4 CHAPTER

Describing and Summarizing Data from One Variable
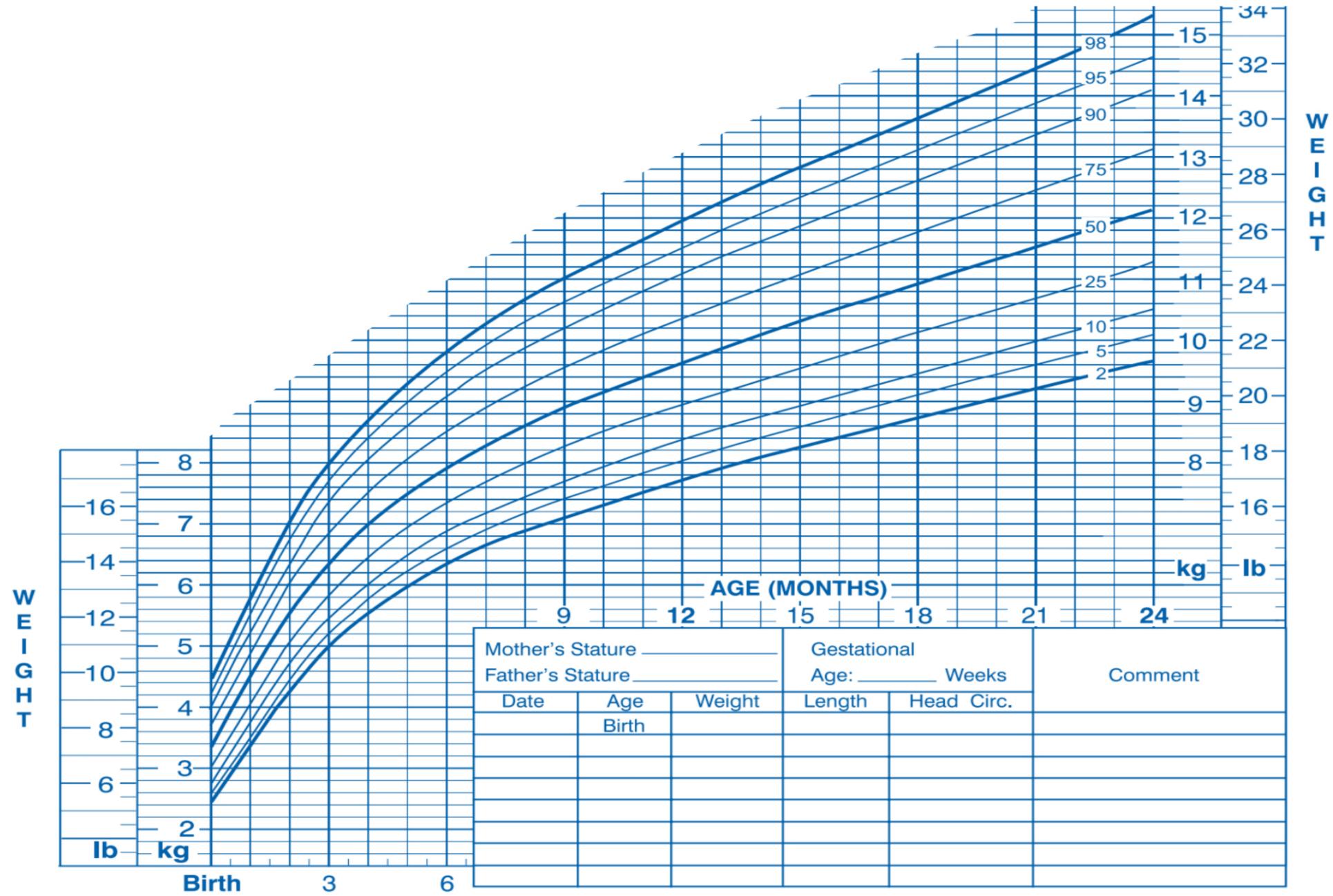
P^{th} Percentile

Given a set of data $x_1, x_2, ..., x_n$, the Pth **percentile** is a value, say x, such that approximately P percent of the data is less than or equal to x and approximately $(100 - P)$ percent of the data is greater than or equal to x.

DEFINITION

WEIGHT

WEIGHT

WEIGHT

WEIGHT

98
95
90
75
50
25
10
5
2

15
14
13
12
11
10
9
8

34
32
30
28
26
24
22
20
18
16

kg   lb

AGE (MONTHS)

9   12   15   18   21   24

8
7
6
5
4
3
2

16
14
12
10
8
6

lb   kg

Birth   3   6

Mother's Stature		Gestational		
Father's Stature		Age: _____ Weeks		Comment
Date	Age	Weight	Length	Head Circ.	
	Birth				

Published by the Centers for Disease Control and Prevention, November 1, 2009

HUMAN SERVICES·USA

3

ascending order

To find the location of the Pth percentile in the ordered data, calculate,

$$\ell = n\left(\frac{P}{100}\right)$$

where n is the number of observations in the ordered data.

No   Is ℓ an integer?   Yes

Round ℓ up to next greatest integer.

Average the data value in the ℓth location with the data value in the $(\ell + 1)^{th}$ location

Find the data value in the

75% → .75

$$\lfloor \ell \rfloor = n(P)$$

Find the 50th percentile for the following data on the number of spelling errors found on 7 pages of a web site.

> 3, 5, 0, 1, 9, 2, 7

$$0, 1, 2, 3, 5, \boxed{7}, 9$$

50th (above 5)   80th (above 7)

$$\ell = 7(.50) = 3, 5 \rightarrow 4$$
$$\ell = 7(.80) = 5, 6 \rightarrow 6$$

Suppose that 40 members of your company are given a screening test for a new position. These scores are reported in the table on the left below. To inform potential employees of their screening test performance you may wish to report various percentiles for the test scores. Find the 10th and 88th percentiles for the test.

$10^{th} \rightarrow 40(.10) = 4$

$28 \leftarrow$

88th

$40(.88) = 35.2$

$= 36$

Ordered Test Scores			
18	43	54	66
21	44	55	67
21	45	55	69
27	45	56	70
29	46	57	71
31	47	58	73
32	48	61	77
33	49	62	80
34	52	63	81
41	54	64	82

Percentile

The **percentile** of some data value x is given by

$$\text{percentile of } x = \frac{\text{number of data values less than or equal to } x}{\text{total number of data values}} \cdot 100.$$

FORMULA

Suppose that 40 members of your company are given a screening test for a new position. These scores are reported in the table on the left below. To inform potential employees of their screening test performance you may wish to report various percentiles for the test scores. Find the 10th and 88th percentiles for the test.

$$\frac{15}{40} = .375$$

$$38\%$$

Ordered Test Scores			
18	43	54	66
21	44	55	67
21	45	55	69
27	45	56	70
29	46	57	71
31	47	58	73
32	48	61	77
33	49	62	80
34	52	63	81
41	54	64	82

Quartiles

The 25th, 50th, and 75th percentiles are known as **quartiles** and are denoted as Q_1, Q_2, and Q_3 respectively.

DEFINITION

Interquartile Range

The **interquartile range** is a measure of dispersion which describes the range of the middle fifty percent of the data. It is calculated as follows.

$$IQR = Q_3 - Q_1$$

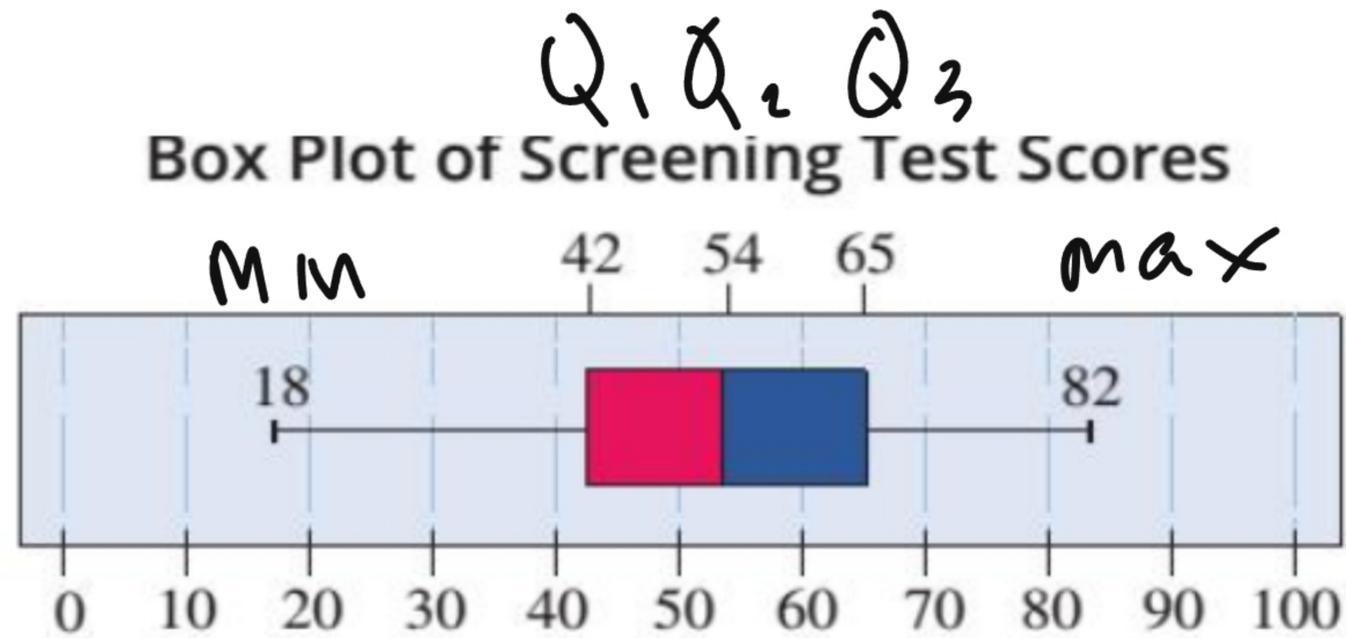**FORMULA**

$25\% \; Q_1$

$50\% \; median \; (Q_2)$

$75\% \; Q_3$

Suppose that 40 members of your company are given a screening test for a new position. These scores are reported in the table on the left below. To inform potential employees of their screening test performance you may wish to report various percentiles for the test scores. Find the 10th and 88th percentiles for the test.

$25\% \rightarrow$ 42 Q_1

$50\% \rightarrow$ 54 median Q_2

$75\% \rightarrow$ 65 Q_3

IQR

$65 - 42 =$

$\boxed{23}$

Ordered Test Scores			
18	43	54	66
21	44	55	67
21	45	55	69
27	45	56	70
29	46	57	71
31	47	58	73
32	48	61	77
33	49	62	80
34	52	63	81
41	54	64	82

$Q_1 \quad Q_2 \quad Q_3$

Box Plot of Screening Test Scores

M M   42   54   65   max

18   82

0   10   20   30   40   50   60   70   80   90   100

5-Number Summary

For a set of data, the 5-number summary consists of the following five values:

1. Minimum

2. First quartile, Q_1

3. Second quartile, Q_2, or the median

4. Third quartile, Q_3

5. Maximum

DEFINITION

Outlier

A data point is considered an **outlier** if it is 1.5 times the interquartile range above the 75th percentile or 1.5 times the interquartile range below the 25th percentile.

DEFINITION

$$\text{Low} \qquad Q_1 - 1.5(IQR)$$

$$\text{High} \qquad Q_3 + 1.5(IQR)$$

position. These scores are reported in the table on the left below. To inform potential employees of their screening test performance you may wish to report various percentiles for the test scores. Find the 10th and 88th percentiles for the test.

$$25\% \rightarrow 42 \quad Q_1$$

$$50\% \rightarrow 54 \text{ median } Q_2$$

$$75\% \rightarrow 65 \quad Q_3$$

Ordered Test Scores			
18	43	54	66
21	44	55	67
21	45	55	69
27	45	56	70
29	46	57	71
31	47	58	73
32	48	61	77
33	49	62	80
34	52	63	81
41	54	64	82

$$IQR$$

$$65 - 42 =$$

$$\boxed{23}$$

$$23 \cdot (1.5) = 34.5$$

Low $Q_1 - 34.5$
$42 - 34.5 = 7.5$

High $Q_3 + 34.5$
$65 + 34.5 = 99.5$

Subjects in a marketing study were shown a film and at the end of the film were given a test to measure their recall. The scores are listed below.

> 97 31 61 49 61 85 35 57 31 26 27 40 86 78 28 61
> 87 62 92 58 38 95 81 68 64 72 45 57 84 100

a. What level of measurement does the data possess?

b. Calculate Q_1, the first quartile.

c. Calculate Q_2, the second quartile.

d. Calculate Q_3, the third quartile.

e. Explain the meaning of these percentiles in the context of the marketing study.

f. Calculate the interquartile range.

$$Q_1 \Rightarrow 30(.25) = 7.5 \rightarrow 8$$
$$Q_2 \Rightarrow 30(.50) = 15$$
$$Q_3 \Rightarrow 30(.75) = 22.5 \rightarrow 23$$

26
27
28
31
31
35
38
40 $\leftarrow Q_1 \ (40)$
45
49
57
57
58
61
61 $\leftarrow Q_2 \ (61)$
61
62
64
68
72
78
81
84 $\leftarrow Q_3 = (84)$
85
86
87
92
95
97
100

14

US Violent Crime Rate (per 100,000 Residents) 2010-2014
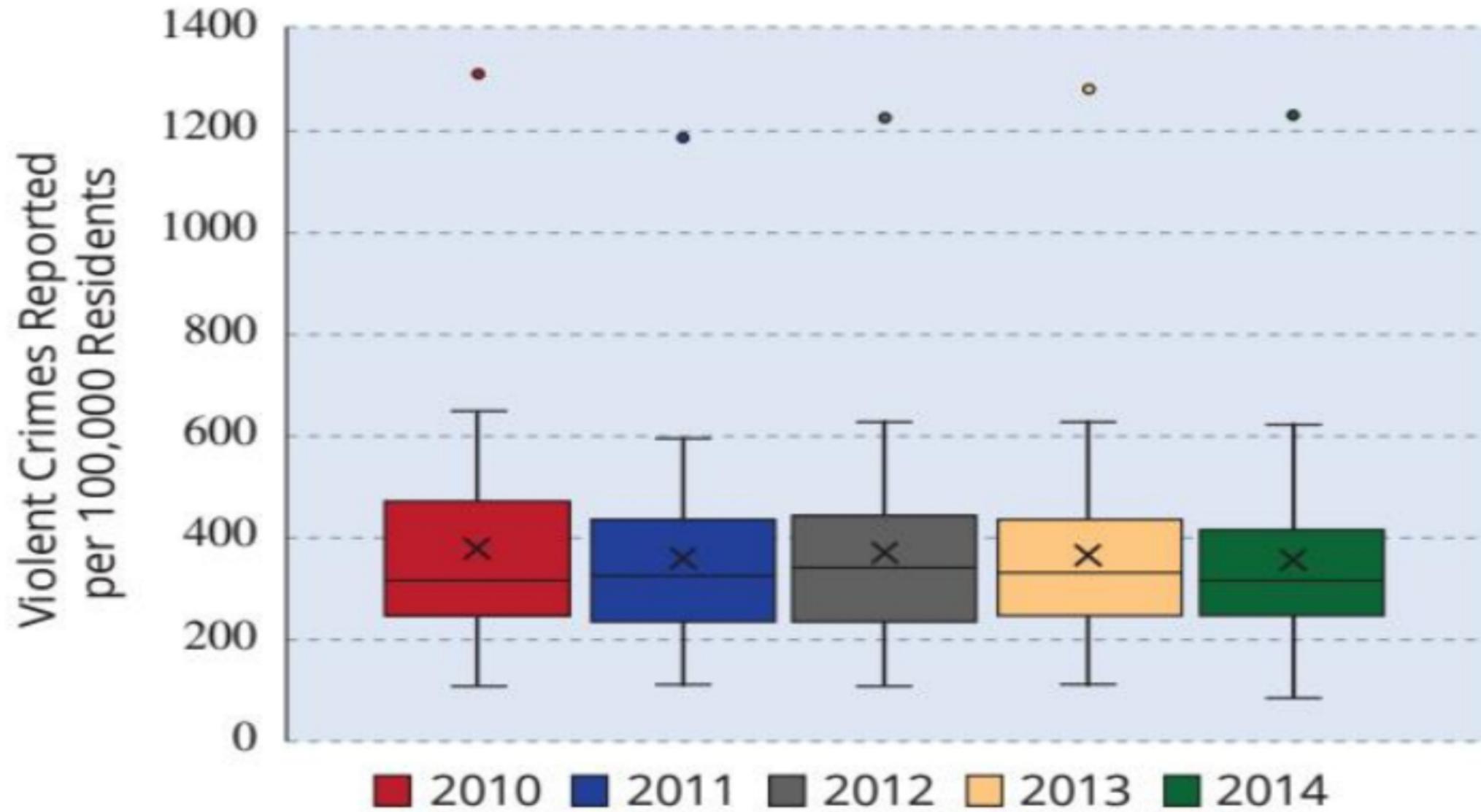
Figure 4.3.5

Education and Earnings 2016
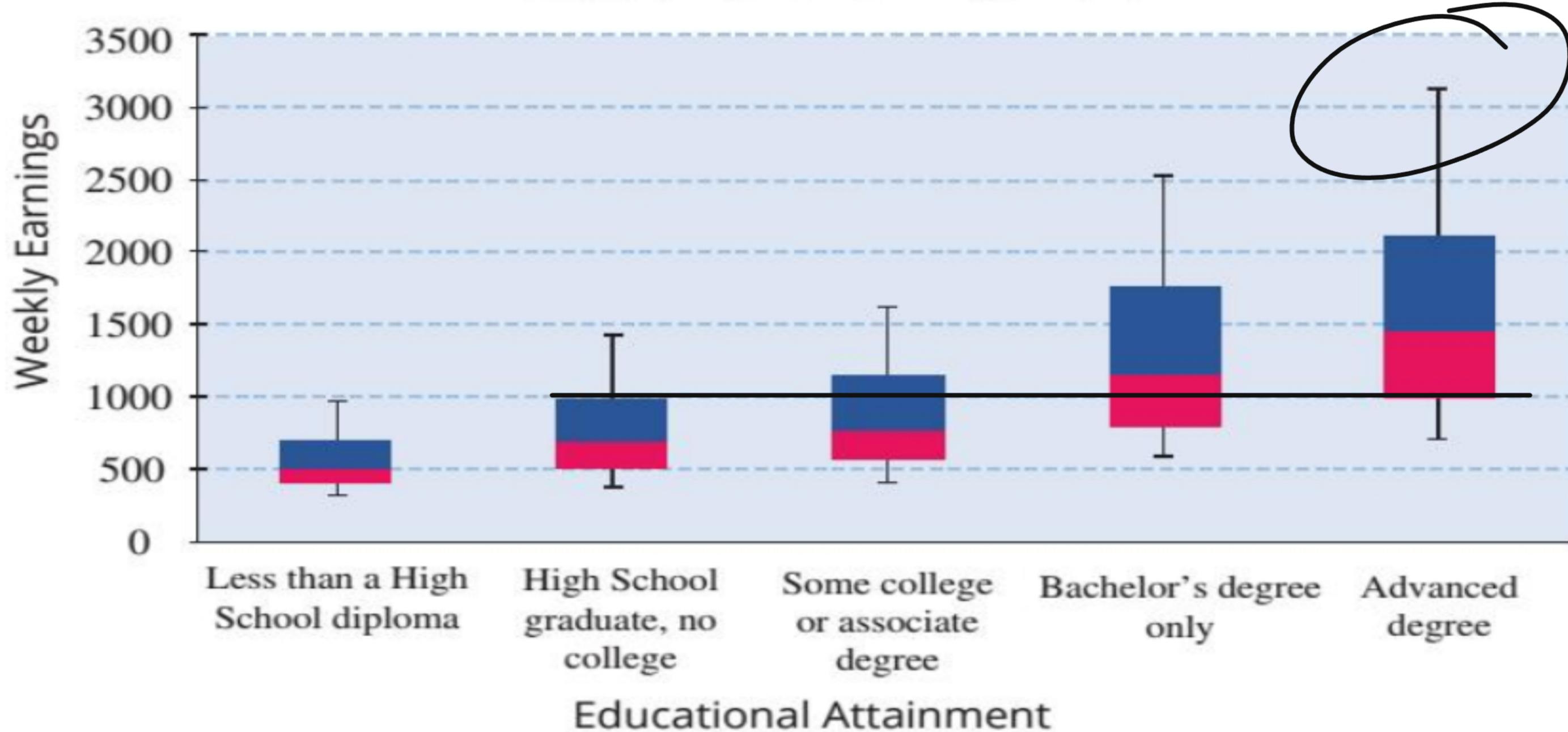
Weekly Earnings vs Educational Attainment

Y-axis: Weekly Earnings (0, 500, 1000, 1500, 2000, 2500, 3000, 3500)

X-axis: Educational Attainment
- Less than a High School diploma
- High School graduate, no college
- Some college or associate degree
- Bachelor's degree only
- Advanced degree

16

Proportion

A **proportion** measures the fraction of a group that possesses some characteristic.

DEFINITION

X = the number that possess the characteristic,

N = the number in the population,

n = the number in the sample, then

$$p = \frac{X}{N}$$ the population proportion, and

$$\hat{p} = \frac{X}{n}$$ the sample proportion.

The symbol $\hat{p}$ is pronounced "p-hat".

Proportion

A **proportion** measures the fraction of a group that possesses some characteristic.

DEFINITION

Suppose your statistics class is composed of 48 students of which 4 are left-handed. What proportion of the class is left-handed?

$$\frac{4}{48} = .08\bar{3} \rightarrow 8.3\%$$

According to a study administered by the National Bureau of Economic Research, half of Americans would struggle to come up with $2000 in the event of a financial emergency. The majority of the 1900 Americans surveyed said they would rely on more than one method to come up with emergency funds if required. In the survey, 532 people said that they "certainly" would not be able to cope with an unexpected $2000 bill if they had to come up with the money in 30 days, and 418 people said they "probably" would not be able to cope.[2]

a. What percentage of Americans "certainly" would not be able to produce $2000 in the event of an emergency according to the study?

$$\frac{532}{1900} = .28 = 28\%$$

b. What percentage of Americans would "probably" not be able to pay a $2000 bill in 30 days if required? 50%

$$\frac{418}{1900} = .22 \approx 72\%$$

c. What does this say about the savings habits of Americans?