A Brief Definition of "Statistics": [Navidi Chapter 1 & Devore Chapter 1]

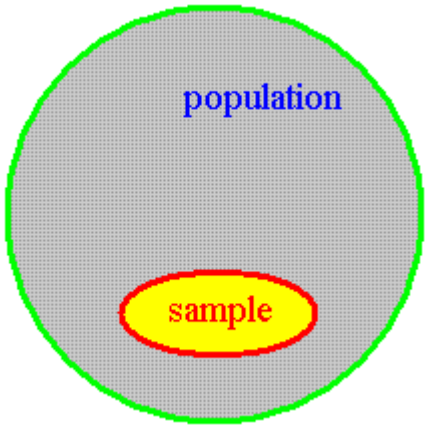
"Statistics" is the science of making decisions in the absence of certainty.

Some more definitions:

A **population** is a set of objects.

A **sample** is a subset of the population.

A **simple random sample** of size n is a sample chosen by a method such that each collection of n items in the population is equally likely to be in the sample as every other collection of n items (as in a fair lottery).



Probability uses knowledge of the population to predict the contents of a sample when the sample is drawn randomly from the population.

Inferential **Statistics** uses knowledge of a random sample to deduce some properties of the population from which the sample was drawn. A good foundation in probability theory is needed before one can study inferential statistics effectively.

The design of this course is:

Descriptive Statistics
Devore Ch. 1; Navidi Ch. 1
Lecture Notes Ch. 1

Probability
Devore Ch. 2-6;
Navidi Ch. 2-4
Notes Ch. 2-11

Inferential Statistics
Devore Ch. 7-9, 12, 14;
Navidi Ch. 5-7
Notes Ch. 11-15

There is also some coverage of decision trees, an introduction to probability through the concept of a "fair bet" and Bayesian confidence intervals. These topics are not in the textbooks. The Devore textbook does not cover the topic of propagation of error.

Random Samples (Navidi textbook: Chapter 1)

Example 1.01

A quality control inspector selects every one hundredth item from a production line for testing, in order to deduce whether or not the production line is operating properly. With a daily production run of 10,000 items, there are 100 items in the sample. Is the sample so drawn a simple random sample?

Example 1.02

An engineer wishes to judge the breaking strength of a set of 10 kg beams that have been delivered in a shipment of 500 beams stacked on a pallet in a large cube. The engineer selects three beams from the top of the stack for testing. Is the sample so drawn a simple random sample?

Example 1.03

The output from one day's run of a production line is labelled from 1 to 10,000. The inspector has a list of 100 distinct numbers between 1 and 10,000, provided from a random number generator and selects the 100 items that match the 100 numbers for the sample to be tested. Is the sample so drawn a simple random sample?

In all of the cases above, the population is **tangible** – it consists of a finite number of real objects.

A random sample of repeated measurements of the same object or process under identical conditions is drawn from a **conceptual population**: the infinite set of all values that could have been observed.

Summarizing Data

Given a set of observations $\{ x_1, x_2, \ldots, x_n \}$, one can summarize the observations using various types of graphical display (time series plot, bar chart, histogram, pie chart, pictogram, etc.) or by numerical measures (stem and leaf diagram, frequency table) or single numbers measuring typical value, spread of values, etc.

Example 1.04 Course text (Navidi, third edition), table 1.2 p. 21

[This data set is available from the course web site, at
`www.engr.mun.ca/~ggeorge/4421/demos/t1/index.html`]

The data set below consists of observations on particulate matter emissions x (in g/gal) for 62 vehicles driven at high altitude.

7.59	6.28	6.07	5.23	5.54	3.46	2.44	3.01	13.63	13.02
23.38	9.24	3.22	2.06	4.04	17.11	12.26	19.91	8.50	7.81
7.18	6.95	18.64	7.10	6.04	5.66	8.86	4.40	3.57	4.35
3.84	2.37	3.81	5.32	5.84	2.89	4.68	1.85	9.14	8.67
9.52	2.68	10.14	9.20	7.31	2.09	6.32	6.53	6.32	2.01
5.91	5.60	5.61	1.50	6.46	5.29	5.64	2.07	1.11	3.32
1.83	7.56								

By itself, this table is not very helpful as we try to grasp the overall picture of emissions.
One way to improve visibility is simply to rearrange these data into ascending order:

1.11	1.50	1.83	1.85	2.01	2.06	2.07	2.09	2.37	2.44
2.68	2.89	3.01	3.22	3.32	3.46	3.57	3.81	3.84	4.04
4.35	4.40	4.68	5.23	5.29	5.32	5.54	5.60	5.61	5.64
5.66	5.84	5.91	6.04	6.07	6.28	6.32	6.32	6.46	6.53
6.95	7.10	7.18	7.31	7.56	7.59	7.81	8.50	8.67	8.86
9.14	9.20	9.24	9.52	10.14	12.26	13.02	13.63	17.11	18.64
19.91	23.38								

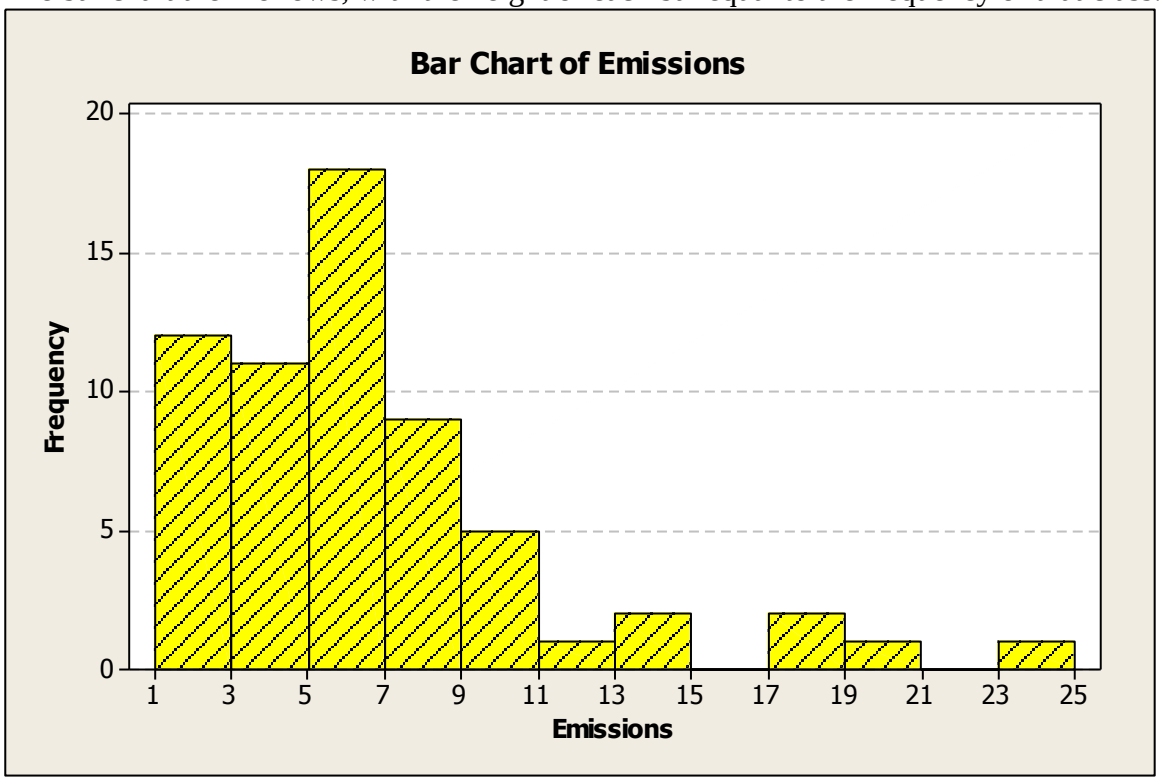
(a) Construct a bar chart of the data, using class intervals of equal width, with the first interval having lower limit 1.0 (inclusive) and upper limit 3.0 (exclusive).

Example 1.04 (continued)

Let us first generate a **frequency table** for emissions manually.

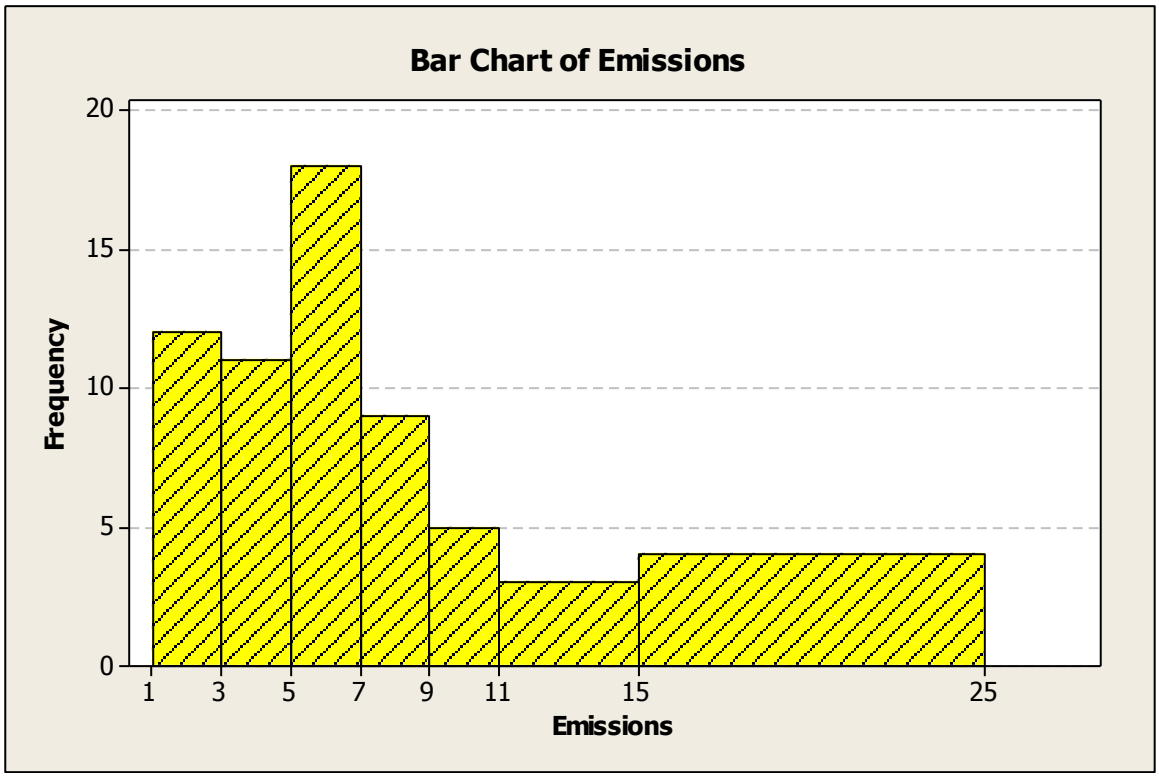
Interval for x	Frequency f
$1 \le x < 3$	12
$3 \le x < 5$	11
$5 \le x < 7$	18
$7 \le x < 9$	9
$9 \le x < 11$	5
$11 \le x < 13$	1
$13 \le x < 15$	2
$15 \le x < 17$	0
$17 \le x < 19$	2
$19 \le x < 21$	1
$21 \le x < 23$	0
$23 \le x < 25$	1
Total:	62

The bar chart then follows, with the height of each bar equal to the frequency of that class:



Example 1.04 (continued)

However, the tail on the right has a rough appearance. It can be smoothed out by combining some adjacent class intervals. However, that would cause a misleading appearance of greater numbers in that tail of the bar chart:



There is a subtle difference between a "bar chart" and a "histogram". A **bar chart** is used for **discrete** (countable) data (such as "number of defective items found in one run of a process") or nonnumeric data (such as "engineering major chosen by students"). The bars are drawn with arbitrary (often equal) width. No two bars should touch each other. The height of each bar is proportional to the frequency.

A **histogram** is used for **continuous** data (such as "shear stress" or "weight" or "time", where between any two possible values another possible value can always be found). [A histogram can also be used for discrete data.] Each bar covers a continuous interval of values and just touches its neighbouring bars without overlapping. Every possible value lies in exactly one interval. Unlike a bar chart, it is the *area* of each bar that is proportional to the frequency in that interval. Only if all intervals are of equal width will the histogram have the same shape as the bar chart.

The **relative frequency** in an interval is the proportion of the total number of observations that fall inside that interval. A relative frequency histogram can then be generated, with bar height given by

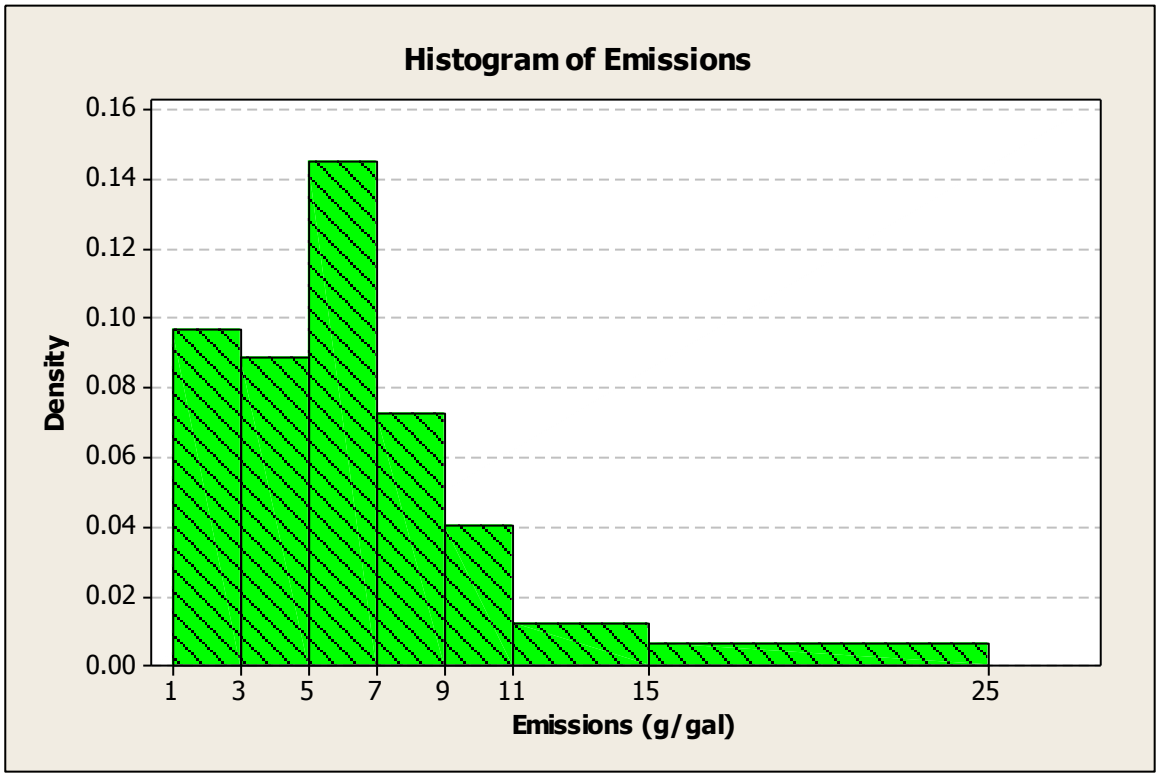
Example 1.04 (continued)

bar height =

The total area of all bars in a relative frequency histogram is always 1. In chapter 6 we will see that the relative frequency histogram is related to the graph of a probability density function, the total area under which is also 1.

Extending the frequency table for the sample of 62 observations of particulate matter emissions:

Interval for x	Frequency f	Rel. Freq. r	Density (= bar heights) d
$1 \le x < 3$	12	.19355	0.09677
$3 \le x < 5$	11	.17742	0.08871
$5 \le x < 7$	18	.29032	0.14516
$7 \le x < 9$	9	.14516	0.07258
$9 \le x < 11$	5	.08065	0.04032
$11 \le x < 15$	3	.04839	0.01210
$15 \le x < 25$	4	.06452	0.00645
Total:	62	1.00000	



These diagrams will be generated during the first tutorial session using Minitab®.
All Minitab sessions are available from the course web site at
" `www.engr.mun.ca/~ggeorge/4421/demos/index.html` ".

Measures of Location

The **mode** is the most common value.

In example 1.04 the mode is

From the frequency tables, the modal class is

A disadvantage of the mode as a measure of location is

The **sample median** $\tilde{x}$ (or the population median $\tilde{\mu}$) is the "halfway value" in an ordered set.
For n data, the median is the $(n + 1)/2$ th value if n is odd.
The median is the semi-sum of the two central values if n is even,
(that is median = [(n/2 th value) + ((n/2 + 1)th value)] / 2).

For the example above,

sample median

In the table of grouped values, the 31st and 32nd values fall in the same class.
The median class is therefore

The **sample arithmetic mean** $\bar{x}$ (or the population mean μ) is the ratio of the sum of the observations to the number of observations.

From individual observations,

and from a frequency table,

For Example 1.04 above, from the 62 raw data (not from the frequency table),

The relative advantages of the mean and the median can be seen from a pair of smaller samples.

Example 1.05

Let $A = \{ 1, 2, 3, 4, 5 \}$ and $B = \{ 1, 2, 3, 23654, 5 \}$.

Then

$\tilde{x} =$ for set A and $\tilde{x} =$ for set B, while

$\bar{x} =$ for set A and $\bar{x} =$ for set B.

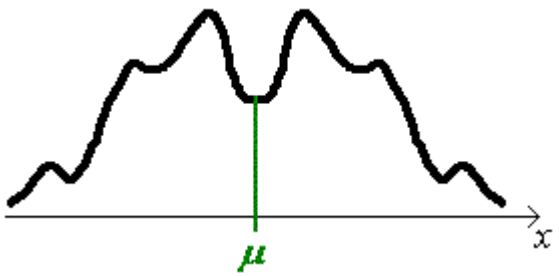
Note that the mode is not well defined for either set.

A disadvantage of the mean as a measure of location is

Advantages of the mean over the median include

- the median uses only the central value(s) while the mean uses all values.
-

For a symmetric population, the mean μ and the median $\tilde{\mu}$ will be equal. If the mode is unique, then it will also be equal to the mean and median of a symmetric population.



The data set in Example 1.04 is far from being symmetric. It shows a strong positive **skew**.

Signs of positive skew include:

Signs of negative skew include:

While conceptual populations can be perfectly symmetric, random samples rarely display perfect symmetry – they may be only approximately symmetric, with mean $\approx$ median and tails of approximately equal length.

Measures of Variation

The simplest measure of variation is the **range** = (largest value – smallest value).
A disadvantage of the sample range is

A disadvantage of the population range is

The effect of outliers can be eliminated by using the distance between the **quartiles** of the data as a measure of spread instead of the full range.

The **lower quartile** q_L is the $\{ (n + 1) / 4 \}$th smallest value.
The **upper quartile** q_U is the $\{ 3(n + 1) / 4 \}$th smallest value.

[Close relatives of the quartiles are the **fourths**.
The lower fourth is the median of the lower half of the data, (including the median if and only if the number n of data is odd).
The upper fourth is the median of the upper half of the data, (including the median if and only if the number n of data is odd).
In practice there is often little or no difference between the value of a quartile and the value of the corresponding fourth.]

The **interquartile range** is $IQR = q_U - q_L$ and
the **semi-interquartile range** is $SIQR = (q_U - q_L) / 2$

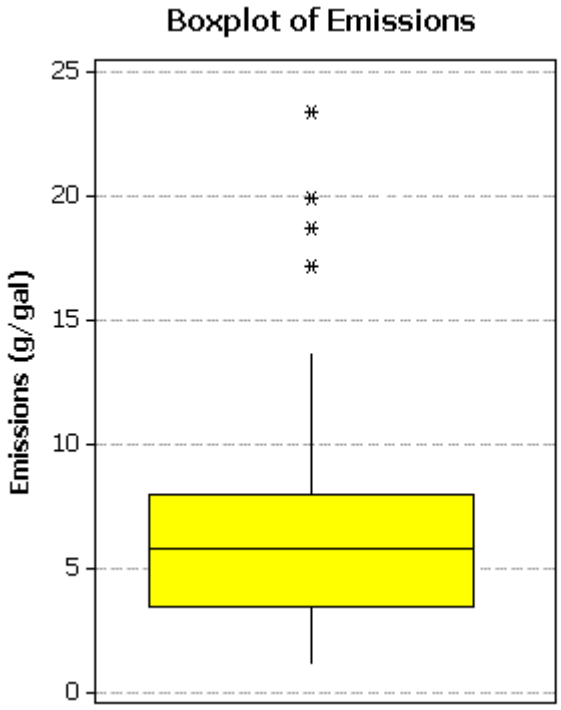
Example 1.04 (again):

$n = 62 \Rightarrow (n + 1) / 4 = 15.75 \Rightarrow q_L =$ value 3/4 of the way from x_{15} to x_{16}

and $3 (n + 1) / 4 = 47.25 \Rightarrow q_U =$ value 1/4 of the way from x_{47} to x_{48}

The semi-interquartile range is then

The **boxplot** illustrates the median, quartiles, outliers and skewness in a compact visual form. The boxplot for example 1.04, as generated by version 14 of MINITAB, is shown below. [and will be generated in the tutorial session.]



Unequal whisker lengths reveal skewness. The whiskers extend as far as the last observation before the inner fence. The fences are *not* plotted by MINITAB.

The inner fences are 1.5 interquartile ranges beyond the nearer quartile, at

$x_L - 1.5\,IQR$ (lower) and $x_U + 1.5\,IQR$ (upper) [–3.411 and +14.819 here]

The outer fences are twice as far away from the nearer quartile, at

$x_L - 3\,IQR$ (lower) and $x_U + 3\,IQR$ (upper) [–9.942 and +21.655 here]

Any observations between inner & outer fences are **mild outliers**, which would be indicated by an open circle (or, in MINITAB, by an asterisk). There are three mild outliers in this example.

Any observations beyond outer fences are **extreme outliers**, which would be indicated by a closed circle (or, in MINITAB, by an asterisk or a zero). There is one extreme outlier in this example.

If you encounter an extreme outlier, then check if the measurement is incorrect or is from a different population. If the observation is genuine from a population that is not strongly skewed, then it is a rare event (< 0.01% in most populations, though not for Example 1.04!).

Measures of variability based on quartiles are not easy to manipulate using calculus methods.

The deviation of the i^{th} observation from the sample mean is $\left(x_i - \bar{x}\right)$. At first sight, one might consider that the sum of all these deviations could serve as a measure of variability. However:

An alternative is the **mean absolute deviation from the mean**, defined as

$$MAD = \frac{1}{n}\sum_{i=1}^{n} \left| x_i - \bar{x} \right|$$

Unfortunately, the function $f(x) = \left| x_i - \bar{x} \right|$ is not differentiable at the one point where the derivative is most needed, at $x = \bar{x}$. Instead, the mean *square* deviation from the mean is used:

The **population variance** σ^2 for a finite population of N values is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \mu\right)^2$$

and the **sample variance** s^2 of a sample of n values is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

The square root of a variance is called the **standard deviation** and is *positive* (unless *all* values are exactly the same, in which case the standard deviation is zero). The reason for the different divisor $(n - 1)$ in the expression for the sample variance s^2 will be explained later.

The MINITAB output for various summary statistics for example 1.04 is shown here:

Variable	N	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum
Emissions	62	6.596	4.519	20.421	1.110	3.425	5.750	7.983	23.380

When calculating a sample variance by hand or on some simpler hand-held calculators, one of the following **shortcut formulæ** may be easier to use:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1} \quad \text{or} \quad s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} \quad \text{or}$$

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} .$$

When all values of x are integers, the last of these three formulæ allows the sample variance to be expressed exactly as a fraction. The formulæ for data taken from a frequency table with m classes are similar:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{m} f_i\left(x_i - \bar{x}\right)^2 \quad \text{or} \quad s^2 = \frac{\sum_{i=1}^{m} f_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{m} f_i x_i\right)^2}{n-1}$$

$$\text{or} \quad s^2 = \frac{\sum_{i=1}^{m} f_i x_i^2 - n\bar{x}^2}{n-1} \quad \text{or} \quad s^2 = \frac{n\sum_{i=1}^{m} f_i x_i^2 - \left(\sum_{i=1}^{m} f_i x_i\right)^2}{n(n-1)}$$

$$\text{where, in each case,} \quad n = \sum_{i=1}^{m} f_i \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^{m} f_i x_i}{\sum_{i=1}^{m} f_i} .$$

However, all of the shortcut formulæ are more sensitive to round-off errors than the definition is.

Example 1.06:

Find the sample variance for the set { 100.01, 100.02, 100.03 } by the definition and by one of the shortcut formulæ, in each case rounding every number that you encounter during your computations to six or seven significant figures, (so that 100.01^2 = 10002.00 to 7 s.f.). The correct value for s^2 in this case is .0001, but rounding errors will cause all three shortcut formulæ to return an incorrect value of zero. (Try it!).

Example 1.07:

Find the sample mean and the sample standard deviation for
x = the number of service calls during a warranty period, from the frequency table below.

x_i	f_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$
0	65		
1	30		
2	3		
3	2		
Sum:	100		

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} =$$

$$s^2 = \frac{n\sum f_i x_i^2 - \left(\sum f_i x_i\right)^2}{n(n-1)} =$$

or

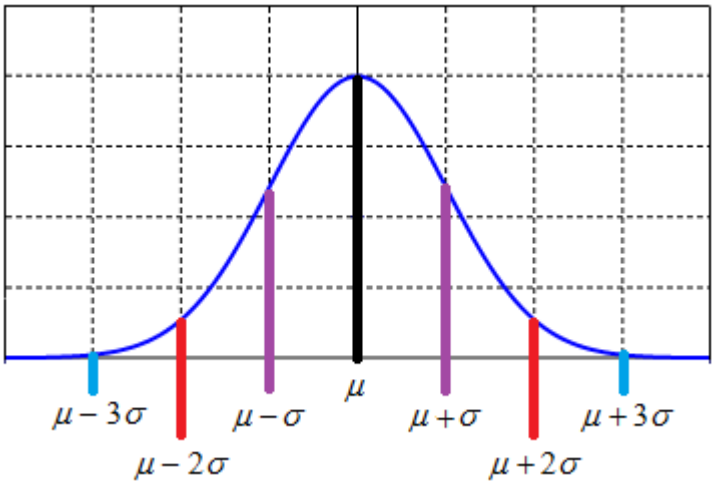
$$s^2 = \frac{1}{n-1}\sum f_i \left(x_i - \bar{x}\right)^2 =$$

For *any* data set:

≥ **3/4** of all data lie within **two** standard deviations of the mean.
≥ **8/9** of all data lie within **three** standard deviations of the mean.

≥ $(1 - 1/\boldsymbol{k}^2)$ of all data lie within ***k*** standard deviations of the mean (Chebyshev's inequality).

For a bell-shaped distribution (for which population mean = population median = population mode and the probability distribution is exactly or nearly Normal [Gaussian]):

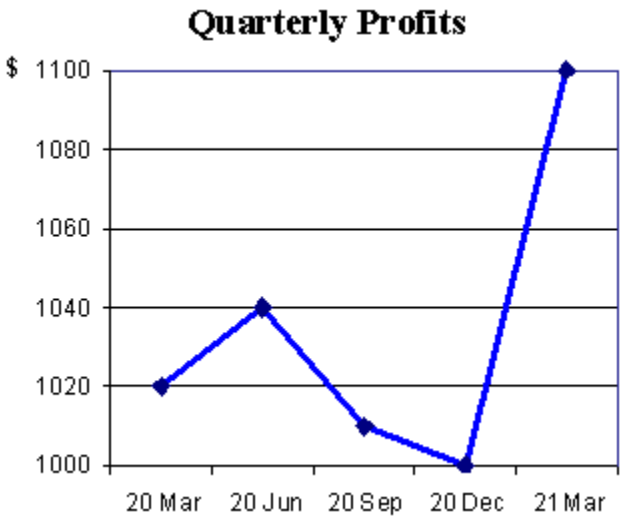


~ **68%** of all data lie within **one** standard deviation of the mean.
~ **95%** of all data lie within **two** standard deviations of the mean.
> **99%** of all data lie within **three** standard deviations of the mean.

Misleading Statistics - Example 1.08

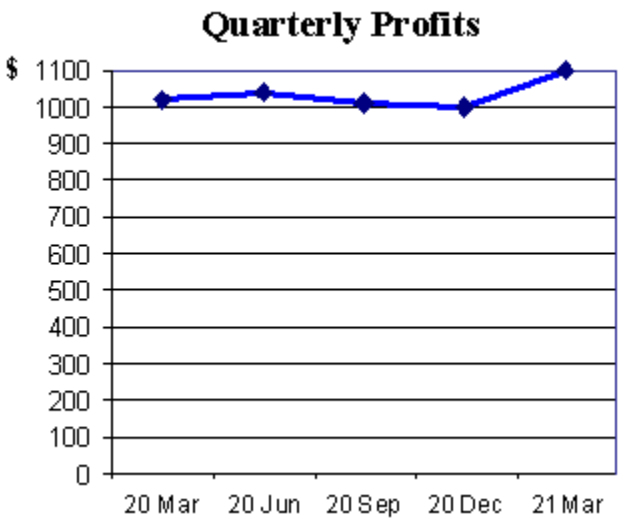
Both graphs below are based on the same information, yet they seem to lead to different conclusions.

"Our profits rose enormously in the last quarter."

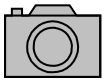


vs.

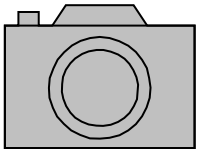
"Our profits rose by only 10% in the last quarter."



Example 1.09 (pictograms)



Number of cameras sold in 2019



Number of cameras sold in 2020

Visual displays can be very misleading. Questions to ask when viewing visual summaries of data include,
for **graphs**:

-
-

for **bar charts / pictograms** :

-

[End of the chapter "Descriptive Statistics"]
