Data Analytics

ECON 1008, Semester 1, 2019

Giulio Zanella
University of Adelaide
School of Economics

Chapter 4

Graphical descriptive techniques – Numerical data

Recall: three types of data

1. Numerical data

The values of **numerical** data are *real numbers*.

E.g. grades, heights, weights, prices, etc.

2. Nominal Data

The values of **nominal** data are *categories*.

E.g. Responses to questions about marital status are categories,
Single = 1, Married = 2, Divorced = 3, Widowed = 4

3. Ordinal Data

Ordinal data: categorical in nature, values have an ***order***:

E.g. University course evaluation system:

Poor = 1, fair = 2, good = 3, very good = 4, excellent = 5

Describing numerical data

There are several graphical methods that are used when the data are ***numerical*** (quantitative).

The most important of these graphical methods is the ***histogram***, which we have already met.

The histogram is

- a powerful graphical technique used to ***summarise*** numerical data
- useful to work with probabilities.

Example 1

(Example 4.1, page 85)

As part of a larger study, an electricity provider wanted to acquire information about the monthly electricity bills of new subscribers in the first month after signing with the company.

The company's marketing manager conducted a survey of 200 new residential subscribers wherein the first month's bills were recorded.

The general manager planned to present his findings to senior executives. What information can be extracted from these data?

Building a Histogram…

Step 1. Look at the range of the variable in your sample,

Largest value = $470.50
Smallest value = $59.50

and define class intervals, e.g., 50-100, 100-150, 150-200…

How do you choose this? Either by "natural appeal" or by choosing a number of classes:

$$\frac{\text{Largest value-Smallest value}}{\text{Number of classes}} = \frac{470.50\text{-}59.50}{9} = \frac{411}{9} = 45.67$$

Building a Histogram…

Step 2. construct the frequency distribution

Table 4.2 Frequency distribution of electricity bills for 200 new Brisbane customers

Class limit*	Tally	Frequency
50 up to 100	𝍸|||	8
100 up to 150	𝍸 𝍸 𝍸 𝍸 ||||	24
150 up to 200	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |	36
200 up to 250	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	60
250 up to 300	𝍸 𝍸 𝍸 𝍸 𝍸 |||	28
300 up to 350	𝍸 𝍸 𝍸 |	16
350 up to 400	𝍸 𝍸	10
400 up to 450	𝍸 |||	8
450 up to 500	𝍸 𝍸	10
Total		200

* Classes contain observations greater than their lower limits (except for the first class) and less than or equal to their upper limits.

Building a Histogram…

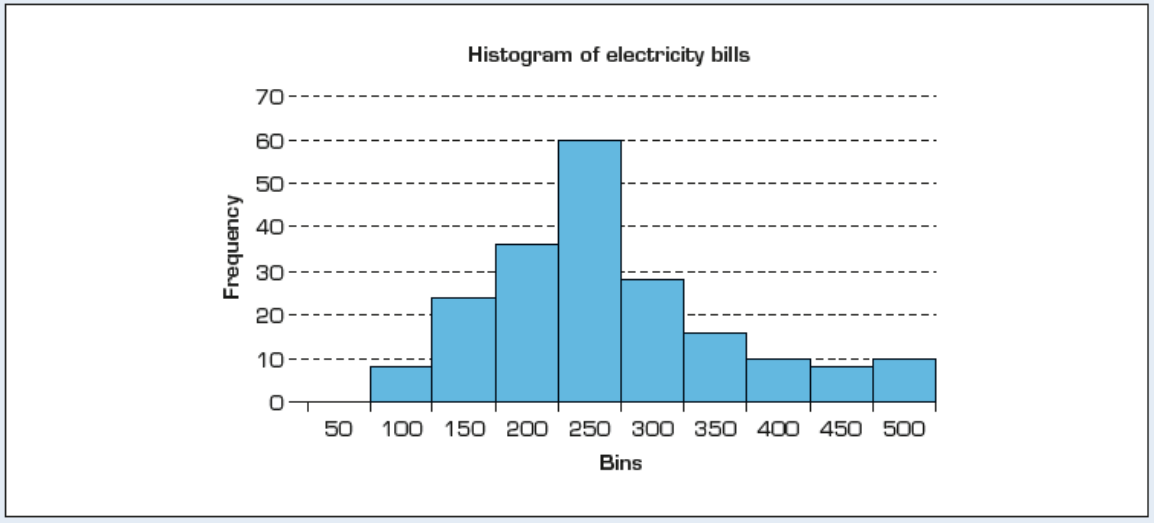
Step 3. Draw a histogram of rectangle bars using the class intervals and the corresponding frequencies.

Table 4.2 Frequency distribution of electricity bills for 200 new Brisbane customers

Class limit*	Tally	Frequency
50 up to 100	𝍸 |||	8
100 up to 150	𝍸 𝍸 𝍸 𝍸 |||	24
150 up to 200	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |	36
200 up to 250	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	60
250 up to 300	𝍸 𝍸 𝍸 𝍸 𝍸 |||	28
300 up to 350	𝍸 𝍸 𝍸	16
350 up to 400	𝍸 𝍸	10
400 up to 450	𝍸 |||	8
450 up to 500	𝍸 𝍸	10
Total		200

* Classes contain observations greater than their lower limits (except for the first class) and less than or equal to their upper limits.

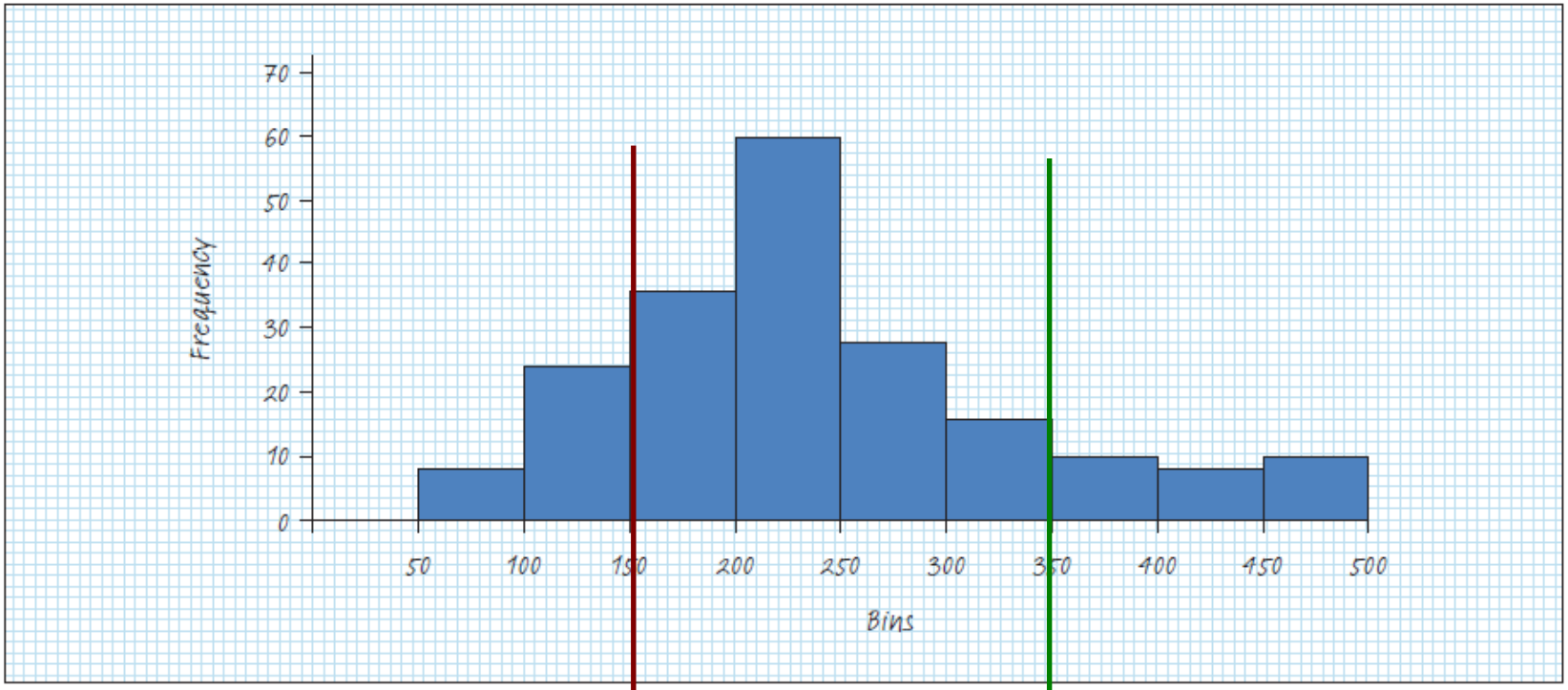
Excel output for Example 4.1



Building a Histogram…

Step 4. Stare at the picture, use pocket calc and interpret

Figure 4.1 Histogram of electricity bills of 200 new Brisbane customers



32 of the bills are 'small', i.e. less than $150.

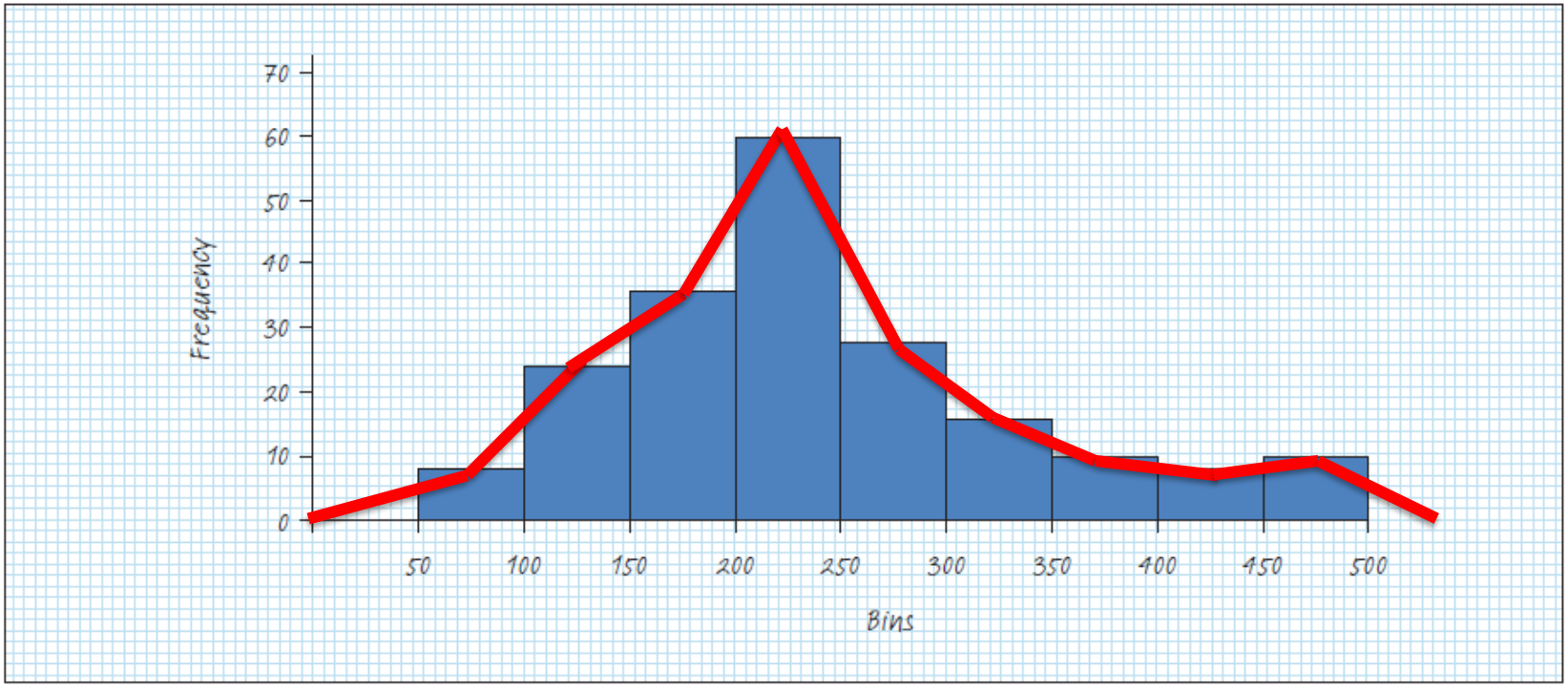
140 of the Electricity bills are in the 'middle range', i.e. between $150 and $350

28 of the electricity bills are 'large' i.e., $350 or more.

Frequency Polygon

By joining the midpoints of each bar with a straight line you can construct a **frequency polygon**, which is a first tool to **smooth the data.**

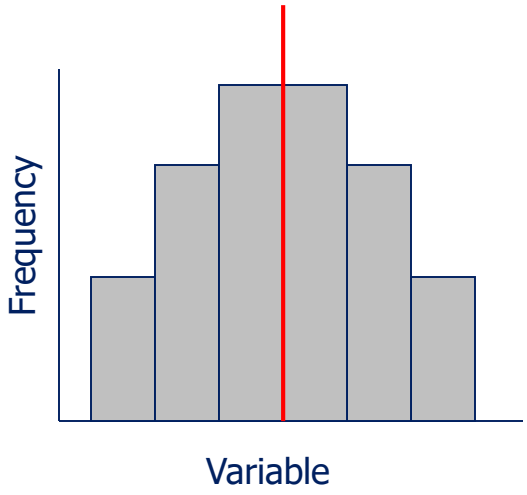
Figure 4.1 Histogram of electricity bills of 200 new Brisbane customers

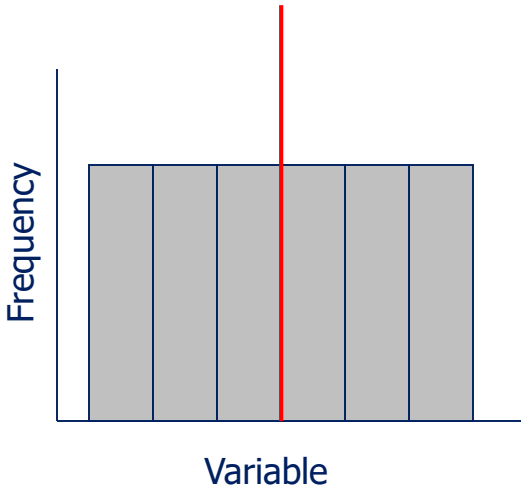


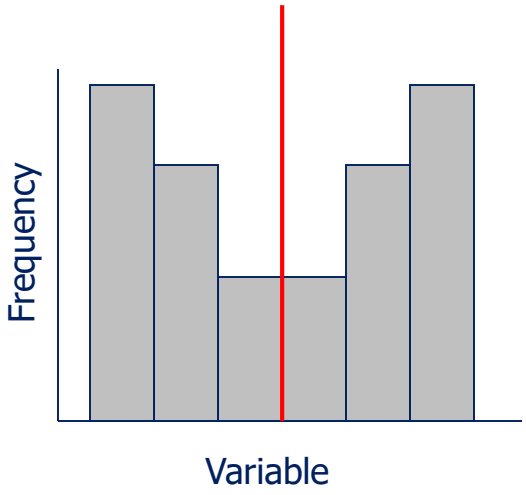
Shapes of Histograms…

Symmetry

A histogram is said to be ***symmetric*** if, when we draw a **vertical line** down the center of the histogram, the two sides are identical in shape and size:







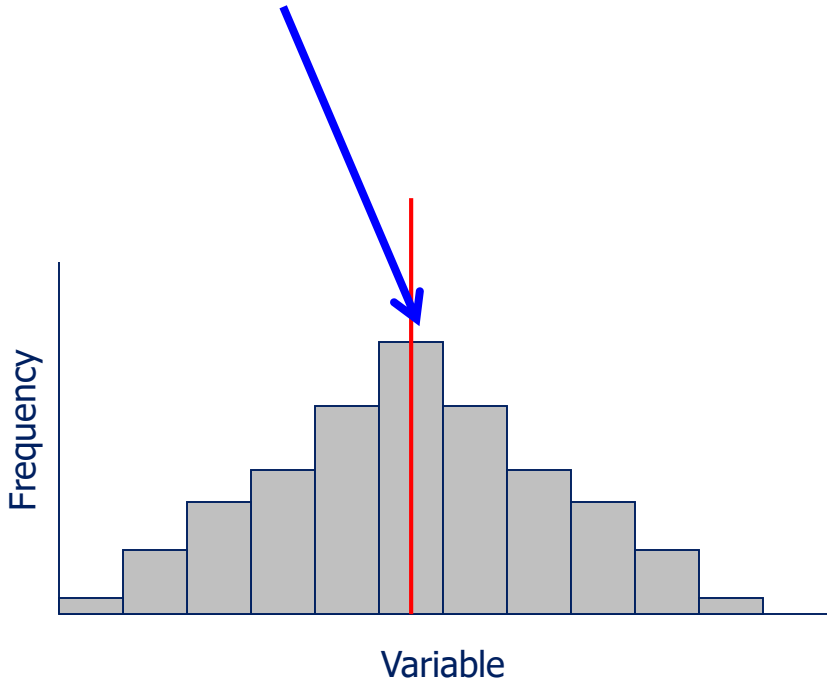
Shapes of Histograms…

Bell Shape

A special type of ***symmetric*** ***unimodal*** histogram is one that is bell shaped:

Many statistical techniques require that the distribution of the population be bell-shaped.

Drawing the histogram helps verify the shape of the population distribution in question.

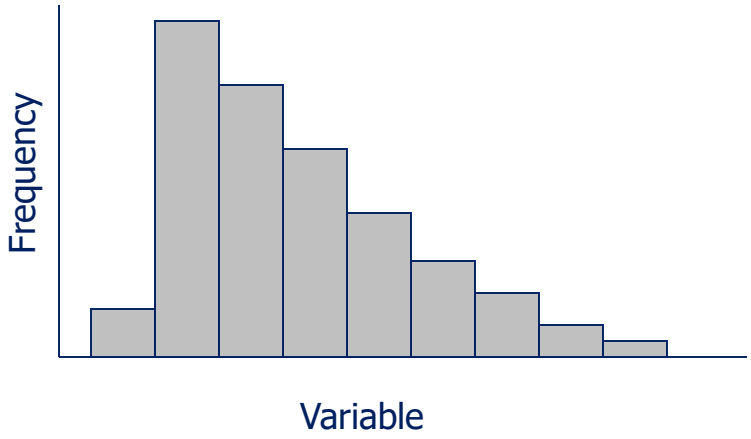


Bell Shaped

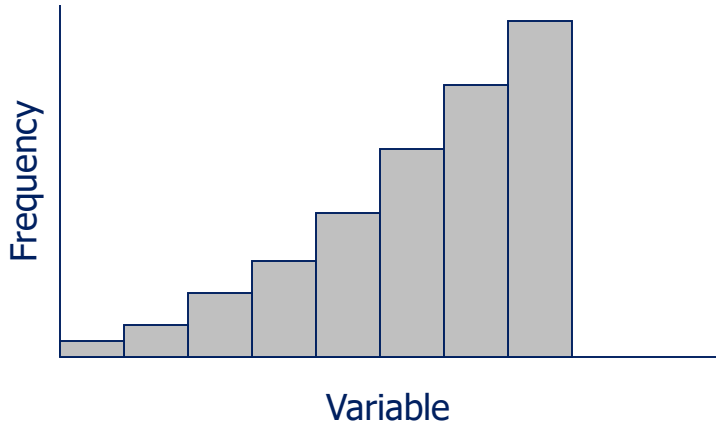
Shapes of Histograms…

Skewness

A skewed histogram is one with a long tail extending either to the right or to the left:



Positively (right) skewed

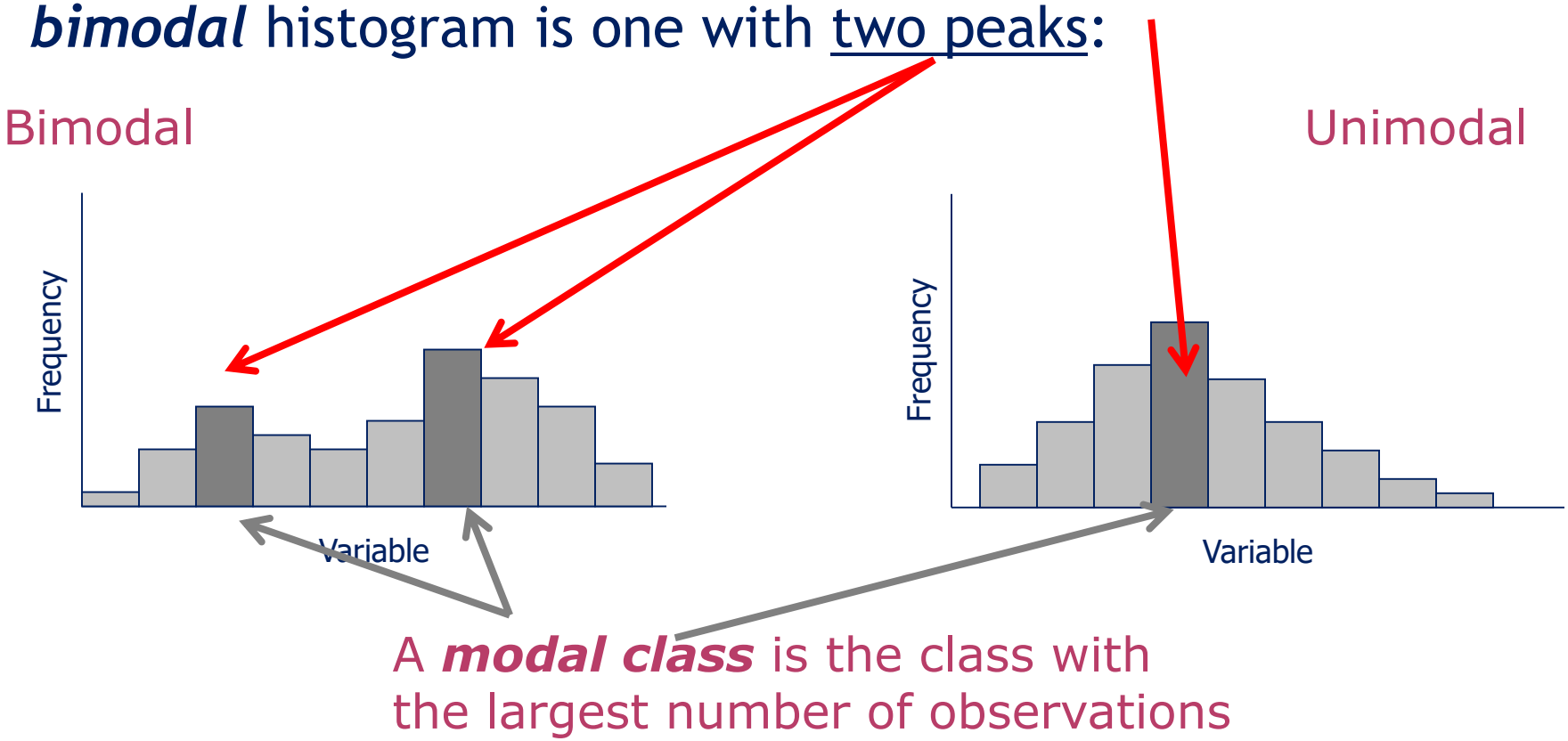


Negatively (left) skewed

Shapes of Histograms…

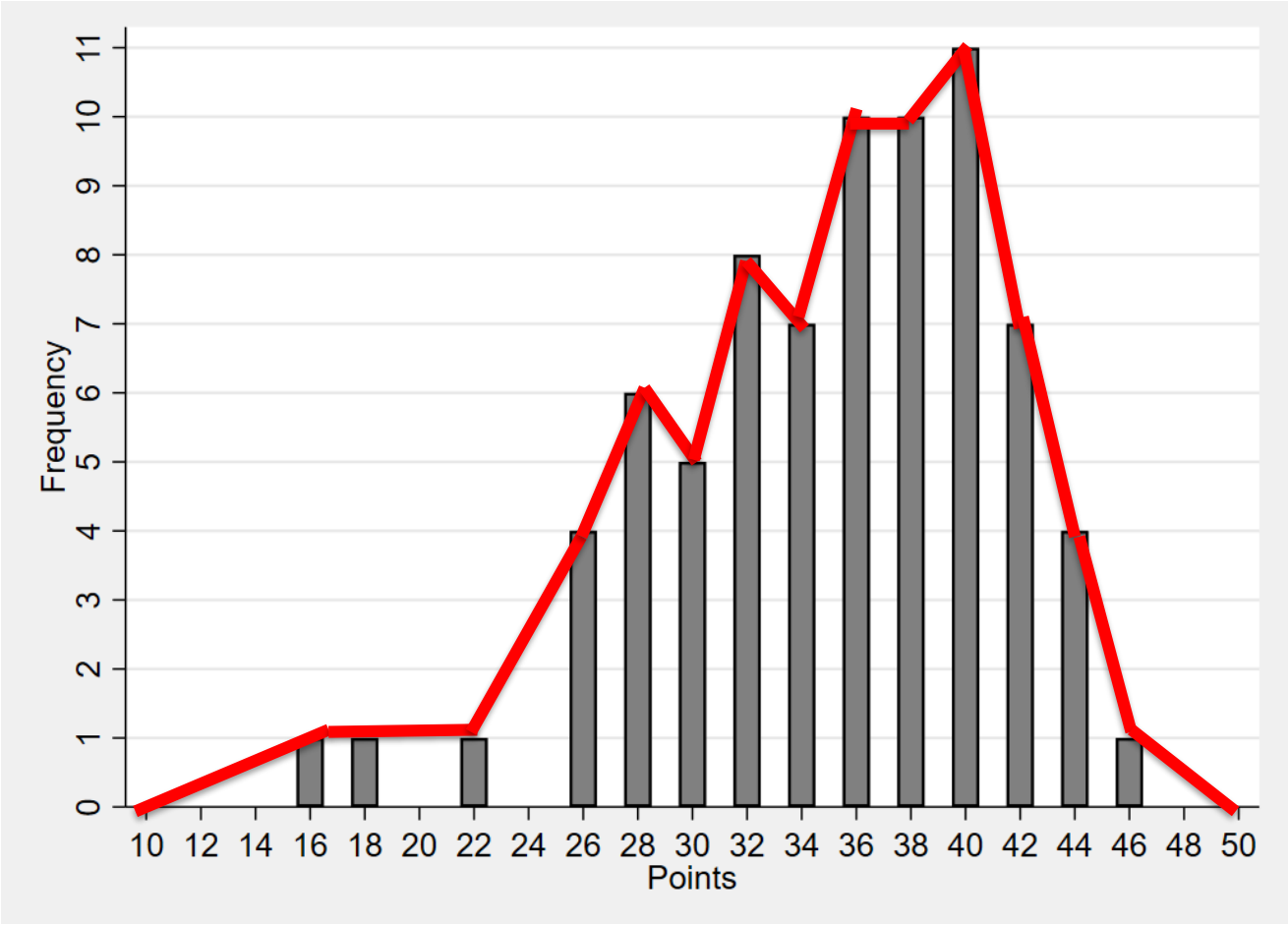
Modality

A ***unimodal*** histogram is one with a single peak, while a ***bimodal*** histogram is one with two peaks:



Shapes of Histograms…

Let's apply these notions to an old histogram (Week 1)



Comparison of Histograms…

Compare and contrast the following histograms based on data from Example 4.3:

Final marks in a statistics course before and after a change of emphasis:

- before: emphasise mathematical proofs, derivations, and manual calculations [manual]

- After: emphasise concepts, ideas, and let a computer do the calculations [computer]

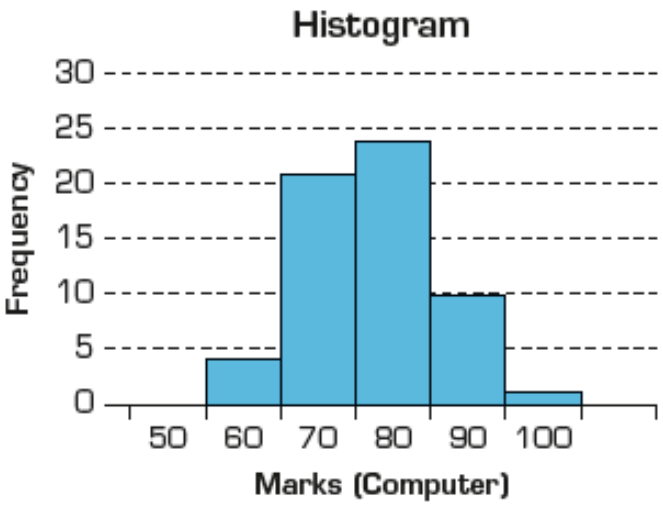
Comparison of Histograms…

Compare and contrast the following histograms based on data from Example 4.3:

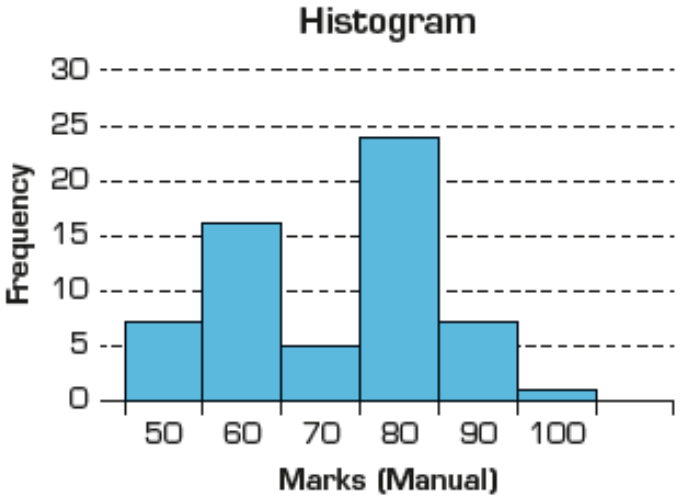
The marks from the computer-based statistics course and the manual statistics course have very different histograms…

Unimodal vs. bimodal

Marks (computer course)



Marks (manual course)



Spread of the marks (narrower | wider)

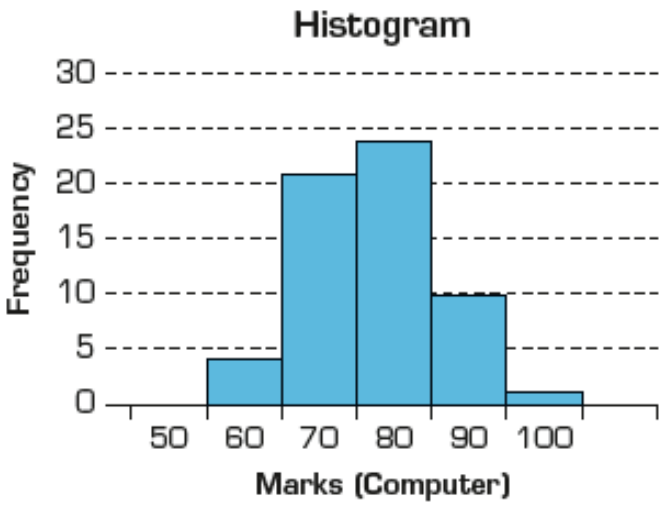
Relative frequency

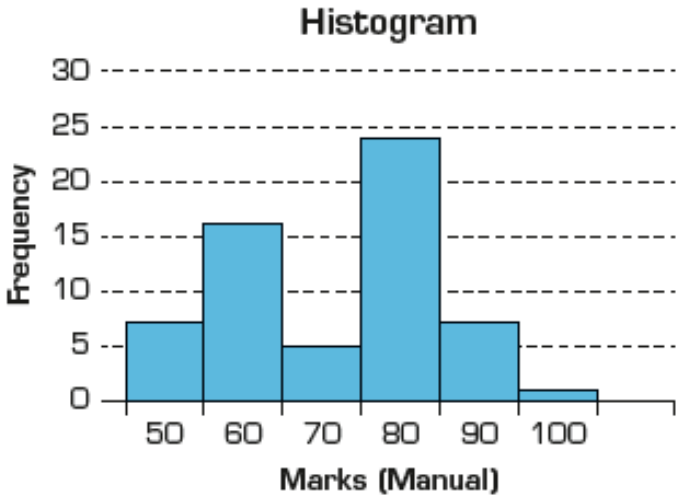
It is often preferable to show the **relative frequency** (proportion) of observations falling into each class, rather than the **absolute frequency (count)** itself.

$$\text{Class relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}$$

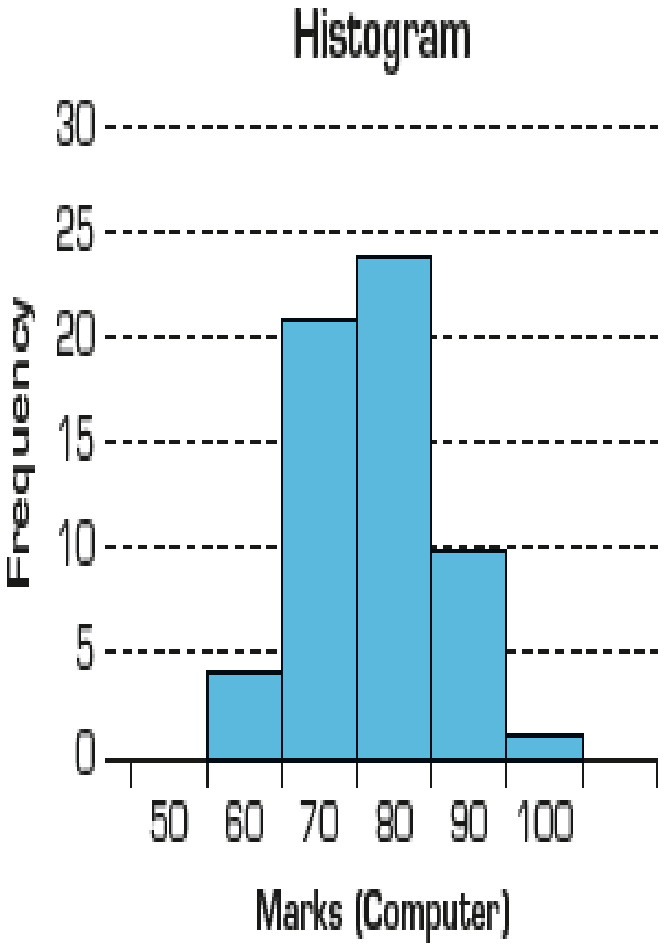
Use relative frequencies when comparing two or more histograms with different numbers of observations. **Why?**

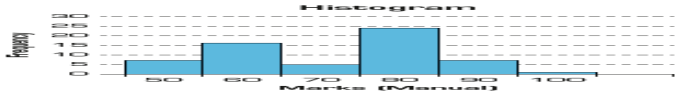
Relative frequencies…





Relative frequencies…





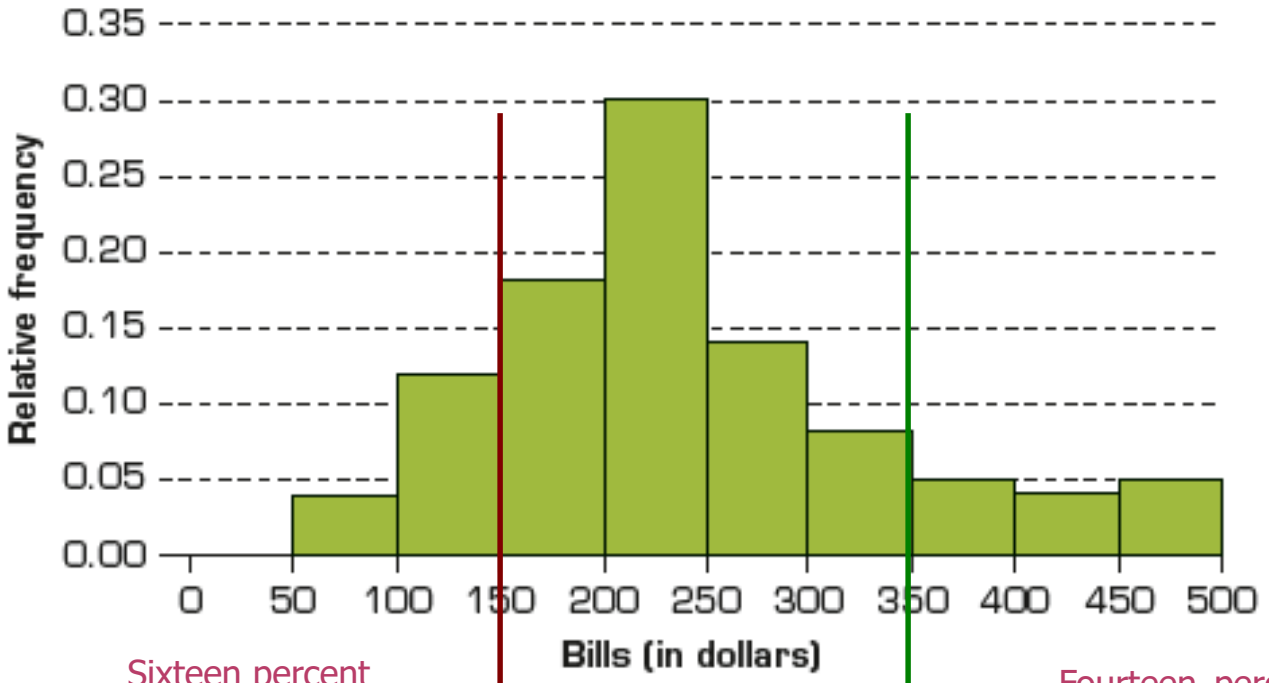
Relative frequencies…

Back to Example 1…

Class limits	Relative frequency
50 up to 100	8/200 = 0.04
100 up to 150	24/200 = 0.12
150 up to 200	36/200 = 0.18
200 up to 250	60/200 = 0.30
250 up to 300	28/200 = 0.14
300 up to 350	16/200 = 0.08
350 up to 400	10/200 = 0.05
400 up to 450	8/200 = 0.04
450 up to 500	10/200 = 0.05
	Total = 1.00

Relative frequencies...

4.2 Relative frequency histogram of electricity bills with equal class width



Sixteen percent of the bills are less than $150.

Seventy percent of the electricity bills are between $150 and $350.

Fourteen percent of the electricity bills are $350 or more.

Cumulative frequency of a class

Cumulative frequency of a class is the number of measurements less than the upper limit of that class.

To obtain the cumulative frequency of a class, we add the frequency of that class with the frequencies of all previous classes.

The **cumulative relative frequency** of a particular class is the proportion of measurements that are less than the upper limit of that class.

Easier to show in an example than to tell in words…

Cumulative frequency of a class

Table 4.8 Cumulative relative frequencies for Example 4.1

Classes	Frequency	Cumulative frequency	Relative frequency	Cumulative relative frequency
50 up to 100	8	8	8/200 = 0.04	8/200 = 0.04
100 up to 150	24	8 + 24 = 32	24/200 = 0.12	32/200 = 0.16
150 up to 200	36	8 + 24 + 36 = 68	36/200 = 0.18	68/200 = 0.34
200 up to 250	60	128	60/200 = 0.30	128/200 = 0.64
250 up to 300	28	156	28/200 = 0.14	156/200 = 0.78
300 up to 350	16	172	16/200 = 0.08	172/200 = 0.86
350 up to 400	10	182	10/200 = 0.05	182/200 = 0.91
400 up to 450	8	190	8/200 = 0.04	190/200 = 0.95
450 up to 500	10	200	10/200 = 0.05	200/200 = 1.00

First class...

Next class: 0.04+0.12=0.16

:

:

Last class: 0.95.+0.05=1.00

This transformation of the data leads to a key notion: **Cumulative Distribution Function (CDF)**, a.k.a. Ogive

Cumulative Distribution Function (CDF)

CDF is a graph of a ***cumulative relative frequency distribution***. We create a CDF in three steps…

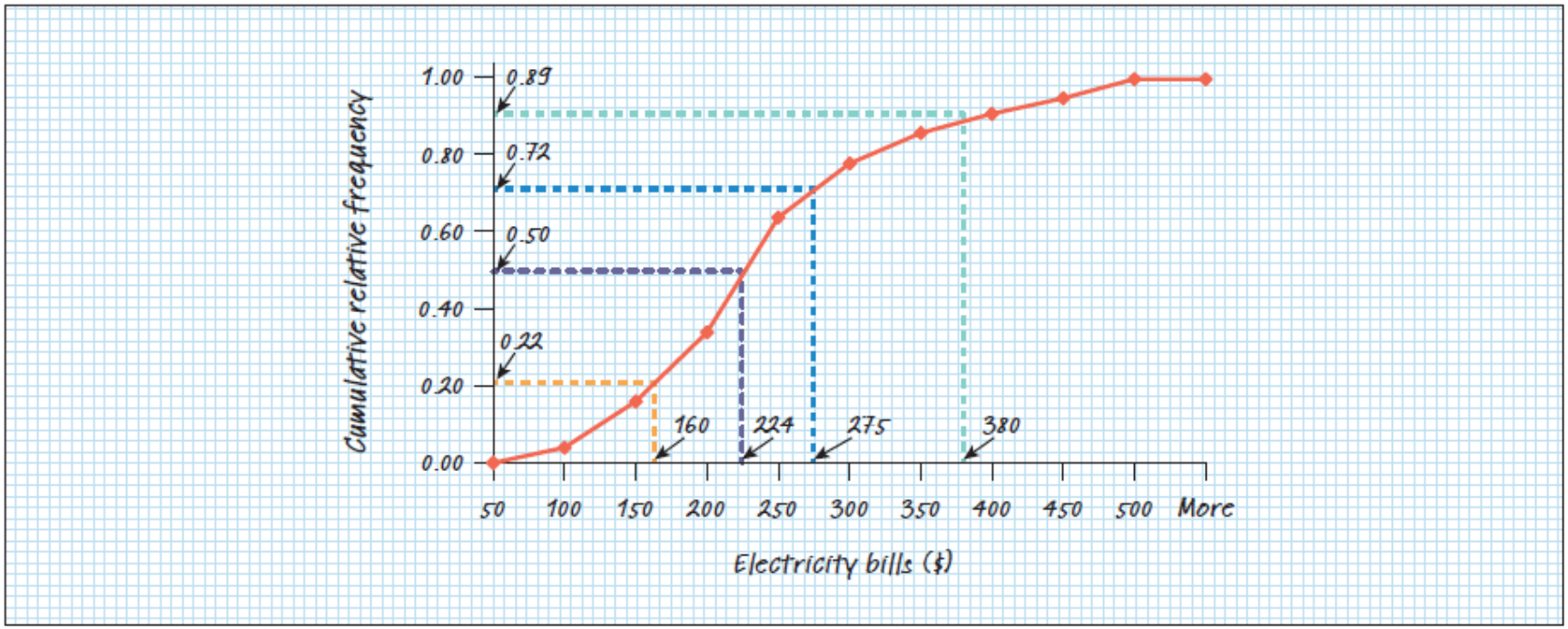
1. Calculate relative frequencies.

2. Calculate ***cumulative relative frequencies*** by summing the current and all previous relative frequencies.
   (of course for the first class the cumulative relative frequency is just its relative frequency.)

3. Graph the cumulative relative frequencies.

CDF...

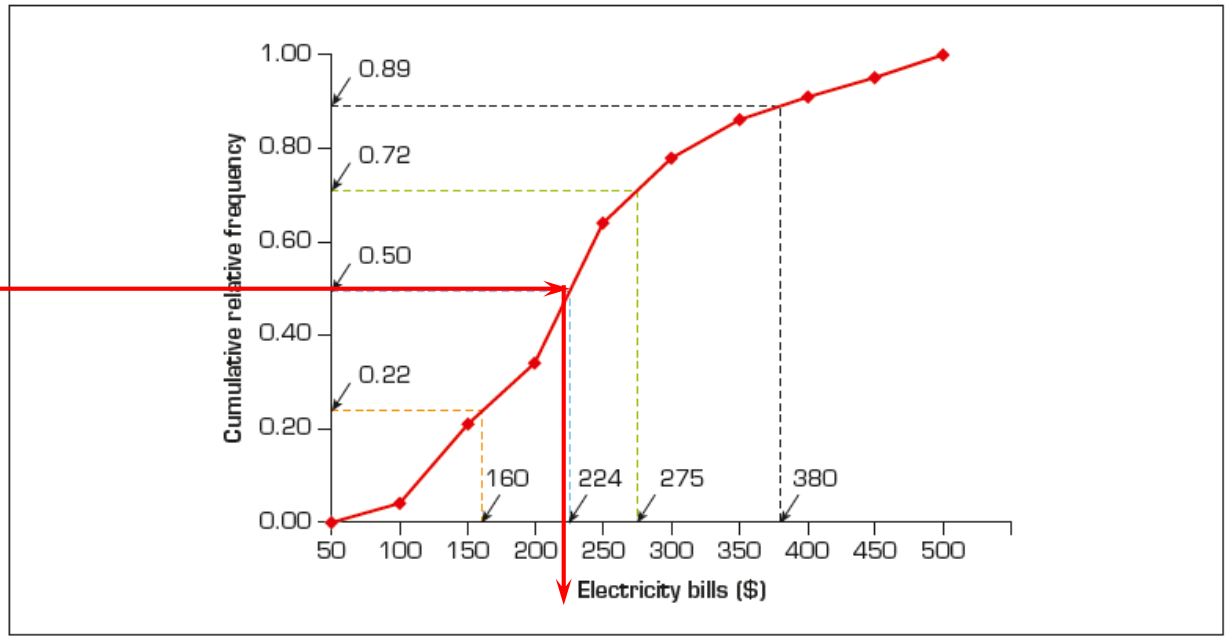
Figure 4.9 Ogive for electricity bills



CDF…

The CDF can be used to answer questions like:
What electricity bill value is at the median?

Figure 4.9 Ogive for electricity bills



We can estimate the electricity bill value that is at the median as approximately $224.

This is why the median is also called
The **50th percentile**.

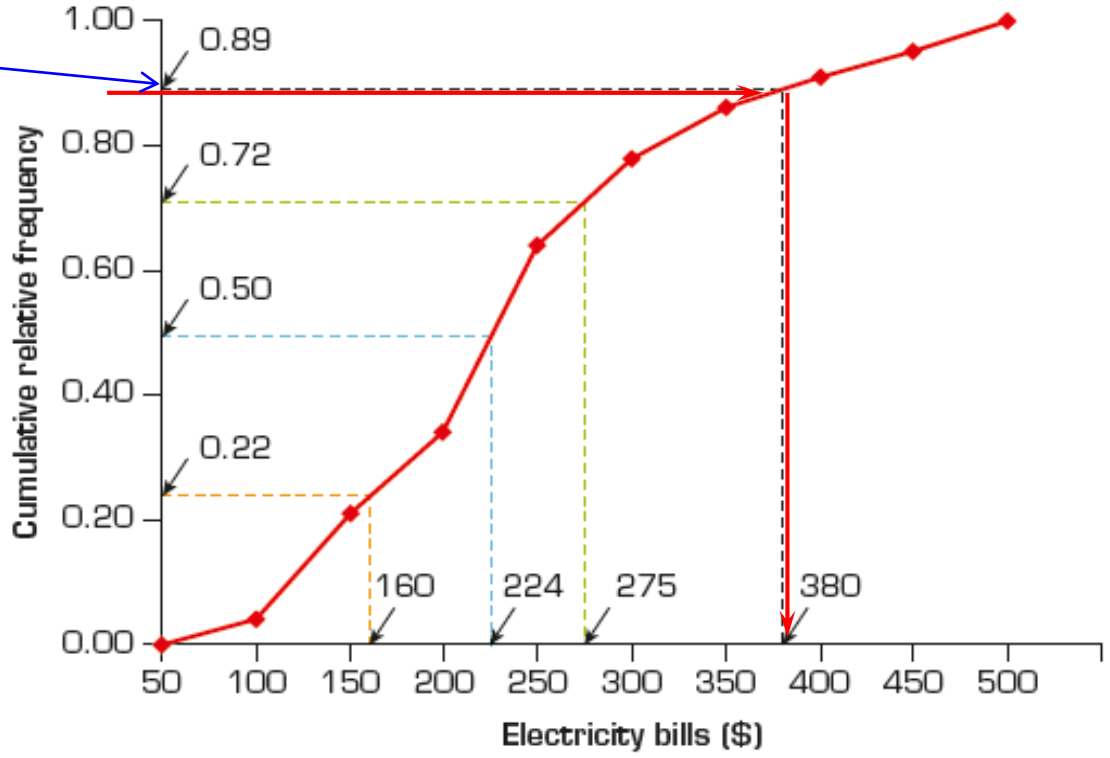
CDF...

What proportion of the electricity bills are less than $380?

around 89%

From the Ogive, we estimate the proportion of electricity bills that are:

- less than $380 is 89%
- greater than $380 is 11%
- less than $275 is 72%
- less than $160 is 22%
- less than $224 is 50%



Describing Time Series Data

Observations measured at the same point in time across individual units are called ***cross-sectional*** data.

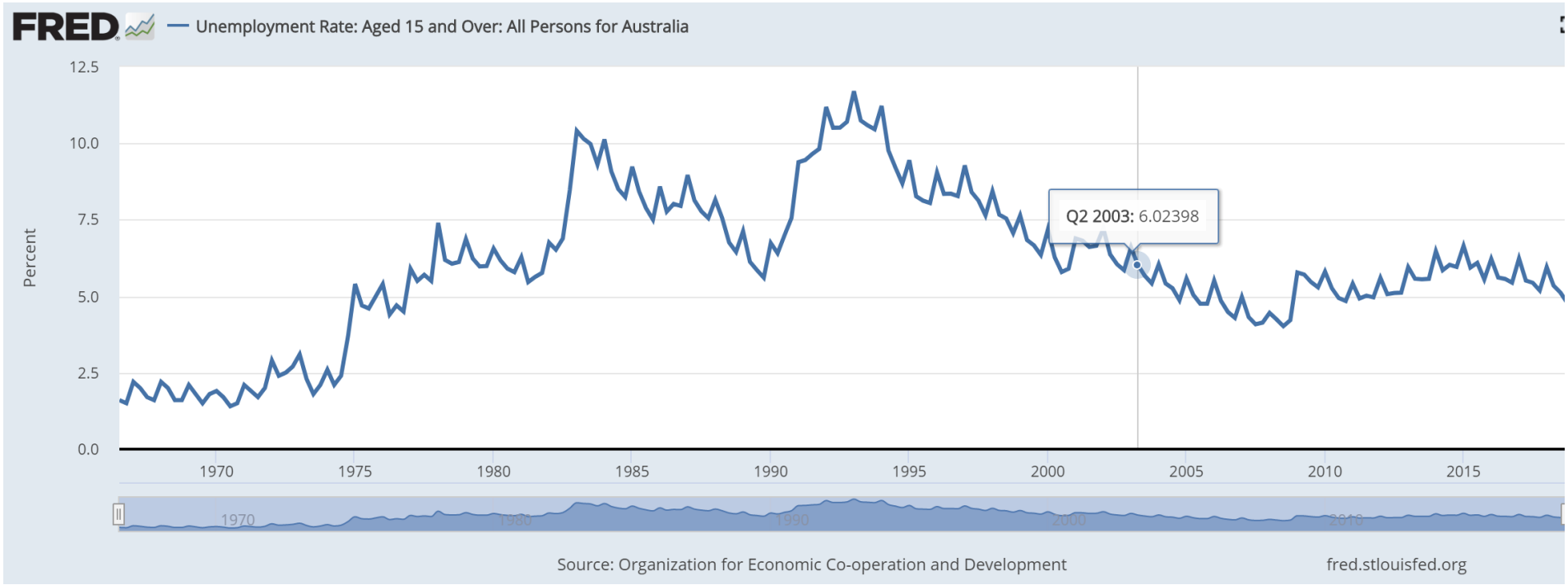
Observations measured at successive points in time on a single unit are called ***time-series*** data.

Time-series data are graphed on a ***line chart***, which plots the value of the variable on the vertical axis against the time periods on the horizontal axis.

Time series data graphed on a line chart is alternatively known as a ***time-series chart***.

Line Chart

Line chart showing Australia's unemployment rate over time



Describing two numerical variables

Often we are interested in the **relationship between two numerical variables.**

For example,

- Advertising and sales
- Rate of unemployment and rate of inflation
- Yield of crops and amount of fertilizer

Example 2

A small-business owner wants to assess the effects of advertising on sales levels.

Paired observation data were collected.

Each pair consisted of monthly advertising expenditure and monthly sales levels (both in millions of dollars).

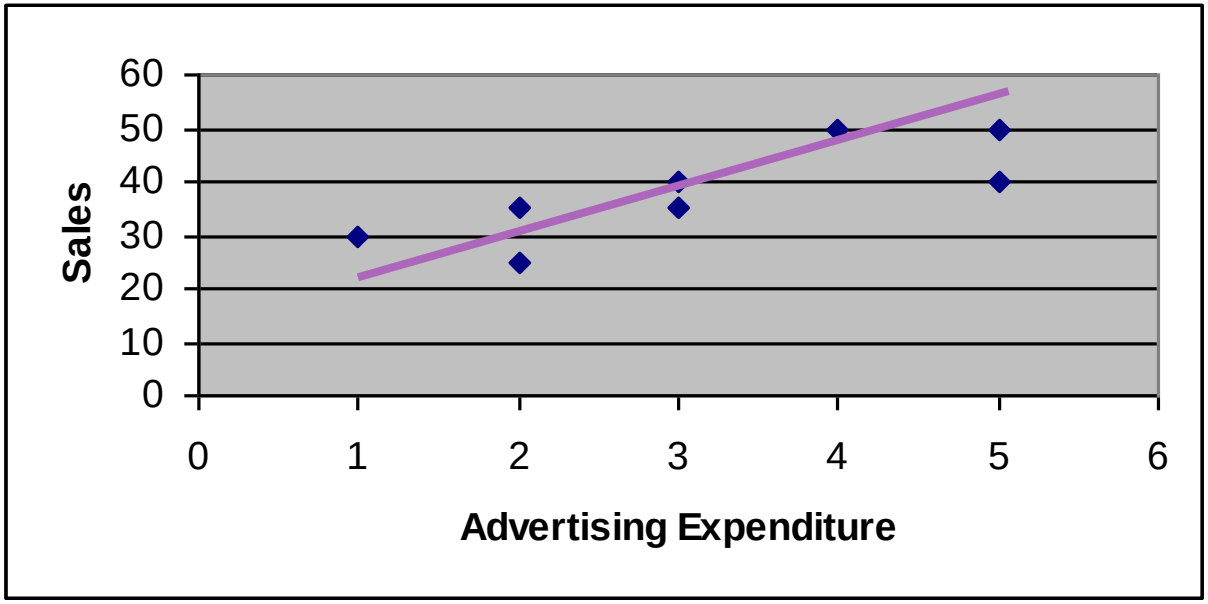
Advert	Sales
1	30
3	40
5	40
4	50
2	35
5	50
3	35
2	25

Scatter diagram

A scatter diagram can describe the relationship between advertising expenditure and sales.

Advert	Sales
1	30
3	40
5	40
4	50
2	35
5	50
3	35
2	25

Excel scatter diagram



Example 3

ARE OIL COMPANIES EXPLOITING CUSTOMERS?

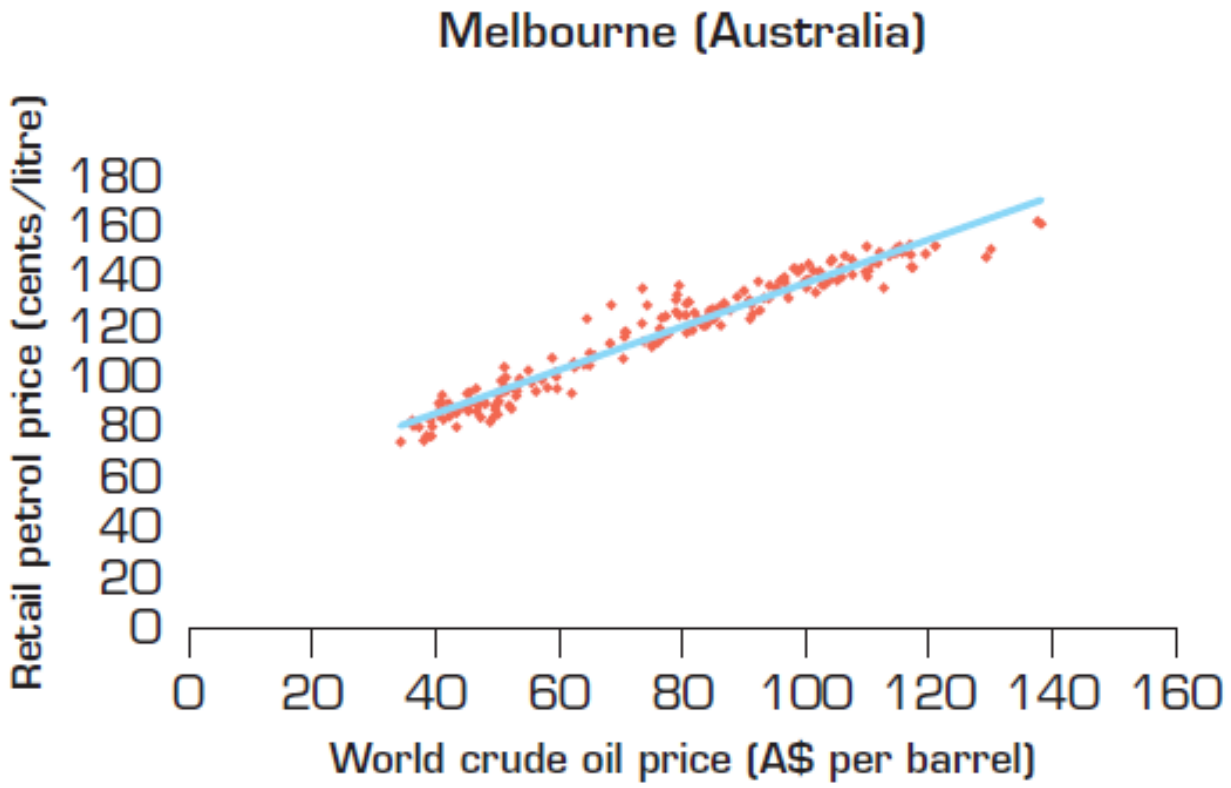
XM04-00 In October 1999, the average retail price of petrol was A$0.74 per litre in Melbourne and the price of oil (Dubai Fetch Crude) was US$34.06 per barrel (1 barrel = 159.18 litres).

Over the next 16 years, the price of both substantially increased. Many drivers complained that the oil companies were guilty of price gouging.

That is, they believed that when the price of oil increased the price of petrol also increased, but when the price of oil decreased, the decrease in the price of petrol seemed to lag behind.

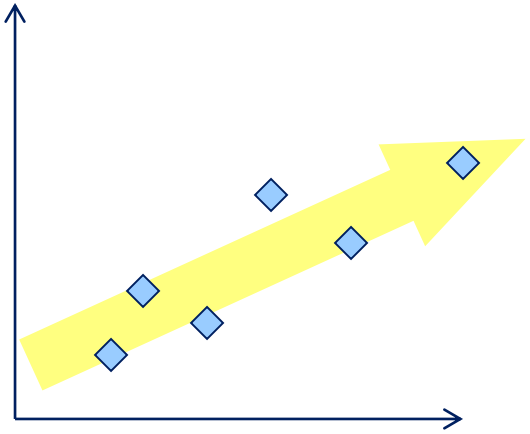
To determine whether this perception is accurate we determined the monthly figures for both commodities. Graphically depict these data and describe the findings.

Petrol prices in Australia vs World crude oil prices, October 1999 – July 2015

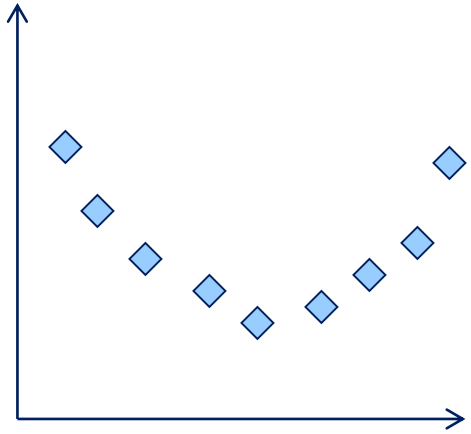


Patterns of Scatter Diagrams…

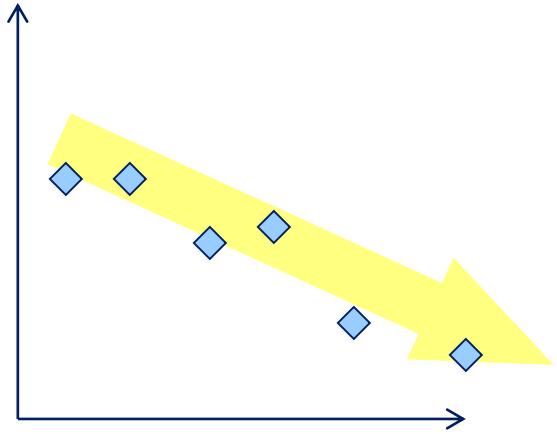
Linearity and direction are two concepts we are interested in.



Positive linear relationship



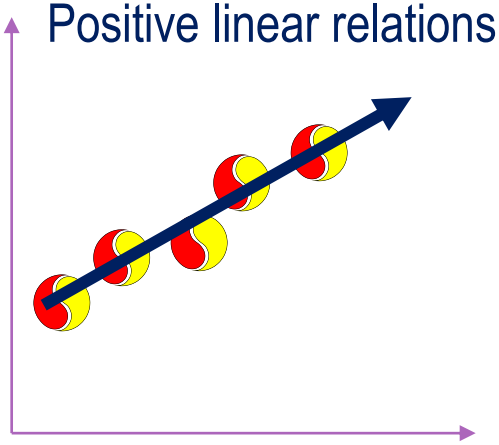
Weak or non-linear relationship



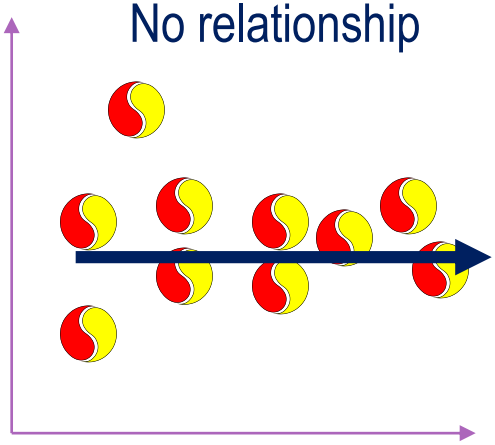
Negative linear relationship

Typical patterns

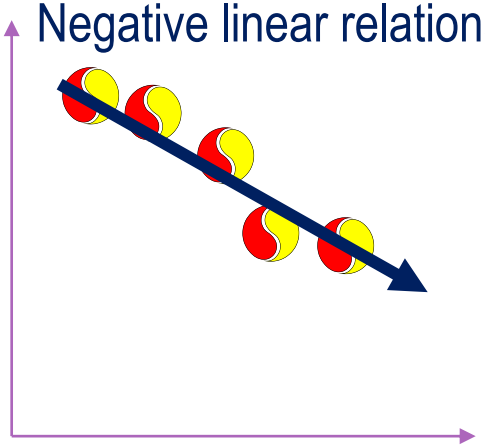
Positive linear relationship



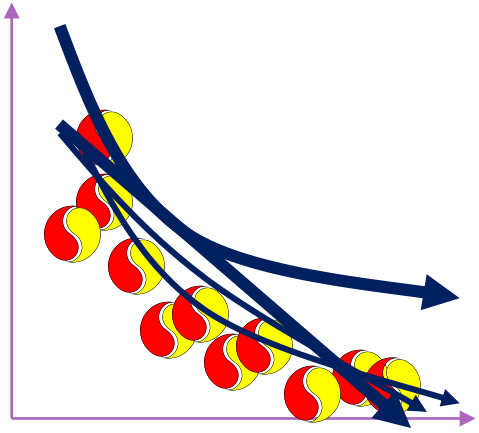
No relationship



Negative linear relationship



Negative nonlinear relationship



Nonlinear (concave) relationship