small cells
>  pie(freq, xlab="(c)")         #pie chart
```

```
>  plot(rSP500, rIBM,xlab="(d)")
>  title("Returns of SP500 vs IBM")
>  dev.off()                                  #turn device off, end printing
```

■Finally, we examine the relationship between SP500 returns and IBM returns via scatter plot. Clearly the returns of IBM depends on those of SP500 (CAPM).

Example 1.8 (Continued). Compute SDs of returns (volatility)

```
> var(rSP500); sd(rSP500)   #compute variance and SD of returns of SP500
[1] 1.570898
[1] 1.253355
> c(sd(rSP500b), sd(rSP500a),  sd(rIBM))  #compute SDs and put them as a vector
[1] 0.8410089 2.8642774 1.6705536
```