Tables & Graphs

Given a set of observations { $x_1, x_2, ..., x_n$ }, one can summarize using:

		Guidelines:
◘	time series plot	
◘	stem and leaf	**5 to 20 stems**
◘	frequency table	**5 to 20 classes , ≈ √ (# observations)**
◘	bar chart	**5 to 20 classes**
◘	histogram	**5 to 20 classes**
◘	pie chart	**3 to 8 sectors**

◘ pictogram (or other methods) **– is the scale by length, area or volume? [poor choice!]**

Example 2.01 Course text (Devore, seventh edition), ex. 1.2 p. 23 q. 24, modified

[This data set is available from the course web site, at
`www.engr.mun.ca/~ggeorge/3423/Minitab/s01DescStat/index.html`]

The data set below consists of observations on shear strengths x (in pounds) of ultrasonic spot welds made on a certain type of alclad sheet.

5434	4948	4521	4570	4990	5702	5241	5112	5015	4659	4806	4637
5670	4381	4820	5043	4886	4599	5288	5299	4848	5378	5260	5055
5828	5218	4859	4780	5027	5008	4609	4772	5133	5095	4618	4848
5089	5518	5333	5164	5342	5069	4755	4925	5001	4803	4951	5679
5256	5207	5621	4918	5138	4786	4500	5461	5049	4974	4592	4173
5296	4965	5170	4740	5173	4568	5653	5078	4900	4968	5248	5245
4723	5275	5419	5205	4452	5227	5555	5388	5498	4681	5076	4774
4931	4493	5309	5582	4308	4823	4417	5364	5640	5069	5188	5764
5273	5042	5189	4986								

(a) Produce a stem and leaf display of the data.

(b) Construct a bar chart of the data, using ten class intervals of equal width, with the first interval having lower limit 4000 (inclusive) and upper limit 4200 (exclusive). [Such a bar chart will be consistent with one that appeared in the paper “Comparison of Properties of Joints Prepared by Ultrasonic Welding and Other Means”, J. *Aircraft,* 1983, pp. 552-556.]

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

4173	4308	4381	4417	4452	4493	4500	4521	4568	4570	4592	4599
4609	4618	4637	4659	4681	4723	4740	4755	4772	4774	4780	4786
4803	4806	4820	4823	4848	4848	4859	4886	4900	4918	4925	4931
4948	4951	4965	4968	4974	4986	4990	5001	5008	5015	5027	5042
5043	5049	5055	5069	5069	5076	5078	5089	5095	5112	5133	5138
5164	5170	5173	5188	5189	5205	5207	5218	5227	5241	5245	5248
5256	5260	5273	5275	5288	5296	5299	5309	5333	5342	5364	5378
5388	5419	5434	5461	5498	5518	5555	5582	5621	5640	5653	5670
5679	5702	5764	5828								

An additional improvement to the visual appearance is the **stem and leaf** display. Part of the output from a `MINITAB` session, using default values, is reproduced on the left hand side below. The left-most column is a cumulative frequency count from the nearer end. Note how `MINITAB` returns only the thousands and hundreds digits in the stem and the tens digit in the leaf. The units digit is truncated (lost altogether). On the right hand side (to the right of the comment markers ###) is shown a manual version that retains both digits of the leaf.

```
Stem-and-leaf of Shear st  N  = 100
Leaf Unit = 10

                           ### Manual version, retaining
                           ### both digits of the leaf:
                           ###   Stem Leaf
    1   41 7               ###    41   73
    1   42                 ###    42
    3   43 08              ###    43   08 81
    6   44 159             ###    44   17 52 93
   12   45 026799          ###    45   00 21 68 70 92 99
   17   46 01358           ###    46   09 18 37 59 81
   24   47 2457788         ###    47   23 40 55 72 74 80 86
   32   48 00224458        ###    48   03 06 20 23 48 48 59 86
   43   49 01234566789     ###    49   00 18 25 31 48 51 65 68 74 86 90
  (14)  50 00124445667789  ###    50   01 08 15 27 42 43 49 55 69 69 76 78 89 95
   43   51 13367788        ###    51   12 33 38 64 70 73 88 89
   35   52 00124445677899  ###    52   05 07 18 27 41 45 48 56 60 73 75 88 96 99
   21   53 034678          ###    53   09 33 42 64 78 88
   15   54 1369            ###    54   19 34 61 98
   11   55 158             ###    55   18 55 82
    8   56 24577           ###    56   21 40 53 70 79
    3   57 06              ###    57   02 64
    1   58 2               ###    58   28
```

The "`(14)`" means that that stem has 14 leaves, *including the median*.

With the "`Increment`" option of "`Graph > Stem-and-leaf`" set to 200 instead of the default value of 100, the number of stems is reduced to the appropriate number, namely $\sqrt{100} = 10$. `MINITAB`'s output is then

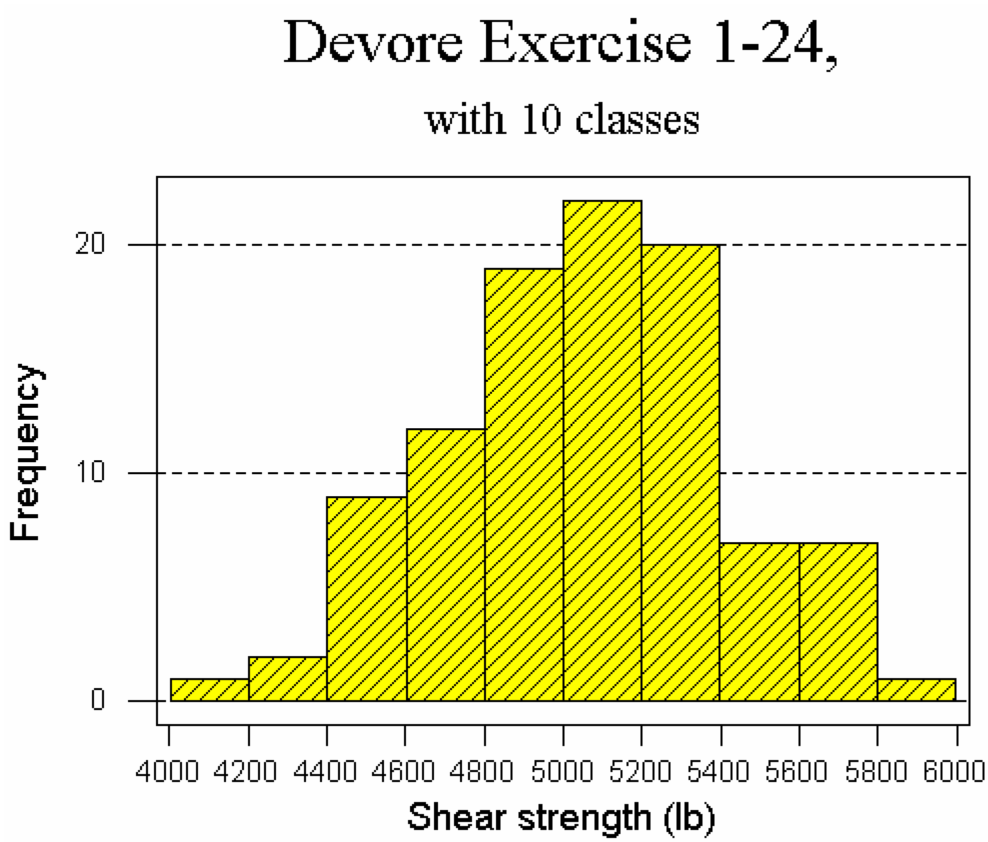
```
Stem-and-leaf of Shear st  N  = 100
Leaf Unit = 100


    1    4 1
    3    4 33
   12    4 444555555
   24    4 666667777777
   43    4 8888888899999999999
  (22)   5 0000000000000011111111
   35    5 22222222222222333333
   15    5 4444555
    8    5 6666677
    1    5 8
```

Notice how the leaf unit is now 100, not 10, so that the last two digits of each value are now lost. **The median class is [5000, 5200) and contains 22 values, as indicated by the "`(22)`" in the first column.**

From this stem and leaf display it is easy to generate a **frequency table** for shear stress manually, in the form required by part (b) of the question.

Interval for x	Frequency f
$4000 \le x < 4200$	1
$4200 \le x < 4400$	2
$4400 \le x < 4600$	9
$4600 \le x < 4800$	12
$4800 \le x < 5000$	19
$5000 \le x < 5200$	22
$5200 \le x < 5400$	20
$5400 \le x < 5600$	7
$5600 \le x < 5800$	7
$5800 \le x < 6000$	1
Total:	100

The bar chart generated by `MINITAB` (which it calls an "histogram") also provides the frequency table:



There is a subtle difference between a "bar chart" and an "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 discipline 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.

An **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). [An 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

$$\text{Bar height} = \frac{\text{Relative Frequency}}{\text{Class Width}} = \frac{\text{Frequency}}{\text{(Total Freq.)*(Class Width)}}.$$

The total area of all bars in a relative frequency histogram is always 1. In chapter 8 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.

For the example above, the total frequency is 100 and the class width is 200, so the height of each bar in the relative frequency histogram is given by

$$\text{Bar height} = \frac{\text{Frequency}}{100 * 200} = \frac{\text{Frequency}}{20\ 000}.$$

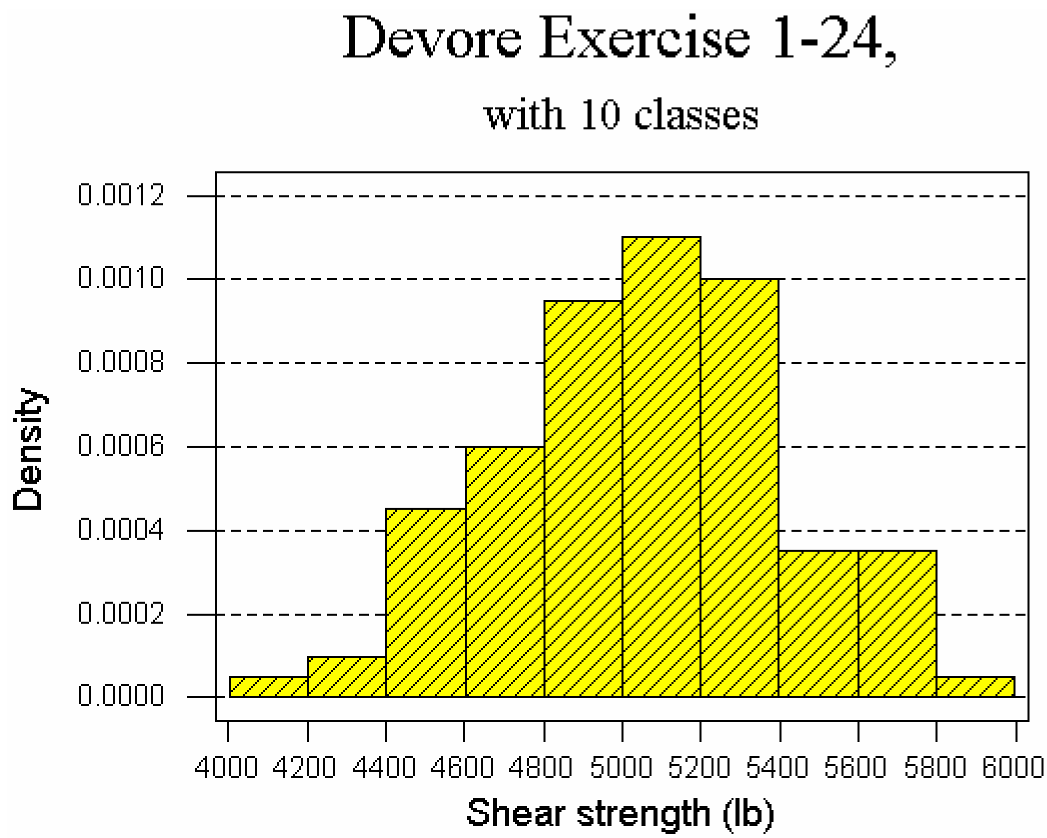
The **cumulative frequency** is the sum of all frequencies up to and including the current class.

Extending the previous table:

Interval for x	Frequency f	Relative Frequency r	Height of histogram bar	Cumulative Frequency c
$4000 \le x < 4200$	1	.01	.00005	1
$4200 \le x < 4400$	2	.02	.00010	3
$4400 \le x < 4600$	9	.09	.00045	12
$4600 \le x < 4800$	12	.12	.00060	24
$4800 \le x < 5000$	19	.19	.00095	43
$5000 \le x < 5200$	22	.22	.00110	65
$5200 \le x < 5400$	20	.20	.00100	85
$5400 \le x < 5600$	7	.07	.00035	92
$5600 \le x < 5800$	7	.07	.00035	99
$5800 \le x < 6000$	1	.01	.00005	100
Total:	100	1.00		

The relative frequency histogram is on the next page.

Relative frequency histogram for the set of 100 observations of shear strengths (in pounds) of ultrasonic spot welds made on a certain type of alclad sheet.



From this diagram, the relative frequency of any class can be recovered by calculating the area of the bar. For example, the relative frequency of the class $4800 \le x < 5000$ is given by

relative frequency = area of bar = $\mathbf{200 \times .00095 = .19}$

Therefore 19% of the 100 data values are in the interval [4800, 5000).

If you are absent from the first Minitab tutorial, then view the web page
`www.engr.mun.ca/~ggeorge/3423/Minitab/s01DescStat/index.html`
carefully.

Measures of Location

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

In example 2.01 the mode is **4848 and 5069 (each occurs twice)**

From the frequency table, the modal class is **$5000 \le x < 5200$ (occurs 22 times)**

A disadvantage of the mode as a measure of location is
that it is not necessarily unique.
For ungrouped continuous data it is not even well defined.

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, **$n = 100$ (even) $\Rightarrow$ $n/2 = 50$**

sample median $\tilde{x} = \dfrac{x_{50} + x_{51}}{2} = \dfrac{5049 + 5055}{2} = \mathbf{5052}$

In the table of grouped values, the 50th and 51st values fall in the same class.
The median class is therefore **$5000 \le x < 5200$ (same as the modal class).**

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, $\bar{x} = \dfrac{\sum x}{n}$ (sample); $\mu = \dfrac{\sum x}{N}$ (pop'ln)

and from a frequency table, $\bar{x} = \dfrac{\sum f \cdot x}{\sum f}$

For the example above, from the 100 raw data (not from the frequency table),

$$\bar{x} = \frac{504916}{100} = \mathbf{5049.16}$$

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

Example 2.02

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

sorted into order: $\boldsymbol{B = \{ 1, 2, 3, 5, 23654 \}}$.

Then

$\tilde{x} = \mathbf{3}$ for set A and $\tilde{x} = \mathbf{3}$ for set B , while

$\bar{x} = \frac{1+2+3+4+5}{5} = \mathbf{3}$ for set A and $\bar{x} = \frac{1+2+3+23654+5}{5} = \mathbf{4733}$ for set B.

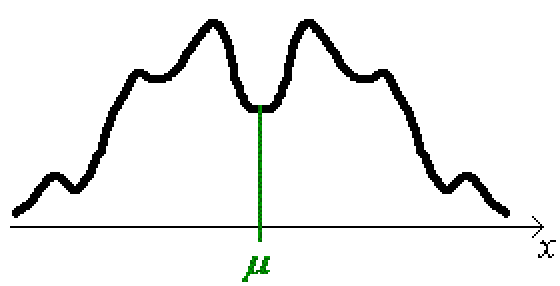
Note that the mode is not well defined for either set.

A disadvantage of the mean as a measure of location is
that it is very sensitive to **outliers** (extreme values).

Advantages of the mean over the median include

- the median uses only the central value(s) while the mean uses all values.
- calculus methods work much better with the mean.

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.



Measures of Variation

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

it often increases as *n* increases.

A disadvantage of the population range is

it may be infinite.

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 2.01:

$n = 100 \Rightarrow (n + 1) / 4 = 25.25 \Rightarrow q_L =$ value 1/4 of the way from x_{25} to x_{26}

$$= \frac{3x_{25} + x_{26}}{4} = \frac{3 \times 4803 + 4806}{4} = \mathbf{4803.75}$$

and $3(n + 1) / 4 = 75.75 \Rightarrow q_U =$ value 3/4 of the way from x_{75} to x_{76}

$$= \frac{x_{75} + 3x_{76}}{4} = \frac{5273 + 3 \times 5275}{4} = \mathbf{5274.50}$$

The semi-interquartile range is then $\dfrac{5274.50 - 4803.75}{2} = \mathbf{235.375}$

The **boxplot** illustrates the median, quartiles, outliers and skewness in a compact visual form. The boxplot for example 2.01, as generated by an older version of MINITAB, is shown below. [See the tutorial session for a more modern version of this output.]

```
MTB > BoxPlot C1.
                         |-- Hinges  --|
                         |             |
                         |             |
         Whisker         ↓     Box     ↓       Whisker
            ↓            ---------------          ↓
     -------------------I       +      I----------------
                         ---------------

------+---------+---------+---------+---------+---------+C1
     4200      4550      4900      5250      5600      5950
                         x      ~      x
                          L     x       U
```

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) [4097.625 and 5980.625 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) [3391.500 and 6686.750 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 no 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).

If you encounter an extreme outlier, then check if the measurement is incorrect or is from a different population. If the observation is genuine, then it is a rare event (< 0.01% in most populations).

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

The deviation of the ith 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:

$$\sum_{i=1}^{n}\left(x - \bar{x}\right) = \sum_{i=1}^{n} x - \bar{x}\sum_{i=1}^{n} 1 = n\bar{x} - n\bar{x} = \mathbf{0}$$

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** $\boldsymbol{\sigma^2}$ 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** $\mathbf{s^2}$ 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 2.01 is shown here:

```
MTB > Describe C1

            N      MEAN    MEDIAN    TRMEAN    STDEV    SEMEAN
C1        100    5049.2    5052.0    5050.5    351.5      35.1

          MIN       MAX        Q1        Q3
C1     4173.0    5828.0    4803.8    5274.5
```

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

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

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

For integer values of x, 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:

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

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

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

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

<u>Example 2.03</u>:

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!).

$\boldsymbol{\Sigma x = 300.06 \;\Rightarrow\; (\Sigma x)^2 = 90036.00\,;}$
$\boldsymbol{\Sigma (x^2) = 10002.00 + 10004.00 + 10006.00 = 30012.00}$

$\boldsymbol{\Rightarrow \quad n\,\Sigma(x^2) - (\Sigma x)^2 = 90036.00 - 90036.00 = 0.00\,!}$

Example 2.04:

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	$f_i \bullet x_i$	$f_i \bullet x_i^2$
0	65	**0**	**0**
1	30	**30**	**30**
2	3	**6**	**12**
3	2	**6**	**18**
Sum:	100	**42**	**60**

[Note that the mode and median of x are both 0.]

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{42}{100} = \mathbf{0.42}$$

$$s^2 = \frac{n \sum f_i x_i^2 - \left(\sum f_i x_i\right)^2}{n(n-1)} = \frac{100 \times 60 - 42 \times 42}{100 \times 99} = \frac{4236}{9900} = \mathbf{0.42787878\ldots}$$

or

$$s^2 = \frac{1}{n-1} \sum f_i (x_i - \bar{x})^2 = \frac{65 \times (0 - 0.42)^2 + \cdots + 2 \times (3 - 0.42)^2}{99} = \cdots = 0.427878\ldots$$

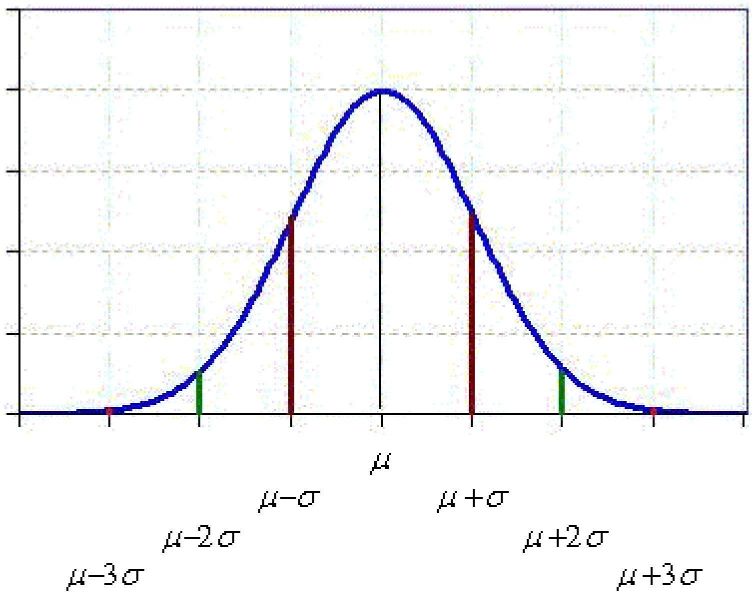
- **tedious, but**
- **less sensitive to round-off errors**

For *any* data set:

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

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

For a bell-shaped distribution (for which population mean = population median = population mode):



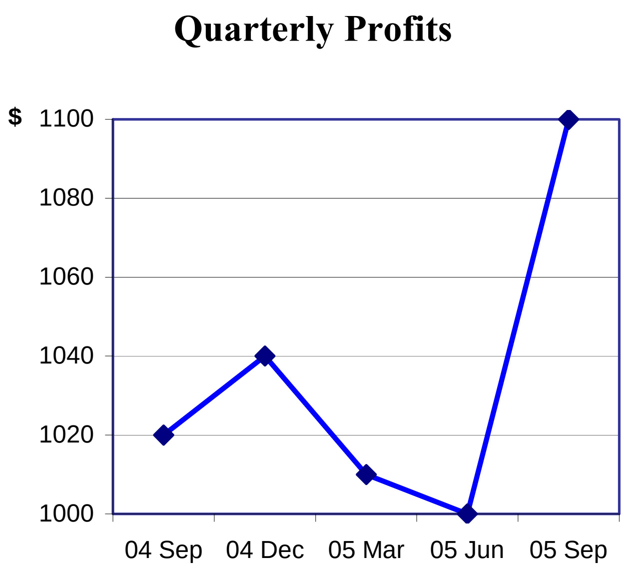
~ **68%** of all data lie within **one** standard deviation of the mean.
~ 95% of all data lie within **two** standard deviations of the mean.
> 99% of all data lie within **three** standard deviations of the mean.

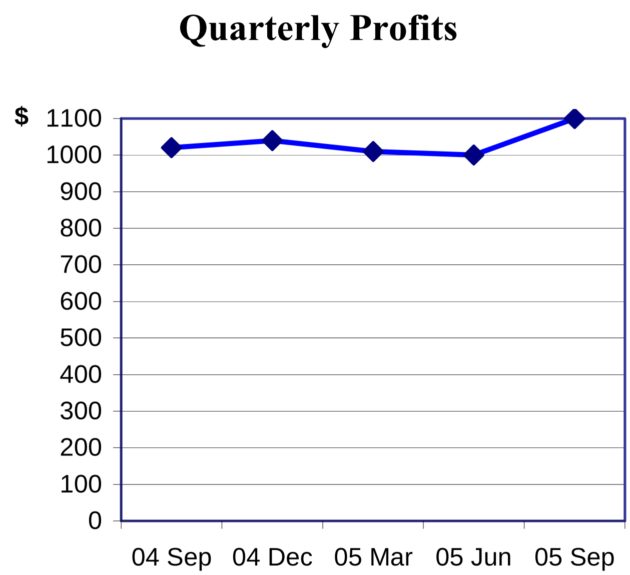
[Note that the points on the normal probability curve where $x = \mu \pm 1\sigma$ are the curve's points of inflection, where the concavity changes sign.]

Misleading Statistics - Example 2.05

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."





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

for **graphs**:

- **Where is the zero?**
- **Are the scales appropriate?**

for **bar charts / pictograms** :

- **Is the frequency proportional to height, area or volume?**

[End of the chapter "Descriptive Statistics"]