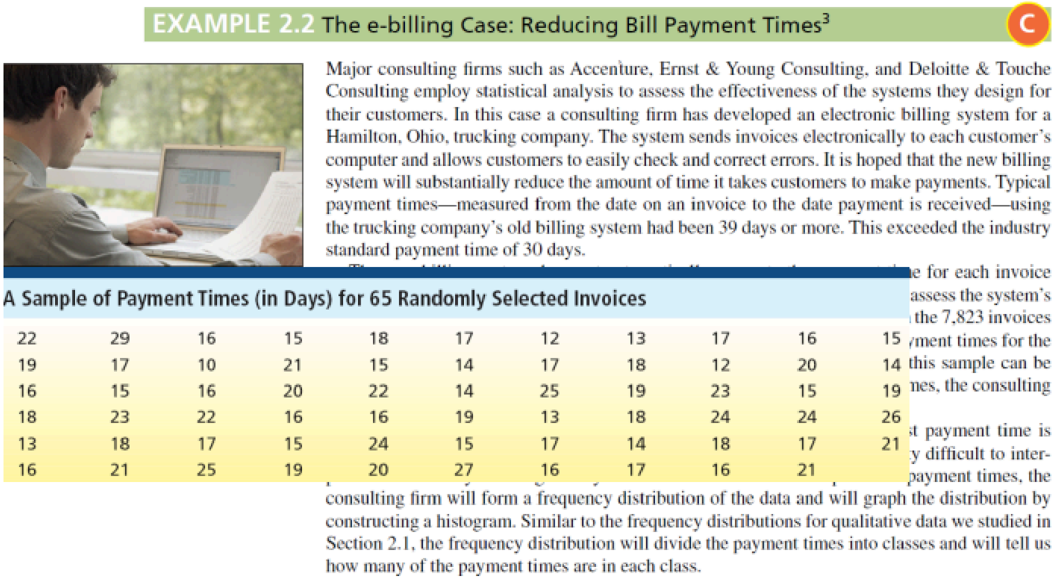
Chapter 2 : Descriptive statistics

Summary of Chapter

2.2 Graphically Summarizing Quantitative Data

Example 2.2 The e-Billing Case



EXAMPLE 2.2 The e-billing Case: Reducing Bill Payment Times[3] C

Major consulting firms such as Accenture, Ernst & Young Consulting, and Deloitte & Touche Consulting employ statistical analysis to assess the effectiveness of the systems they design for their customers. In this case a consulting firm has developed an electronic billing system for a Hamilton, Ohio, trucking company. The system sends invoices electronically to each customer's computer and allows customers to easily check and correct errors. It is hoped that the new billing system will substantially reduce the amount of time it takes customers to make payments. Typical payment times—measured from the date on an invoice to the date payment is received—using the trucking company's old billing system had been 39 days or more. This exceeded the industry standard payment time of 30 days.

e for each invoice
assess the system's
the 7,823 invoices
ment times for the
this sample can be
nes, the consulting

t payment time is
y difficult to inter-
payment times, the consulting firm will form a frequency distribution of the data and will graph the distribution by constructing a histogram. Similar to the frequency distributions for qualitative data we studied in Section 2.1, the frequency distribution will divide the payment times into classes and will tell us how many of the payment times are in each class.

A Sample of Payment Times (in Days) for 65 Randomly Selected Invoices

22	29	16	15	18	17	12	13	17	16	15
19	17	10	21	15	14	17	18	12	20	14
16	15	16	20	22	14	25	19	23	15	19
18	23	22	16	16	19	13	18	24	24	26
13	18	17	15	24	15	17	14	18	17	21
16	21	25	19	20	27	16	17	16	21	

Number of Classes

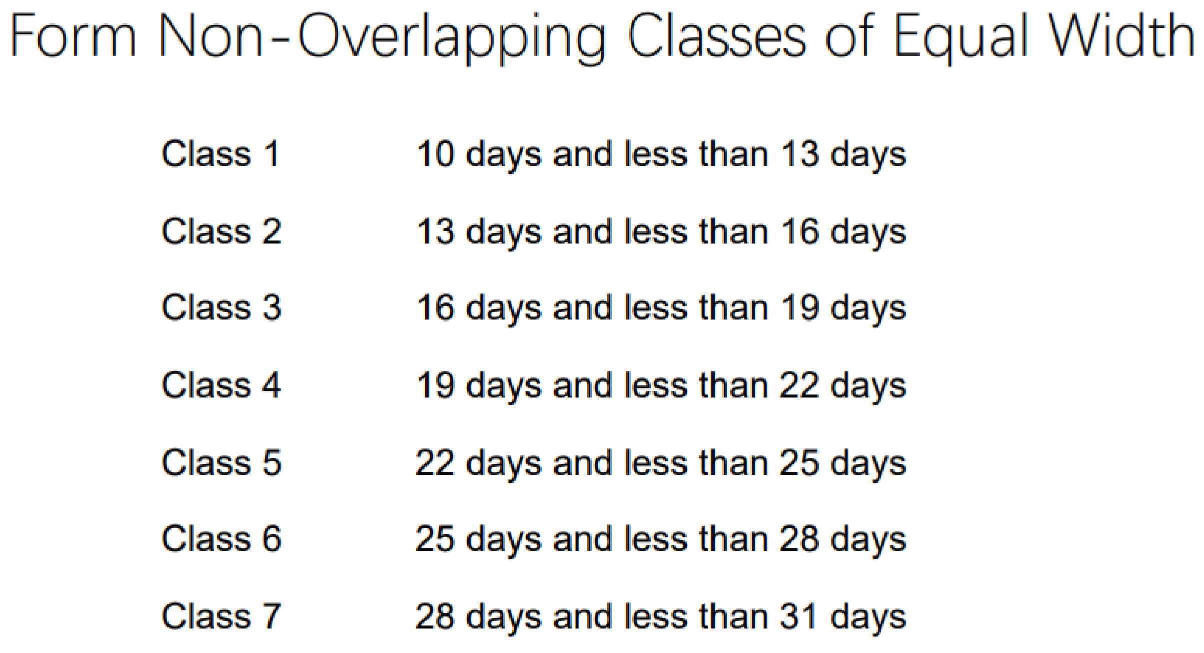
- Group all of the n data into K number of classes
- K is the smallest whole number for which $2^k >= n$
- In the examples 2.2 n = 65

  - for K = 6,2^6 = 64, <n
  - For K = 7,2^7 = 128, > n
  - So use K = 7 classes

Class Length

- Find the length of each class as the largest measurements minus the smallest divided by the number of classes found earlier (K)
- For example 2.2 (29-10)/7 = 2.7143
  - Because payments measured in days, round to three days

$$\text{approximate class length} = \frac{\text{largest measurement} - \text{smallest measurement}}{\text{number of classes}}$$



Form Non-Overlapping Classes of Equal Width

Class 1	10 days and less than 13 days
Class 2	13 days and less than 16 days
Class 3	16 days and less than 19 days
Class 4	19 days and less than 22 days
Class 5	22 days and less than 25 days
Class 6	25 days and less than 28 days
Class 7	28 days and less than 31 days

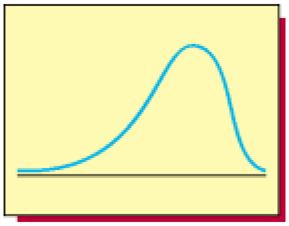
Skewness

Skewed distributions are not symmetrical about their center. Rather, they are lop-sided with a longer tail on one side or the other

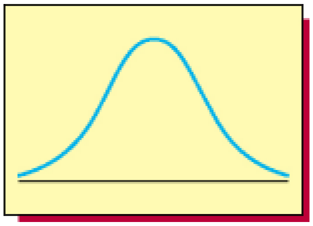
- A population is distributed according to its relative frequency curve

- The skew is the side with the longer tail

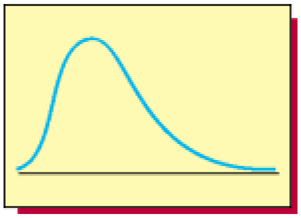
Left Skewed



Symmetric

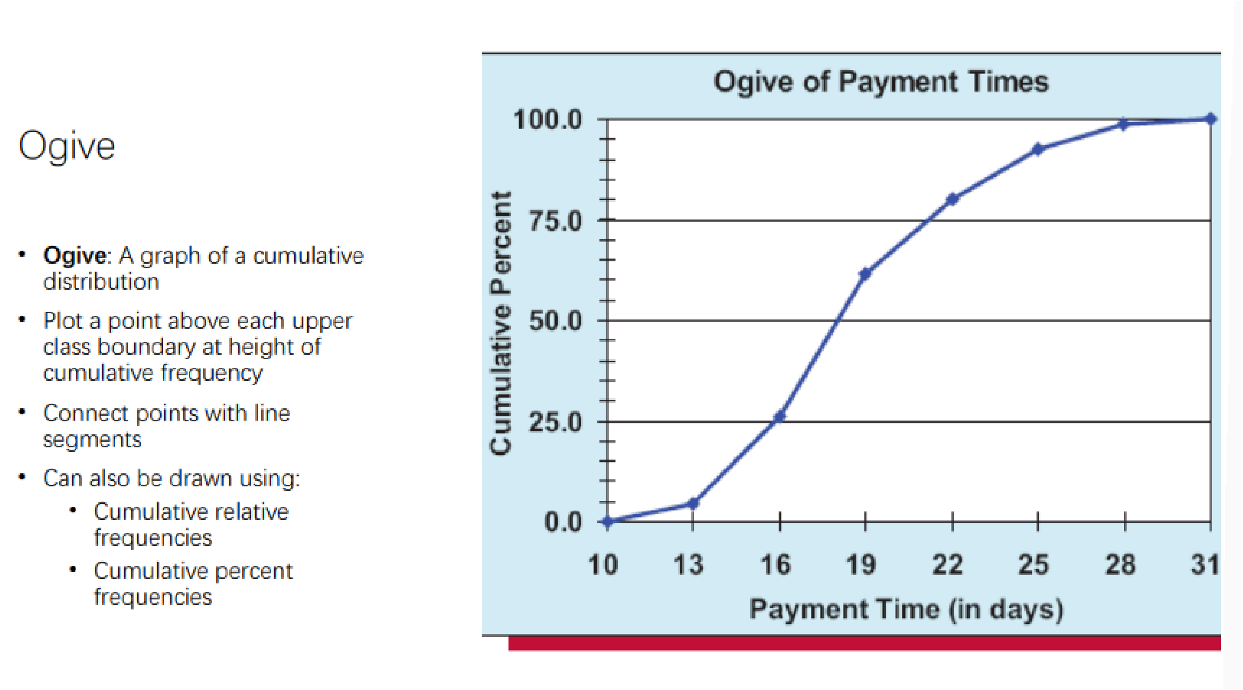


Right Skewed



Cumulative Distributions

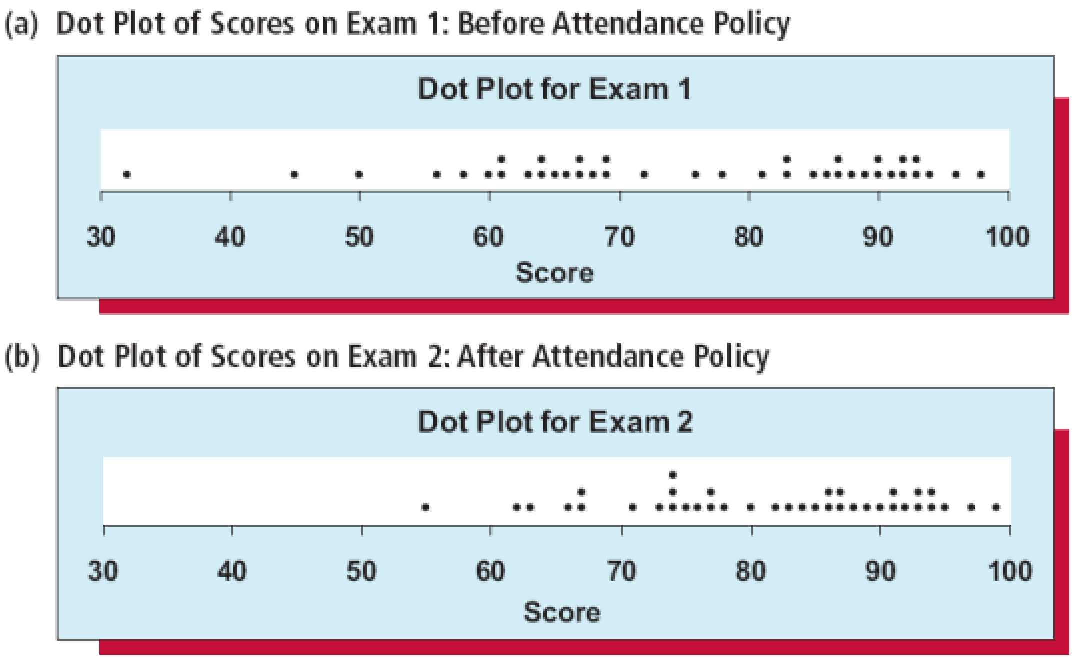
- Another way to summarize a distribution is to construct a cumulative distribution
- To do this, use the same number of classes, class lengths, and class boundaries used for the frequency distribution
- Rather than a count, we record the number of measurements that are less than the upper boundary of that class
  - In other words, a running total



Ogive

- **Ogive**: A graph of a cumulative distribution
- Plot a point above each upper class boundary at height of cumulative frequency
- Connect points with line segments
- Can also be drawn using:
  - Cumulative relative frequencies
  - Cumulative percent frequencies

2.3 Dot Plots


(a) Dot Plot of Scores on Exam 1: Before Attendance Policy

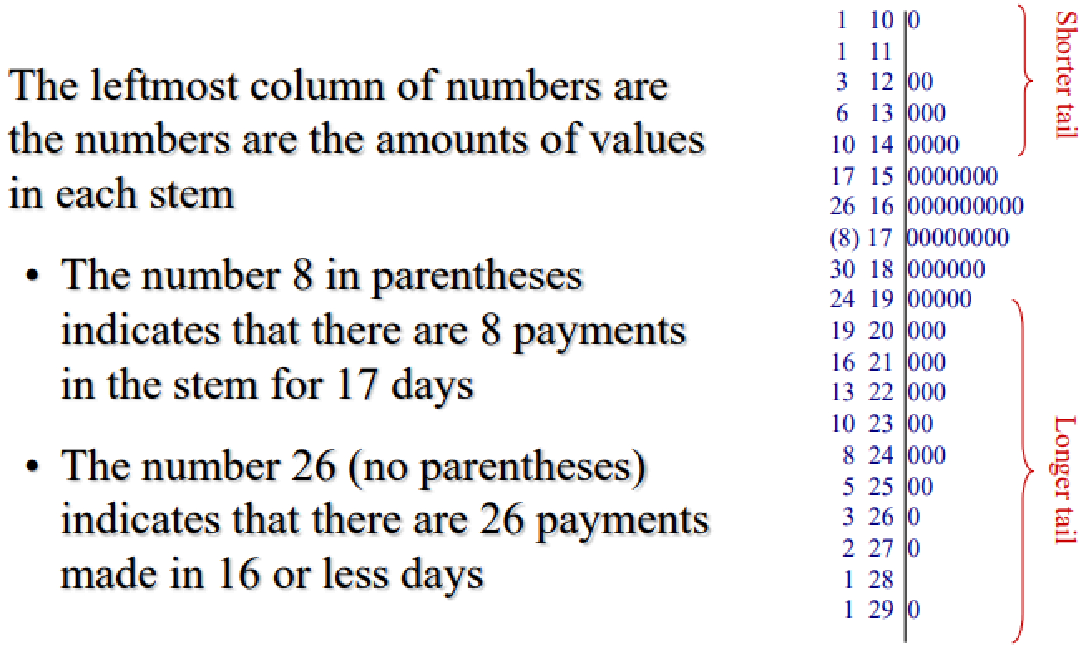
(b) Dot Plot of Scores on Exam 2: After Attendance Policy

2.4 Stem-and-leaf Display

- The purpose of a stem-and-leaf display is to see the overall pattern of the data, by grouping the data into classes
    - To see:
        - The variation from class to class
        - The amount of data in each class
        - The distribution of the data within each class
    - Best for small to moderately sized data distribution

Constructing a Stem-and-Leaf Display

- Decided what units will be used for the stems and the leaves. As a general rule, choose units for the stems so that there will be somewhere between 5 and 20 stems.
- Place the stems in a column with the smallest stem at the top of the column and the largest stem at the bottom.
- Enter the leaf for each measurement into the row corresponding to the proper stem. The leaves should be single digit numbers (rounded values)
- If desired, rearrange the leaves so that they are in increasing order from left to right.



The leftmost column of numbers are the numbers are the amounts of values in each stem

- The number 8 in parentheses indicates that there are 8 payments in the stem for 17 days
- The number 26 (no parentheses) indicates that there are 26 payments made in 16 or less days

1	10	0
1	11	
3	12	00
6	13	000
10	14	0000
17	15	0000000
26	16	000000000
(8)	17	00000000
30	18	000000
24	19	00000
19	20	000
16	21	000
13	22	000
10	23	00
8	24	000
5	25	00
3	26	0
2	27	0
1	28	
1	29	0

Chapter Summary from Gemini

The chapter begins by explaining **descriptive statistics**, which is the science of summarizing the important characteristics of a data set. It introduces two main ways to do this: **tabular and graphical methods**.

Summarizing Qualitative Data

For qualitative data (data with categories or names), the main methods are:

- **Frequency Distribution:** A table that shows the count (or frequency) of items in each category.
- **Relative and Percent Frequency Distributions:** These tables show the proportion or percentage of items in each category.
- **Bar Charts:** These use vertical or horizontal rectangles to represent the frequency of each category. You can find an example of a bar chart on

  page 12.
- **Pie Charts:** A circle divided into slices, where the size of each slice represents its relative or percent frequency. An example is on

  page 13.
- **Pareto Charts:** A special bar chart used for quality control, where the bars are arranged in decreasing height from left to right to highlight the most frequent problems. The example for this is on

  page 15.

Summarizing Quantitative Data

For quantitative data (numerical data), the chapter focuses on:

- **Frequency Distributions and Histograms:** A **histogram** is a graphical representation of a frequency distribution. It uses rectangles to show how many data points fall into specific numerical classes. An example can be seen on

  pages 18, 25, and 67. The chapter also explains the step-by-step process to construct a frequency distribution and histogram, including how to find the number of classes (

  $2K \geq n$) and the class length.
- **Distribution Shape:** Histograms help us understand the shape of a distribution. A distribution can be symmetric, left-skewed, or right-skewed depending on where the longer tail is located. The different shapes are illustrated on

  page 27.

- **Frequency Polygons and Ogives:** A frequency polygon is a line graph of a frequency distribution. An

  ogive is a graph of a cumulative frequency distribution, which shows a running total of the data. You can see an ogive on

  page 31.

- **Dot Plots and Stem-and-Leaf Displays:** These are useful for summarizing small to moderately sized data sets, as they let you see every individual data point while also showing the overall shape of the distribution. Examples of dot plots are on

  page 32, and stem-and-leaf displays are on **pages 38, 39, and 43**.

Studying Relationships Between Variables

The chapter also introduces methods for analyzing the relationship between two variables:

- **Contingency Tables (Crosstabulation):** These tables classify data on two different dimensions and are used to investigate possible relationships between variables.
- **Scatter Plots:** This is a graphical method for studying the relationship between two numerical variables by plotting paired coordinates. You can see examples of positive, negative, and no linear relationships on

  page 55.

▼ Chapter Summary from Notebook llm

I. Graphically Summarizing Qualitative Data

Qualitative data, also known as categorical data, uses names to identify different categories. This type of data can be effectively summarized using frequency distributions, relative frequency distributions, bar charts, and pie charts.

- Frequency Distribution: "A table that summarizes the number (or frequency) of items in each of several non-overlapping classes." This provides a count of how many observations fall into each category.

- **Relative Frequency**: This is the "proportion or fraction of data that is contained in that class." It's calculated by dividing the class frequency by the total number of data values and can be expressed as a decimal or a percentage (percent frequency).
- **Formula**: "Relative frequency = (frequency of a class) / (total number of observations)"
- Multiplying by 100 yields the percent frequency.
- **Bar Charts**: These graphical representations use "a vertical or horizontal rectangle [to represent] the frequency for each category." The height of the bar can represent frequency, relative frequency, or percent frequency.
- **Pie Charts**: A pie chart "is a circle divided into slices where the size of each slice represents its relative frequency or percent frequency." It's useful for showing parts of a whole.
- **Pareto Chart**: While not detailed, it is listed as a method for summarizing qualitative data, typically used for categorizing and displaying the causes of problems in order of their frequency.

II. Graphically Summarizing Quantitative Data

Quantitative data, which consists of numerical measurements, often requires different summarization techniques to understand the shape of its distribution. Histograms, dot plots, and stem-and-leaf displays are crucial for this.

A. Frequency Distributions and Histograms

- **Purpose**: To "group the measurements into classes of a frequency distribution and then displaying the data in the form of a histogram." This helps visualize the shape of the distribution.
- **Frequency Distribution (Quantitative)**: Similar to qualitative data, it's "a list of data classes with the count or “frequency” of values that belong to each class." The process involves "Classify and count."
- **Histogram**: "A picture of the frequency distribution." It's "a graph in which rectangles represent the classes." The base of each rectangle represents

the class length, and the height represents the frequency (in a frequency histogram) or relative frequency (in a relative frequency histogram).

Steps in Constructing a Frequency Distribution and Histogram:

1. **Find the number of classes (K):** Use "Sturges rule," where "K is the smallest whole number for which 2^k≥n " (n is the total number of data points). For example, if n=65, K=7 is used because 2^6=64 < 65, and 2^7=128 > 65.
2. **Find the class length (L):** Calculated as "(Largest value - Smallest value) / K." The result is often rounded up to a convenient whole number. For example, (29-10)/7 = 2.7143, rounded to 3 days/class.
3. **Form non-overlapping classes of equal width:** Classes start at the smallest data value. The lower limit of the first class is the smallest value, and the upper limit is "smallest value + (L − 1)." Subsequent classes follow this pattern (e.g., 10 to 12 days, 13 to 15 days).
4. **Tally and count the number of measurements in each class:** This forms the frequency for each class.
5. **Graph the histogram.**

Skewed Distributions: These are "not symmetrical about their center."

- **Right-Skewed:** Has a "longer tail on one side," specifically the right side (higher values).
- **Left-Skewed:** Has a "longer tail" on the left side (lower values).
- **Symmetric:** Evenly distributed around the center.

Cumulative Distributions and Ogives:

- **Cumulative Distribution:** Summarizes data by recording "the number of measurements that are less than the upper boundary of that class," essentially a running total.

- Ogive: "A graph of a cumulative distribution." Points are plotted above each upper class boundary at the height of the cumulative frequency and connected with line segments. Can also use cumulative relative or percent frequencies.

B. Dot Plots

- A simple way to visualize the distribution of quantitative data, especially for smaller datasets. Not detailed further in the provided text, but mentioned as a graphical method.

C. Stem-and-Leaf Displays

- **Purpose**: To "see the overall pattern of the data, by grouping the data into classes." It allows observation of "the variation from class to class," "the amount of data in each class," and "the distribution of the data within each class." Best for "small to moderately sized data distributions."
- **Construction Steps**:

1. **Decide on stem and leaf units**: Stems should result in 5 to 20 stems. Leaves are typically single-digit numbers.
2. **Place stems**: Stems are listed in a column, smallest at the top.
3. **Enter leaves**: Each measurement's leaf is placed in the row corresponding to its stem.
4. **Rearrange leaves (optional)**: Leaves can be ordered in increasing value from left to right.

- **Example (Car Mileage Case)**: Shows how a stem (e.g., 29 for 29.x) and leaves (e.g., 8 for 29.8) are used. The display can reveal distribution shape, such as "slightly 'left-skewed'" if the lower portion has slightly more data.
- **Example (Payment Time Case)**: Demonstrates how the display can quickly show that sampled payment times are "substantially less than the 39-day typical payment time of the former billing system." Also, it identified a "right-skewed" distribution because "the 'tail' of the distribution consisting of the higher payment times is longer than the 'tail' of the distribution consisting of the smaller payment times."

III. Contingency Tables (Crosstabulation)

- **Purpose**: "Classifies data on two dimensions" to investigate relationships between variables.
- **Structure**: Rows classify data according to one variable, and columns classify according to a second variable. A third variable is counted within the cells.
- **Components**: "Row totals provide a frequency distribution for the different fund types," and "Column totals provide a frequency distribution for the different satisfaction levels."
- **Analysis**: Relationships are investigated by computing "row and column percentages."
- **Row percentages**: "dividing each cell’s frequency by its row total."
- **Column percentages**: "dividing by the column total."
- **Variable Types**: Can be used for two qualitative variables, a quantitative and a qualitative variable, or two quantitative variables (where quantitative variables are often categorized).

IV. Scatter Plots

- **Purpose**: "Used to study relationships between two variables (numerical)."
- **Construction**: One variable is placed on the x-axis, and the second on the y-axis, with a dot marking "pair coordinates."
- **Types of Relationships Identified**:
- **Linear**: A straight-line relationship.
- **Positive**: "When one variable goes up, the other variable goes up."
- **Negative**: "When one variable goes up, the other variable goes down."
- **No Linear Relationship**: "There is no coordinated linear movement between the two variables."
- Useful for visualizing data patterns and trends.

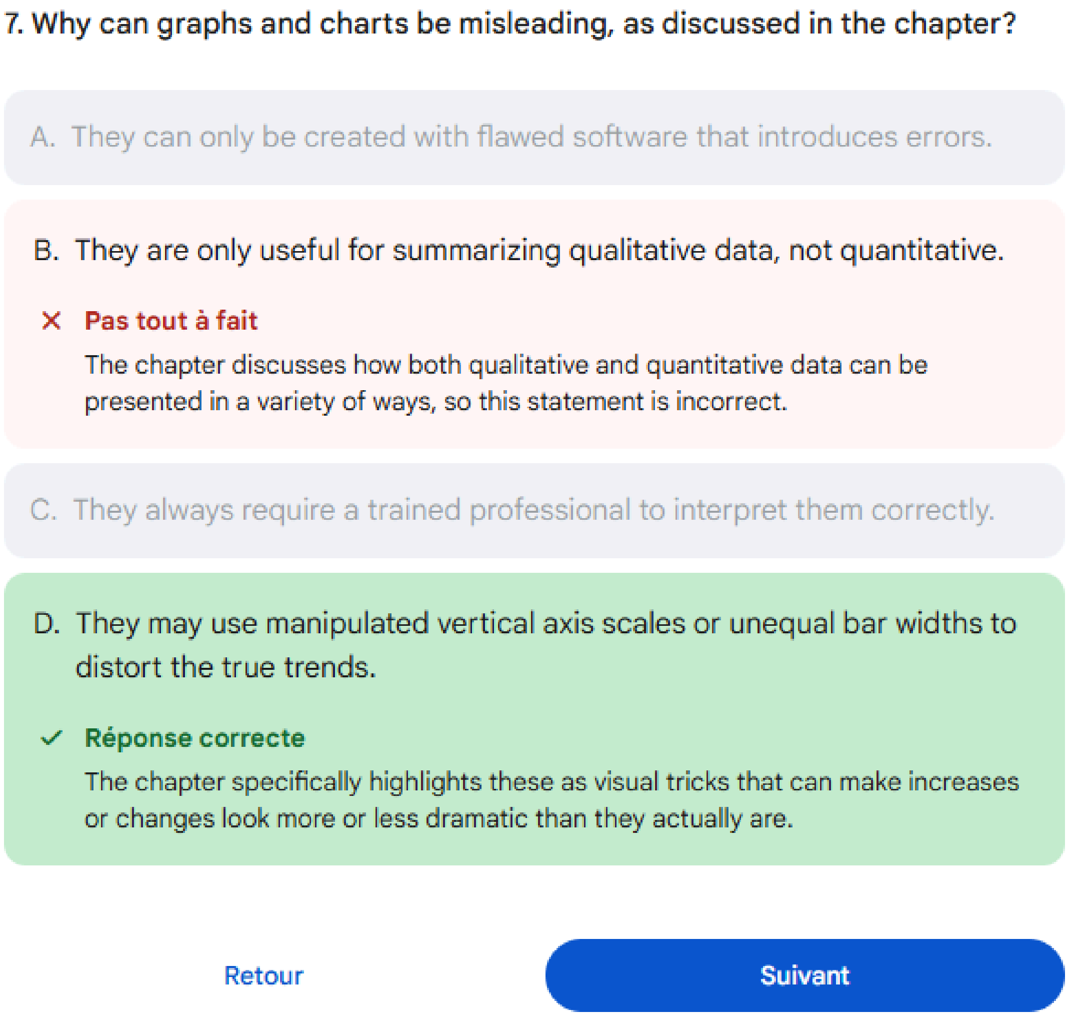
V. Misleading Graphs and Charts

- This section highlights the potential for graphs and charts to be manipulated or designed in a way that misrepresents the data, particularly mentioning "Horizontal Scale Effects." This emphasizes the importance of critical evaluation of visual data presentations.

In summary, the provided sources detail essential descriptive statistical methods for organizing and visualizing data. These techniques range from simple frequency counts and bar/pie charts for qualitative data to more complex histograms, stem-and-leaf displays, and scatter plots for quantitative data, all aimed at revealing patterns, distributions, and relationships within datasets. The importance of constructing accurate and non-misleading visualizations is also underscored.

▼ Remember

- The pareto chart ranks problems from the highest to lowest frequency, allowing you to focus on the most common problems.
- Cumulative distribution is a table that shows a cumulative total of frequencies or percentages for each class.
- A dot plot allows you to see each data point while providing insight into the overall distribution



7. Why can graphs and charts be misleading, as discussed in the chapter?

A. They can only be created with flawed software that introduces errors.

B. They are only useful for summarizing qualitative data, not quantitative.

✗ **Pas tout à fait**

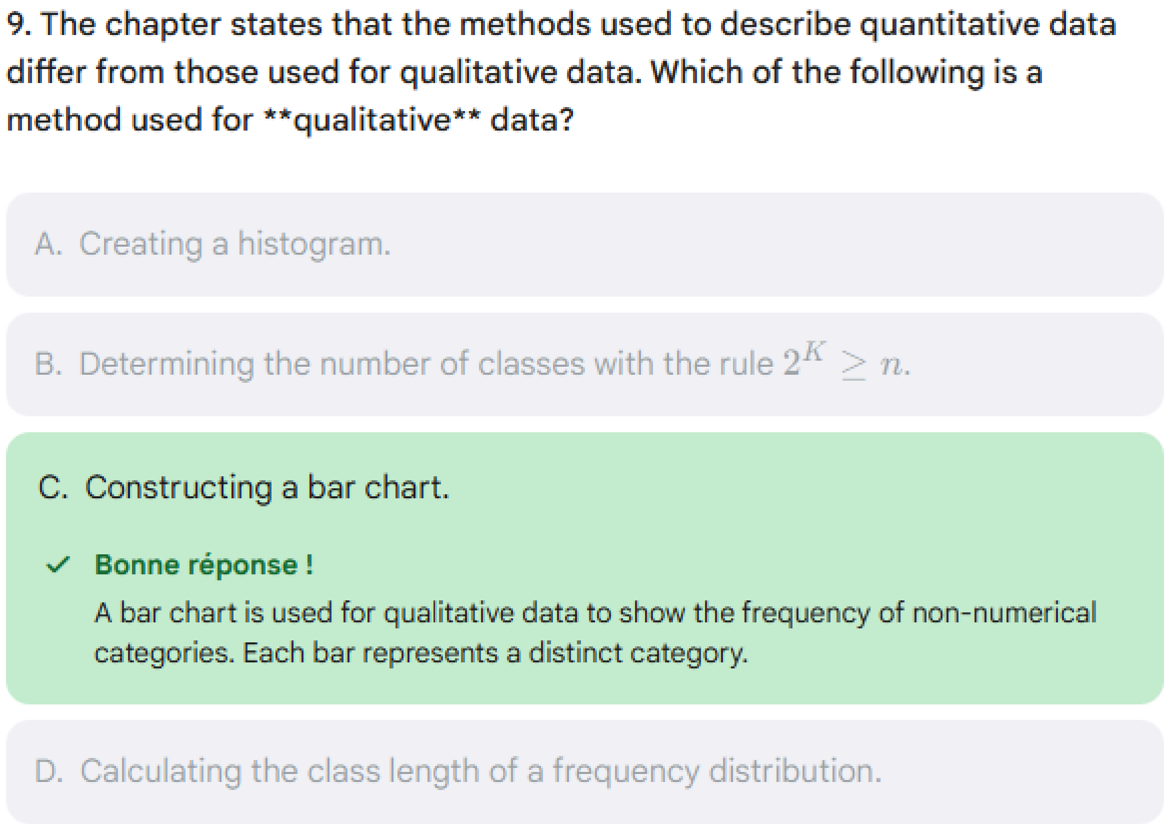
The chapter discusses how both qualitative and quantitative data can be presented in a variety of ways, so this statement is incorrect.

C. They always require a trained professional to interpret them correctly.

D. They may use manipulated vertical axis scales or unequal bar widths to distort the true trends.

✓ **Réponse correcte**

The chapter specifically highlights these as visual tricks that can make increases or changes look more or less dramatic than they actually are.



9. The chapter states that the methods used to describe quantitative data differ from those used for qualitative data. Which of the following is a method used for **qualitative** data?

A. Creating a histogram.

B. Determining the number of classes with the rule $2^K \geq n$.

C. Constructing a bar chart.

✓ Bonne réponse !

A bar chart is used for qualitative data to show the frequency of non-numerical categories. Each bar represents a distinct category.

D. Calculating the class length of a frequency distribution.

▼ FAQ

1. How are qualitative and quantitative data summarized graphically?

Qualitative data, which involves names or categories, is effectively summarized using frequency distributions, relative frequency distributions, bar charts, and pie charts. Frequency distributions show the count of items in each category, while relative frequency distributions display the proportion or percentage. Bar charts use vertical or horizontal rectangles to represent the frequency or relative frequency of each category, and pie charts divide a circle into slices whose size corresponds to the relative or percent frequency of each category.

Quantitative data, on the other hand, often requires understanding the shape of its distribution. This is achieved by grouping measurements into classes of a frequency distribution and then visualizing the data with a histogram. Histograms use rectangles to represent classes, where the base represents the class length and the height represents the frequency or relative frequency within that class. Other methods for quantitative data include dot plots and stem-and-leaf displays.

2. What are frequency and relative frequency distributions, and how are they calculated?

A frequency distribution is a table that summarizes the number (or frequency) of items falling into each of several non-overlapping categories or classes. For qualitative data, these classes are the distinct categories. For quantitative data, measurements are grouped into classes of equal width.

The relative frequency of a class is the proportion or fraction of the total data contained in that class. It is calculated by dividing the class frequency by the total number of data values. Relative frequency can be expressed as a decimal or a percentage (percent frequency, obtained by multiplying the relative frequency by 100). A relative frequency distribution lists all data classes and their associated relative frequencies.

3. What are the key steps involved in constructing a frequency distribution and histogram for quantitative data?

Constructing a frequency distribution and histogram for quantitative data involves five main steps:

1. **Find the number of classes (K):** This is determined by finding the smallest whole number K such that 2^K is greater than or equal to the total number of data points (n). For example, if n=65, then K=7 because 2^6=64 < 65, and 2^7=128 > 65.
2. **Find the class length (L):** Calculated by dividing the range of the data (largest measurement minus smallest measurement) by the number of classes (K). This value is often rounded up to a convenient number, like rounding 2.7143 days to 3 days/class.
3. **Form non-overlapping classes of equal width:** The first class starts at the smallest data value. Its upper limit is calculated as the smallest value + (L - 1). Subsequent classes start at the upper limit of the previous class + 1 and continue for a length of (L - 1), ensuring all classes have equal width and do not overlap.

4. **Tally and count the number of measurements in each class:** Go through the data and count how many values fall into each defined class. This gives the frequency for each class.
5. **Graph the histogram:** Create a bar-like graph where the base of each rectangle represents a class interval (class length) and the height represents the frequency or relative frequency of that class.

4. What is a stem-and-leaf display, and what information does it convey?

A stem-and-leaf display is a method for visualizing the overall pattern of a data set, particularly useful for small to moderately sized distributions. It groups data into classes while retaining the individual data values.

To construct it, data points are divided into a "stem" (typically the leading digit(s)) and a "leaf" (typically the trailing digit). Stems are listed in a column, and the leaves for each measurement are entered into the row corresponding to their proper stem. The leaves are then often rearranged in increasing order. This display allows you to see:

- The overall pattern and shape of the data distribution.
- The variation from class to class.
- The amount of data in each class.
- The distribution of data within each class.

It's a quick way to get a sense of the data's central tendency, spread, and skewness, as it preserves the original data values unlike a histogram.

5. What are skewed distributions, and how can you identify them graphically?

Skewed distributions are distributions that are not symmetrical about their center; they are "lop-sided" with a longer "tail" on one side. The "skew" refers to the side with the longer tail.

- **Right-skewed (positively skewed) distribution:** The longer tail extends to the right (higher values). This means there are a few unusually large values pulling the average to the right. Graphically, the bulk of the data will be on the left, and the tail will stretch out towards the right.

- **Left-skewed (negatively skewed) distribution:** The longer tail extends to the left (lower values). This indicates a few unusually small values pulling the average to the left. Graphically, the bulk of the data will be on the right, and the tail will stretch out towards the left.
- **Symmetric distribution:** The distribution is balanced, and both sides are roughly mirror images of each other.

You can identify skewness by observing the shape of histograms, frequency polygons, or stem-and-leaf displays.

6. What is a cumulative distribution and an ogive?

A cumulative distribution is another way to summarize a data distribution. It uses the same classes, class lengths, and class boundaries as a frequency distribution, but instead of just counting the frequencies within each class, it records the number of measurements that are *less than* the upper boundary of each class. In essence, it provides a running total of frequencies.

An **ogive** is a graphical representation of a cumulative distribution. To create an ogive, a point is plotted above each upper class boundary at the height corresponding to its cumulative frequency. These points are then connected with line segments. Ogives can also be drawn using cumulative relative frequencies or cumulative percent frequencies, providing a visual representation of how data accumulates across the classes.

7. What are contingency tables and scatter plots used for in data analysis?

Contingency Tables (Crosstabulation): Contingency tables classify data on two or more dimensions, typically showing the frequency of observations for each combination of categories for two or more variables. Rows classify data according to one variable, and columns classify it according to a second variable. They are particularly useful for:

- Summarizing the frequency distribution for individual variables (using row and column totals).

- Investigating possible relationships or associations between two or more categorical (qualitative) variables. By computing row and column percentages, one can observe how the distribution of one variable changes across the categories of another.

Scatter Plots: Scatter plots are used to study relationships between two numerical (quantitative) variables. One variable is placed on the x-axis, and the second variable is placed on the y-axis. Each pair of coordinates for the two variables is represented by a dot on the graph. Scatter plots help visualize:

- **Linear relationships:** Whether the dots tend to form a straight line.
- **Direction of relationship: Positive linear relationship:** As one variable increases, the other also tends to increase.
- **Negative linear relationship:** As one variable increases, the other tends to decrease.
- **No linear relationship:** The dots show no coordinated linear movement.
- **Strength of relationship:** How closely the points cluster around a potential line.

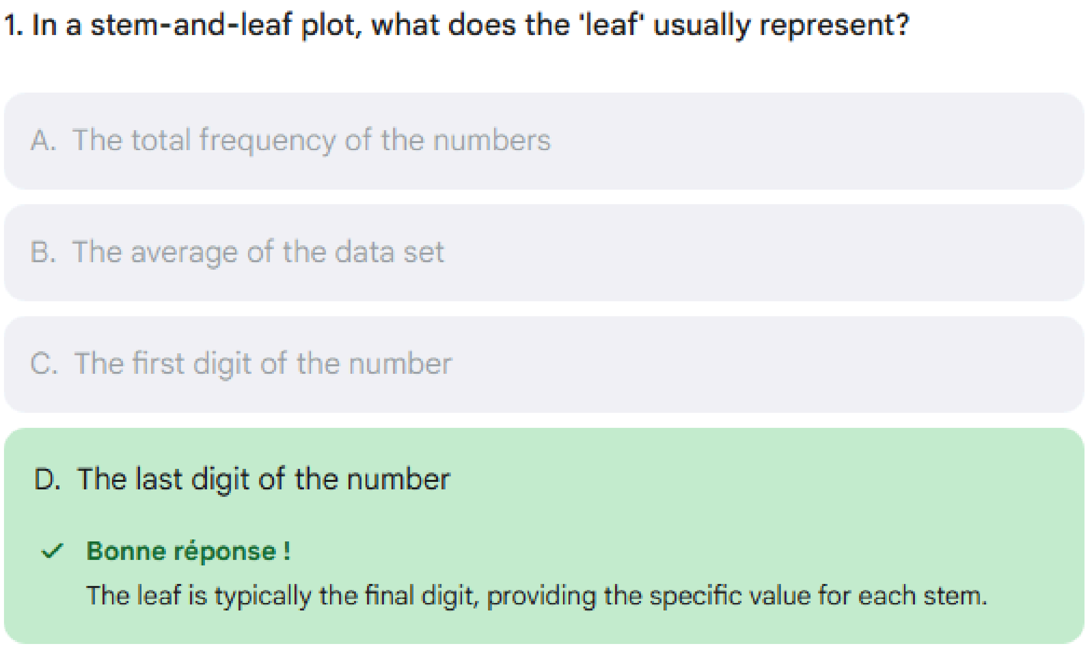
8. What are some common ways graphs and charts can be misleading?

Graphs and charts, while powerful for data visualization, can be intentionally or unintentionally misleading. The source specifically mentions "Horizontal Scale Effects" as one way. While not elaborating, common methods of misleading include:

- **Truncated Y-axis:** Starting the Y-axis at a value greater than zero can exaggerate small differences, making them appear more significant than they are.
- **Inconsistent scales:** Using different intervals or non-linear scales on axes can distort perceptions of change or magnitude.
- **Omitting data:** Not including all relevant data points or time periods can paint an incomplete or biased picture.

- **Improper use of chart types:** Using a chart type that is not appropriate for the data can lead to misinterpretation (e.g., using a 3D pie chart where slice sizes are harder to compare).
- **Visual distortions:** Using disproportionate visual elements (like 3D effects, varying bar widths, or images that don't scale proportionally to the data) can create a false impression.
- **Cherry-picking data:** Presenting only data that supports a particular agenda while ignoring contradictory evidence.
- **Lack of clear labels or titles:** Ambiguous or missing labels can lead to confusion and incorrect conclusions.

▼ Stem and leaf Quiz



1. In a stem-and-leaf plot, what does the 'leaf' usually represent?

A. The total frequency of the numbers

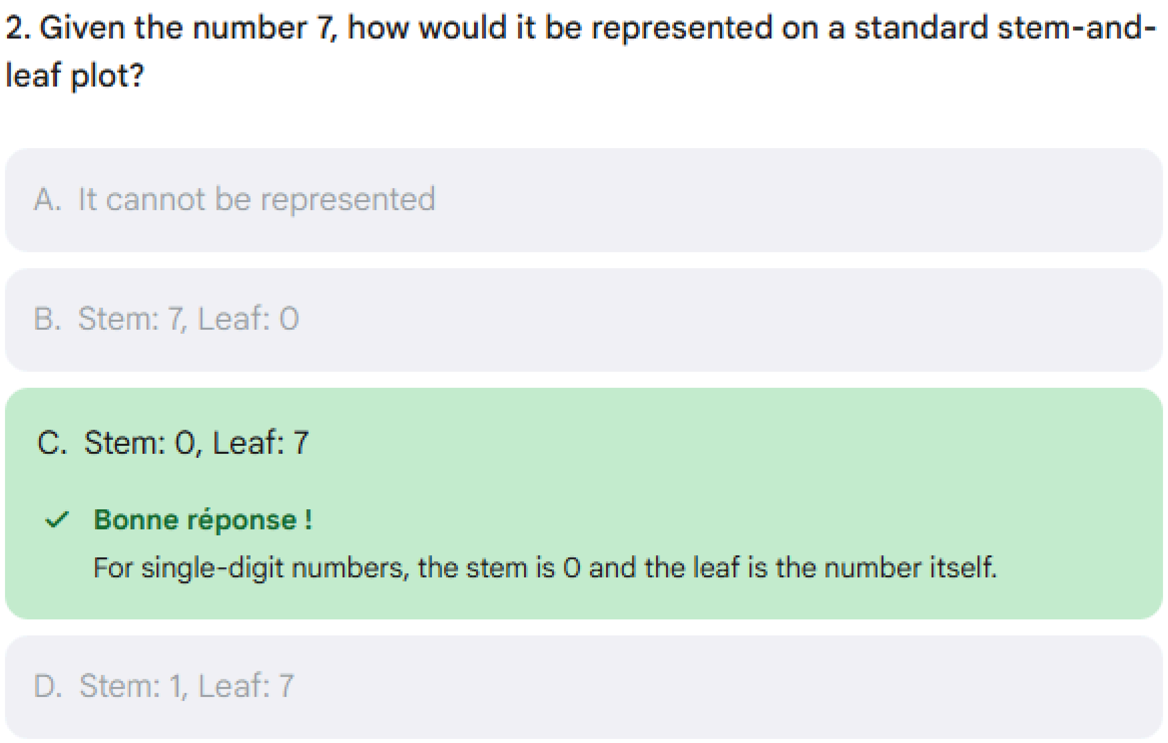
B. The average of the data set

C. The first digit of the number

D. The last digit of the number

✓ **Bonne réponse !**

The leaf is typically the final digit, providing the specific value for each stem.



2. Given the number 7, how would it be represented on a standard stem-and-leaf plot?

A. It cannot be represented

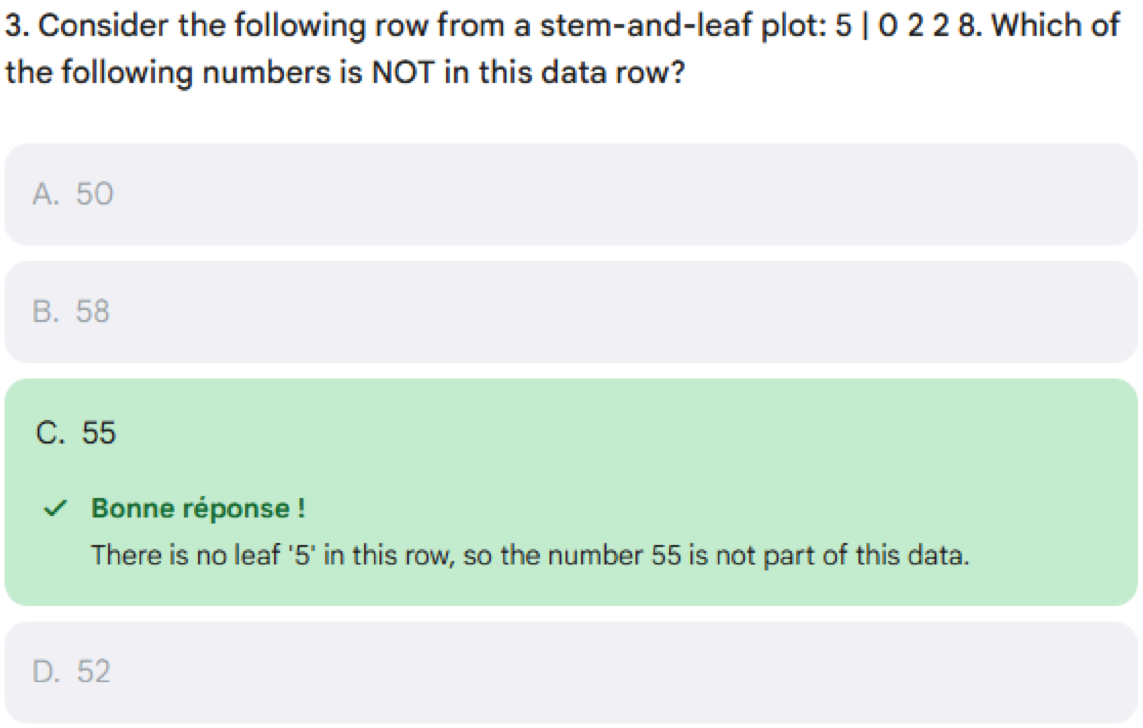
B. Stem: 7, Leaf: 0

C. Stem: 0, Leaf: 7

✓ **Bonne réponse !**

For single-digit numbers, the stem is 0 and the leaf is the number itself.

D. Stem: 1, Leaf: 7



3. Consider the following row from a stem-and-leaf plot: 5 | 0 2 2 8. Which of the following numbers is NOT in this data row?

A. 50

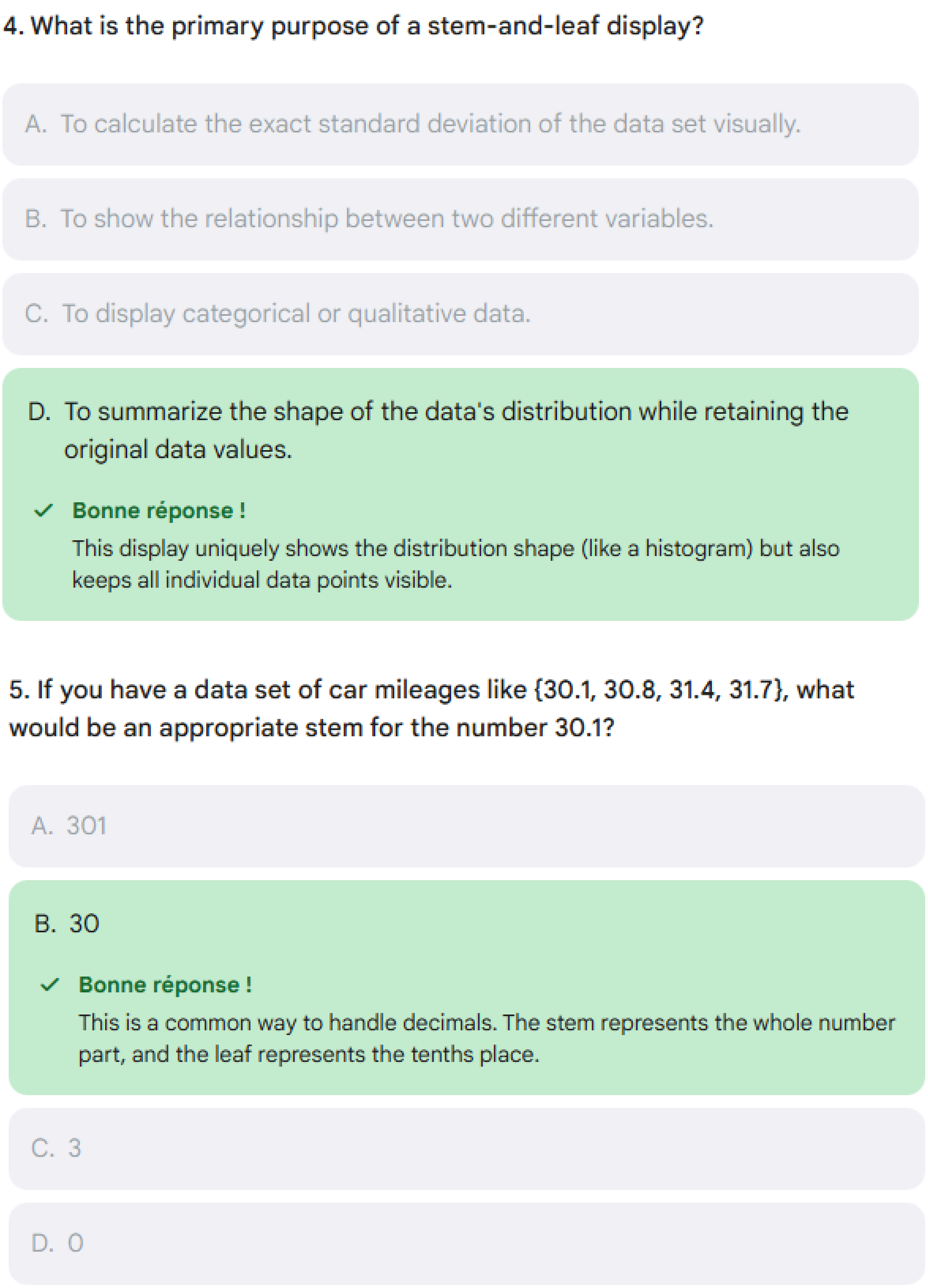
B. 58

C. 55

✓ **Bonne réponse !**

There is no leaf '5' in this row, so the number 55 is not part of this data.

D. 52



4. What is the primary purpose of a stem-and-leaf display?

A. To calculate the exact standard deviation of the data set visually.

B. To show the relationship between two different variables.

C. To display categorical or qualitative data.

D. To summarize the shape of the data's distribution while retaining the original data values.

✓ **Bonne réponse !**

This display uniquely shows the distribution shape (like a histogram) but also keeps all individual data points visible.

5. If you have a data set of car mileages like {30.1, 30.8, 31.4, 31.7}, what would be an appropriate stem for the number 30.1?

A. 301

B. 30

✓ **Bonne réponse !**

This is a common way to handle decimals. The stem represents the whole number part, and the leaf represents the tenths place.

C. 3

D. 0