STATISTICS for Lawyers & Social Sciences

CHAPTERS 2&3

SUMMARIZING DATA: LISTING & GROUPING

Dot Diagram

2.3 An automobile dealer complains to a manufacturer's representative that there are numerous defects in the paint of 25 automobiles received in a shipment. The manufacturer's representative checked each car and found the following defects:

1, 3, 0, 2, 1, 2, 2, 1, 2, 1, 1, 1, 2, 2, 3,

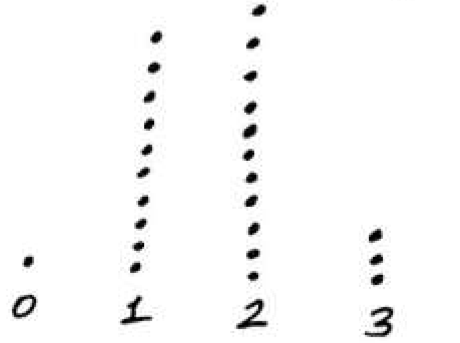
1, 1, 2, 1, 2, 2, 1, 3, 2, and 2

Construct a dot diagram that shows the number of automobiles that have 0, 1, 2, and 3 paint defects.

2.3) (i) Count the data

Defects	Tally	Frequency
0	/	1
1	𝍸 𝍸	10
2	𝍸 𝍸 /	11
3	///	3

(ii) Construct dot diagram



Pareto Diagram

2.5 Forty-two people attending a conference were offered their choice of nonalcoholic beverages and selected the following:

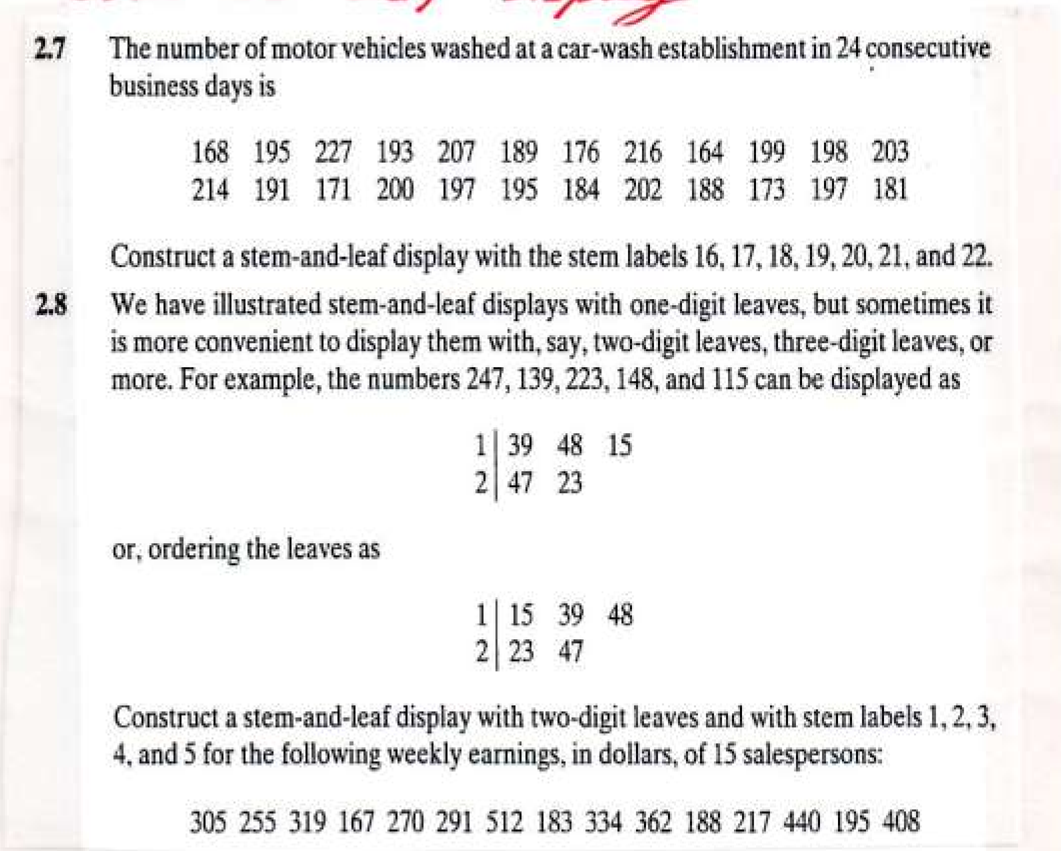
Beverage	*Number*
Coffee	15
Soda pop	10
Juice	7
Tea	6
None	4
Total	42

Construct a Pareto diagram like that of Figure 2.2 with the largest value plotted on top and the remaining values plotted below in descending order.

Coffee	● ● ● ● ● ● ● ● ● ● ● ● ● ● ●
Sooda pop	● ● ● ● ● ● ● ● ● ●
Juice	● ● ● ● ● ● ●
Tea	● ● ● ● ● ●
None	● ● ● ●

Note that, the choices are already displayed in descending order. If it weren't, we would order them.

Stem and leaf Display



2.7 The number of motor vehicles washed at a car-wash establishment in 24 consecutive business days is

168	195	227	193	207	189	176	216	164	199	198	203
214	191	171	200	197	195	184	202	188	173	197	181

Construct a stem-and-leaf display with the stem labels 16, 17, 18, 19, 20, 21, and 22.

2.8 We have illustrated stem-and-leaf displays with one-digit leaves, but sometimes it is more convenient to display them with, say, two-digit leaves, three-digit leaves, or more. For example, the numbers 247, 139, 223, 148, and 115 can be displayed as

```
1| 39 48 15
2| 47 23
```

or, ordering the leaves as

```
1| 15 39 48
2| 23 47
```

Construct a stem-and-leaf display with two-digit leaves and with stem labels 1, 2, 3, 4, and 5 for the following weekly earnings, in dollars, of 15 salespersons:

305 255 319 167 270 291 512 183 334 362 188 217 440 195 408

2.7) (i) Label steams and leaves

```
16| 8 4
17| 6 1 3
18| 9 4 8 1
19| 5 3 9 8 1 7 5 7
20| 7 3 0 2
21| 6 4
22| 7
```

(ii) Make each row in ascending order

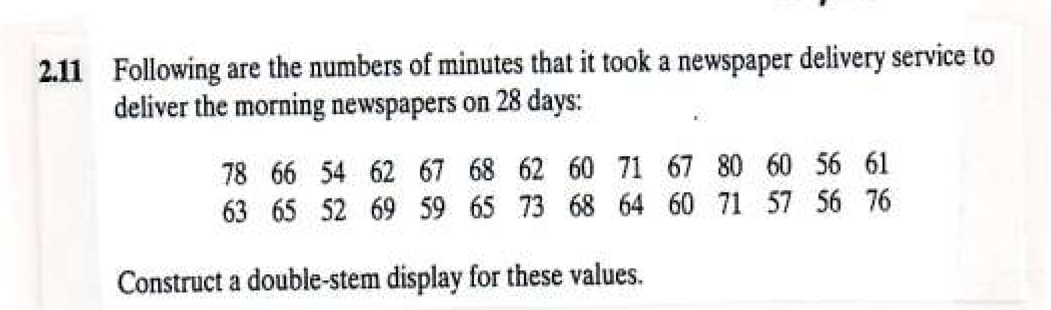
```
16| 4 8
17| 1 3 6
18| 1 4 8 9
19| 1 3 5 5 7 7 8 9
20| 0 2 3 7
21| 4 6
22| 7
```

2.8) (i)

1	67 83 88 95
2	55 70 91 17
3	05 19 34 62
4	40 08
5	12

(ii)

1	67 83 88 95
2	17 55 70 91
3	05 19 34 62
4	08 40
5	12



2.11 Following are the numbers of minutes that it took a newspaper delivery service to deliver the morning newspapers on 28 days:

78 66 54 62 67 68 62 60 71 67 80 60 56 61
63 65 52 69 59 65 73 68 64 60 71 57 56 76

Construct a double-stem display for these values.

2.11) we double the number of stem positions by cutting in half the interval covered by each ten digits. Namely, for example 5* will include 50, 51, 52, 53 and 54 wheras 5. will include the remaining 55, 56, 57, 58 and 59.

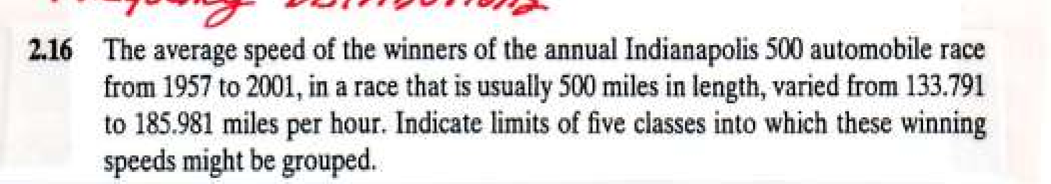
(i)

5*	4 2
5.	6 9 7 6
6*	2 2 0 0 1 3 4 0
6.	6 7 8 7 5 9 5 8
7*	1 3 1
7.	8 6
8*	0

(ii)

5*	2 4
5.	6 6 7 9
6*	0 0 0 1 2 2 3 4
6.	5 5 6 7 7 8 8 9
7*	1 1 3
7.	6 8
8*	0

Frequency Distributions



2.16 The average speed of the winners of the annual Indianapolis 500 automobile race from 1957 to 2001, in a race that is usually 500 miles in length, varied from 133.791 to 185.981 miles per hour. Indicate limits of five classes into which these winning speeds might be grouped.

2.16) $\text{Min}(X_i) = 133{,}791$

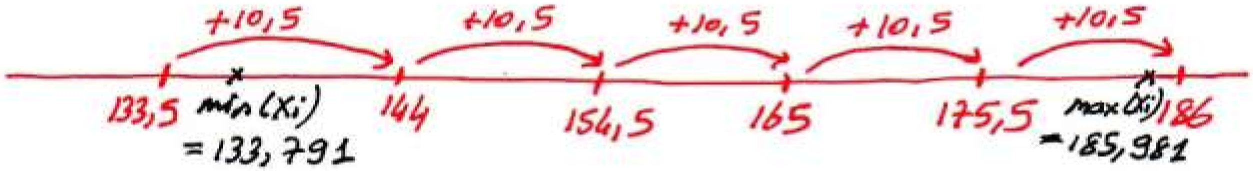
$\text{Max}(X_i) = 185{,}981$

$\text{Range} = \text{max}(X_i) - \text{Min}(X_i) = 185{,}981 - 133{,}791 = 52{,}19$

k = Number of classes = 5

To cover all the data values, we Round UP range and start with a number less than Min(X_i). The class length is found by;

$$\text{Class length} = \frac{\text{Range (UP)}}{k} = \frac{52,5}{5} = 10,5$$



Note that the solution is NOT exact. The important thing is, "Make sure that you cover all the data!" For example, one may use class length = $\frac{53}{5}$ = 10,6 and start with 133, that's also correct.

Finally, the classes are;

Average Speed	Tally	Frequency
$133,5 \leq X_i < 144,0$		
$144,0 \leq X_i < 154,5$		
$154,5 \leq X_i < 165,0$		
$165,0 \leq X_i < 175,5$		
$175,5 \leq X_i < 186$		

* The next step of frequency distribution is, we count the data by "Tally" and write the frequency of each class. That's what we will do in the next example.

2.30 Following are the numbers of years of service of all 55 justices of the United States Supreme Court who were appointed from the years 1900 to 2002. The data includes the nine justices who were members of the court in 2002 and all other appointees since 1900.

29	5	23	16	16
19	8	36	16	15
3	15	9	33	27
4	16	5	5	21
5	7	1	23	16
26	16	12	31	16
5	11	6	3	14
10	15	13	4	12
10	6	7	24	11
26	34	18	17	9
22	19	7	24	8

(a) Group the years of service into a table having the classes 0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, and 35–39.

(b) Convert the distribution obtained in part (a) into a cumulative "or more" distribution.

2.31 Convert the distribution obtained in part (a) of Exercise 2.30 into a

(a) percentage distribution;

(b) cumulative "less than or equal to" percentage distribution;

(c) cumulative "or more" distribution.

2.31)

2.30) Class	Tally	Frequency	(a) Parantage	(b) Cumm. or less	(c) Cumm. or more
0-4	𝍸	5	$\left(\frac{5}{55}\right)\cdot 100 = 9,09\%$	5	55
5-9	𝍸 𝍸 ||||	14	$\frac{14}{55}\cdot 100 = 25,45\%$	19	50
10-14	𝍸 |||	8	14,55%	27	36
15-19	𝍸 𝍸 ||||	14	25,45%	41	28
20-24	𝍸 |	6	10,91%	47	14
25-29	||||	4	7,27%	51	8
30-34	|||	3	5,45%	54	4
35-39	/	+ 1	$\frac{1}{55}\cdot 100 =$ + 1,83%	55 = n	1
		n=55	100%		

* Cum. or less distribution is obtained by summing frequencies downward (starting with 0-4) and Cum. or more distribution is obtained likewise but summing upward (starting with 35-39)

2.35 In the sport of baseball the winner of the World Series is the first team to win four of the seven or fewer games that are played. The losing team can therefore win 0, 1, 2, or 3 games. For the last forty of the World Series played, ending with the year 2000, the losing team won only the following number of games:

3	1	3	0	3	3	0	3	3	1
1	3	3	3	1	3	0	2	2	3
2	2	3	1	1	3	3	3	1	0
0	3	2	2	2	2	3	0	0	1

Construct a distribution showing how many games, 0, 1, 2, or 3, were won by the defeated team of the World Series.

2.36 A survey made at a resort city showed that 50 tourists arrived by the following means of transportation:

car, train, plane, plane, plane, bus, train, car, car, car, plane, car, plane, train, car, car, bus, car, plane, plane, train, train, plane, plane, car, car, train, car, car, plane, car, car, plane, bus, plane, bus, car, plane, car, car, train, train, car, plane, bus, plane, car, car, train, bus

Construct a categorical distribution showing the frequencies corresponding to the different means of transportation.

2.35)

Games	Tally	Frequency
0	卌 ||	7
1	卌 |||	8
2	卌 |||	8
3	卌 卌 卌 ||	17

2.36)

Transportation	Tally	Frequency
Car	卌 卌 卌 卌	20
Train	卌 ||||	9
Plane	卌 卌 卌	15
Bus	卌 |	6

Histogram, Frequency Polygon & O-give

2.40 Following is the distribution of salaries of the fifty governors of the United States for the year 2001.

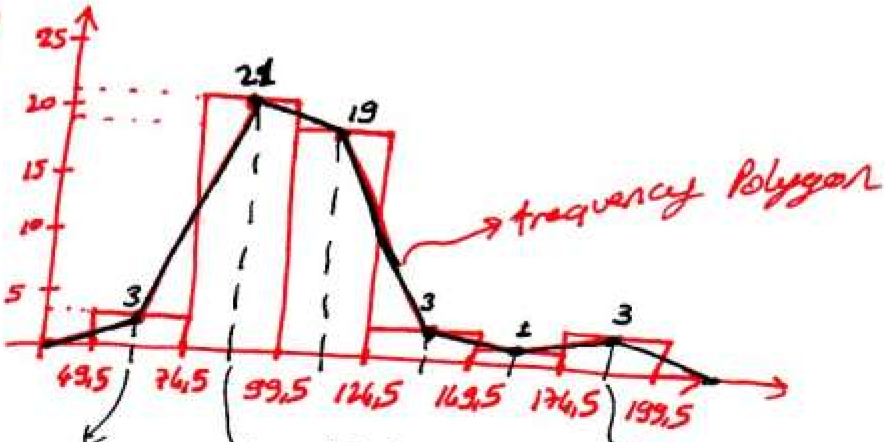
Salaries (in thousands of dollars)	*Frequency*
50– 74	3
75– 99	21
100–124	19
125–149	3
150–174	1
175–199	3
Total	50

(a) Draw a histogram of this distribution.

(b) Draw a frequency polygon of this distribution. Class marks for the frequency polygon are 62, 87, . . . , 187.

2.41 Convert the distribution of Exercise 2.40 into a cumulative "less than" distribution and draw an ogive.

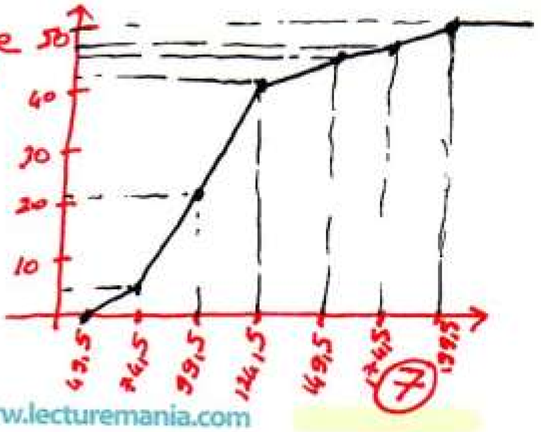
2.40)



$\frac{50+74}{2}=62$ $\quad 87=\frac{75+99}{2}$ - - - $\quad \frac{175+199}{2}=187$

2.41)

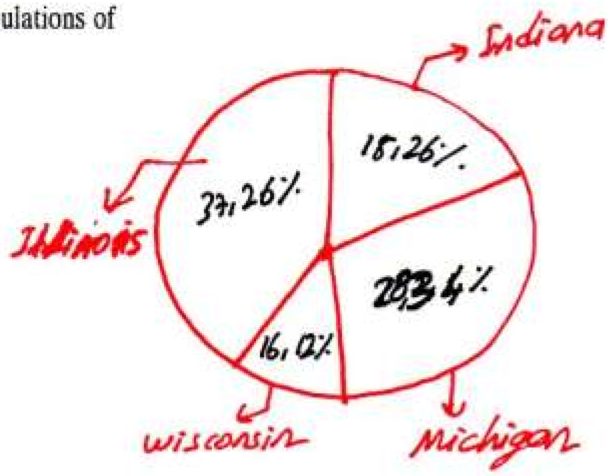
Salary	Frequency	Cumulative "less than"
50-74	3	3
75-99	21	24
100-124	19	43
125-149	3	46
150-174	1	47
175-199	3	50



Pie Chart

2.44 Data from the U.S. Bureau of the Census shows that the resident populations of the East North Central states are as follows:

State	*Population (in thousands)*
Indiana	6,080
Illinois	12,419
Michigan	9,438
Wisconsin	5,364
Total	33,301



Construct a pie chart of this categorical distribution.

2.44)

State	Population	Percentage	Angle
Indiana	6080	$\frac{6080}{33301} \times 100 = 18,26\%$	$\frac{18,26}{100} \times 360 = 65,7°$
Illinois	12419	$\frac{12419}{33301} \times 100 = 37,26\%$	$\frac{37,26}{100} \times 360 = 134,1°$
Michigan	9438	ı 28,34%	ı 102,1°
Wisconsin	5364	ı 16,12%	ı 58,1°
TOTAL	33301	100%	360°

SUMMARIZING DATA: STATISTICAL DESCRIPTIONS

Measures of Location

(i) Mean (or average or expected value)

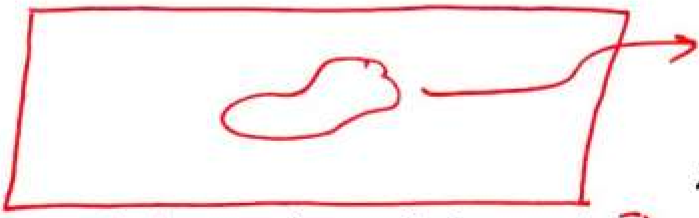
(ii) Median → the number in the middle of the data

(iii) Mode (or Modal) → most frequent number or category

(iv) Quartiles: Q_1, Q_2, Q_3 → numbers that divide the data into 4; 25th percentile: Q_1, 50th percentile: Q_2 (= median), 75th percentile: Q_3

(v) Box and whisker plot: A graph showing $\min(X_i) < Q_1 < Q_2 < Q_3 < \max(X_i)$

Population versus Sample



Random Sample (i.e. $n=50$)
Randomly selected units from the population.
Using their values of X_i, we'll make "Inference" for population parameters

Population: The whole units that a study is made

Ex: Bilkent University Students (i.e. $N=12000$)

Random Variable: X

let's say, X: Weekly food expenditure of a student at Bilkent.

	Population Parameters (Unknown Constants)	Sample Statistics (Known Variables)
Mean	$\mu = \frac{\Sigma X_i}{N}$	$\overline{X} = \frac{\Sigma X_i}{n}$
Variance	$\sigma^2 = \frac{\Sigma (X_i - \mu)^2}{N}$	$s^2 = \frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n-1}$
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

* The idea is that, if we had the oportunity to look all the X_i values of population, we would reach constant values of Population Parameters. However, it is practically impossible; then, population parameters are "Unknown Constants"

To estimate or test population parameters, we take a random sample and calculate Sample Statistics. They are "Known Variables" because we don't know if they truly reflect the population. Another samples would yield different values.

(i) Mean $\bar{X} = \frac{\sum X_i}{n} = \frac{\text{Sum of the data values}}{\text{Sample Size}}$

(ii) Median Follow these steps to find the median;

(i) Rank the data

(ii) Find "Position of the median" $= \frac{n+1}{2}$

(iii) Find median, the number in the middle.

(iii) Mode The most frequent number or category. Note that, mean and median always exist and unique. However, mode may NOT exist, or may be more than one.

(iv) Quartiles Similar to median, we find quartiles by,

(i) Rank the data

(ii) Position of quartiles; $Q_1 \rightarrow 0{,}25 \cdot (n+1)$

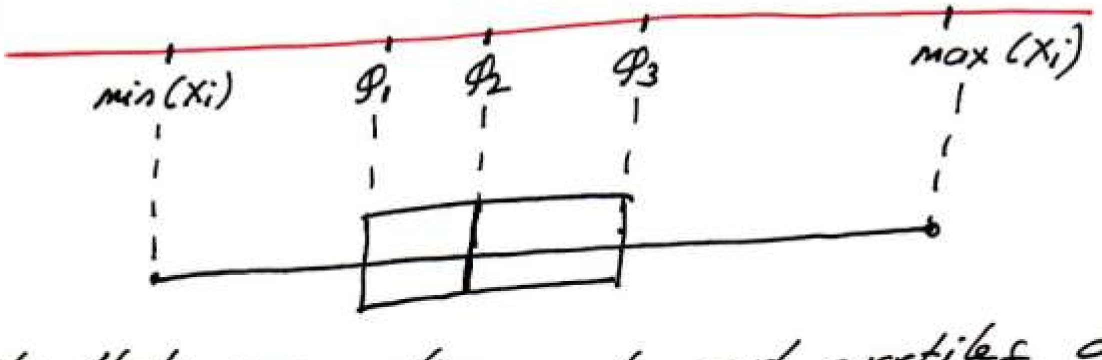
$Q_2 = \text{Median} \rightarrow 0{,}50 \cdot (n+1)$

$Q_3 \rightarrow 0{,}75 \cdot (n+1)$

(iii) Find Quartiles

(v) Box and Whisker plot: Gives an idea about the "shape" of the population.

Ex;



Note that; mean, median, mode and quartiles are "Location Statistics" because they measure "Where the data is"

3.8 According to the Bureau of Labor Statistics, the average pay (in thousands of dollars) paid in the top 15 metropolitan areas in the United States in 1999 was

51.4	50.4	45.6	44.8	44.6
42.9	41.6	40.6	40.1	40.1
39.8	39.5	38.6	38.5	38.3

Find their mean annual pay in thousands of dollars.

3.9 The records of 15 persons convicted of various crimes showed that, respectively, 4, 3, 0, 0, 2, 4, 4, 3, 1, 0, 2, 0, 2, 1, and 4 of their grandparents were foreign born. Find the mean and discuss whether it can be used to support the contention that the "average criminal" has two foreign-born grandparents.

3.8) $$\bar{X} = \frac{\sum X_i}{n} = \frac{51,4+50,4+45,6+ \cdots + 38,5+38,3}{15} = \frac{636,8}{15} = 42,45$$

3,9) $$\bar{X} = \frac{4+3+0+\cdots+1+4}{15} = \frac{30}{15} = 2$$

Consider the frequency distribution of the data

Value	Tally	Frequency (Weight: w_i)
0	////	4
1	//	2
2	///	3
3	//	2
4	////	4

If we consider frequencies as weight, we find the "Weighted Mean", which is the same result;

$$\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i} = \frac{4\cdot0+2\cdot1+3\cdot2+2\cdot3+4\cdot4}{4+2+3+2+4} = \frac{30}{15} = 2$$

3.19 Previously, we stated that in 2000 the production of milk, in thousands of pounds per year per cow, was 19.0 in Michigan, 21.2 in Wisconsin, and 17.8 in Minnesota. Given that there were 300,000 cows in Michigan, 1,340,000 in Wisconsin, and 534,000 cows in Minnesota, what is the average annual milk production for the three states combined?

3.20 An instructor counts the final examination in a course three times as much as each of the three one-hour examinations. What is the average grade of a student who received grades of 75, 77, and 58 on the three one-hour examinations and 82 on the final examination?

3.19)

	Milk Production	Weight
Michigan	19,0	300 000
Wisconsin	21,2	1 340 000
Minnestota	17,8	534 000

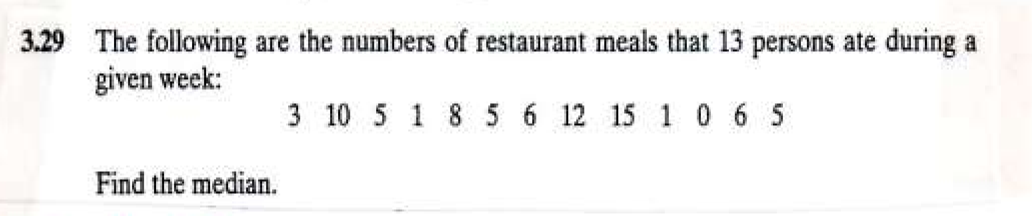
$$\overline{X}_w = \frac{\sum w_i \cdot X_i}{\sum w_i} = \frac{300\,000 \cdot 19{,}0 + 1340000 \cdot 21{,}2 + 534\,000 \cdot 17{,}8}{300000 + 1340000 + 534000}$$

$$= \frac{43632200}{2\,174\,000} = 20{,}07$$

3.20)

Exam	1	2	3	Final
Grade	75	77	58	82
Weight	1	1	1	3

$$\overline{X}_w = \frac{1.75 + 1.77 + 1.58 + 3.82}{1+1+1+3} = 76$$


3.29 The following are the numbers of restaurant meals that 13 persons ate during a given week:

3 10 5 1 8 5 6 12 15 1 0 6 5

Find the median.

3.29) (i) 0 1 1 3 5 5 5 6 6 8 10 12 15

(ii) Position of Median $= \frac{13+1}{2} = \frac{14}{2} = 7$

(iii) Median: $\tilde{X} = 5 \rightarrow$ 7th number

Let, we have an additional value 18 in the data. Then;

(i) 0 1 1 3 5 5 5 6 6 8 10 12 15 18 ; $n=14$

(ii) Position of Median $= \frac{14+1}{2} = \frac{15}{2} = 7{,}5 \rightarrow$ Mean of 7th and 8th numbers

(iii) Median $= \frac{5+6}{2} = 5{,}5$

3.34 The automatic telephone answering machine of a large office transferred the following numbers of incoming calls to various personnel during 25 business days.

147	123	135	154	146
182	132	128	144	177
150	142	169	137	193
129	198	156	154	140
138	107	146	125	117

Find the median number of these telephone calls

(a) directly;

(b) by first constructing a stem-and-leaf display.

3.34) (i)

10	7
11	7
12	3 8 9 5
13	5 2 7 8
14	7 6 4 2 0 6
15	4 0 6 4
16	9
17	7
18	2
19	3 8

$\xRightarrow{\text{Rank}}$

10	7
11	7
12	3 5 8 9
13	2 5 7 8
14	0 2 4 6 6 7
15	0 4 4 6
16	9
17	7
18	2
19	3 8

(ii) Position of median $= \frac{25+1}{2} = 13$ (iii) $\tilde{X} = 144$

3.46 Selected major bodies of water in the Gulf Coast are Tampa Bay, Galveston Bay, Atchafalaya Bay, Matagorda Bay, Lake Borgne, Mobile Bay, Breton Sound, Florida Bay, Lake Pontchartrain, Laguna Madre, Mississippi Sound; and their areas in square miles are 212, 236, 245, 253, 271, 310, 511, 616, 631, 733, and 813.

(a) Calculate the values of Q_1, Q_3, and the median.

(b) Construct a box-and-whisker plot for the areas of the water bodies.

3.46) a) (i) 212 236 245 253 271 310 511 616 631 733 813

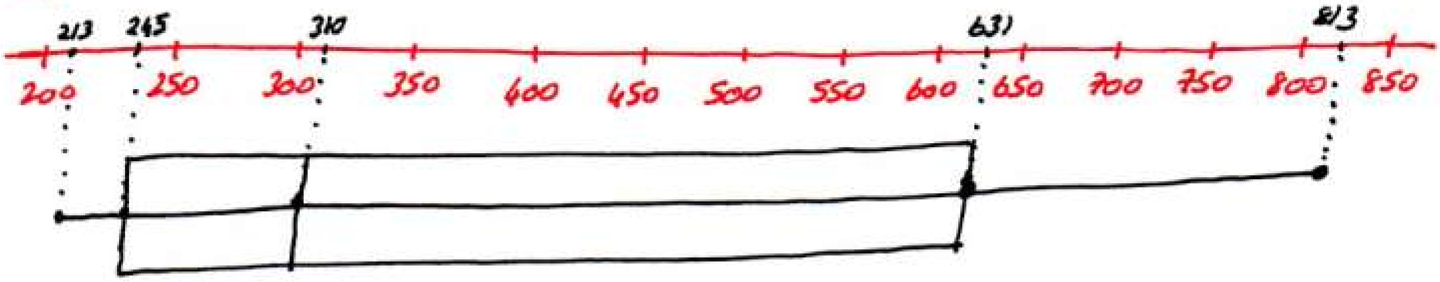
(ii) $n = 11$; Quartile positions: $Q_1 \to 0{,}25.(n+1) = 0{,}25.12 = 3$

$\text{median} = Q_2 \to 0{,}50.(n+1) = 0{,}50.12 = 6$

$Q_3 \to 0{,}75.(n+1) = 0{,}75.12 = 9$

(iii) $Q_1 = 245$; $Q_2 = 310$; $Q_3 = 631$

b)



c) Let, we have an additional data value 200 and $n=12$.

Then; (i) 200 212 236 245 253 271 310 511 616 631 733 813

(ii) Quartile Positions; $Q_1 \rightarrow 0{,}25.(n+1) = 0{,}25.13 = 3{,}25 \cong 3$

$Q_2 \rightarrow 0{,}50.(n+1) = 0{,}50.13 = 6{,}5$ → mean of 6th and 7th

$Q_3 \rightarrow 0{,}75.(n+1) = 0{,}75.13 = 9{,}75 \cong 10$

(iii) $Q_1 = 236$; $Q_2 = \frac{271+310}{2} = 290{,}5$; $Q_3 = 631$

Mid Quartile and Midrange

$$\text{Midquartile} = \frac{Q_1 + Q_3}{2} \; ; \quad \text{Midrange} = \frac{\text{Min}(X_i) + \text{Max}(X_i)}{2}$$

are other measures of central location.

For Exercise 3.46 (original data), let

d) Find the midquartile and the midrange

Answer: $\text{Midquartile} = \frac{245+631}{2} = 438$

$\text{Midrange} = \frac{213+813}{2} = 513$

3.51 The following data provided by the Current Population Survey is the percent of total unemployment (as a percent of the civilian labor force) for the year 2000. The data refers to the eastern states and the District of Columbia that are specified in Exercise 3.50.

4 3 3 3 4 2 5 4 4 4 4 6 2 6 4 4 4 4

(a) Find the mode;

(b) find the median;

(c) find the mean.

3.51)

Unemployement	Tally	frequency	Cum. Frequency
2	//	2	2
3	///	3	5
4	𝍸 𝍸	10	15
5	/	1	16
6	//	+ 2	18
		n = 18	

a) Mode = 4

b) Pos. of median $= \frac{n+1}{2} = \frac{18+1}{2} = 9,5$

$5 < 9 < 10 \leq 15$ so, 9th and 10th numbers are both 4

Median $= \frac{4+4}{2} = 4$

c) $\bar{X} = \frac{2.2+3.3+10.4+1.5+2.6}{18} \cong 3,89$

3.56 Asked for their favorite color, 50 persons gave the following responses:

red	blue	blue	green	yellow
blue	brown	red	blue	red
red	green	white	blue	red
green	blue	red	green	green
purple	white	yellow	blue	blue
blue	red	red	brown	orange
white	green	blue	blue	blue
red	blue	red	yellow	green
yellow	blue	blue	orange	red
green	white	purple	blue	red

What was their modal choice?

3.56)

Color	Tally	Frequency
Red	𝍸 𝍸 //	12
Blue	𝍸 𝍸 𝍸 /	16
Green	𝍸 ///	8
Yellow	////	4
Orange	//	2
White	////	4
Purple	//	2
Brown	//	2

Modal Choice is Blue.

Measures of Variation

Range $R = Max(X_i) - Min(X_i)$

Interquartile Range $IQR = Q_3 - Q_1$

Variance & Standard Deviation

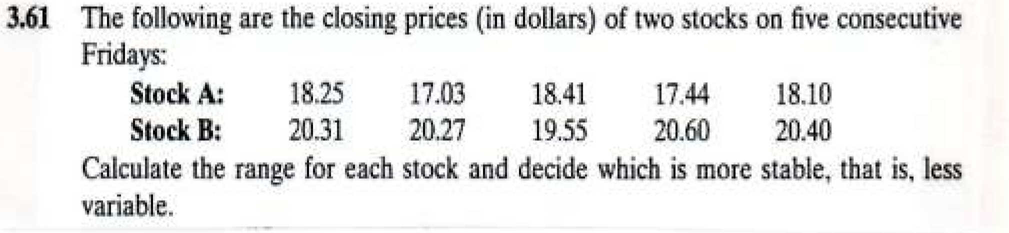
* Population: $\underbrace{\sigma^2}_{\text{Variance}} = \frac{\sum(x_i-\mu)^2}{N}$; $\underbrace{\sigma}_{\text{Standard Deviation}} = \sqrt{\sigma^2}$

* Sample ; $s^2 = \underbrace{\frac{\sum(x_i-\bar{X})^2}{n-1}}_{\text{Definition Formula}} = \underbrace{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}}_{\text{Calculation Formula}}$; $s = \sqrt{s^2}$

Coefficient of Variation: $CoV = \frac{s}{\bar{X}} \cdot 100$

Range, IQR, Variance and Standard Deviation are "Dispersion Statistics" because they measure how wide the data is, independent from their location

When we compare two data whose means are NOT equal, standard deviation does not make sense. To compare dispersion of such two data, we compare their CoV's.



3.61 The following are the closing prices (in dollars) of two stocks on five consecutive Fridays:

Stock A:	18.25	17.03	18.41	17.44	18.10
Stock B:	20.31	20.27	19.55	20.60	20.40

Calculate the range for each stock and decide which is more stable, that is, less variable.

3.61) $R_A = 18{,}41 - 17{,}03 = 1{,}38$

$R_B = 20{,}60 - 19{,}55 = 1{,}05$

Stock B prices are more stable because its range is less than that of Stock A.

3.65 According to the Insurance Information Institute, per capita fire losses (excluding losses from forest fires and government property) in the United States from 1990 to 1999 were, respectively, 38, 45, 53, 44, 49, 45, 47, 48, 46, and 42 dollars.

(a) Calculate s using the formula that defines s.

(b) Rework this exercise, using the computing formula.

(c) Verify that the solutions of parts (a) and (b) are identical.

X_i	$X_i - \bar{X}$	$(X_i-\bar{X})^2$	X_i^2
$38-45,7=-7,7$		$(-7,7)^2$	38^2
$45-45,7=-0,7$		$(-0,7)^2$	45^2
$53-45,7=7,3$		$(7,3)^2$	53^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$46-45,7=0,3$		$(0,3)^2$	46^2
$+\ 42-45,7=-3,7$		$+\ (-3,7)^2$	$+\ 42^2$
$\Sigma X_i=457$		$\Sigma(X_i-\bar{X})^2=148,1$	$\Sigma X_i^2=21033$

$\boxed{n=10}$

$$\bar{X}=\frac{457}{10}=45,7$$

3.65) a) $$s^2=\frac{\Sigma(X_i-\bar{X})^2}{n-1}=\frac{148,1}{9}=16,46$$

$$s=\sqrt{16,46}=4,06$$

b) $$s^2=\frac{\Sigma X_i^2-\frac{(\Sigma X_i)^2}{n}}{n-1}=\frac{21033-\frac{457^2}{10}}{9}=\frac{148,1}{9}=16,46$$

$$s=\sqrt{16.46}=4,06$$

3.71 In 1999 the electric energy generated in units of billions of kilowatt-hours in the seven west north-central states of North Dakota, South Dakota, Nebraska, Kansas, Minnesota, Iowa, and Missouri was 31.3, 10.6, 30.0, 42.0, 44.0, 37.0, and 73.5. Modifv the computing formula for the standard deviation so that it applies to populations (that is, replace the $n-1$ by n and then replace each n by N) and then use it to calculate σ for the given data.

X_i	X_i^2
$31,3$	$31,3^2$
$10,6$	$10,6^2$
$\vdots$	$\vdots$
$+\ 73,5$	$73,5^2$
$\Sigma X_i=268,4$	$\Sigma X_i^2=12643,3$

$N=7$

$$\sigma^2=\frac{\Sigma X_i^2-\frac{(\Sigma X_i)^2}{N}}{N}=\frac{12643,3-\frac{268,4^2}{7}}{7}=310,3$$

$$\sigma=\sqrt{310,3}=17,6$$

3.89 A sample of the ages of five women in an aerobics class revealed the following ages: 22, 18, 26, 20, and 24. Their weights were 115, 159, 141, 137, and 130 pounds.

(a) Find the mean and standard deviation of the ages and use these values to determine the coefficient of variation.

(b) Find the mean and standard deviation of the weights and use these values to determine the coefficient of variation.

(c) Using the coefficients of variation of parts (a) and (b), determine which of the two sets of data are more variable.

Ages		Weights	
X_i	X_i^2	Y_i	Y_i^2
22	22^2	115	115^2
18	18^2	159	159^2
26	26^2	$\vdots$	$\vdots$
20	20^2		
$+\ 24$	24^2	$+\ 130$	130^2
$\Sigma X_i=110$		$\Sigma Y_i=682$	
	$\Sigma X_i^2=2460$		$\Sigma Y_i^2=94056$

3.89) a) $\overline{X} = \frac{110}{5} = 22$; $s_x^2 = \frac{2460 - \frac{110^2}{5}}{4} = 10$; $s_x = \sqrt{10} = 3,16$

$$CoV_x = \frac{s_x}{\overline{X}} \cdot 100 = \frac{3,16}{22} \cdot 100 = 14,37\%$$

b) $\overline{y} = \frac{682}{5} = 136,4$; $s_y^2 = \frac{94056 - \frac{682^2}{5}}{4} = 257,8$

$s_x = \sqrt{257,8} = 16,06$

$$CoV_y = \frac{s_y}{\overline{y}} \cdot 100 = \frac{16,06}{136,4} \cdot 100 = 11,77\%$$

c) Data set of Ages are more varicable, $CoV_x > CoV_y$

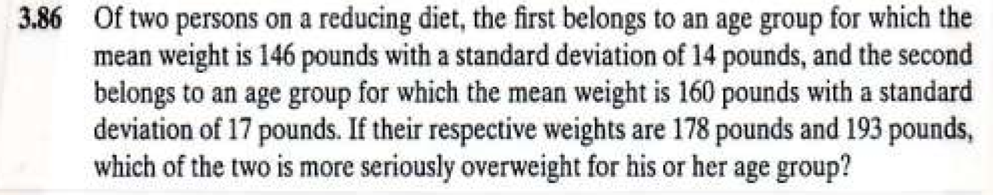
Standardized Score

Consider the idea of "Curve." Your success is determined by the average score of the class and its standard deviation.

Standardized score Z gives "How many standard deviations is the unit away from the mean."

$$Z = \frac{X - \mu}{\sigma}$$

When we compare two values from different populations, we compare their standardized scores.



3.86 Of two persons on a reducing diet, the first belongs to an age group for which the mean weight is 146 pounds with a standard deviation of 14 pounds, and the second belongs to an age group for which the mean weight is 160 pounds with a standard deviation of 17 pounds. If their respective weights are 178 pounds and 193 pounds, which of the two is more seriously overweight for his or her age group?

3.86)

Group 1	Group 2
$\mu_1 = 146$	$\mu_2 = 160$
$\sigma_1 = 14$	$\sigma_2 = 17$
$X_1 = 178$	$X_2 = 193$

$z_1 = \frac{178 - 146}{14} = \underline{\underline{2,286}}$ } More seriously overweight

$z_2 = \frac{193 - 160}{17} = 1,941$

Approximate Statistics for grouped data (or Frequency Distributions)

Mean and Variance:

$$\bar{X} = \frac{\sum f_i x_i}{n} \qquad s^2 = \frac{\sum f_i x_i^2 - \frac{(\sum f_i x_i)^2}{n}}{n-1}$$

where x_i are class marks (middle points) and $n = \sum f_i$

Median, Quartiles and Percentiles

$$\tilde{X} = L + \frac{j}{f} \cdot c$$

L: Lower Bound of the class
f: Frequency
j: Number of items to reach position of median
c: Interval of the Class

Quartiles and Percentiles are found likewise, as we'll see in the next example;

*3.101 If a share of common stock has a market price of \$30.00, and its earnings per share is \$5.00, then its price-earnings (PE) ratio is $\frac{30.00}{5.00} = 6.0$. If the earnings of this common stock had been \$2.00, \$3.00, or \$4.00, then the PE ratio would have been 15.0, 10.0, and 7.5, respectively. Suppose an investor's stock portfolio consisted of 40 stocks with the following PE ratios:

PE ratio	*Number of stocks*
0–4	3
5–9	10
10–14	12
15–19	9
20–24	4
25–29	2
Total	40

(a) Find the mean;

(b) find the standard deviation;

(c) find the median;

(d) find quartile 1 and quartile 3;

(e) find the eightieth percentile.

3.101)

PE Ratio	x_i	f_i	$f_i \cdot x_i$	x_i^2	$f_i \cdot x_i^2$	Cum. f_i
0-4	2	3	6	4	12	3
5-9	7	10	70	49	490	13
10-14	12	12	144	144	1728	25
15-19	17	9	153	289	2601	34
20-24	22	4	88	484	1936	38
25-29	27	+ 2	+ 54	729	+ 1458	40
		$n=\sum f_i = 40$	$\sum f_i x_i = 515$		$\sum f_i x_i^2 = 8225$	

a) $\overline{X} = \frac{515}{40} = 12,875$

b) $s^2 = \dfrac{8225 - \frac{515^2}{40}}{39} = 40,88$

$s = \sqrt{40,88} = 6,39$

c) Position of median $= \frac{40}{2} = 20$ (we used $\frac{n}{2}$ for simplicity)

looking at the Cum. fi values, we passed 20 first time at class 10-14, and we need $20-13=7$ values to reach median. Then; -0,5 $L=9,5$; $j=7$; $f=12$; $c=14,5-9,5=5$ so;

$$\tilde{X} = 9,5 + \frac{7}{12} \cdot 5 = 12,417$$

d) Position of $Q_1 = \frac{40}{4} = 10$; Position of $Q_3 = 3 \cdot \frac{40}{4} = 30$

$L=4,5$; $j=10-3=7$; $f=10$; $c=5$

$L=14,5$; $j=30-25=5$; $f=9$; $c=5$

$Q_1 = 4,5 + \frac{7}{10} \cdot 5 = 8$

$Q_3 = 14,5 + \frac{5}{9} \cdot 5 = 17,278$

e) Position of $P_{18} = 0,18 \cdot 40 \cong 7$

$L=4,5$; $j=7-3=4$; $f=10$; $c=5$

$P_{18} = 4,5 + \frac{4}{10} \cdot 5 = 6,5$

Chebyshev's Theorem & Empirical Rule

For any population; AT LEAST $\left(1-\frac{1}{k^2}\right).100\%$ of the data lies within the interval $\mu \pm k.\sigma$

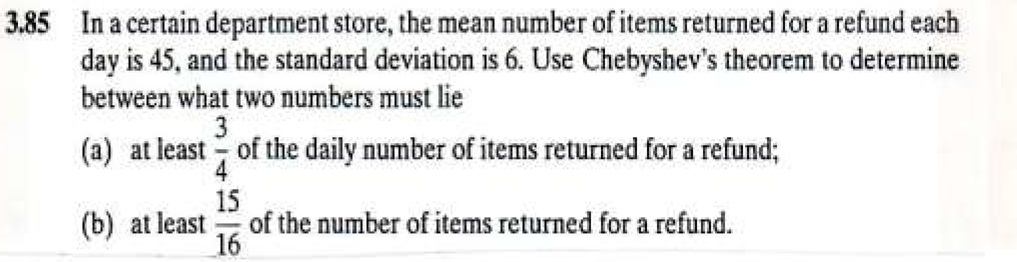
If the data is mounded (bell-shaped) and symmetric, APROXIMATELY;

$\mu \pm 1.\sigma$ interval contains 68%.

$\mu \pm 2\sigma$ interval contains 95%.

$\mu \pm 3\sigma$ interval contains 99,7% of the data.

(We'll see that these approximations follow Normal Distribution)



3.85 In a certain department store, the mean number of items returned for a refund each day is 45, and the standard deviation is 6. Use Chebyshev's theorem to determine between what two numbers must lie

(a) at least $\frac{3}{4}$ of the daily number of items returned for a refund;

(b) at least $\frac{15}{16}$ of the number of items returned for a refund.

3.85) $\mu = 45$; $\sigma = 6$ (Note that μ and σ can be replaced by Sample Statistics $\overline{x}$ and s respectively)

a) $1-\frac{1}{k^2}=\frac{3}{4}$

$\frac{1}{k^2}=\frac{1}{4}$

$k=2$

The interval is;

$\mu \pm k.\sigma$

$$\underset{\swarrow -}{45} \pm \underset{+ \searrow}{2.6}$$

$(33 ; 57)$

b) $1-\frac{1}{k^2}=\frac{15}{16}$

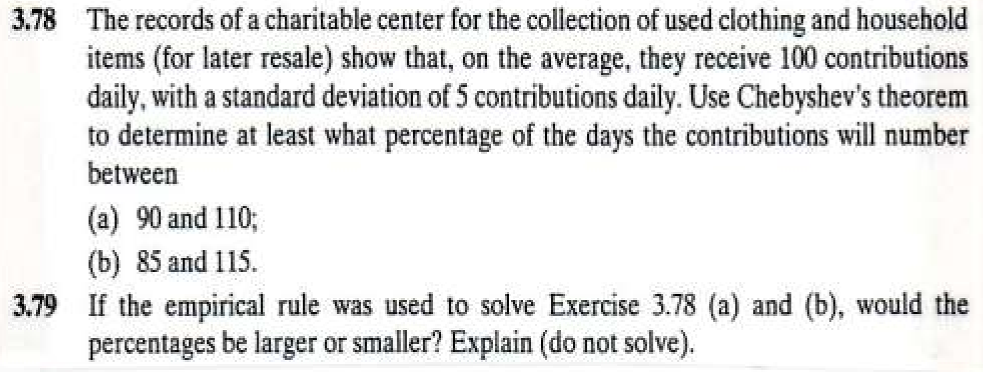
$\frac{1}{k^2}=\frac{1}{16}$

$k=4$

The interval is;

$$\underset{\swarrow -}{45} \pm \underset{+ \searrow}{3.6}$$

$(27 ; 63)$



3.78 The records of a charitable center for the collection of used clothing and household items (for later resale) show that, on the average, they receive 100 contributions daily, with a standard deviation of 5 contributions daily. Use Chebyshev's theorem to determine at least what percentage of the days the contributions will number between

(a) 90 and 110;

(b) 85 and 115.

3.79 If the empirical rule was used to solve Exercise 3.78 (a) and (b), would the percentages be larger or smaller? Explain (do not solve).

3.78) $\mu = 100 ; \sigma = 5$

a) Interval; $(90 ; 110)$ $\hookrightarrow \mu + k.\sigma = 110$

$\left(1 - \frac{1}{2^2}\right).100 = 75\%$

$100 + 5k = 110$

$k = 2$

b) Interval: $(85; 115)$ $\hookrightarrow \mu + k\sigma = 115$

$\left(1 - \frac{1}{3^2}\right).100 = 89\%$

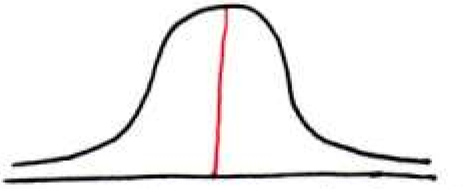
$100 + 5k = 115$

$k = 3$

3.79) (a) Contains approximately 95%

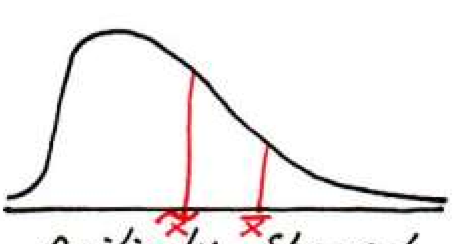
(b) Contains approximately 99,7%

Skewness



Symmetrical (Bell Shaped) Distribution

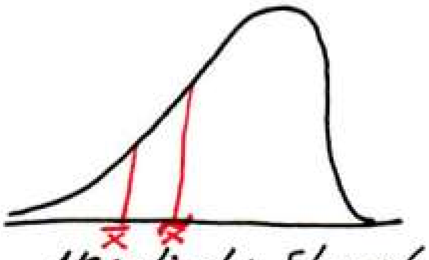
Mean $\approx$ Median



Positively Skewed Distribution

Mean > Median

(Right Skewed)



Negatively Skewed Distribution

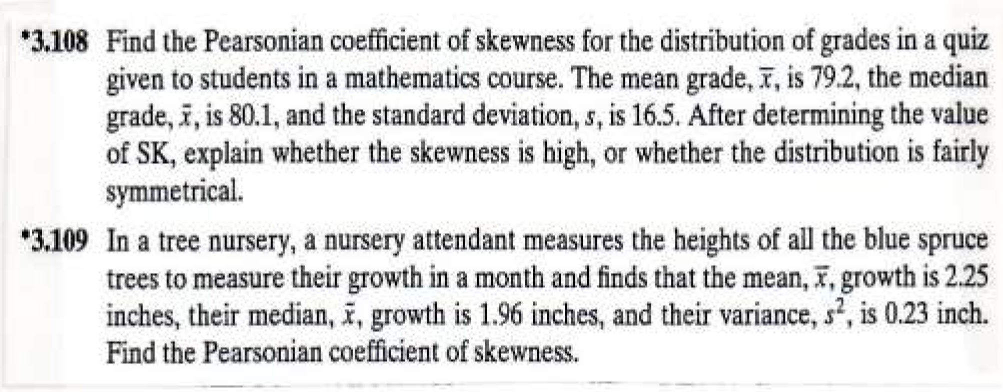
Mean < Median

(Left Skewed)

Pearsonian Coefficient of Skewness

$$SK = \frac{3.(\text{mean} - \text{median})}{\text{standard deviation}}$$

For a perfectly skewed distribution, the value of SK is 0.



***3.108** Find the Pearsonian coefficient of skewness for the distribution of grades in a quiz given to students in a mathematics course. The mean grade, $\bar{x}$, is 79.2, the median grade, $\tilde{x}$, is 80.1, and the standard deviation, s, is 16.5. After determining the value of SK, explain whether the skewness is high, or whether the distribution is fairly symmetrical.

***3.109** In a tree nursery, a nursery attendant measures the heights of all the blue spruce trees to measure their growth in a month and finds that the mean, $\bar{x}$, growth is 2.25 inches, their median, $\tilde{x}$, growth is 1.96 inches, and their variance, s^2, is 0.23 inch. Find the Pearsonian coefficient of skewness.

3.108) $\bar{X} = 79{,}2$; $\tilde{X} = 80{,}1$; $s = 16{,}5$

$$SK = \frac{3.(79{,}2 - 80{,}1)}{16{,}5} = -0{,}164$$

Distribution of the grades are fairly symmetrical, slightly skewed to the left.

3.109) $\bar{X} = 2{,}25$; $\tilde{X} = 1{,}96$; $s^2 = 0{,}23 \Rightarrow s = \sqrt{0{,}23} = 0{,}48$

$$SK = \frac{3.(2{,}27 - 1{,}96)}{0{,}48} = 1{,}94$$

Distribution of the heights are skewed to the right.