Chapter 1:Descriptive statistics

- Descriptive statistics summarises a mass of information.
- We may use graphical and/or numerical methods
- Examples of the former are the bar chart and XY chart, examples of the latter are averages and standard deviations

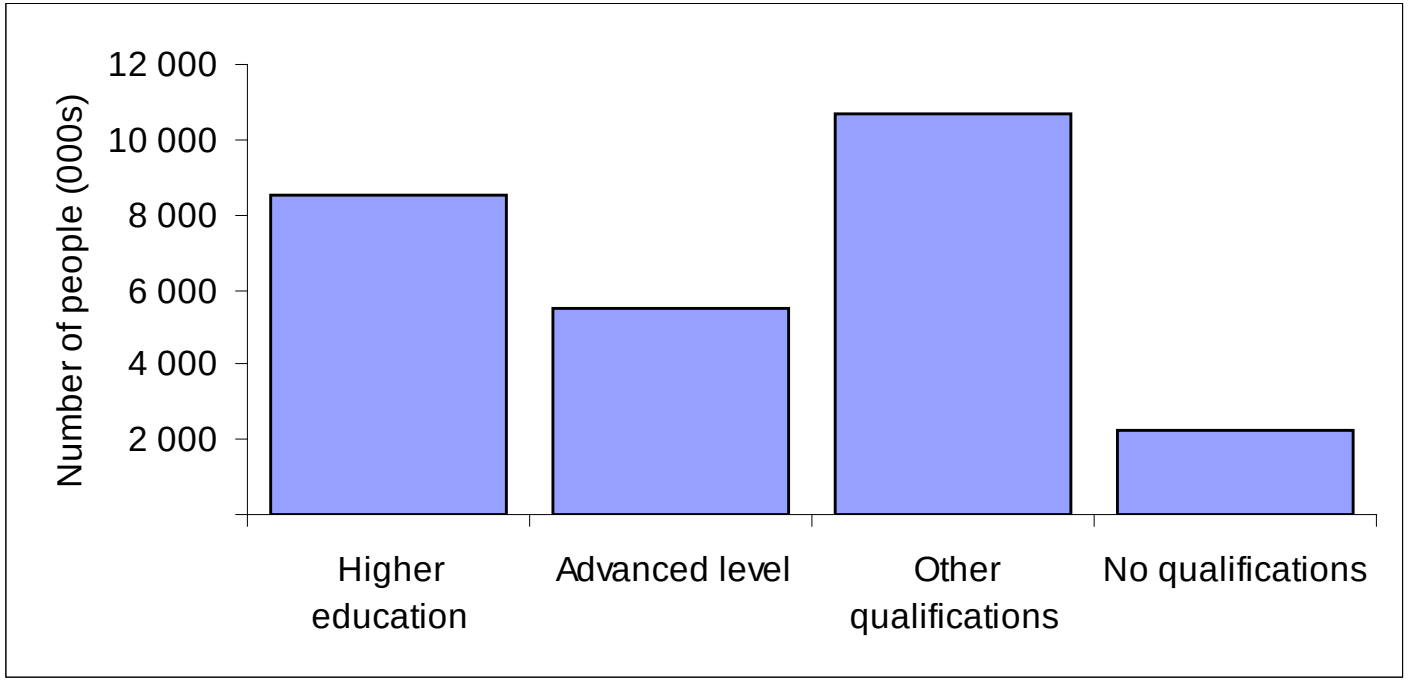
Graphical techniques

- Education and employment data

	Higher education	Advanced level	Other qualifi-cations	No qualifi-cations	Total
In work	8 541	5 501	10 702	2 260	27 004
Unemployed	232	247	758	309	1 546
Inactive	1 024	1 418	3 150	2 284	7 876
	9 797	7 166	14 610	4 853	36 426

Slide 1.3

The bar chart



Note: the height of each bar is determined by the associated frequency. The first bar is 8541 units high, the second is 5501, and so on. The ordering of the bars could be reversed ('no qualifications' becoming the first category) without altering the message.

Figure 1.1 Educational qualifications of people in work in the UK, 2006

A multiple bar chart

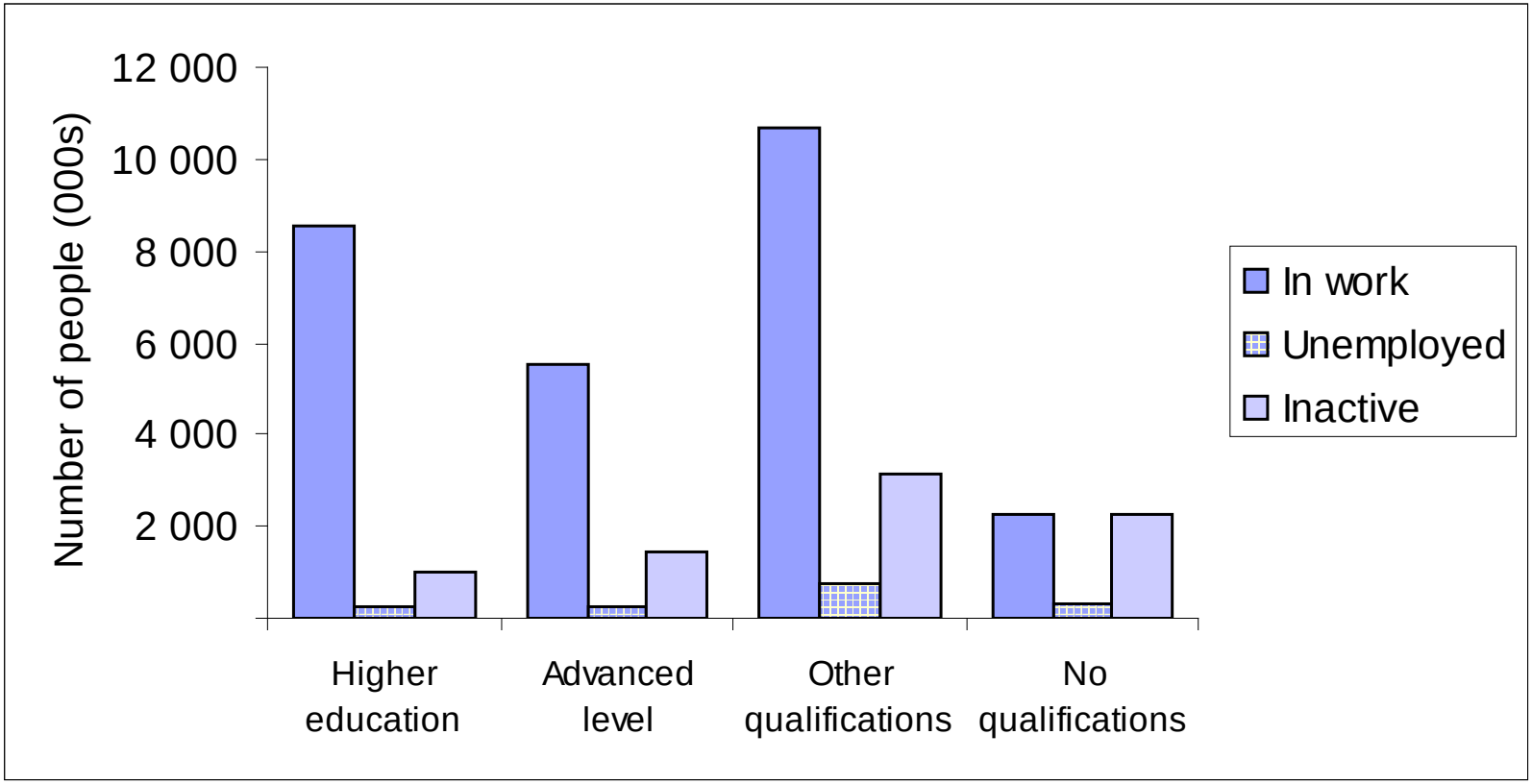


Figure 1.2 Educational qualifications by employment category

Slide 1.5

The stacked bar chart

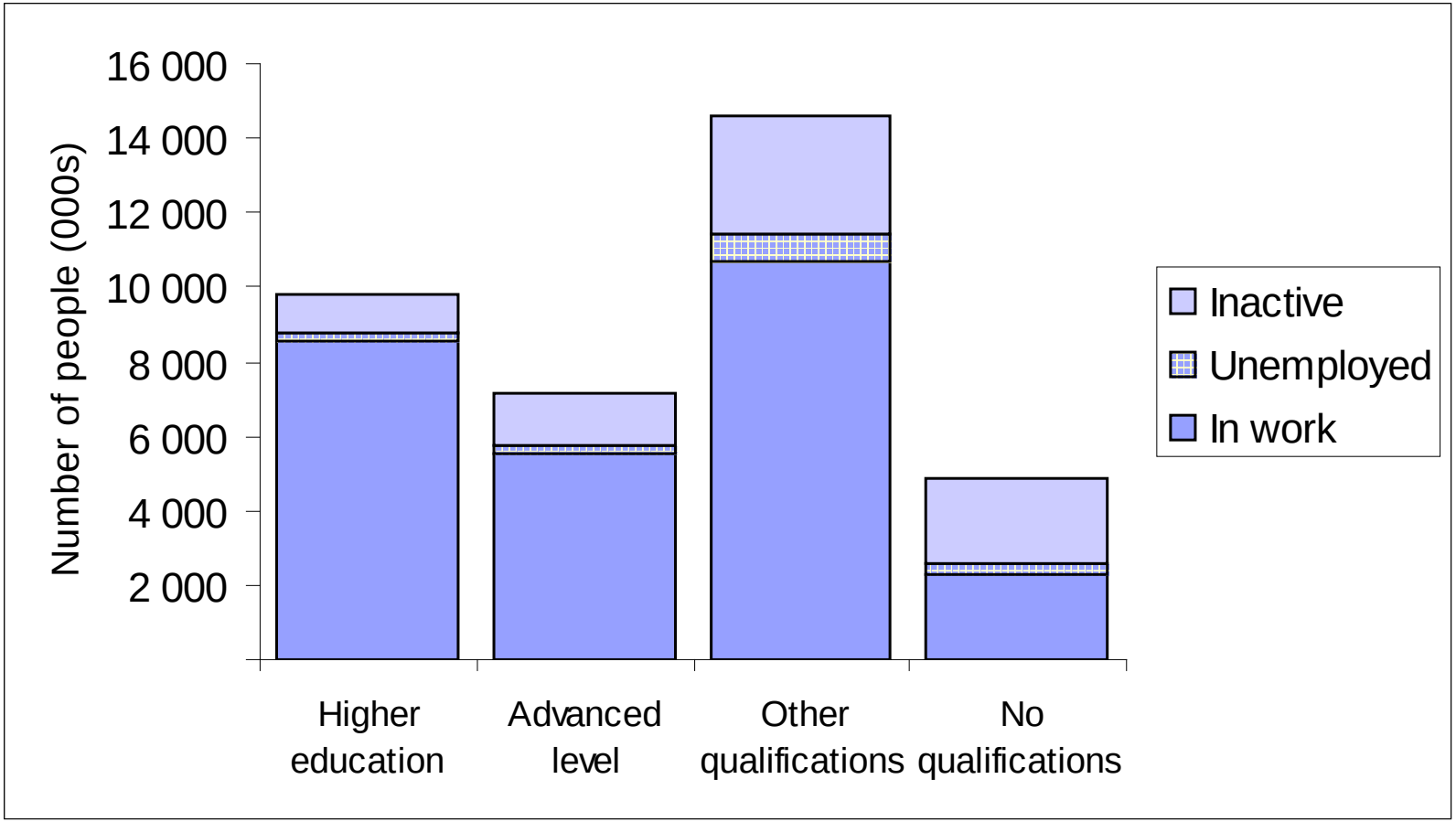


Figure 1.3 Stacked bar chart of educational qualifications and employment status

A stacked bar chart (percentages)

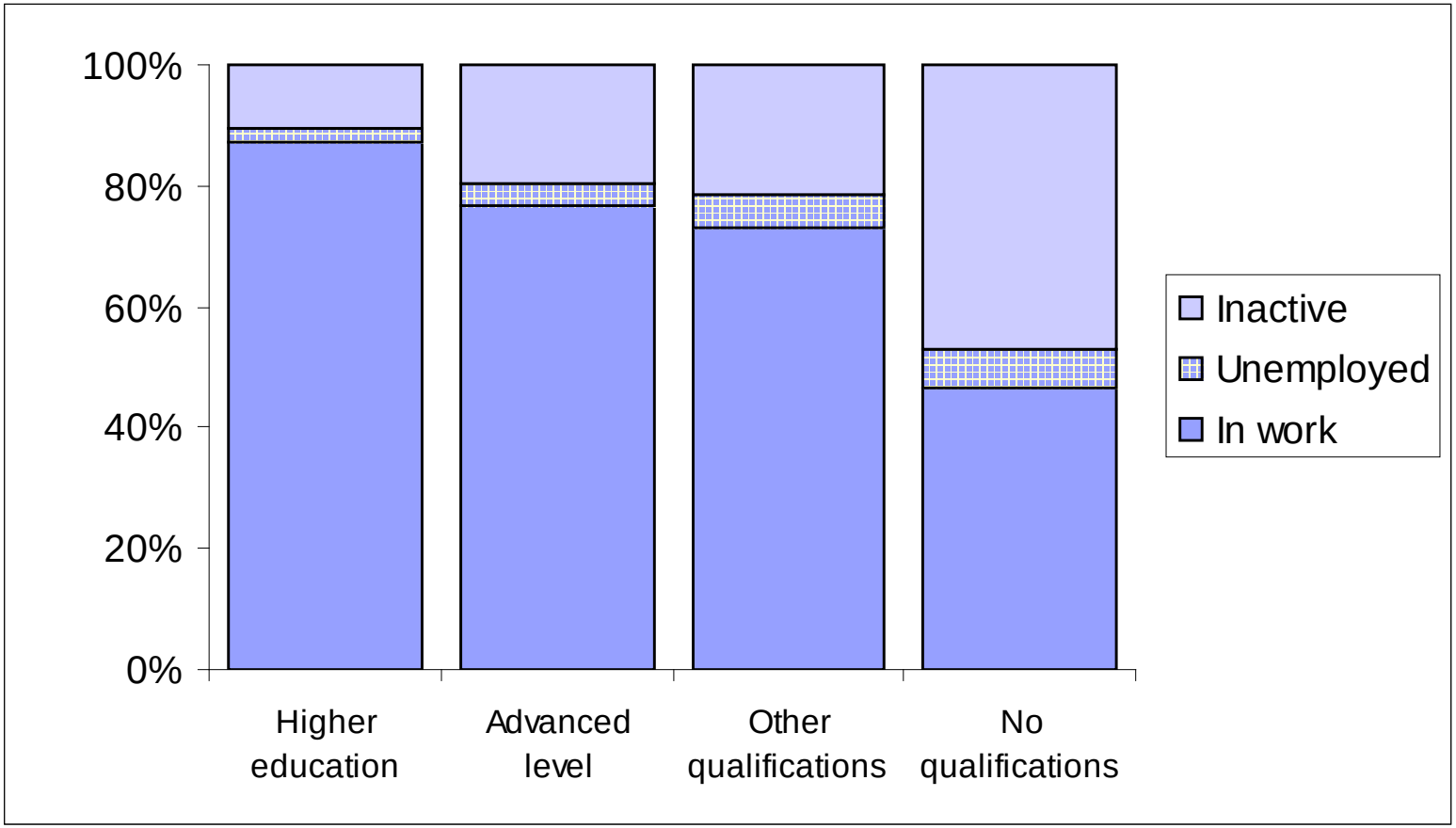


Figure 1.4 Percentages in each employment category, by educational qualification

The pie chart

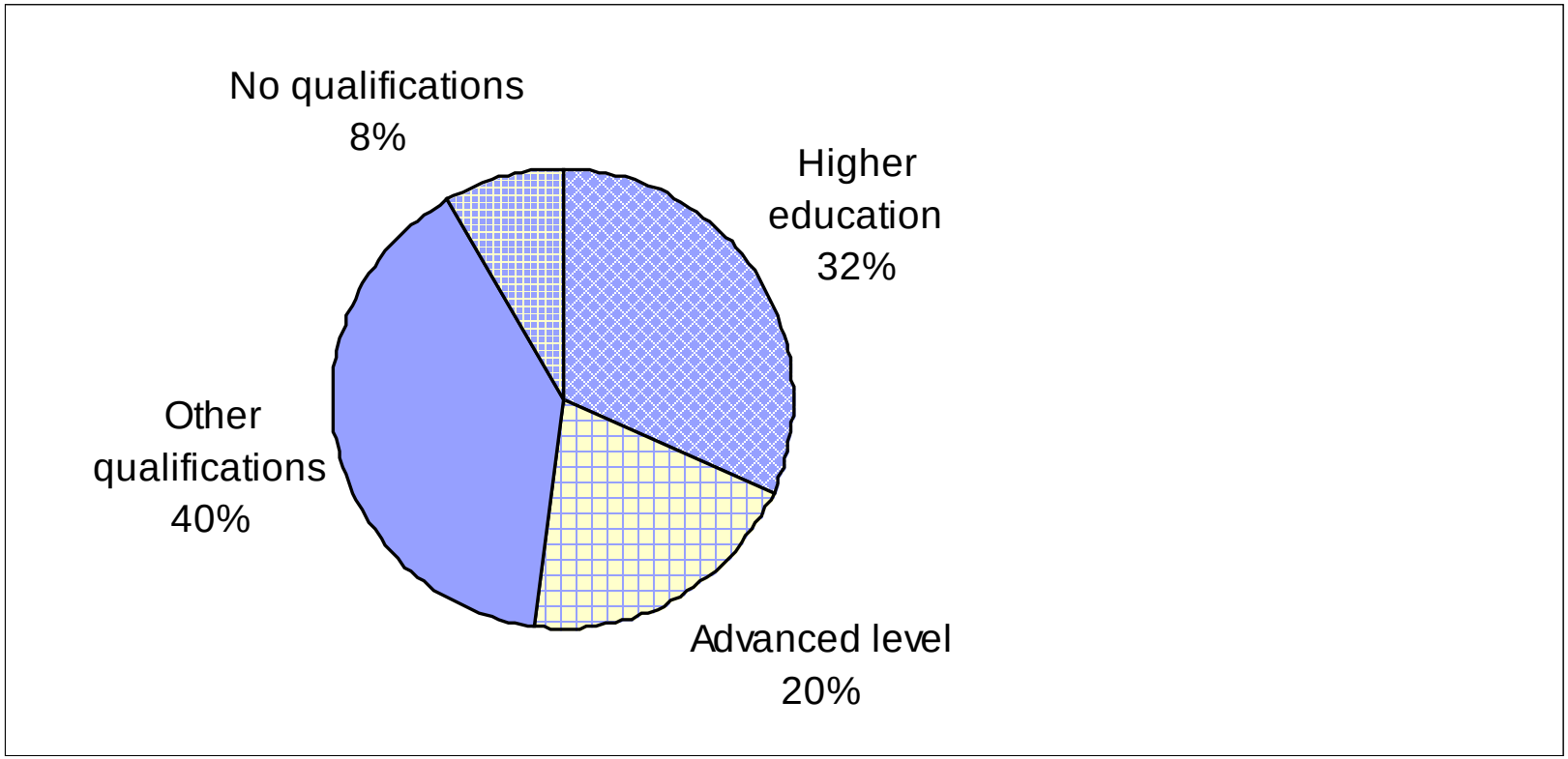


Figure 1.5 Educational qualifications of those in work

Data on wealth in the UK

Class interval (£)	Number (000s)
0–9,999	2,448
10,000–24,999	1,823
25,000–39,999	1,375
40,000–49,999	480
50,000–59,999	665
60,000–79,999	1,315
80,000–99,999	1,640
100,000–149,999	2,151
150,000–199,000	2,215
200,000–299,000	1,856
300,000–499,999	1,057
500,000–999,999	439
1,000,000–1,999,999	122
2,000,000 or more	50
Total	17,636

Table 1.3 The distribution of wealth, UK, 2003

A (misleading!) bar chart

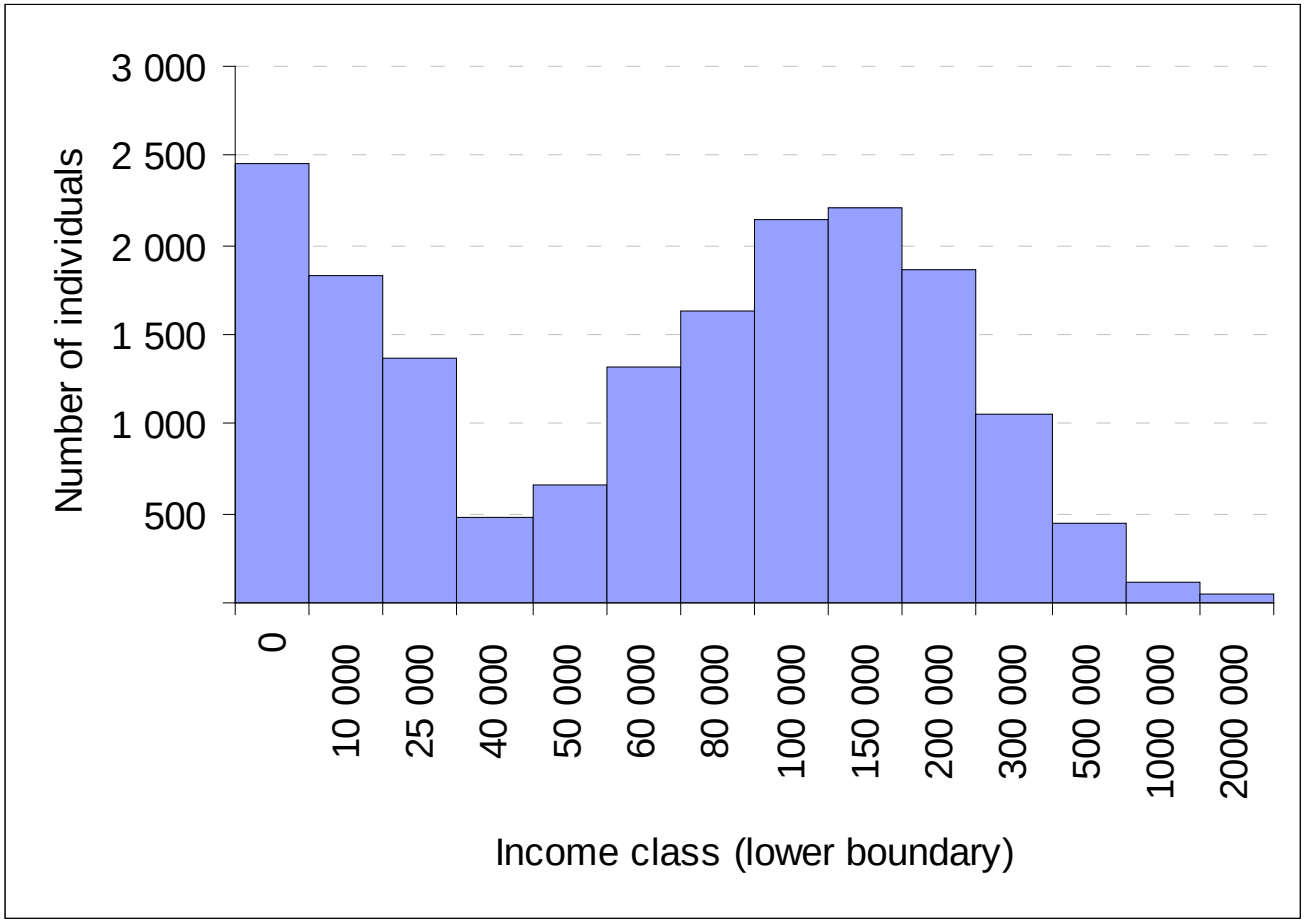


Figure 1.7 Bar chart of the distribution of wealth in the UK, 2003

The histogram – the correct picture

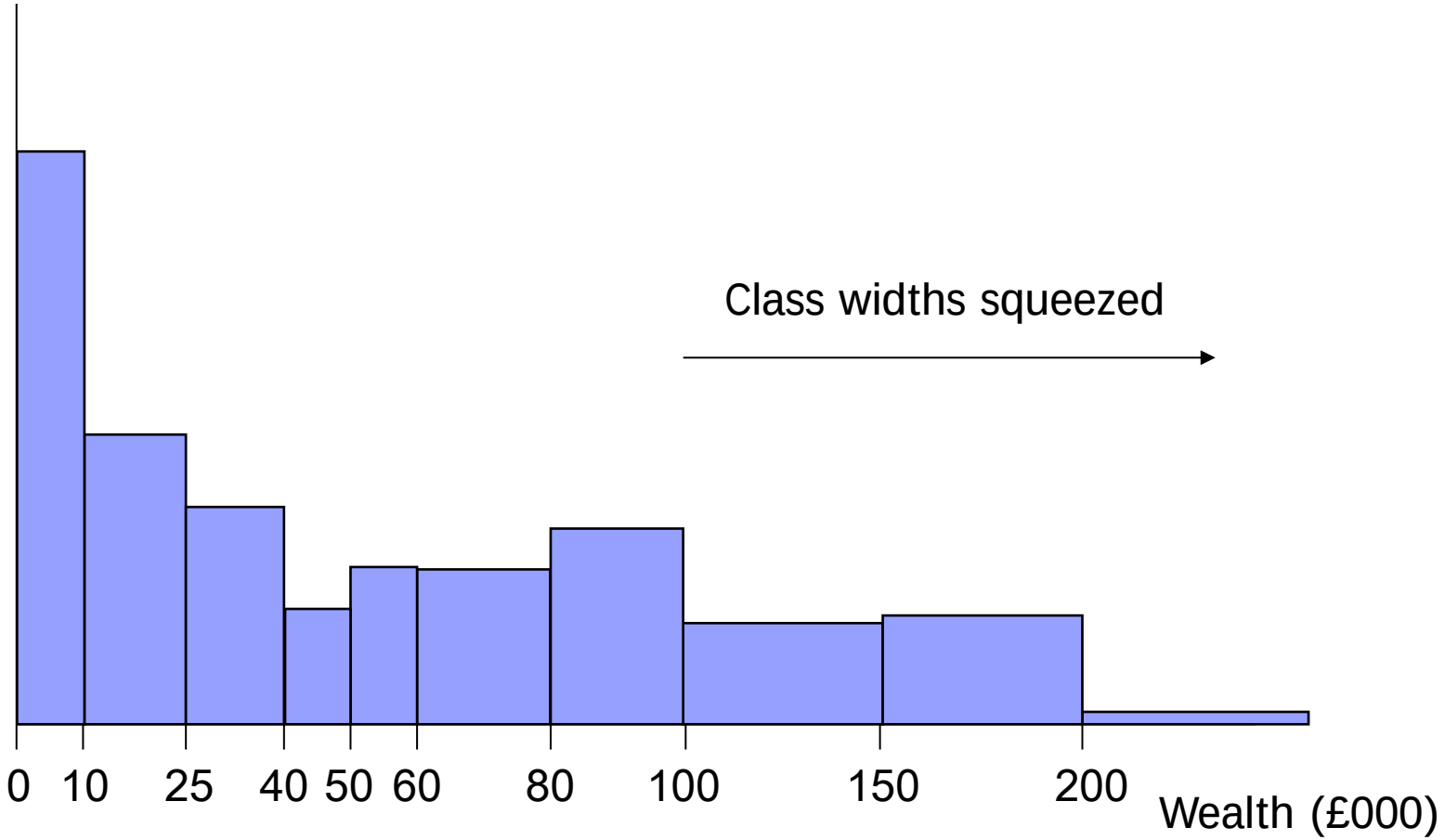


Figure 1.9 Histogram of the distribution of wealth in the UK, 2003

Histogram vs bar chart

- The bar chart gives the wrong picture because of varying class widths

Class interval £)	Numbers (thousands)
10,000–24,999	1,823
:	:
200,000–299,000	1,856

- These two classes have similar frequencies (similar heights for bar chart) but the second is over six times wider. Adjusting for width, its frequency should be 278 (1,856 $\times$ 15/100)

The frequency density

- Applying this principle leads to calculation of the frequency densities:

Class interval (£)	Number (000s)	Frequency density
0–9,999	2,448	0.2448
10,000–24,999	1,823	0.1215
25,000–39,999	1,375	0.0917
40,000–49,999	480	0.0480
50,000–59,999	665	0.0665

Numerical techniques

- We examine measures of
  - Location
  - Dispersion
  - Skewness

Measures of location

- Mean – strictly the arithmetic mean, the well known ‘average’
- Median – the wealth of the person in the middle of the distribution
- Mode – the level of wealth that occurs most often
- These different measures can give different answers…

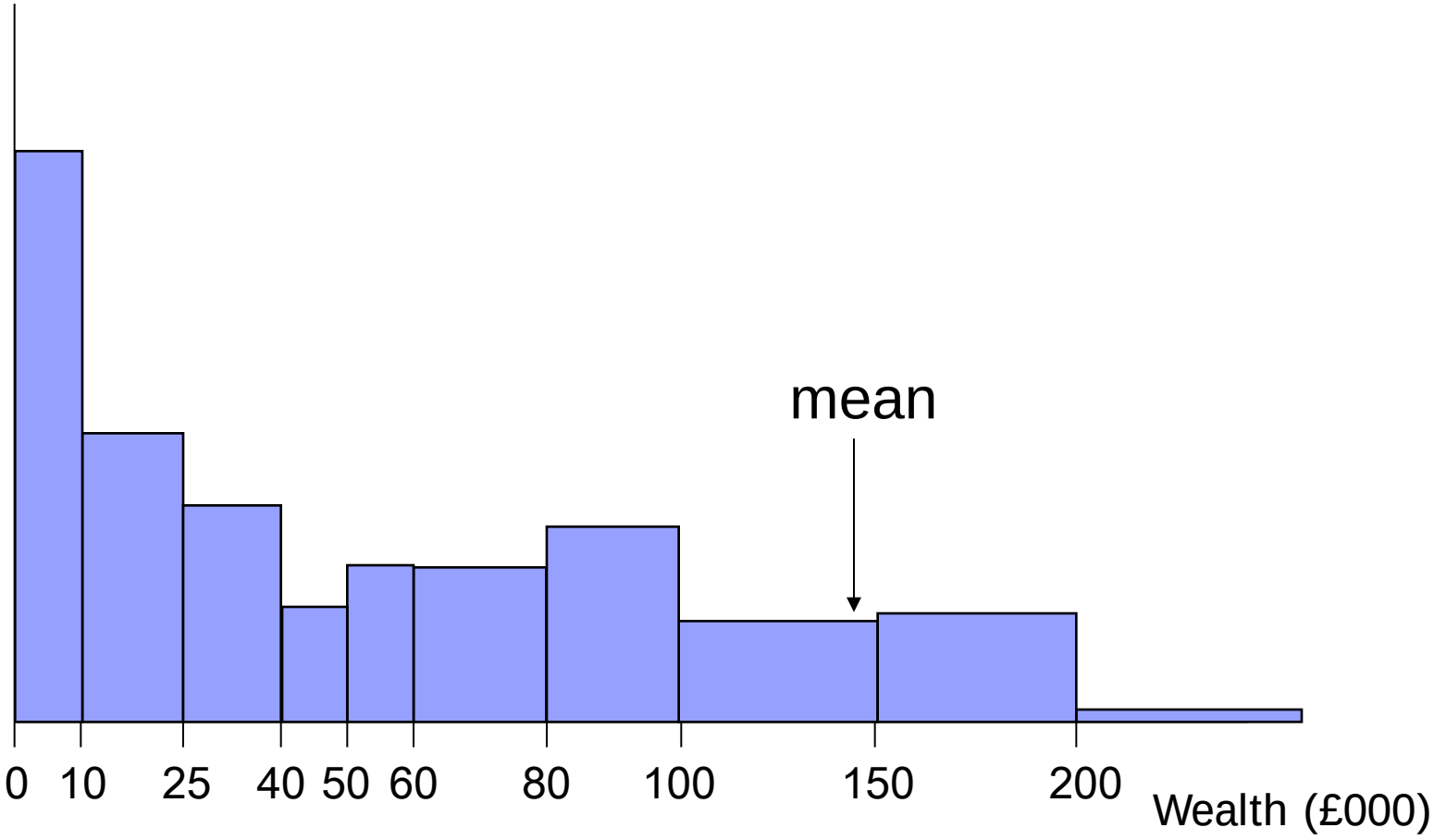
The mean of the wealth distribution

Range	x	f	fx
0-	5.0	2,448	12,240.0
10000-	17.5	1,823	31,902.5
25000-	32.5	1,375	44,687.5
40000-	45.0	480	21,600.0
50000-	55.0	665	36,575.0
60000-	70.0	1,315	92,050.0
80000-	90.0	1,640	147,600.0
100000-	125.0	2,151	268,875.0
150000-	175.0	2,215	387,625.0
200000-	250.0	1,856	464,000.0
300000-	400.0	1,057	422,800.0
500000-	750.0	439	329,250.0
1000000-	1500.0	122	183,000.0
2000000-	3000.0	50	150,000.0
Total		17,636	2,592,205.0

$$\mu = \frac{\sum fx}{\sum f} = \frac{2{,}592{,}205}{17{,}636} = 146.984$$

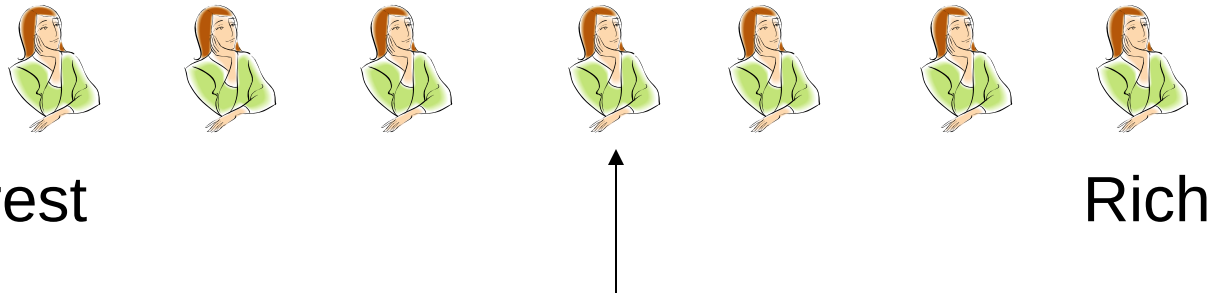
Slide 1.16

Locating the mean



The median

- The wealth of the ‘middle person’ – i.e. the one located halfway through the distribution



- The median is little affected by outliers, unlike the mean

Calculating the median

- 17,636 observations, hence person 8,818 in rank order has the median wealth
- This person is somewhere in the £60-80k interval

Range	Frequency	Cumulative frequency
0-	2,448	2,448
10000-	1,823	4,271
25000-	1,375	5,646
40000-	480	6,126
50000-	665	6,791
60000-	1,315	8,106
80000-	1,640	9,746
100000-	2,151	11,897
150000-	2,215	14,112

Number with wealth less than £80k

Number with wealth less than £100k

Calculating the median (continued)

- To find the precise median, use

$$x_L + \left(x_U - x_L\right)\left\{\frac{\frac{N}{2} - F}{f}\right\}$$

$$= 80 + \left(100 - 80\right) \times \left\{\frac{\frac{17{,}636{,}000}{2} - 8{,}106{,}000}{1{,}640{,}000}\right\} = 90.829$$

- Median wealth is £90,829

The mode

- The mode is the observation with the highest frequency

Size	Sales
8	7
10	25
12	36
14	11
16	3
18	1

Modal dress size = 12

The mode (continued)

- For grouped data, the mode corresponds to the interval with greatest frequency density

Class interval (£)	Number (000s)	Frequency density
0–9,999	2,448	0.2448
10,000–24,999	1,823	0.1215
25,000–39,999	1,375	0.0917
40,000–49,999	480	0.0480
50,000–59,999	665	0.0665

Modal class

Mode = £0–10,000

Slide 1.22

Differences between mean, median and mode

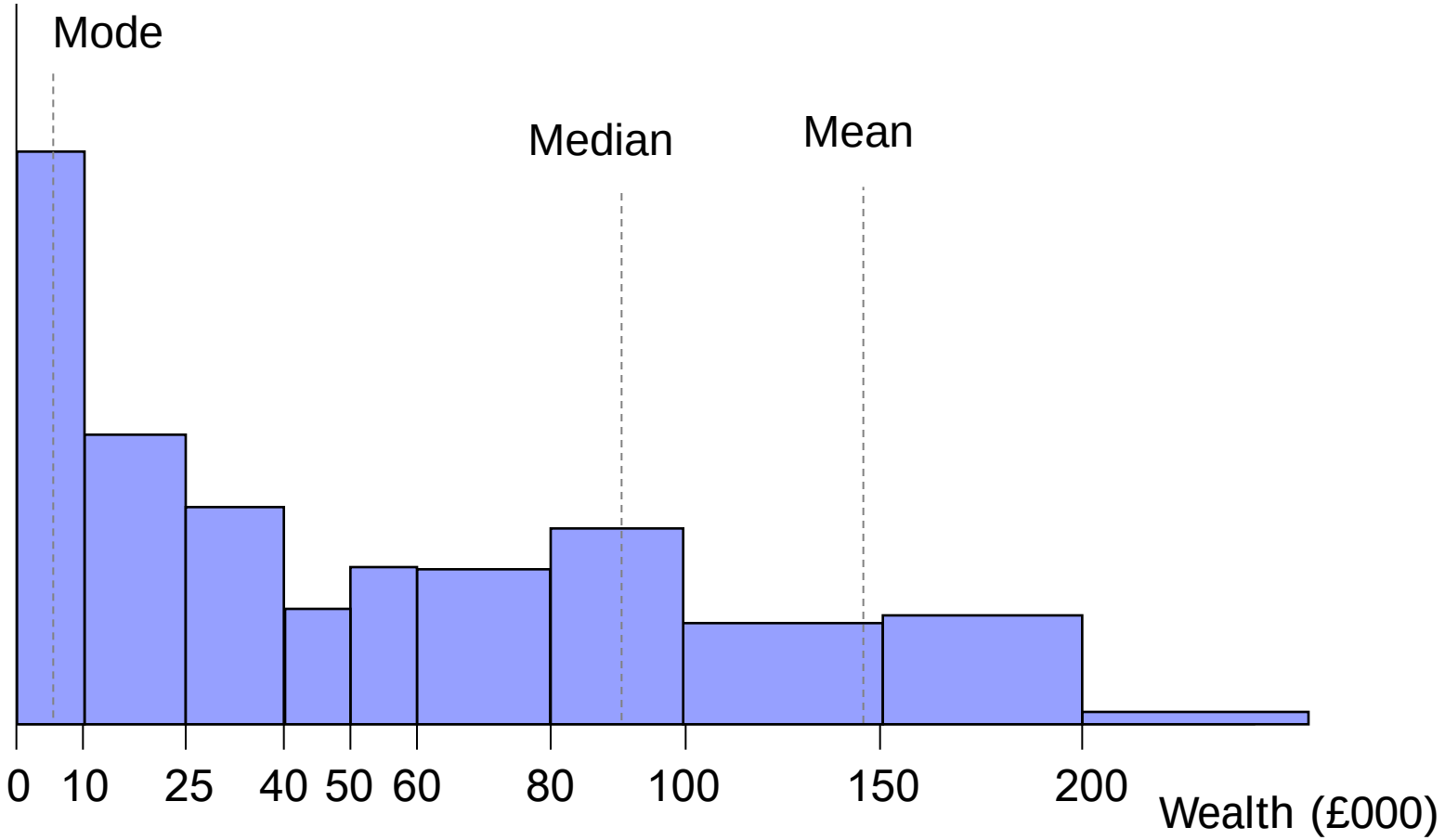


Figure 1.12 The histogram with the mean, median and mode marked

Measures of dispersion

- The range – the difference between smallest and largest observation. Not very informative for wealth
- Inter-quartile range – contains the middle half of the observations
- Variance – based on all observations in the sample

Inter-quartile range

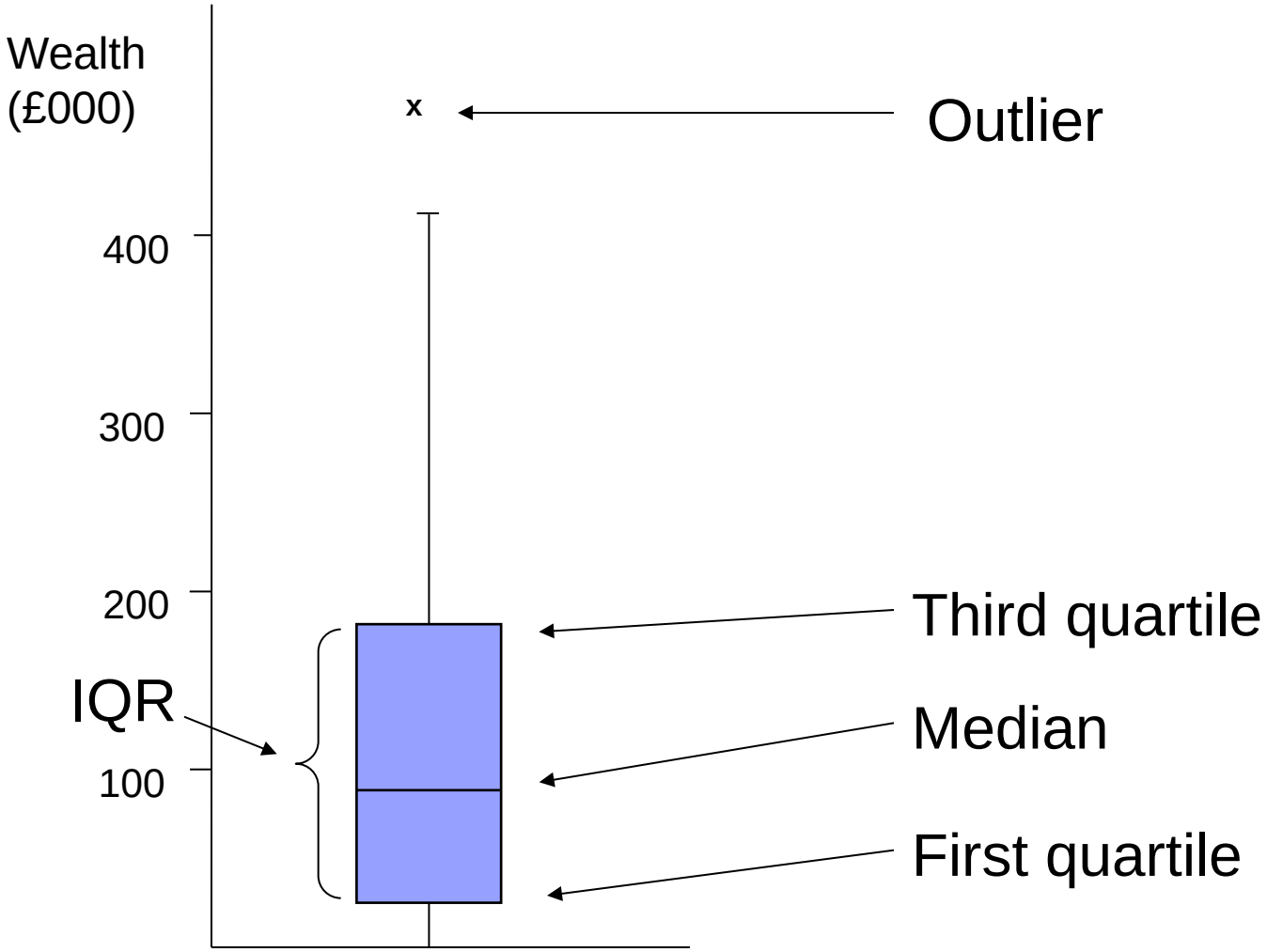
- First quartile – one quarter of the way through the distribution, person ranked 4,409

$$Q_1 = 25{,}000 + (40{,}000 - 25{,}000)\left\{\frac{4{,}409 - 4{,}271}{1{,}375}\right\} = 26{,}505.5$$

- Third quartile – three quarters of the way through the distribution, person ranked 13,227… hence Q_3 = £180,022.6

- IQR = $Q_3 - Q_1$ = 180,023 – 26.505 = 153,517

Box and whiskers plot



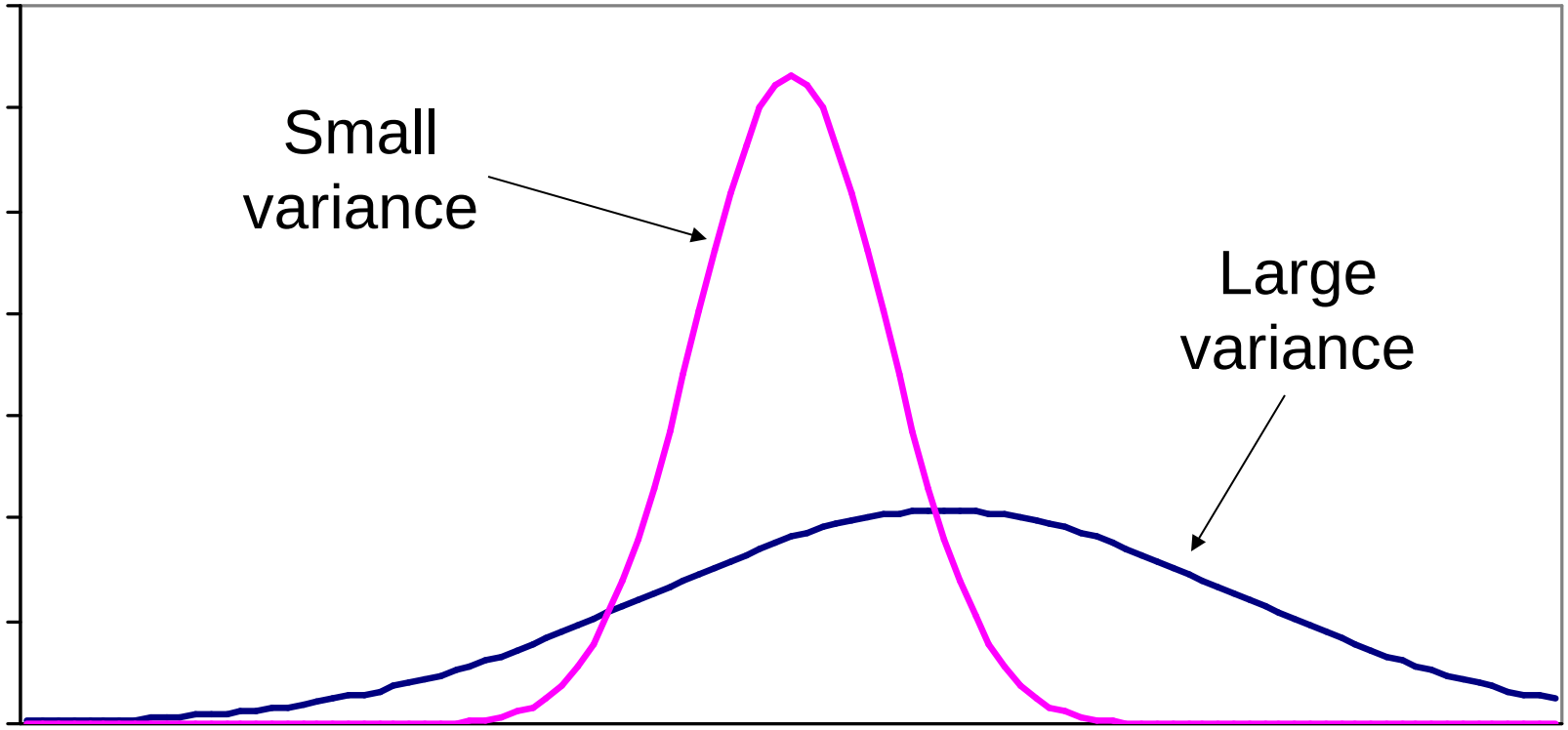
The variance

- The variance is the average of all squared deviations from the mean:

$$\sigma^2 = \frac{\sum f(x-\mu)^2}{\sum f}$$

- The larger this value, the greater the dispersion of the observations

The variance (continued)



Calculation of the variance

Range	midpoint	f	$x-\mu$	$(x-\mu)^2$	$f(x-\mu)^2$
0	5.0	2,448	- 142.0	20,159.38	49,350,158.77
10,000	17.5	1,823	- 129.5	16,766.04	30,564,482.57
25,000	32.5	1,375	- 114.5	13,106.52	18,021,469.99
40,000	45.0	480	- 102.0	10,400.68	4,992,326.62
50,000	55.0	665	- 92.0	8,461.01	5,626,568.95
60,000	70.0	1,315	- 77.0	5,926.49	7,793,339.80
80,000	90.0	1,640	- 57.0	3,247.15	5,325,317.93
100,000	125.0	2,151	- 22.0	483.28	1,039,544.38
150,000	175.0	2,215	28.0	784.91	1,738,579.16
200,000	250.0	1,856	103.0	10,612.35	19,696,526.45
300,000	400.0	1,057	253.0	64,017.23	67,666,217.05
500,000	750.0	439	603.0	363,628.63	159,632,966.88
1,000,000	1500.0	122	1,353.0	1,830,653.04	223,339,670.45
2,000,000	3000.0	50	2,853.0	8,139,701.86	406,985,092.85
Totals		17,636			1,001,772,261.83

$$\sigma^2 = \frac{\sum f(x-\mu)^2}{\sum f} = \frac{1{,}001{,}772{,}261.83}{17{,}636} = 56{,}802.69$$

The standard deviation

- The variance is measured ‘squared £s’ (because we used squared deviations)
- Hence take the square root to get back to £s
  This gives the standard deviation

$$\sigma = \sqrt{56{,}802.69} = 238.333$$

- or £238,333

Sample measures

- For sample data, use

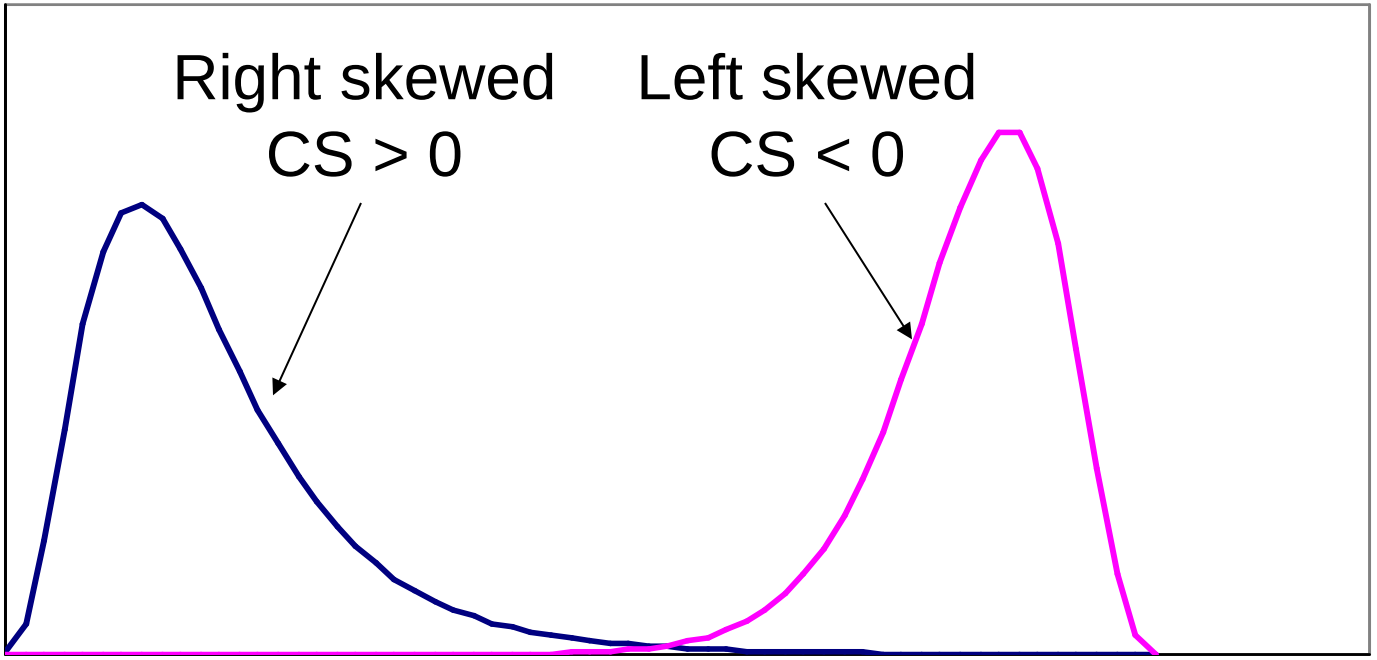
$$s^2 = \frac{\sum f(x - \bar{x})^2}{n - 1}$$

To calculate the sample variance

- This gives an unbiased estimate of the population variance

- Take the square root of this for the sample standard deviation

Measuring skewness



$$Coefficient\ of\ skewness = \frac{\sum f(x-\mu)^2}{N\sigma^3}$$

Skew of the wealth distribution

Range	x	f	$x-\mu$	$(x-\mu)^3$	$f(x-\mu)^3$
0	5.0	2,448	-142.0	-2,862,304	-7,006,919,444
10,000	17.5	1,823	-129.5	-2,170,929	-3,957,603,101
25,000	32.5	1,375	-114.5	-1,500,484	-2,063,165,040
40,000	45.0	480	-102.0	-1,060,700	-509,136,073
50,000	55.0	665	-92.0	-778,275	-517,552,779
60,000	70.0	1,315	-77.0	-456,244	-599,960,339
80,000	90.0	1,640	-57.0	-185,034	-303,456,461
100,000	125.0	2,151	-22.0	-10,624	-22,853,059
150,000	175.0	2,215	28.0	21,990	48,708,509
200,000	250.0	1,856	103.0	1,093,245	2,029,062,756
300,000	400.0	1,057	253.0	16,197,402	17,120,654,081
500,000	750.0	439	603.0	219,273,979	96,261,276,819
1,000,000	1500.0	122	1,353.0	2,476,903,349	302,182,208,638
2,000,000	3000.0	50	2,853.0	23,222,701,860	1,161,135,092,991
Total		17,636	4,457.2	25,927,167,232	1,563,796,357,499

$$\frac{\sum f(x-\mu)^3}{N\sigma^3} = \frac{1{,}563{,}796{,}357{,}499}{17{,}636 \times 13{,}537{,}964} = 6.550$$

Summary

- We can use graphical and numerical measures to summarise data
- The aim is to simplify without distorting the message
- Measures of location, dispersion and skewness provide a good description of the data