

Data on a continuous variable

We now consider tabular and graphical presentations of data sets that contain numerical measurements on a virtually **continuous scale**. Of course, the recorded measurements are always rounded.

In contrast with the discrete case, a data set of measurements on a continuous variable may contain many distinct values. Then, a table or plot of all distinct values and their frequencies will **not** provide a condensed or informative summary of the data.

The two main graphical methods used to display a data set of measurements are the dot diagram and the histogram.

Dot diagrams are employed when there are relatively few observations (less than 20 or 25), **histograms** are used with a larger number of observations.

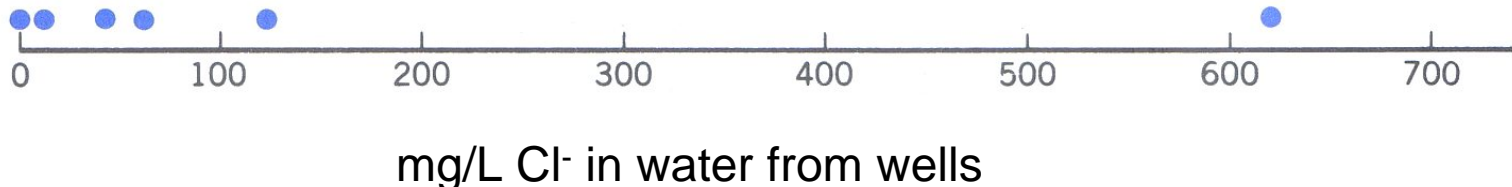
Dot diagram

When the data consist of a small set of numbers, they can be graphically represented by drawing a line with a scale covering the range of values of the measurements. Individual measurements are plotted above this line as prominent dots. The resulting diagram is called a **dot diagram**.

Example 4.

The mg/L of Cl^- in six water samples from wells in a country environment are given in increasing order by: 3, 15, 46, 64, 126, 623.

The data extend from 3 to 623. Drawing a line segment from 0 to 700, we can plot the data as shown in the Figure. The diagram shows a cluster of small values and a single rather large value.



Frequency distributions on intervals

When the data consist of a large number of measurements, a dot diagram may be quite tedious to construct. More seriously, overcrowding of the dots will cause them to smear and mar the clarity of the diagram. In such cases, it is convenient to **condense** the data by grouping the observations according to intervals and recording the frequencies of the intervals. The main steps in this process are outlined as follows.

Constructing a Frequency Distribution for a Continuous Variable

1. Find the minimum and the maximum values in the data set.
2. Choose intervals or cells of equal length that cover the range between the minimum and the maximum without overlapping. These are called **class intervals**, and their end points **class boundaries**.
3. Count the number of observations in the data that belong to each class interval. The count in each class is the **class frequency** or **cell frequency**.
4. Calculate the **relative frequency** of each class by dividing the class frequency by the total number of observations in the data:

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

The choice of the number and position of the class intervals is primarily a **matter of judgement** guided by the following considerations:

❑ The number of classes usually ranges from 5 to 15, depending on the number of observations in the data.

❑ Grouping the observations sacrifices information concerning how the observations are distributed within each cell. With too few cells, the loss of information is serious. On the other hand, if one chooses too many cells and the data set is relatively small, the frequencies from one cell to the next would jump up and down in a chaotic manner and no overall pattern would emerge.

As an initial step, frequencies may be determined with a large number of intervals that can later be combined as desired in order to obtain a smooth pattern of the distribution.

Example 5

Content of Na^+ in water (mg/L) in 40 samples from different wells in an urban area.

TABLE

3.20	11.70	13.64	15.60	15.89	28.44	29.07	37.34
41.81	43.35	43.94	49.51	49.82	51.20	51.43	52.47
53.72	53.92	54.03	56.89	63.80	66.40	68.64	70.15
70.98	74.52	76.68	77.84	80.91	84.04	85.70	86.48
88.92	89.28	91.36	91.62	98.79	102.39	104.21	124.27

Construct a frequency distribution of the sales data.

To construct a frequency distribution, we first notice that the minimum is 3.20 and the maximum is 124.27. We choose class intervals of length 25 as a matter of convenience.

The selection of class boundaries is more complicated. Because the data have two decimal places, we could add a third decimal figure to avoid the possibility of any observation falling exactly on the boundary. Alternatively we could write 0-25 and make the endpoint convention that the left-hand limit is included but no the right

TABLE Frequency distribution for Na⁺ (mg/L) data

Class Interval	Frequency	Relative Frequency
0–25	5	$\frac{5}{40} = .125$
25–50	8	$\frac{8}{40} = .200$
50–75	13	$\frac{13}{40} = .325$
75–100	11	$\frac{11}{40} = .275$
100–125	3	$\frac{3}{40} = .075$
Total	40	1.000

Remark: the rule requiring equal class intervals is inconvenient when the data are spread over a wide range, but are highly concentrated in a small part of the range with relatively few numbers elsewhere.

Using smaller intervals where the data are highly concentrated and larger intervals where the data are sparse helps to reduce the loss of information due to grouping.

Histograms

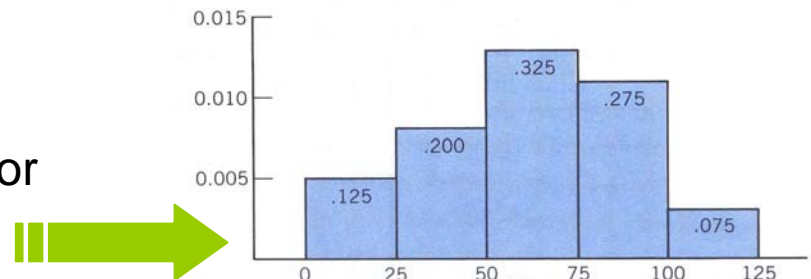
A frequency distribution can be graphically presented as a **histogram**.

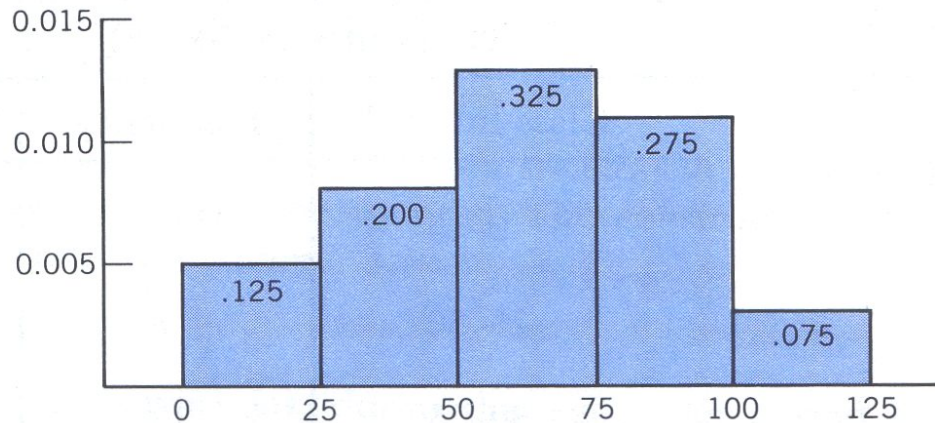
To draw a histogram, we first mark the class intervals on the horizontal axis. On each interval, we then draw a **vertical rectangle** whose area represents the relative frequency, that is the proportion of the observations occurring in that class interval.

The total area of all rectangles equals the sum of the relative frequencies, which is 1.

The total area of a histogram is 1.

Histogram for the frequency distribution for
Na⁺ (mg/L) data





Histogram for the frequency distribution for Na⁺ (mg/L) data

For example the rectangle drawn on the class interval 0-25 has its area = $0.005 \times 25 = 0.125$, which is the relative frequency of this class.

Actually, we determined the height 0.005 as:

$$\text{Height} = \frac{\text{Relative frequency}}{\text{Width of interval}} = \frac{0.125}{25} = 0.005$$

Key Idea Guidelines for Number of Intervals

Sample Size	Number of Intervals
Fewer than 50	5-6
50 to 100	6-8
Over 100	8-10

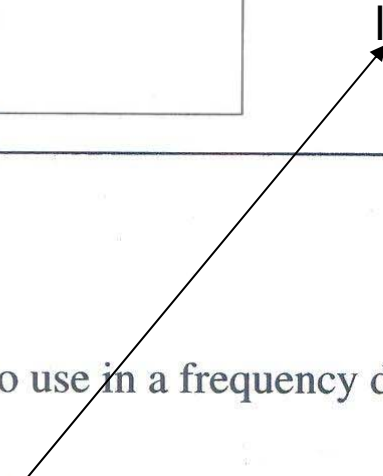
Key Idea Sturges's Rule

Let k = approximate number of classes to use in a frequency distribution

n = number of observations

$$k = 1 + 3.322(\log n)$$

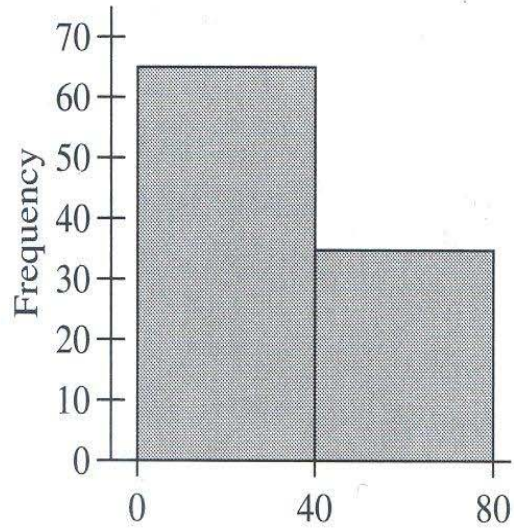
\log_{10}



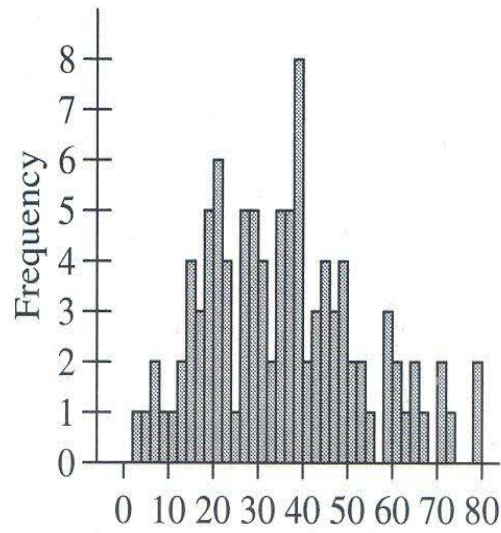
Key Idea Interval Width

$$\text{Interval width} = \frac{\text{largest number} - \text{smallest number}}{\text{number of intervals}}$$

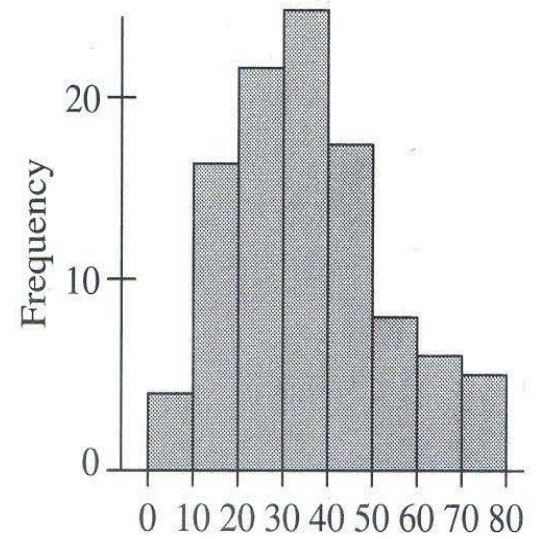
Intervals normally are of equal size.



(a) Too few bars.



(b) Too many bars.



(c) Follows Key Ideas

Cumulative frequency distribution

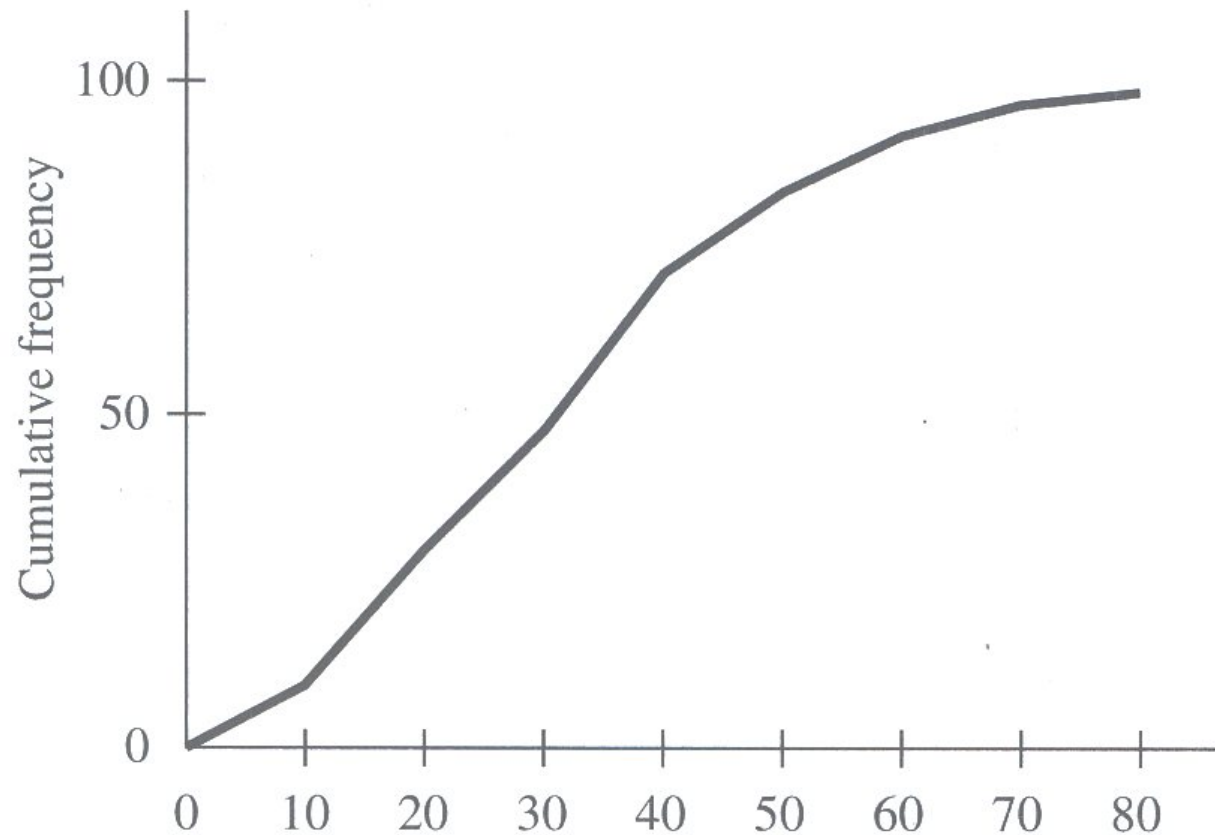
A cumulative frequency distribution contains the total number of observations whose values are less than the upper limit of each interval.

It is constructed by adding the frequencies of all frequency distribution intervals up to and including the present interval.

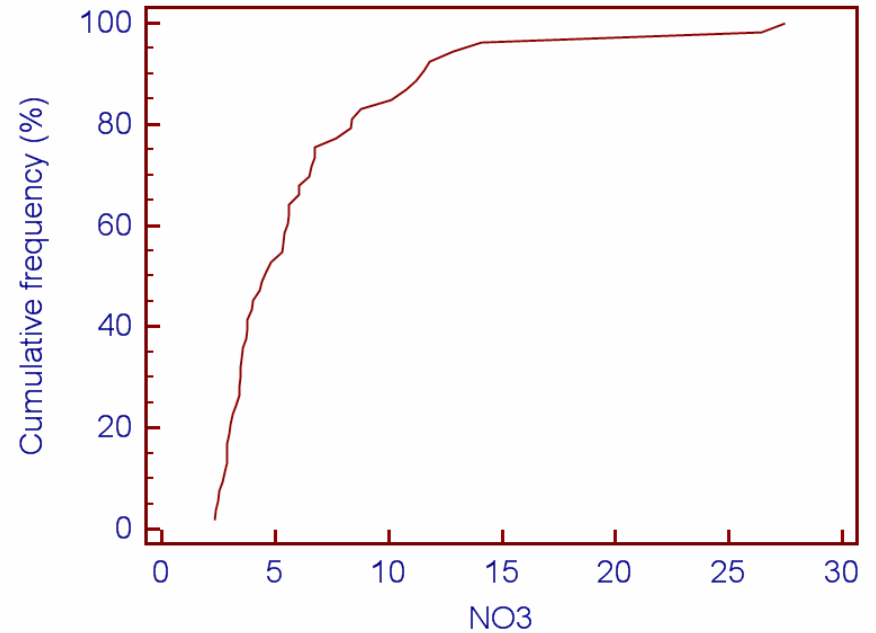
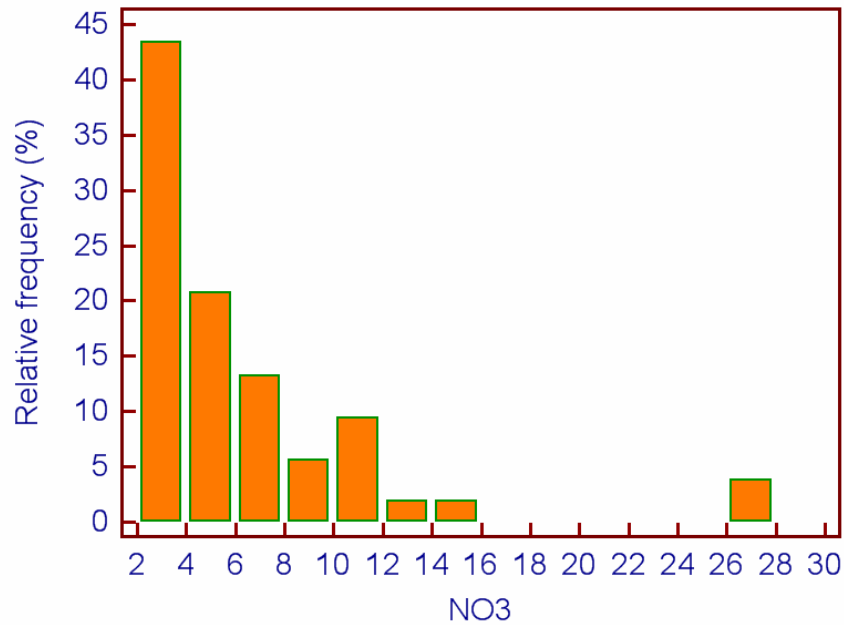
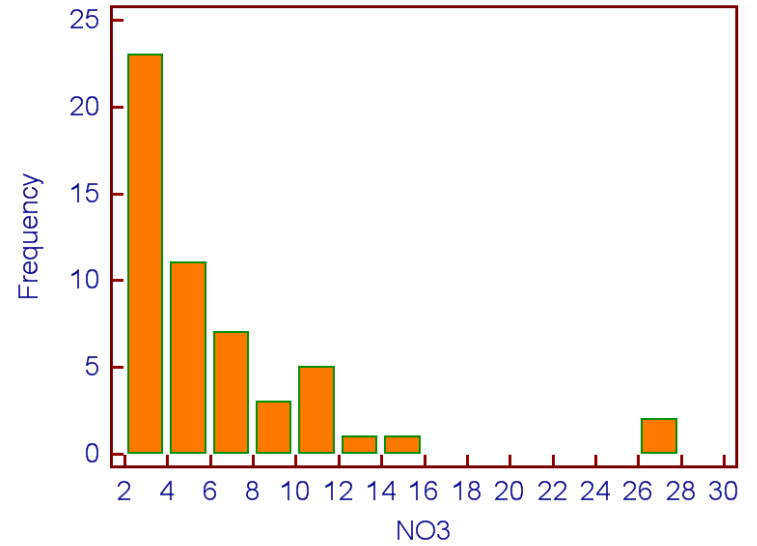
variable	Cumulative frequency %	Relative cumulative frequency
< 10	5	0.05
< 20	20	0.20
< 30	41	0.41
< 40	65	0.65
< 50	81	0.81
< 60	89	0.89
< 70	95	0.95
< 80	100	1.00

It is important to note that the final interval in a cumulative frequency distribution must contain the total number of observations in the sample and a cumulative percentage of 100%.

An **ogive** is a line graph connecting points that are the cumulative percentage of observations below the upper limit of each class in a cumulative frequency distribution.



NO₃ (mg/L) in alpine rivers



Stem-and-leaf display

A stem-and-leaf display provides a more efficient variant of the histogram for displaying data, especially when the observations are two-digit numbers.

TABLE Na⁺ (mg/L) in water

75	98	42	75	84	87	65	59	63
86	78	37	99	66	90	79	80	89
68	57	95	55	79	88	76	60	77
49	92	83	71	78	53	81	77	58
93	85	70	62	80	74	69	90	62
84	64	73	48	72				

This plot is obtained by sorting the observations into rows according to their leading digit.

The stem-and-leaf display for the data of the table is shown in the figure.

TABLE Stem-and-Leaf Display for the Examination Scores

0	
1	
2	
3	7
4	289
5	35789
6	022345689
7	01234556778899
8	00134456789
9	0023589

Stem-and leaf diagram: variable with a one-digit stem

1.8, 1.9, 2.4, 2.6, 2.8,....



Stem	Leaf
1	8 9
2	4 6 8
3	0 0 2 3 4 7 8
4	0 1 2 2 5 5 5 6 7 9
5	
6	1 3 4

Cumulative Frequency	Stem	Leaf
1	21	2
3	22	2 9
7	23	3 4 5 9
13	24	0 1 3 4 7 9
19	25	1 2 3 5 5 7
24	26	1 1 1 2 6
30	27	1 2 3 5 6 8
40	28	0 2 3 4 4 4 5 6 9 9
51	29	0 1 2 2 4 4 4 5 7 7 7
(10)	30	1 1 1 2 6 7 8 8 8 9
51	31	0 1 1 1 2 4 5 6 8
42	32	1 1 4 5 6 8 9
35	33	1 2 3 5 7 8 8 9
27	34	0 0 1 1 1 3 3 3 4 6
17	35	1 6 7 7
13	36	0 1 2 5 5 6 6 8 8
4	37	2 3
2	38	0 7



21.2, 22.2, 22.9, 23.3, 23.4,....

The diagram can be associate with the cumulative frequency (the first column)

Measures of center

The described graphic procedures help us to visualise the pattern of a data set of measurements. To obtain a more objective summary description and a comparison of data sets, we must go one step further and obtain numerical values for the location or **center** of the data and the amount of **variability** present.

Because data are normally obtained by sampling from a large population, our discussion of numerical measures is restricted to data arising in this context.

Moreover, when the population is finite and completely sampled, the same arithmetic operations can be carried out to obtain numerical measures for the population.

To effectively present the ideas and associated calculations, it is convenient to represent a data set by symbols.

The **sample mean** of a set of n measurements x_1, x_2, \dots, x_n is the sum of these measurements divided by n . The sample mean is denoted by \bar{x} .

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{or} \quad \frac{\sum x_i}{n}$$

Example

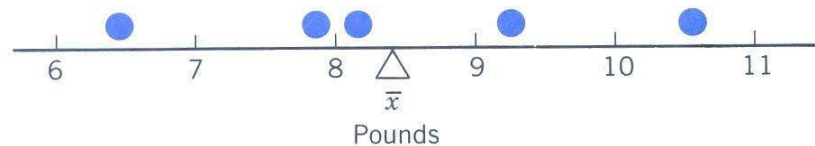
The birth weights in pounds of five babies born in a hospital on a certain day are 9.2, 6.4, 10.5, 8.1, and 7.8. Obtain the sample mean and create a dot diagram.

SOLUTION

The mean birth weight for these data is

$$\bar{x} = \frac{9.2 + 6.4 + 10.5 + 8.1 + 7.8}{5} = \frac{42.0}{5} = 8.4 \text{ pounds}$$

The dot diagram of the data appears in Figure where the sample mean (marked by Δ) is the balancing point or center of the picture.



Dot diagram and the sample mean for the birth-weight data.

Key Idea Arithmetic Mean

The **arithmetic mean**, or **average**, is the sum of the data values divided by the number of observations. If the data set is a sample, the sample mean, \bar{X} , is

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where n is the sample size and \sum means “to add.” If the data set is a population, the **population mean**, μ , is

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

where N is the population size.

Key Idea Statistic and Parameter

A **statistic** is a numerical measure computed from a sample. A **parameter** is a numerical measure computed from a population.

Another measure of center is the middle value.

The **sample median** of a set of n measurements x_1, \dots, x_n is the middle value when the measurements are arranged from smallest to largest.

Roughly speaking, the median is the value that divides the data into two equal halves. In other words, 50% of the data lie below the median and 50% above it. If n is an odd number, there is a unique middle value and it is the median. If n is an even number, there are two middle values and the median is defined as their average. For instance, the data 3, 5, 7, 8 have two middle values 5 and 7, so the median = $(5 + 7)/2 = 6$.

Example

SOLUTION

Find the median of the birth-weight data

The measurements, ordered from smallest to largest, are

6.4, 7.8, 8.1, 9.2, 10.5

The middle value is 8.1, and the median is therefore 8.1 pounds.

The median M

To find the **median** of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n + 1)/2$ observations up from the start of the list.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the start of the list.

The median M is **not affected** by a few very small or very large observations, whereas the presence of such extremes will have a considerable effect on the mean. For extremely **asymmetrical distributions**, the median is likely to be a **more sensible measure** of center than the mean.

Example



Calculate the median of the following values:
46, 15, 3, 126, 623, 64

To find the median, first we order the data. The ordered values are:

3, 15, 46, 64, 126, 623

There are two middle values, so:

$$\text{Median} = \frac{46 + 64}{2} = 55$$

The sample mean is:

$$\bar{x} = \frac{3 + 15 + 46 + 64 + 126 + 623}{8} = \frac{877}{8} = 146.2$$

One large value greatly inflates the mean. Here the median appears to be a better indicator of the center than the mean.

If the number of observations is quite large (greater than, say, 25 or 30), it is sometimes useful to extend the notion of the median and divide the ordered data set into **quarters**.

Just as the point for division into halves is called the median, the points for division into quarters are called **quartiles**.

Thus, the points of division into more general fractions are called **percentiles**.

The sample $100p$ -th percentile is a value such that after the data are ordered from smallest to largest, at least $100p\%$ of the observations are at or below this value and at least $100(1 - p)\%$ are at or above this value.

If we take $p = .5$, the above conceptual description of the sample $100(.5) = 50$ th percentile specifies that at least half the observations are equal or smaller and at least half are equal or larger. If we take $p = .25$, the sample $100(.25) = 25$ th percentile has proportion one-fourth of the observations that are the same or smaller and proportion three-fourths that are the same or larger.

Calculating the Sample $100p$ -th Percentile

1. Order the data from smallest to largest.
2. Determine the product (*sample size*) \times (*proportion*) = np .

If np is not an integer, round it up to the next integer and find the corresponding ordered value.

If np is an integer, say, k , calculate the average of the k th and $(k + 1)$ st ordered values.

Sample Quartiles

Lower (first) quartile	$Q_1 = 25$ th percentile
Second quartile (or median)	$Q_2 = 50$ th percentile
Upper (third) quartile	$Q_3 = 75$ th percentile

Example

The data from 50 measurements of the traffic noise level at an intersection are already ordered from smallest to largest in Table . Locate the quartiles and also compute the 10th percentile.

TABLE Measurements of Traffic Noise Level in Decibels

52.0	55.9	56.7	59.4	60.2	61.0	62.1	63.8	65.7	67.9
54.4	55.9	56.8	59.4	60.3	61.4	62.6	64.0	66.2	68.2
54.5	56.2	57.2	59.5	60.5	61.7	62.7	64.6	66.8	68.9
55.7	56.4	57.6	59.8	60.6	61.8	63.1	64.8	67.0	69.4
55.8	56.4	58.9	60.0	60.8	62.0	63.6	64.9	67.1	77.1

Courtesy of J. Bollinger.

SOLUTION

To determine the first quartile, we take $p = .25$ and calculate the product $50 \times .25 = 12.5$. Because 12.5 is not an integer, we take the next larger integer 13. In Table we see that the 13th-ordered observation is 57.2. So, the first quartile is $Q_1 = 57.2$.

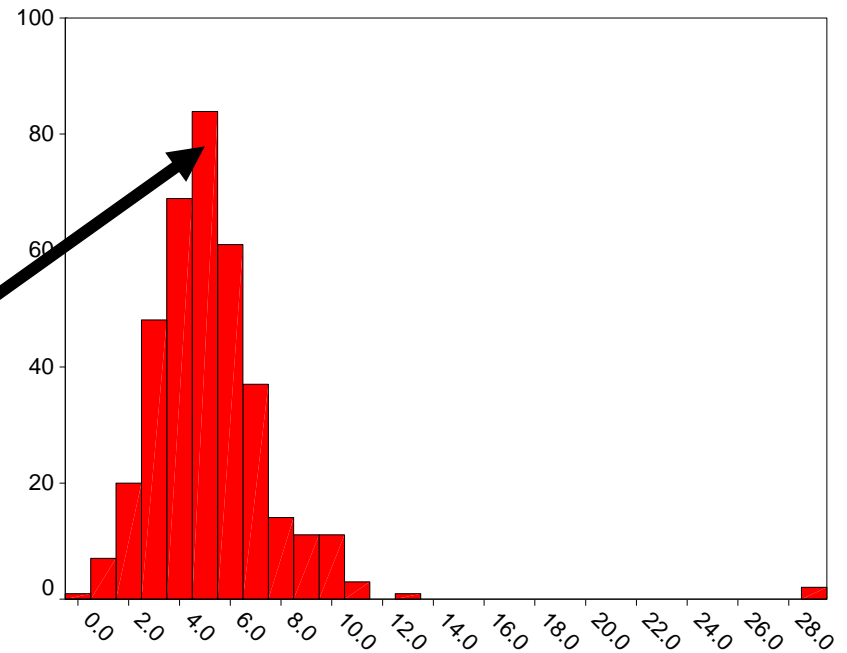
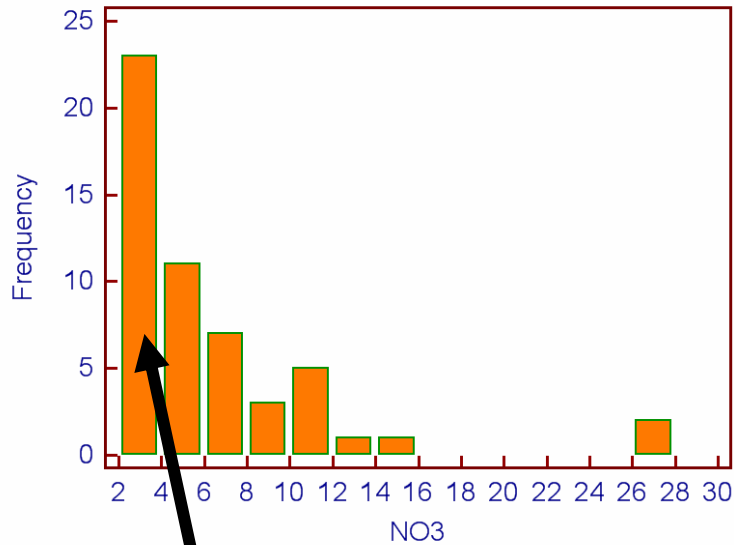
We confirm that this observation has 13 values *at or below* it and 38 observations *at or above* so that it does satisfy the conceptual definition.

For the median, we take $p = .5$ and calculate $50 \times .5 = 25$. Because this is an integer, we average the 25th and 26th smallest observations to obtain the median $= (60.8 + 61.0)/2 = 60.9$.

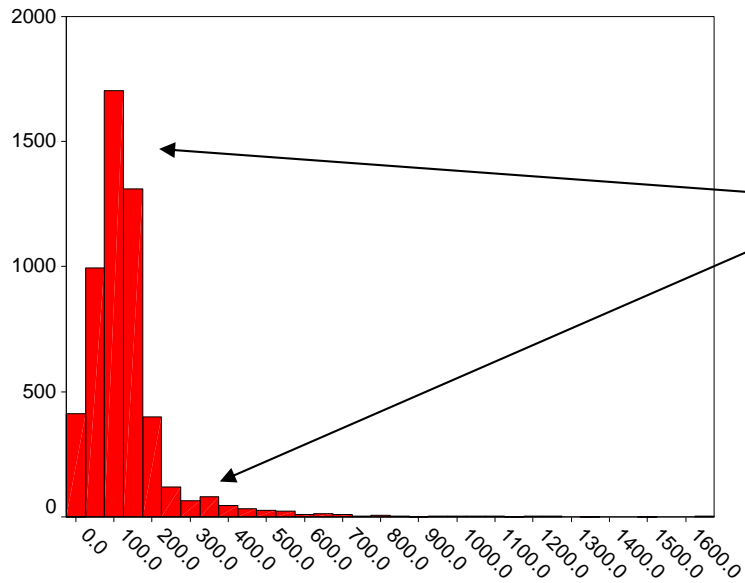
Next, to determine the third quartile, we take $p = .75$ and calculate $50 \times .75 = 37.5$. The next larger integer is 38 and the third quartile $Q_3 = 64.6$. More simply, we could mimic the calculation of the first quartile, but now count down 13 observations from the top.

For the 10th percentile, we determine $50 \times .10 = 5$. The product is an integer so we take the average of the 5th and 6th observations, $(55.8 + 55.9)/2 = 55.85$ as the 10th percentile. Only 10% of the 50 measurements of noise level were quieter than 55.85 decibels.

The **mode** is any value occurring most frequently in the set of observations. It is convenient to arrange observations in increasing order as an aid to seeing how often each value occurs.



modal class



Bimodal?

The application of the logarithmic transformation allow us to identify the presence of a bimodality (two set of data with a different statistics of central tendency)

