DESCRIPTIVE STATISTICS

GABRIEL A. OKYERE (PhD)

KNUST

goasare17@gmail.com

January 31, 2017

Tabular Representation of Data

The data gathered from a survey/experiment are usually summarized or organized numerically in tabular form using a frequency distribution table and its related forms. The distribution is said to be ungrouped if it shows the distinct observations and their corresponding occurrences, called frequencies. If the number of observations is too large then they are put into groups, called classes or categories. The number of classes is usually chosen between 5 and 20, inclusive.

Tabular Representation of Data

The general rule is to use small number of classes for small amount of data and large number of classes for large amount of data. The best choice of number of classes (k) is suggested by the following:

- The number of classes is the smallest integer value, k such that $2^k \geq n$, or
- Sturges' (Approximation) Rule: A rule for determining the desirable number of groups into which a distribution of observations should be classified,

$$K = 1 + 3.322 \log_{10} n$$

and class width,

$$C = \frac{Range(R)}{K}$$

where n is the total number of observations.

Tabular Representation of Data

Another useful technique for summarizing data is the **relative frequency** or **cumulative frequency distribution table**. The relative frequency indicates the proportion of occurrence of the observations while cumulative frequency distribution shows the total number of occurrences above or below certain key observations or classes.

Tabular Representation of Data

Example

The table below gives the number of children per family for 20 families selected from Kumasi, a city in Ghana. The data, presented in the form in which it was collected is called raw data.

1	3	0	0	2	1	2	3	3	3
1	3	2	1	3	3	0	3	1	2

Tabular Representation of Data

Approach to solving the previous example

The frequencies, relative and cumulative frequencies for the above data are shown in the distribution below:

No. of children (x)	No. of families (f)	Relative Frequency	Cumulative Frequency
0	3	0.15	3
1	5	0.25	8
2	4	0.20	12
3	8	0.40	20
Total	n= 20	1.00	

It is observed from this data that a greater number (8) or proportion (40%) of the families have three (3) children. Three (3) families have no children.

Tabular Representation of Data

Example 2

The data below show the age distribution of cases of malaria reported during a year at a hospital.

34	17	25	37	19	19	27	19	44	24	24
22	32	12	13	16	18	14	12	16	14	17
10	16	22	20	15	15	10	10	14	17	20
18	13	32	13	13	18	30	24	34	44	31
43	40	28	31	15	22	15	31	18	27	35
35	20	32	38	32						

Organize the data into a grouped frequency distribution table.

Tabular Representation of Data

Steps to solving Example 2

The given data is grouped into a number of classes in a frequency distribution as follows:

a. The number of classes, k, (since it is not given) using the sturges'(Approximation) Rule,
$k = 1 + 3.322 \log_{10} n$ where n $= 60$
$k = 1 + 3.322 \log_{10} 60 = 6.907 \simeq 7$

b. The range (R) and the class width (C) are computed as

$$R = \text{Maximum observation} - \text{Minimum observation (weight)}$$
$$= 44 - 10 = 34$$
$$C = \frac{Range}{k} = \frac{34}{7} = 4.8571 \simeq 5$$

Tabular Representation of Data

Steps to solving Example 2 Cont.

1. **Class Boundaries**:
   We determine first $LB_1 - UB_1$ as follows:
   $LB_1 = (\text{Least observed value}) - \frac{1}{2}(\text{smallest unit of the measurement})$

$$LB_1 = 10 - \frac{1}{2}(1) = 9.5$$
$$UB_1 = LB_1 + C = 9.5 + 5 = 14.5$$

The subsequent class boundaries are obtained by adding C to the class limits or boundaries as shown in the distribution table below:

Tabular Representation of Data

The required frequency distribution is

Age	Class Mark (x_i)	Frequency (f_i)	Cumulative Frequency
9.5-14.5	12	11	11
14.5-19.5	17	19	30
19.5-24.5	22	9	39
24.5-29.5	27	4	43
29.5-34.5	32	9	52
34.5-39.5	37	4	56
39.5-44.5	42	4	60
Total		60	

NB: Class Mark(Midpoint): The number in the middle of the class. It is found by adding the upper and lower limits and dividing by two. It can also be found by adding the upper and lower boundaries and dividing by two.

Graphical Representation of Data

The data represented on frequency distribution and its related forms are further summarized using graphs or charts for stronger visual impact. These diagrams are very useful in interpreting data when quick analysis of data is needed. The diagrams are categorized for **quantitative** and **qualitative** data.
Quantitative data are represented graphically using Histogram/Frequency Polygon, Cumulative Frequency Curve, Dot plots, The Box-and-Whisker plot and The Stem-and-Leaf Plots

Histogram and Frequency Polygon

Histogram

A histogram consists of rectangle with:

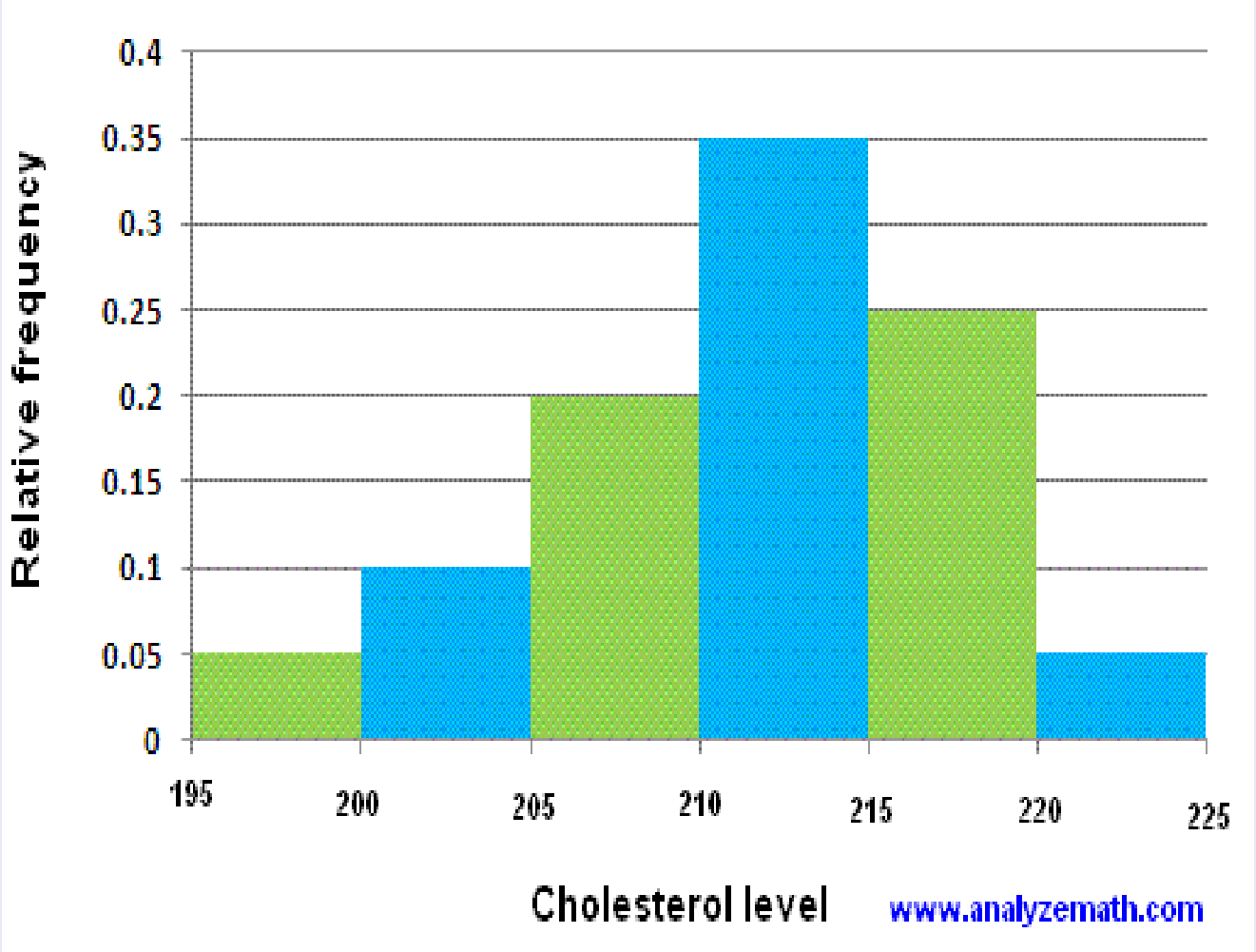
- bases on a horizontal axis, centres at the class marks, and lengths equal to the class widths,
- areas proportional to class frequencies.

If the class intervals are of equal size, then the heights of the rectangles are proportional to the class frequencies and it is then customary to take the heights numerically equal to the class frequencies. If the class intervals are of different widths, then the heights of the rectangles are proportional to

$$\textbf{Frequency Density} = \frac{\text{Class Frequency}}{\text{Class width}}$$

Histogram and Frequency Polygon

An example of a Histogram

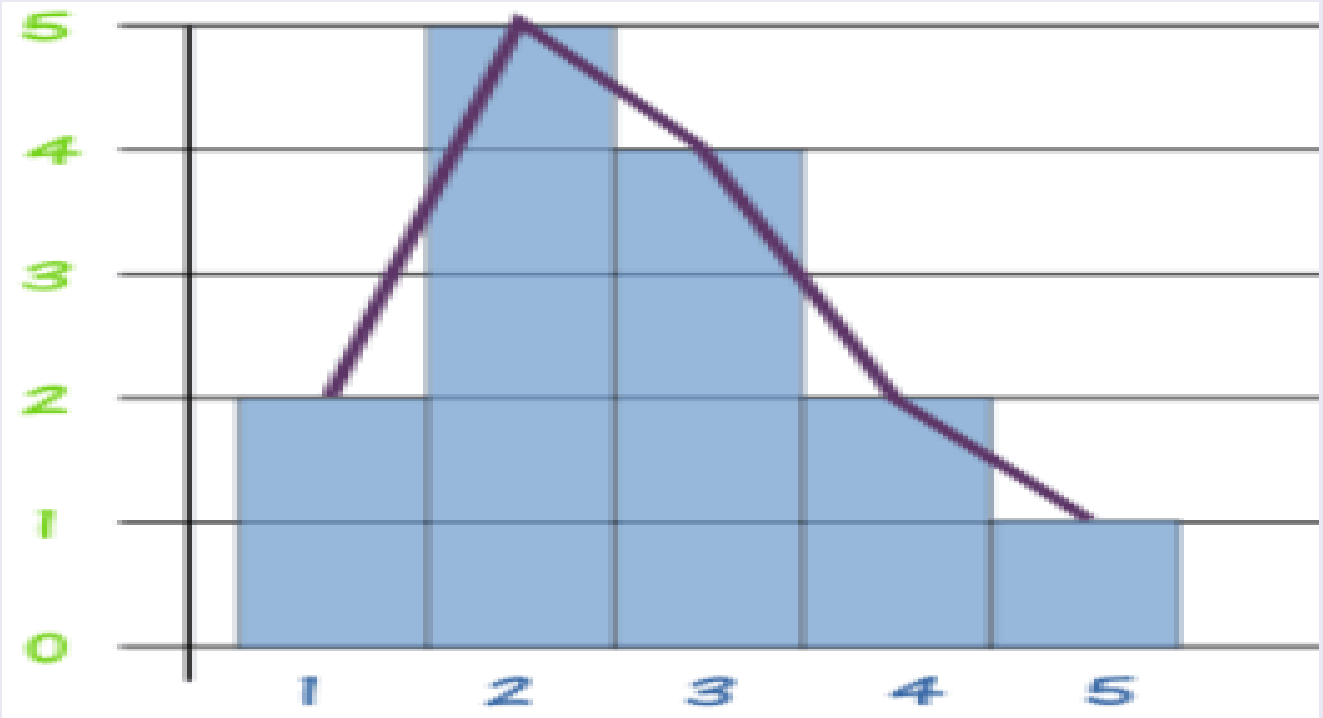


Histogram and Frequency Polygon

Frequency Polygon

Another diagram closely associated with histogram is the frequency polygon. It is drawn by joining the mid-points of tops of rectangular bars in a histogram.

An example of a Frequency Polygon

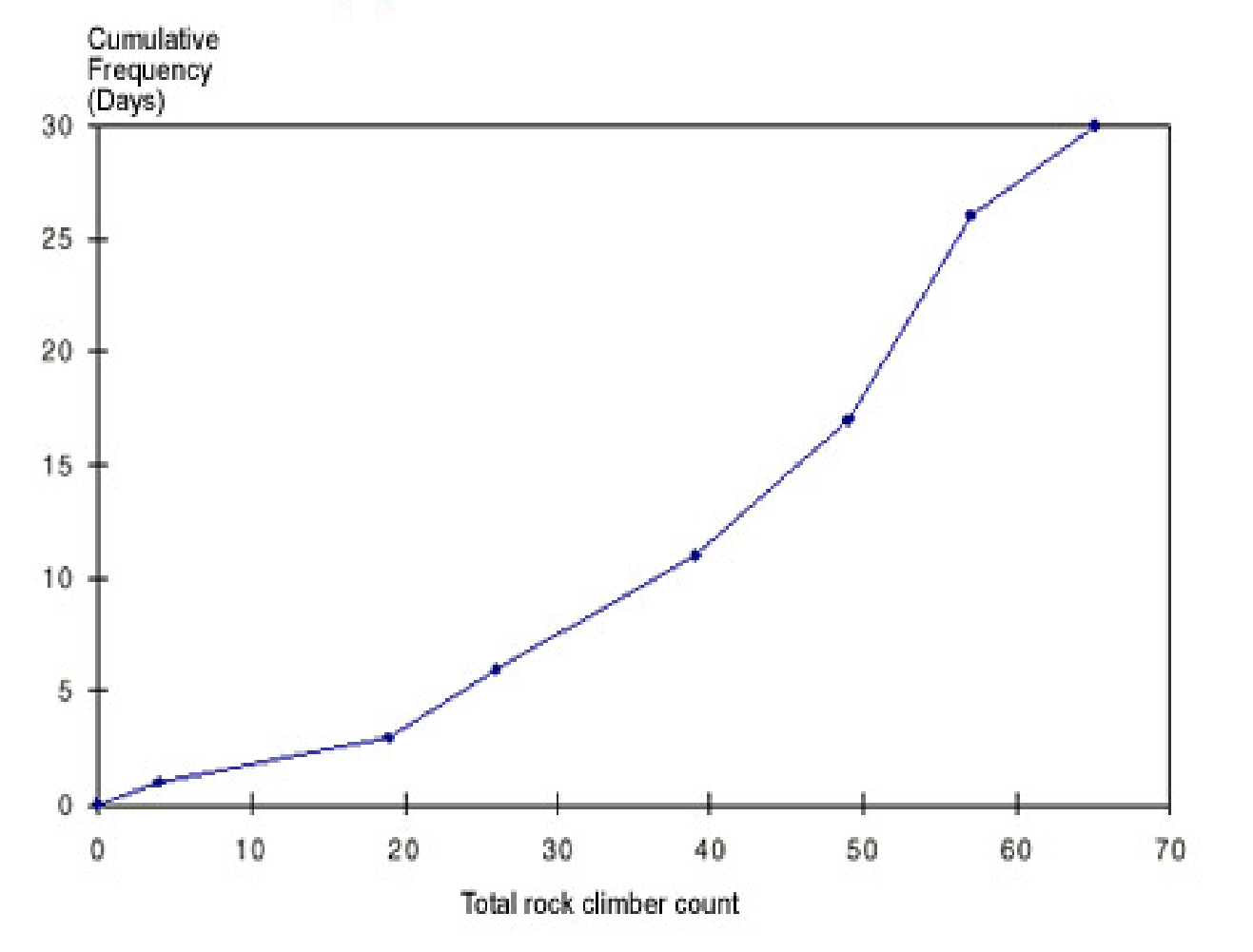


Cumulative Frequency Curve

The cumulative frequency distribution shows the number of observations that fall above or below a specified value of observation. The cumulative frequency of a class is observed by cumulating (or summing) all frequencies up to the class. A graph obtained by plotting the cumulative points by smooth curve is called **cumulative frequency curve** or **Ogive**.

Cumulative Frequency Curve

An example of a cumulative frequency curve

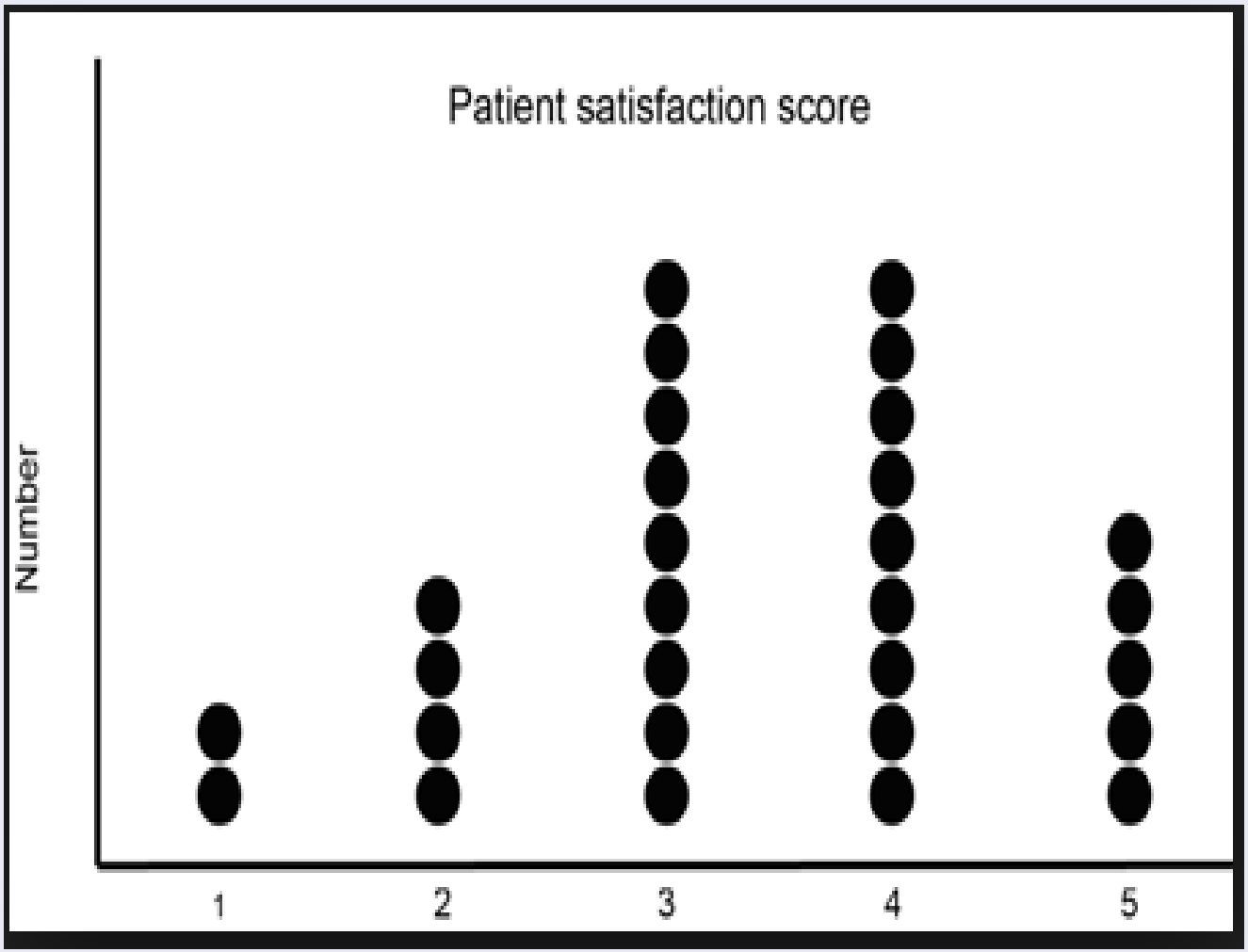


Dot Plots

A dot plot is a graphical display used in statistics that uses dots to represent data. Dot plots can be used for univariate data; that is, data with only one variable that is being measured.If there are multiple occurrences of a specific value, then the dots will be stacked vertically.

Cumulative Frequency Curve

An example of a Dot Plot



Box and Whisker Plot

Steps to drawing a box and whisker plot

1. Represent the variable of interest on a horizontal line(or sometimes on a vertical line)
2. Draw a box in the space above the horizontal axis in such a way that the left end of the box aligns with the first quartile (Q1) and the right end of the box aligns with the third quartile (Q3).
3. Divide the box into two parts by a vertical line that aligns with the median, (Q2).
4. Draw a horizontal line, called a whisker, from the left end of the box to a point that aligns with the smallest measurement in the data set.
5. Draw another horizontal line(or whisker) from the right end of the box to a point that aligns with the largest measurement in the data set.

Box and Whisker Plot

It can be seen that a box plot gives a visual summary of five keys numbers that are associated with a set of data. These are the minimum value, the lower quartile, the median, the upper quartile and the maximum value. Examination of a box-and-whisker plot for a set of data reveals information regarding the amount of spread, location of concentration, and summary of the data.

Stem-and-Leaf Plot

The stem-and-leaf plot was originally developed by John Tukey. It is extremely useful in summarizing reasonably sized data sets(usually under 100), and unlike histograms, results in no loss of information. The stem-and-leaf plot is constructed by first separating each observed value in the data set into two parts, called stem and leaf. The stems are then arranged vertically in ascending order of magnitude and the leaves are recorded against their corresponding stems. A stem and leaf plot has an advantage over a grouped frequency distribution since the stem and leaf plot retains the actual data by showing them in graphic form.

Graphical Representation of Qualitative Data

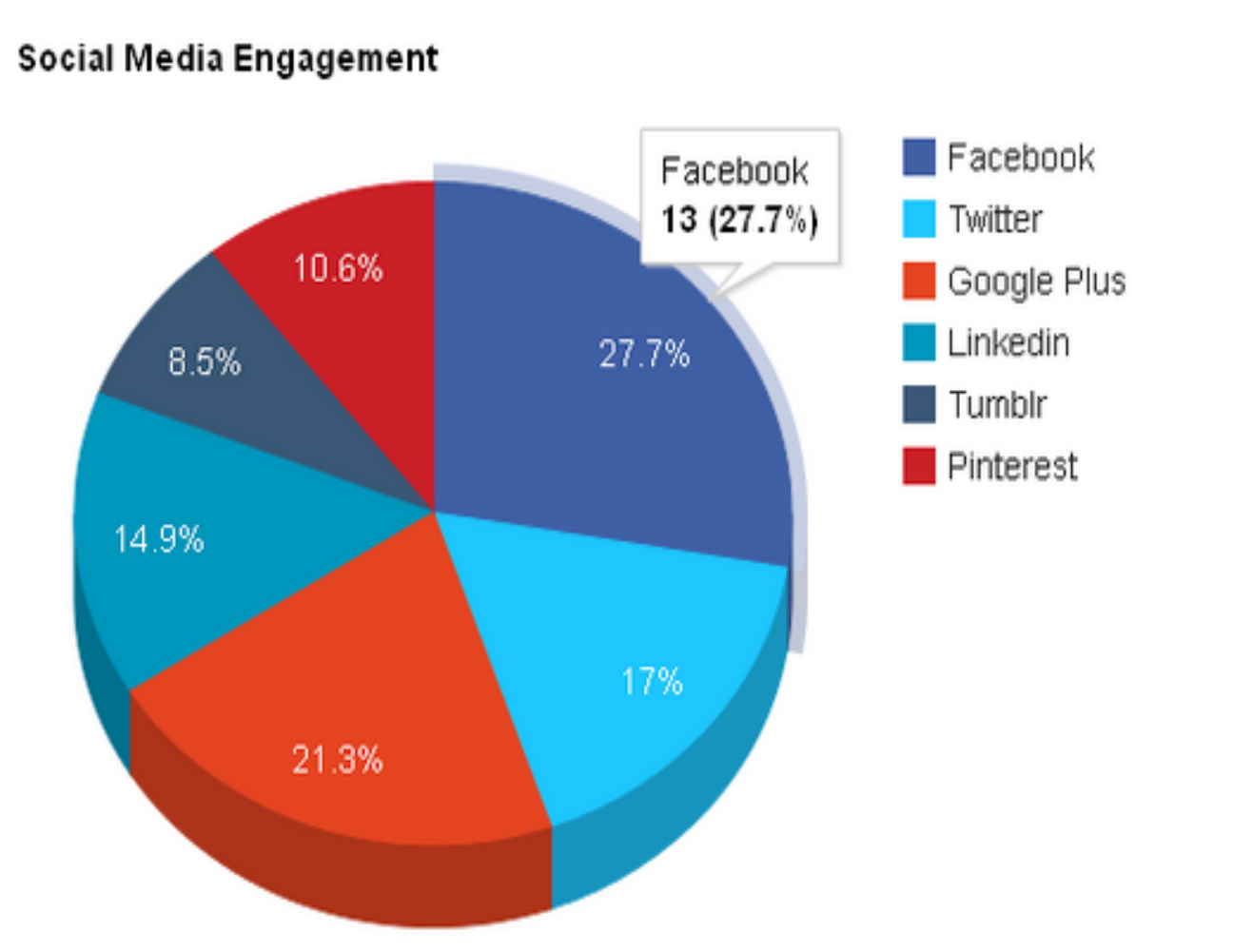
The most commonly used graphical representation of qualitative data are **bar charts** and **pie charts**.

Pie Chart

A pie chart is a circular diagram giving various fractions of section of a given data. The total number of observations of the data is represented by a pie which is denoted by a circle. The pie is then cut into slices (sectors) where each slice represents a category of the data. The size of a slice is proportional to the relative frequency of a category. A pie chart is often used in newspapers, magazines and articles to depict budgets and other economic information. In constructing a pie chart, we represent the total number of observations by a circle of an angle of 360°. The angle of a slice (sector) at center of a pie (circle) is given by the product:
Relative Frequency $\times$ 360

Graphical Representation of Qualitative Data

An example of a Pie Chart



Pie Chart

Example

The table below shows the US Education ratings by four hundred Educators.

Rating	Frequency
A	35
B	260
C	93
D	12

Construct a table showing the ratings along with the frequencies, relative frequencies, percentages, and sector angles necessary to construct the pie chart, and hence draw the pie chart for the ratings.

Pie Chart

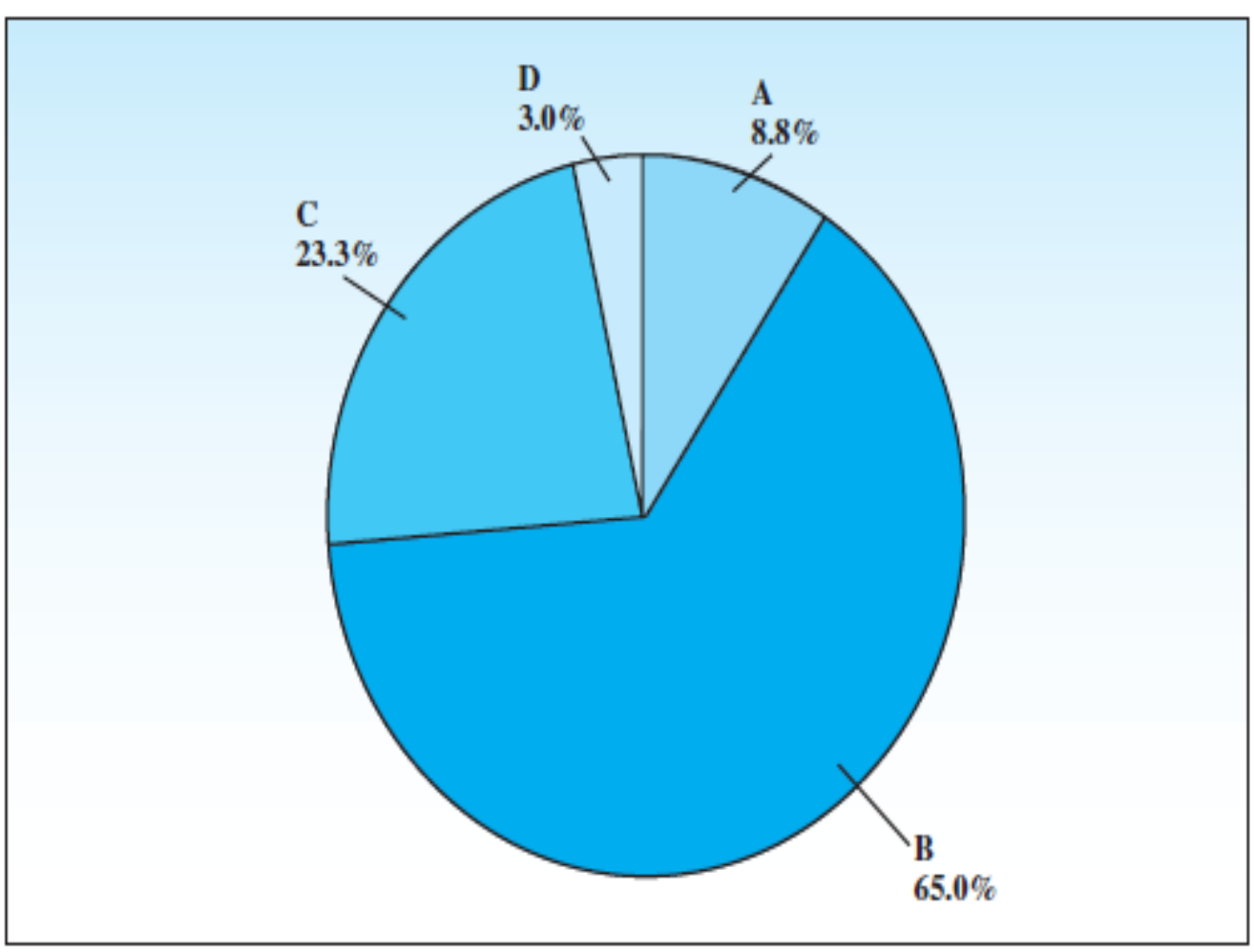
Solution

The table below shows the US Education ratings by four hundred Educators.

Rating	**Frequency**	**Rel. Freq.**	**Percent**	**Angle**
A	35	$\frac{35}{400} = 0.09$	9%	$0.09 \times 360^\circ = 32.4$
B	260	$\frac{260}{400} = 0.65$	65%	$0.65 \times 360^\circ = 234.0$
C	93	$\frac{93}{400} = 0.23$	23%	$0.23 \times 360^\circ = 82.8$
D	12	$\frac{12}{400} = 0.3$	3%	$0.03 \times 360^\circ = 10.8$

Pie Chart

Solution Cont'd



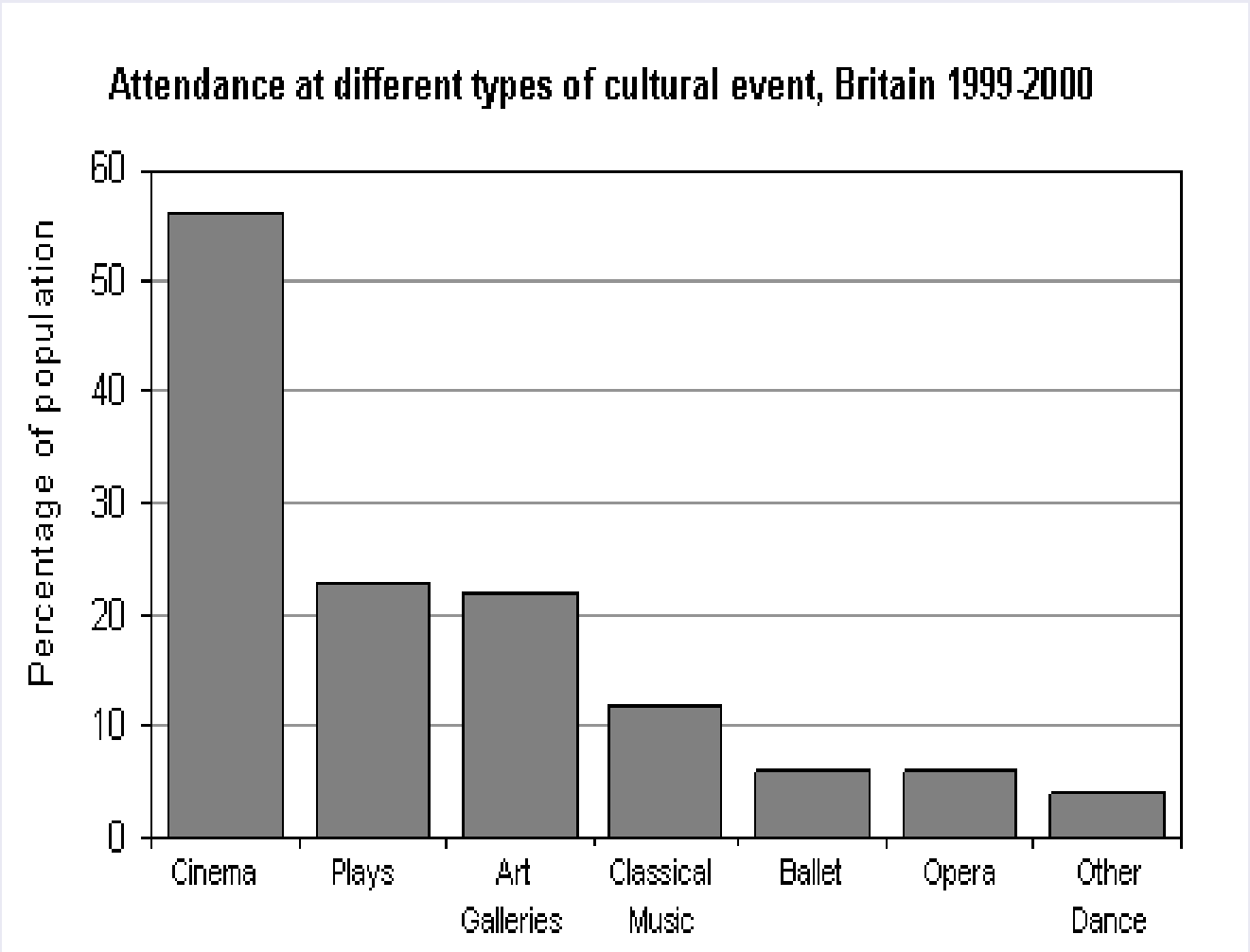
Graphical Representation of Qualitative Data

Bar Charts

Bar charts are a type of graph that are used to display and compare the number, frequency or other measure (e.g. mean) for different discrete categories of data.Bar charts are one of the most commonly used types of graph because they are simple to create and very easy to interpret. They are also a flexible chart type and there are several variations of the standard bar chart including horizontal bar charts, grouped or component charts, and stacked bar charts.

Graphical Representation of Qualitative Data

An example of a Bar Chart



Graphical Representation of Qualitative Data

Bar Chart Continued

In the example above, which shows the percentage of the British population who attended different types of cultural events during 1999-2000, the types of event are the discrete categories of data. The chart is constructed such that the lengths of the different bars are proportional to the size of the category they represent. The x-axis represents the different categories and so has no scale. In order to emphasize the fact that the categories are discrete, a gap is left between the bars on the x-axis. The y-axis does have a scale and this indicates the units of measurement.

Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.
Two graphical techniques for identifying outliers are **scatter plots** and **box plots**.
A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median. The following quantities (called fences) are needed for identifying extreme values in the tails of the distribution:

1. Lower inner fence
2. Upper inner fence
3. Lower outer fence
4. Upper outer fence

Outliers

How to find the fences

1. Lower inner fence $= Q_1 - 1.5\text{IQ}$
2. Upper Inner fence $= Q_3 + 1.5\text{IQ}$
3. Lower outer fence $= Q_1 - 3\text{IQ}$
4. Upper outer fence $= Q_3 + 3\text{IQ}$

A point beyond an inner fence on either side is considered a **mild outlier**.
A point beyond an outer fence is considered an **extreme outlier**.

Outliers

Example

The data set of N = 90 ordered observations as shown below is examined for outliers:

30	171	184	201	212	250	265	270	272	289	305
306	322	322	336	346	351	370	390	404	409	411
436	437	439	441	444	448	451	453	470	480	482
487	494	495	499	503	514	521	522	527	548	550
559	560	570	572	574	578	585	592	592	607	616
618	621	629	637	638	640	656	668	707	709	719
737	739	752	758	766	792	792	794	802	818	830
832	843	858	860	869	918	925	953	991	1000	1005
1068	1441									

Outliers

Solution

- Median = $\frac{(n+1)}{2}$ largest data point = the average of the 45th and 46th ordered points = (559 + 560)/2 = 559.5
- Lower quartile = 0.25(N+1)th ordered point = 22.75th ordered point = 411 + 0.75(436-411) = 429.75
- Upper quartile = .75(N+1)th ordered point = 68.25th ordered point = 739 +0.25(752-739) = 742.25
- Interquartile range = 742.25 - 429.75 = 312.5
- Lower inner fence = 429.75 - 1.5 (312.5) = -39.0
- Upper inner fence = 742.25 + 1.5 (312.5) = 1211.0

From an examination of the fence points and the data, one point (1441) exceeds the upper inner fence and stands out as a mild outlier.

Outliers

A histogram with an overlaid box plot are shown below.