IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: MSc

Subject : Probability and Statistics

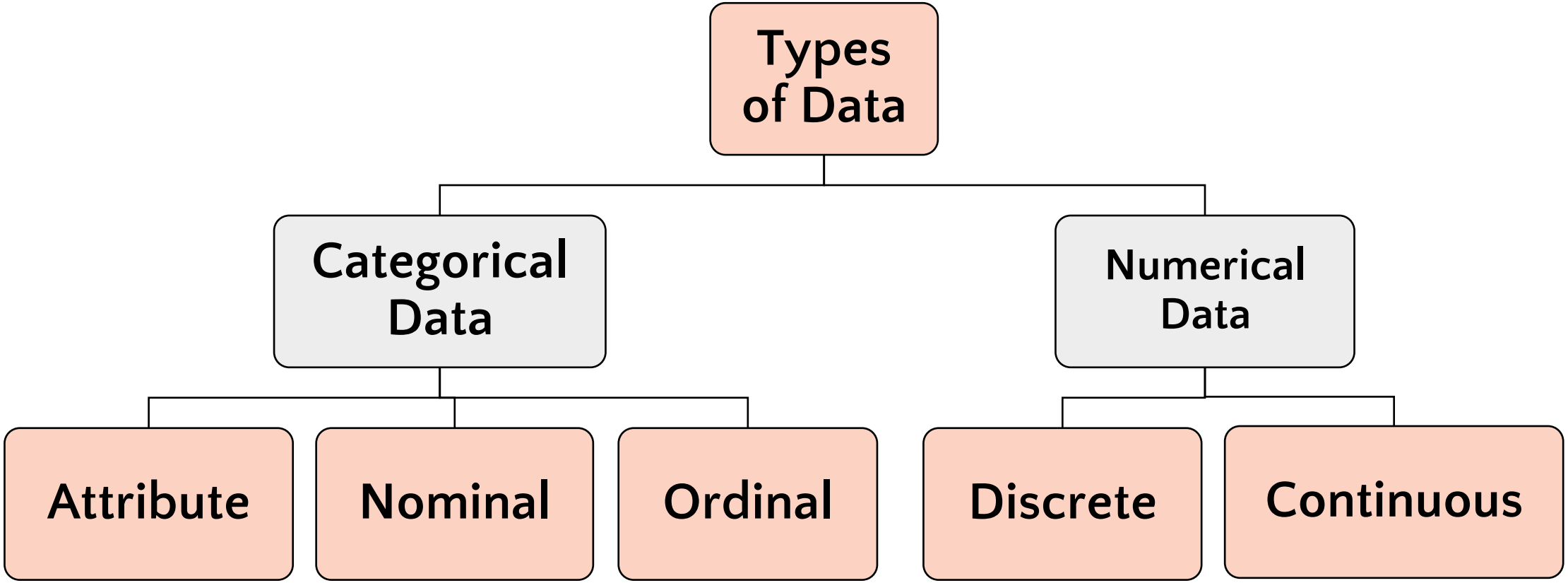
Chapter: Unit 1 Chapter 1

Chapter Name: Summarizing Data

Topics to be covered

1. Types of Data
2. Categorical Data
3. Numerical Data
4. Frequency Distribution
5. Representing Discrete Distribution
6. Dealing with Grouped Data
7. Histogram
8. How to make a histogram?
9. Stem & Leaf Diagram
10. Lineplots
11. Introducing Cumulative Frequency
12. Box and Whisker Plot

1 Types of Data



2 Categorical Data

- Categorical variables represent types of data which may be divided into groups. Examples of categorical variables are race, sex, age group, and educational level. While the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of groups.
- Categorical data types are attributes treated as distinct symbols or just names. The color of the iris of the human eye is a categorical data type because it takes a value like black, green, blue, gray, etc. There is no direct relationship among the data values, and hence, mathematical operators except the logical or “is equal” operator cannot be applied. They are also called a nominal or polynominal data type, derived from the Latin word for name.
- An ordered nominal data type is a special case of a categorical data type where there is some kind of order among the values. An example of an ordered data type is temperature expressed as hot, mild, cold.

2 Categorical Data

Attribute (or *dichotomous) data* have only two categories, *eg* yes/no, male/female, claim/no claim.

Nominal data have several unordered categories, eg type of policy, nature of claim.

Ordinal data have several ordered categories, eg questionnaire responses such as “strongly in favor / … / strongly against”.

3 Numerical Data

- Numerical data is a data type expressed in numbers, rather than natural language description. Sometimes called quantitative data, numerical data is always collected in number form. Numerical data differentiates itself from other number form data types with its ability to carry out arithmetic operations with these numbers.
- For example, numerical data of the number of male students and female students in a class may be taken, then added together to get the total number of students in the class. This characteristic is one of the major ways of identifying numerical data.
- Numerical data can take 2 different forms, namely; discrete data, which represents countable items and continuous data, which represents data measurement. The continuous type of numerical data is further sub-divided into interval and ratio data, which is known to be used for measuring items.

Categorical Data

- Categorical data represents groups or categories.
- Examples:
  1. Car brands: Audi, BMW and Mercedes.
  2. Answers to yes/no questions: yes and no

Numerical

- Numerical data represents numbers. It is divided into two groups: discrete and continuous. Discrete data can be usually counted in a finite matter, while continuous is infinite and impossible to count.

- Examples:
  Discrete: # children you want to have, SAT score
  Continuous: weight, height

Concept Checker

Answer the following dating agency questionnaire and state what type of data is required in each question:

1. *How old are you? (Give your age last birthday.)*
2. *How tall are you? (State as accurately as you can.)*
3. *What sex are you?*
4. *What color are your eyes?*
5. *Do you smoke?*
6. *How would you rate your looks? (10 =Drop-dead gorgeous, 1= Seen better days)*

4 Frequency Distribution

- The data from a discrete data set can be summarized using a frequency distribution, that is, by counting the number of 0's, 1's, 2's, etc.
- A frequency distribution gives a summary of the data from a discrete data set. For eg:

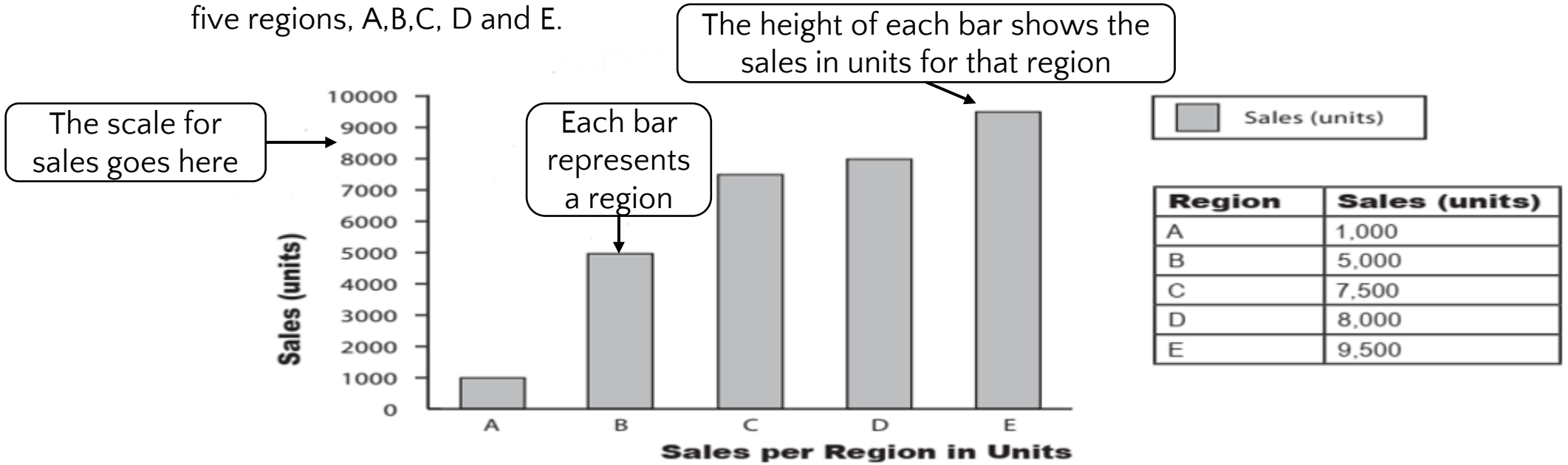
# of claims	# of policies
0	500
1	125
2	20

5 Representing Discrete Distribution

- A bar chart can be used to display a frequency distribution.

Vertical bar charts

- Vertical bar charts show categories on the horizontal axis and either frequency or percentage on the vertical axis. The height of each bar indicates the value of it's category. Here's an example showing the sales figures for five regions, A,B,C, D and E.



Region	Sales (units)
A	1,000
B	5,000
C	7,500
D	8,000
E	9,500

6 Dealing with grouped data

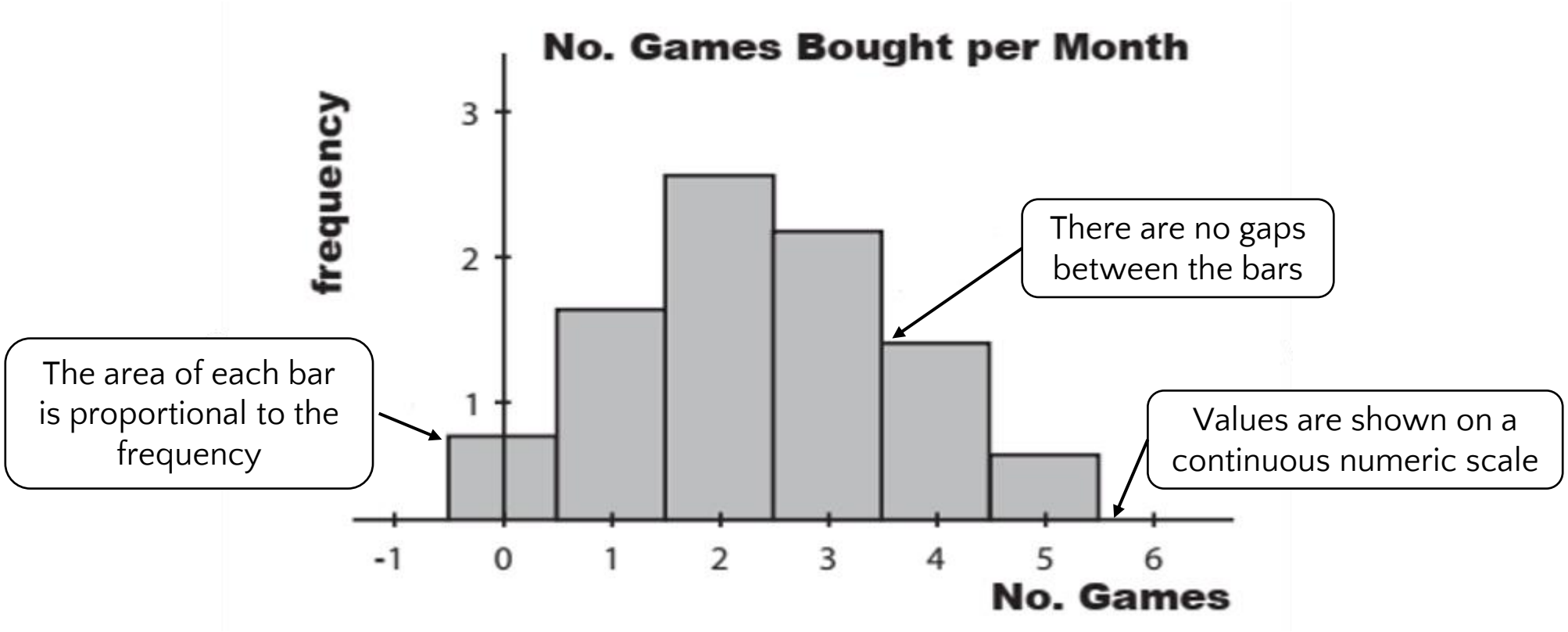
- When the set of data is numeric and, what's more, the scores are grouped into intervals. So what's the best way of charting data like this?

Score	Frequency
0-199	5
200-399	29
400-599	56
600-799	17
800-999	3

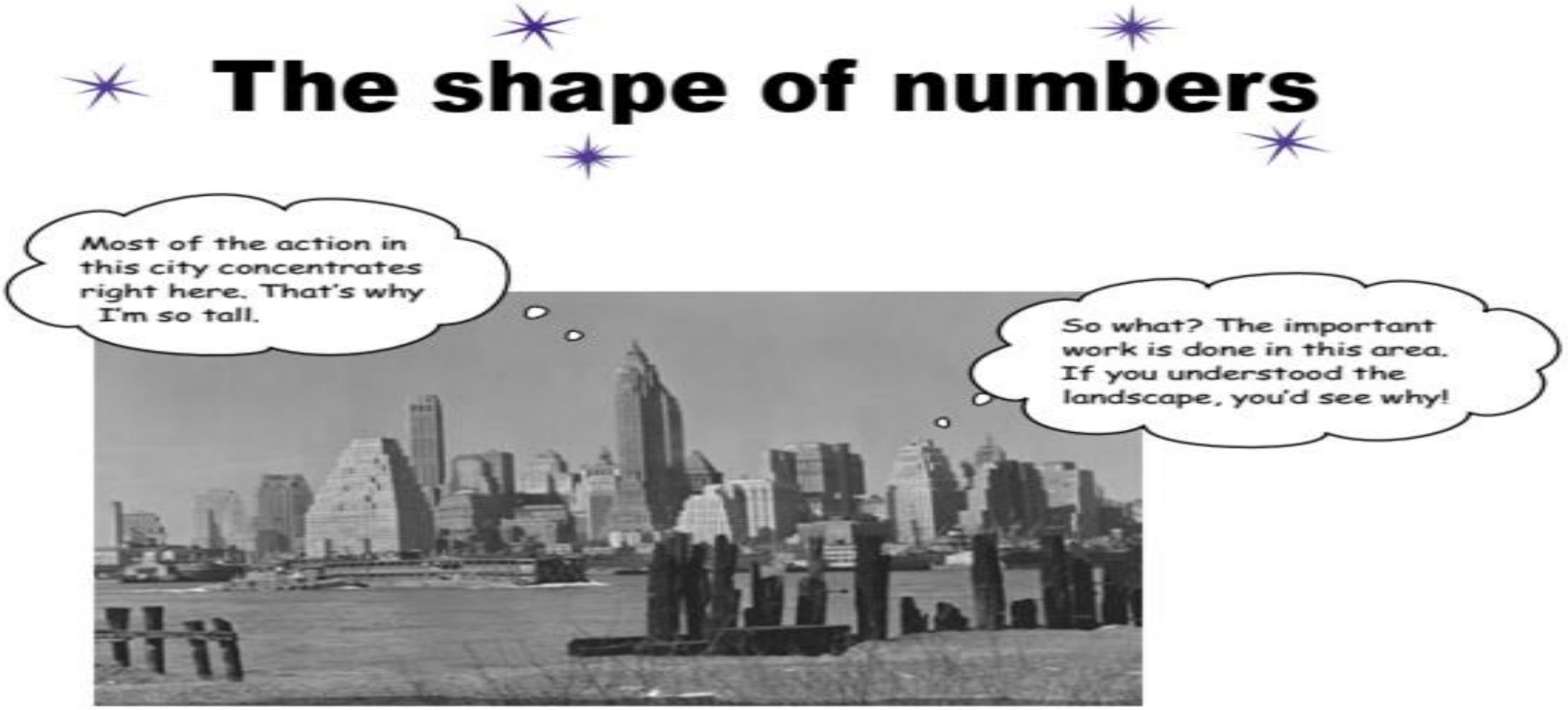
- Rather than treat each range of scores as a separate category, we can take advantage of the data being numeric and present the data using a continuous numeric scale instead
- This means that instead of using bars to represent a single item, we can use each bar to represent a range of scores.

7 Histogram

- To do this, we can create a histogram
- Histograms are like bar charts but with two key differences. The first is that the area of each bar is proportional to the frequency, and the second is that there are no gaps between the bars on the chart.



7 Histogram


The shape of numbers
Most of the action in this city concentrates right here. That's why I'm so tall.
So what? The important work is done in this area. If you understood the landscape, you'd see why!

8 How to make a Histogram?

- Step 1: To make a histogram, start by finding bar widths
- Step 2: Find the bar heights

$$Area\ of\ bar = Frequency\ of\ group$$
$$Frequency = Width\ of\ bar \times Height\ of\ bar$$
$$Height\ of\ Bar = \frac{Frequency}{Width\ of\ Bar}$$

- Step 3: Draw your chart – a histogram

This is the number of hours people play for

This is frequency with which people play for this length of time

Hours	Frequency
0–1	4,300
1–3	6,900
3–5	4,900
5–10	2,000
10–24	2,100

8 How to make a Histogram?

Step 1: Find Bar Widths

- We find how wide our bars need to be by looking at the range of values they cover. In other words, we need to figure out how many full hours are covered by each group.

Hours	Frequency	Width
0-1	4,300	1
1-3	6,900	2
3-5	4,900	2
5-10	2,000	5
10-24	2,100	14

- Let's take the 1–3 group. This group covers 2 full hours, 1–2 and 2–3.
- This means that the width of the bar needs to be 2, with boundaries of 1 and 3.

8 How to make a Histogram?

Step 2: Find Bar Heights

- Now that we have the widths of all the groups, we can use these to find the heights the bars need to be. Remember, we need to adjust the bar heights so that the overall area of each bar is proportional to the group's frequency.

$$\text{Area of bar} = \text{Frequency of group}$$

$$\text{Frequency} = \text{Width of bar} \times \text{Height of bar}$$

$$\text{Height of Bar} = \frac{\text{Frequency}}{\text{Width of bar}}$$

Concept Checker

- What should the height of each bar be? Complete the Table

Hours	Frequency	Width	Height (Frequency Density)
0–1	4,300	1	4,300 ÷ 1 = 4,300
1–3	6,900	2	
3–5	4,900	2	
5–10	2,000	5	
10–24	2,100	14	

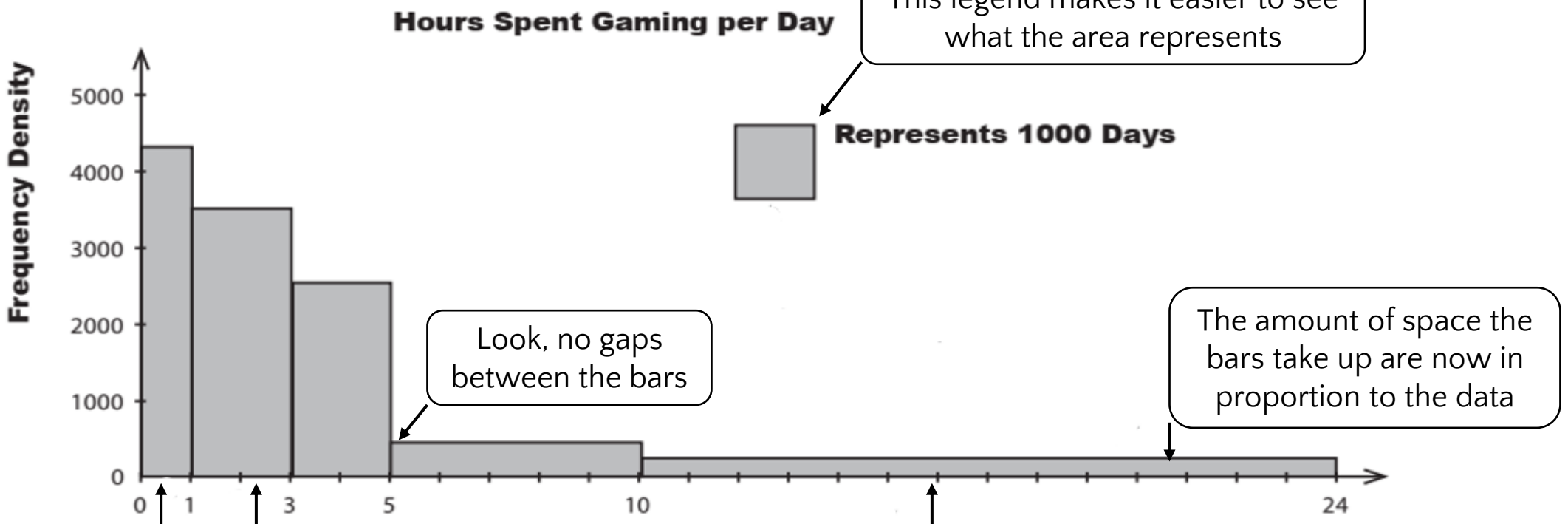
Concept Checker

Solution:

Hours	Frequency	Width	Height (Frequency Density)
0–1	4,300	1	4,300 ÷ 1 = 4,300
1–3	6,900	2	6,900 ÷ 2 = 3,450
3–5	4,900	2	4,900 ÷ 2 = 2,450
5–10	2,000	5	2,000 ÷ 5 = 400
10–24	2100	14	2,100 ÷ 14 = 150

8 How to make a Histogram?

Step 3: Draw your chart—a histogram



This legend makes it easier to see what the area represents

Look, no gaps between the bars

The amount of space the bars take up are now in proportion to the data

Even though the bar for 1-3 has a higher frequency, the frequency density is actually lower than for 0-1

The bars are drawn on the continuous numeric scale

7 Histograms: Summary

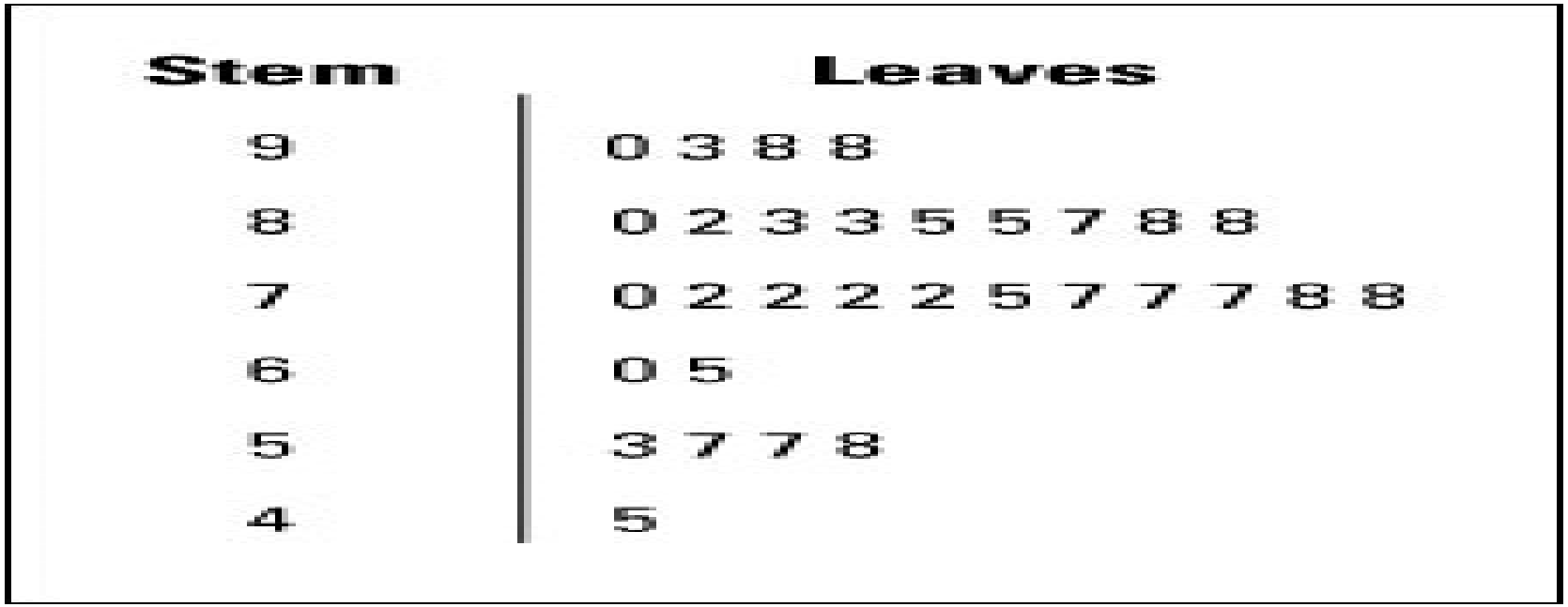
- **Frequency density** relates to how concentrated the frequencies are for grouped data. It's calculated using:

$$Frequency\ density = \frac{Frequency}{Group\ width}$$

- A histogram is a chart that specializes in grouped data. It looks like a bar chart but the height of each bar equates to frequency density rather than frequency.
- When drawing histograms, the width of each bar is proportional to the width of it's group. The bars are shown on a continuous numeric scale.
- In a histogram, the frequency of a group is given by the area of it's bar.
- A histogram has no gaps between it's bars.

9 Stem & Leaf Diagram

- Stem-and-leaf displays organize data so that an entire distribution of scores is quickly and easily comprehensible. The display breaks each score into two components: a leaf, which is usually the last digit of the score, and a stem, which is everything else. The objective is to create a layout that looks like this:



Stem	Leaves
9	0 3 8 8
8	0 2 3 3 5 5 7 8 8
7	0 2 2 2 2 5 7 7 7 8 8
6	0 5
5	3 7 7 8
4	5

9 Stem & Leaf Diagram

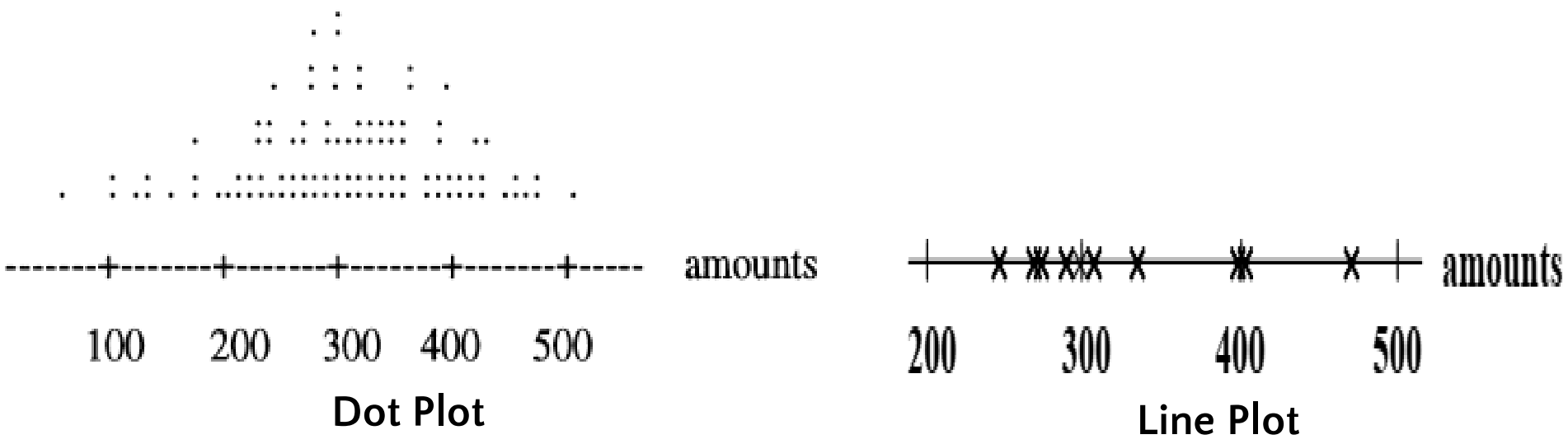
The stems are on the left with units of 10 and the leaves are on the right with units of 1.

An alternative to the histogram is the stem and leaf display.

It gives a visual representation similar to the histogram but does not lose the detail of the individual data points in the grouping.

10 Lineplots

- For smaller data sets another alternative diagram is the dotplot or lineplot.
- In a lineplot, the data points are plotted as “dots” or “crosses” along a line with a scale.
- Here is a computer generated dotplot for the water leakage claim amount data.



11 Introducing cumulative frequency

- The cumulative frequency of a value is the sum of the frequencies up to and including that value. It tells you the total frequency up to that point. As an example, suppose you have data telling you how old people are. The cumulative frequency for value 27 tells you how many people there are up to and including age 27.

Hours	Frequency
0–1	4,300
1–3	6,900
3–5	4,900
5–10	2,000
10–24	2,100

11 Introducing cumulative frequency

- So what are the cumulative frequencies?

Hours	Frequency	Upper limit	Cumulative frequency
0	0	0	0
0–1	4,300	1	4,300
1–3	6,900	3	4,300+6,900 = 11,200
3–5	4,900	5	4,300+6,900+4,900 = 16,100
5–10	2,000	10	4,300+6,900+4,900+2,000 = 18,100
10–24	2,100	24	4,300+6,900+4,900+2,000+2,100 = 20,200

11 Introducing cumulative frequency

Drawing the cumulative frequency graph:

You plot this point by plotting 10 hrs against a cumulative frequency of 18,100

Always start the chart where the cumulative frequency is 0

- If your cumulative frequency decreases at any point, check your answers.
- The cumulative frequency table or diagram is commonly used to find the median or interquartile range.

11 Introducing cumulative frequency

- Cumulative frequency is the total frequency up to a particular value. It is a running total of the frequencies.
- Use a cumulative frequency graph to plot the upper limit of each group of data against cumulative frequency.
- Use a line chart if you want to show trends, for example, over time.
- You can show more than one set of data on a line chart. Use one line for each set of data and make sure it's clear which line is which.
- You can use line charts to make basic predictions as it's easy to see the shape of the trend. Just extend the trend line, trying to keep the same basic shape.
- Don't use line charts to show categorical data unless you're showing trends for each category. For example, over time. If you do this, draw one line per category.

12 Box and Whisker Plot

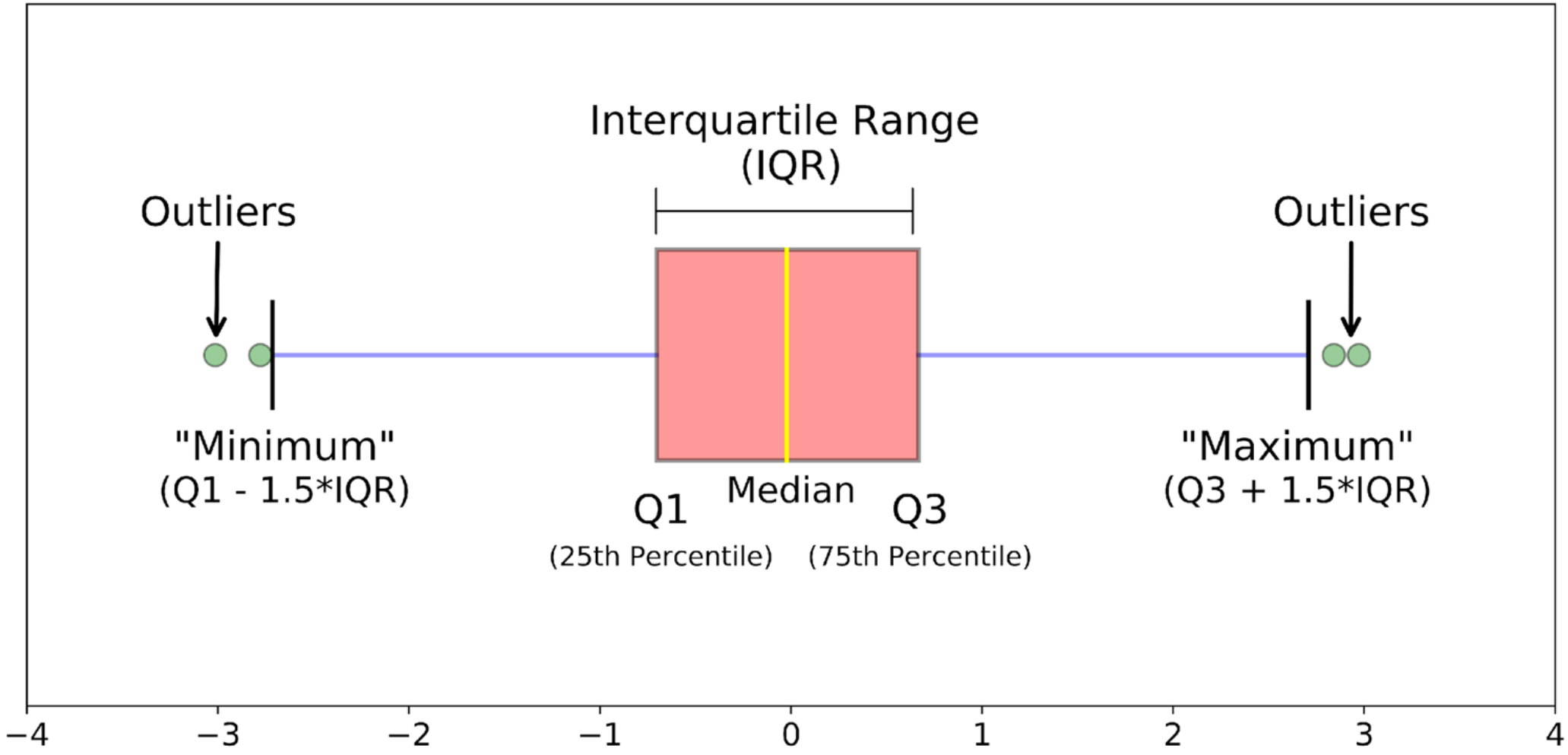
Box and whisker plots let you visualize ranges.

We've talked a lot about different sorts of ranges, and it would be useful to be able to compare the ranges of different sets of data in a visual way.

There's a chart that specializes in showing different types of ranges: the ***box and whisker*** diagram, or ***box plot***.

A box and whisker diagram shows the range, interquartile range, and median of a set of data.

12 Box and Whisker Plot



12 Box and Whisker Plot

- If your data has outliers, the range will be wider.
- On a box and whisker diagram, the length of the whiskers increases in line with the upper and lower bounds.
- You can get an idea of how data is skewed by looking at the whiskers on the box and whisker diagram.
- If the box and whisker diagram is symmetric, this means that the underlying data is likely to be fairly symmetric, too.

Summary

- Numerical data can be discrete or continuous
- Categorical data can be dichotomous (attribute), nominal or ordinal.
- Data can be presented either in tabular form (using a frequency table, a cumulative frequency table or a stem and leaf diagram) or in graphical form (using a *line-plot, a dot-plot, a boxplot, a bar chart or a histogram*).