INSTITIÚID TEICNEOLAÍOCHTA CHEATHARLACH

INSTITUTE OF TECHNOLOGY CARLOW

BASIC STATISTICS AND LINEAR REGRESSION

1 Some Basic Statistics

Statistics is the science of collecting, studying and analysing numerical data. The numerical data could be for example official statistics on employment, or on imports and exports, or monthly meteorological records for a particular region. The subject divides into two branches. *Descriptive Statistics* is mainly concerned with collecting, summarising and interpreting data. *Inferential Statistics* is concerned with methods for obtaining and analysing data to make inferences applicable in a wider context (e.g., from sample to population). It is concerned also with the precision and reliability of such inference insofar as this involves probabilistic considerations. In this context statistics may be described as a branch of mathematics based on probability theory.

We introduce basic statistical analysis by considering measures of central tendency and measures of dispersion. We assume that the data given is in the form of a *frequency distribution*, or a frequency distribution has been constructed from the *raw data*.

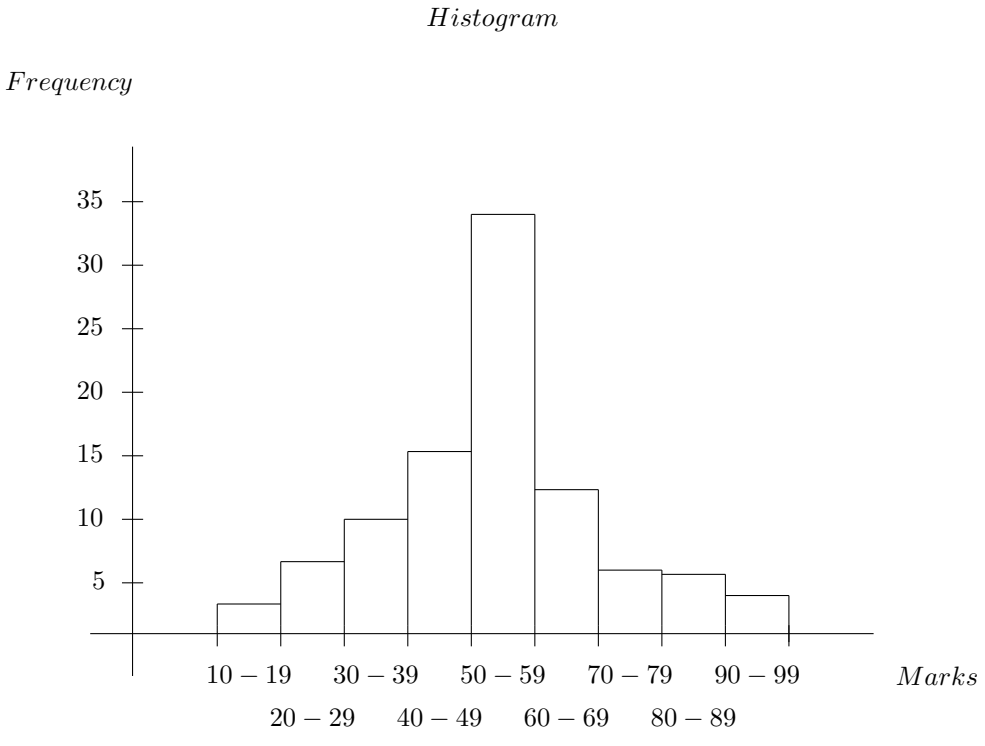
Example The following is a record of the percentage marks gained by candidates in an examination:

65	57	57	55	20	54	52	49	58	52
86	39	50	48	83	71	66	54	51	27
30	44	34	78	36	63	67	55	40	56
63	75	55	15	96	51	54	52	53	42
50	25	85	27	75	40	37	46	42	86
16	45	12	79	50	46	46	59	57	50
56	74	50	68	52	61	40	38	57	31
35	93	54	26	67	62	51	52	54	61
93	84	28	66	62	57	45	43	47	33
45	25	77	80	91	67	53	55	51	36

Looking at the figures there are 100 marks ranging from 12 to 96. Say, we set up 9 classes each of width 10 for our frequency distribution.

Marks Awarded	*Number of Candidates*
$10-19$	3
$20-29$	7
$30-39$	10
$40-49$	16
$50-59$	34
$60-69$	13
$70-79$	7
$80-89$	6
$90-99$	4

With the data in this form we can represent it pictorially using a *histogram*.



From the histogram we can determine the mode. We can easily construct a cumulative frequency curve and determine from this graph the median of the data set as well as other measures of location. Also, with the data in this form we can easily determine the mean and the standard deviation. We must define each of the terms mentioned.

Firstly, a *frequency distribution* will take the form

x_1	x_2	x_3	.....	.....	x_n
f_1	f_2	f_3	.....	.....	f_n

A *group frequency distribution* summarises data into groups of values and takes the form

$x_1 - x_2$	$x_2 - x_3$	$x_3 - x_4$	.....	.....	.....
f_1	f_2	f_3	.....	.....	.....

Statistics under the heading of **measures of central tendency** include mode, median and mean.

Definition The *mode* of a dataset is the observation that occurs most frequently in the dataset.

Definition The *median* of a dataset is the middle number when the observations are arranged in ascending order.

Definition The *mean* of a dataset is the sum of the observations in the dataset divided by the number of observations in the dataset.

$$\bar{x} = \frac{1}{n}\sum fx \qquad where \qquad n = \sum f$$

The most common **measure of dispersion** is the standard deviation.

Definition The *standard deviation* of a dataset is the average of the deviations from the mean given by the following formula

$$s = \left[\frac{1}{n-1}\sum fd^2\right]^{\frac{1}{2}} \qquad where \qquad d = x - \bar{x}$$

Remark For *raw data*, we have

$$\bar{x} = \frac{1}{n}\sum x \quad and \quad s = \left[\frac{1}{n-1}\sum d^2\right]^{\frac{1}{2}}$$

where n is the number of observations and $d = x - \bar{x}$. These formula are similar to the above with frequency f removed.

Example Consider the following raw data

$$23\ ,\ 44\ ,\ 34\ ,\ 24\ ,\ 61\ ,\ 45\ ,\ 53\ ,\ 39$$

Firstly, we can get the required totals as follows

x	$x - \bar{x}$	$(x - \bar{x})^2$
23	-17.375	301.89
44	3.625	13.14
34	-6.375	40.64
24	-16.375	268.14
61	20.625	425.39
45	4.625	21.39
53	12.625	159.39
39	-1.375	1.89
$\overline{\mathbf{323}}$		$\overline{\mathbf{1231.87}}$

Now

$$\bar{x} = \frac{1}{n}\sum x$$

Hence

$$\bar{x} = \frac{1}{8}\Big(323\Big) = 40.375$$

Also

$$s = \left[\frac{1}{n-1}\sum d^2\right]^{\frac{1}{2}} \qquad where \qquad d = x - \bar{x}$$

Hence

$$s = \left[\frac{1}{7}(1231.87)\right]^{\frac{1}{2}} = 13.266$$

Exercise Find the mean $\bar{x}$ and standard deviation s of the following raw data

$$150\ ,\ 488\ ,\ 600\ ,\ 125\ ,\ 179\ ,\ 315\ ,\ 208\ ,\ 82\ ,\ 263\ ,\ 859$$

Example The number of defective items produced each day by a production process is recorded in the table below:

$Number\ of\ Defectives$	$Number\ of\ Days$
$40-44$	14
$45-49$	29
$50-54$	44
$55-59$	38
$60-64$	25
$65-69$	10

i Calculate the mean $\bar{x}$ and standard deviation s of the distribution.

ii Plot a graph of the *cumulative frequency curve.*

iii Estimate from this graph the median of the data.

i Let x = mid-interval value.

x	f	fx	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
42	14	588	-11.9	141.61	1982.54
47	29	1363	-6.9	47.61	1380.69
52	44	2288	-1.9	3.61	158.84
57	38	2166	3.1	9.61	365.18
62	25	1550	8.1	65.61	1640.25
67	10	670	13.1	171.61	1716.1
	$\overline{\mathbf{160}}$	$\overline{\mathbf{8625}}$			$\overline{\mathbf{7243.6}}$

Now

$$\bar{x} = \frac{1}{n}\sum fx \qquad where \qquad n = \sum f$$

Hence

$$\bar{x} = \frac{1}{160}(8625) = 53.9$$

Also

$$s = \left[\frac{1}{n-1}\sum fd^2\right]^{\frac{1}{2}} \qquad where \qquad d = x - \bar{x}$$

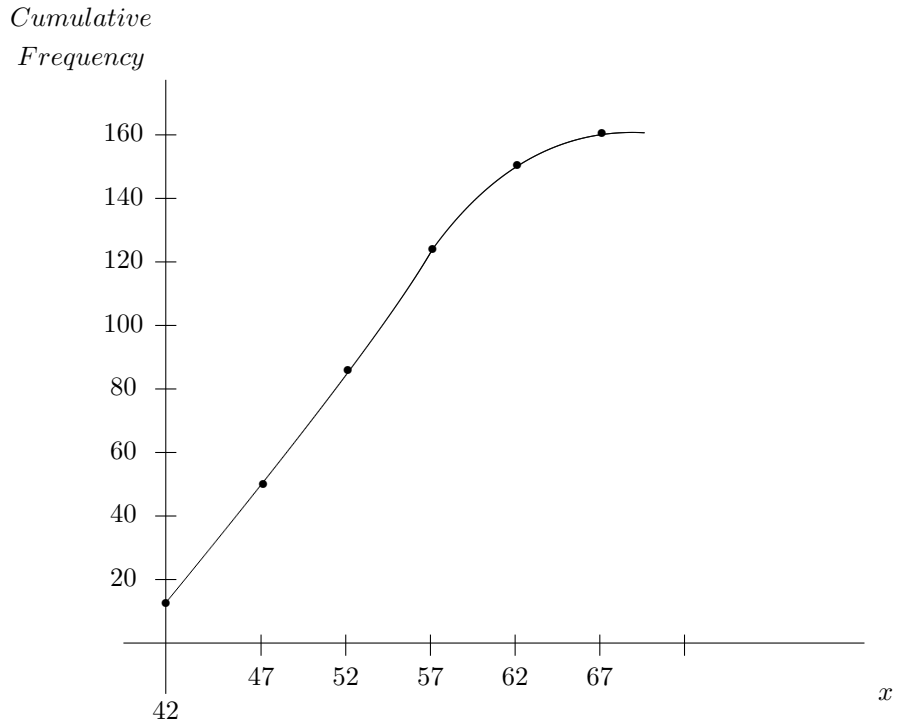
Hence

$$s = \left[\frac{1}{159}(7243.6)\right]^{\frac{1}{2}} = 6.7636$$

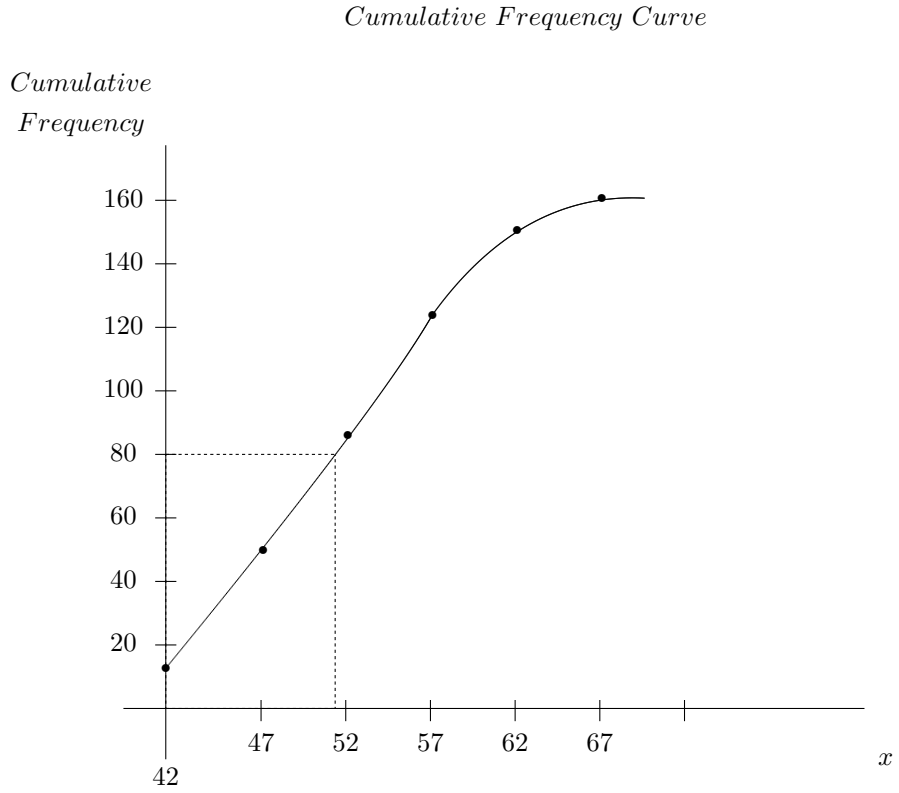
ii To plot a *cumulative frequency curve* for this example we tabulate as follows

x	f	$Cumulative\ frequency$
42	14	14
47	29	43
52	44	87
57	38	125
62	25	150
67	10	160

Cumulative Frequency Curve



iii To estimate the median from a cumulative frequency curve we estimate the middle-most measurement i.e., 50% through the distribution, as follows



The median is approximately 51.

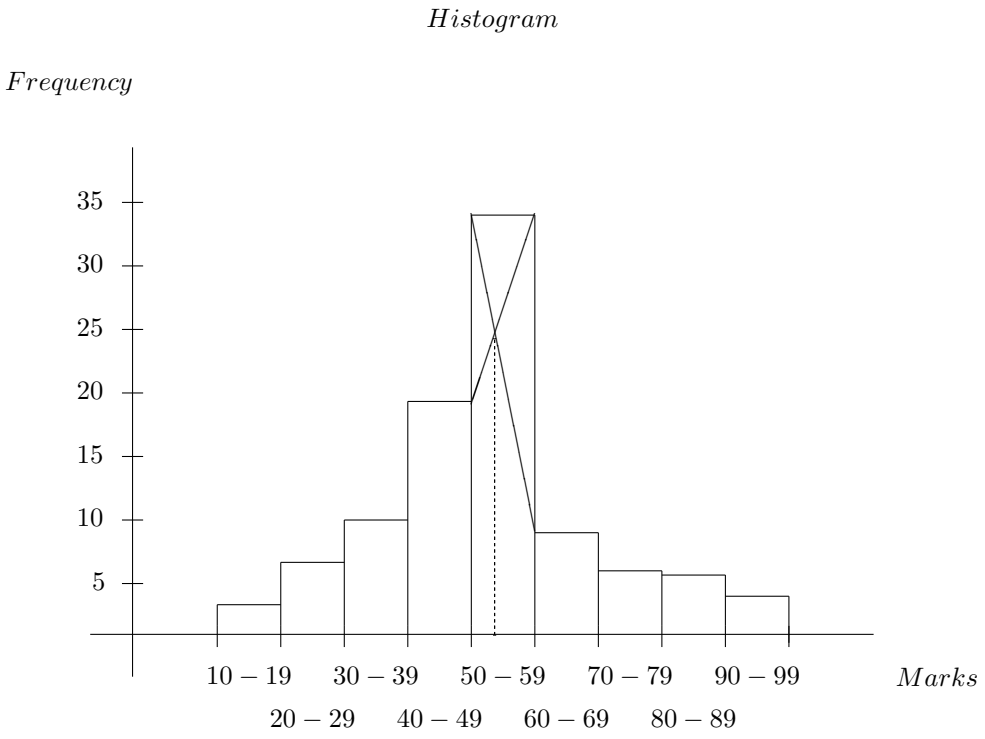
Remark Some other measures of location can be estimated from the cumulative frequency graph, namely the first and third quartiles.

The three quartiles of raw data or a frequency distribution are those numbers that lie one-quarter, one-half and three-quarters of the way along the group, and are called the lower quartile, middle and upper quartiles.

The three quartiles are denoted by Q_1, Q_2 and Q_3. The median is Q_2.

A further measure of dispersion is the *inter-quartile range* $Q_3 - Q_1$.

Remark Finally, we can estimate the *mode* (the observation that occurs most frequently) of any grouped frequency distribution by firstly constructing its *histogram* and using the following construction.



From this construction the *mode* is estimated to be 54.

Exercise Expenditure on monthly mobile phone bills is recorded in the table below:

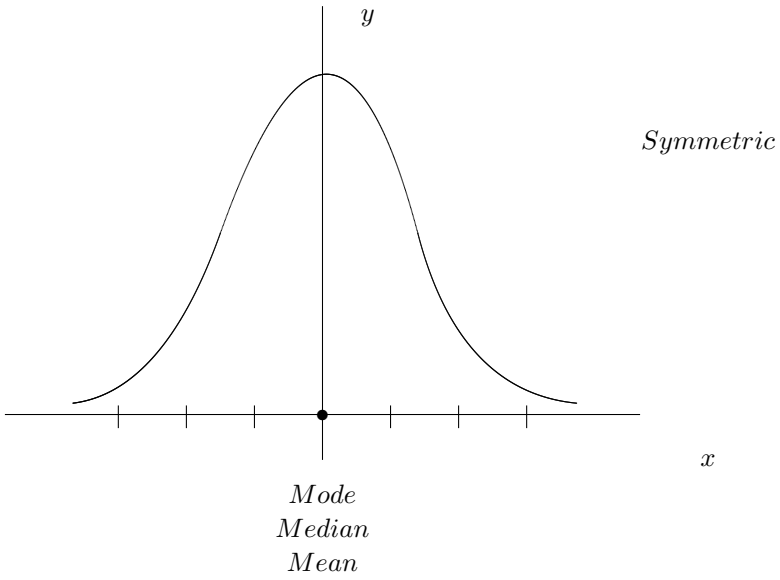
Monthly Bill	*Number of People*
$10 < 20$	12
$20 < 30$	27
$30 < 40$	47
$40 < 50$	33
$50 < 60$	21
$60 < 70$	10

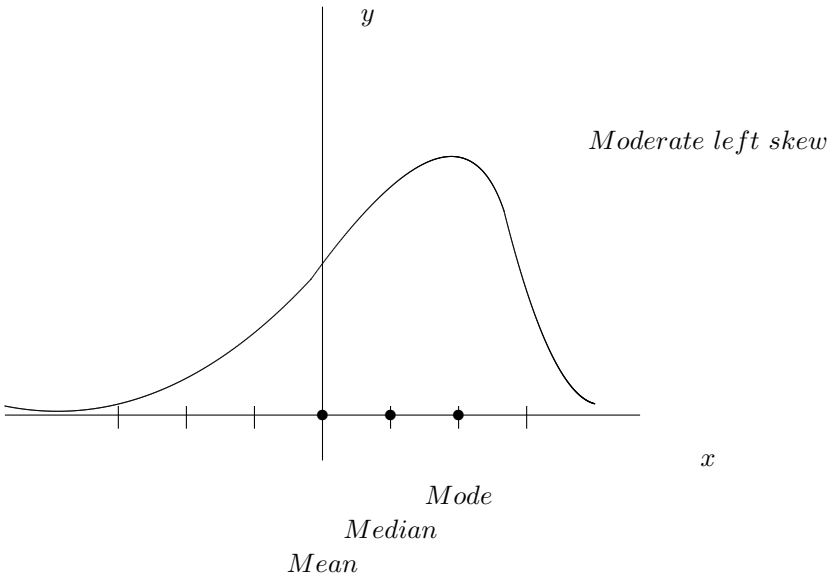
i Calculate the mean $\bar{x}$ and standard deviation s of the distribution.

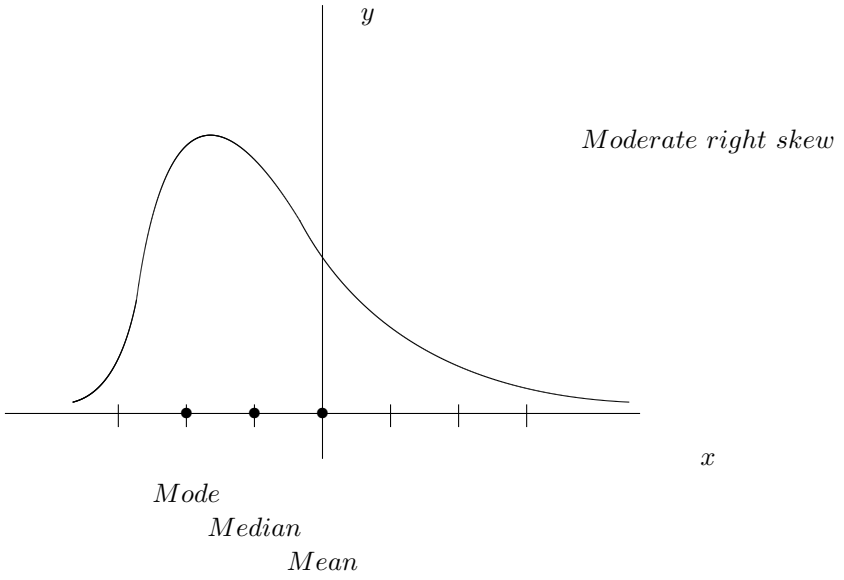
ii Plot a graph of the *cumulative frequency curve*.

iii Estimate form this graph the median of the data and the first and third quartiles.

Frequency curves of distributions may be relatively symmetric, but more often are skewed to some extent. The following frequency curves indicate symmetric, moderately left-skewed and moderately right-skewed distributions.







Consider the results of the Irish Leaving Certificate Mathematics Examination. Which frequency curve would indicate that the mathematics paper had been set at a satisfactory standard?

1.1 The Normal Distribution

From samples of data we now move to **large data sets** which we refer to as a population. To distinguish between sample and population we have a change in notation for the mean and standard deviation.

For a sample of size n: $\bar{x} = mean\ ,\ s = standard\ deviation$

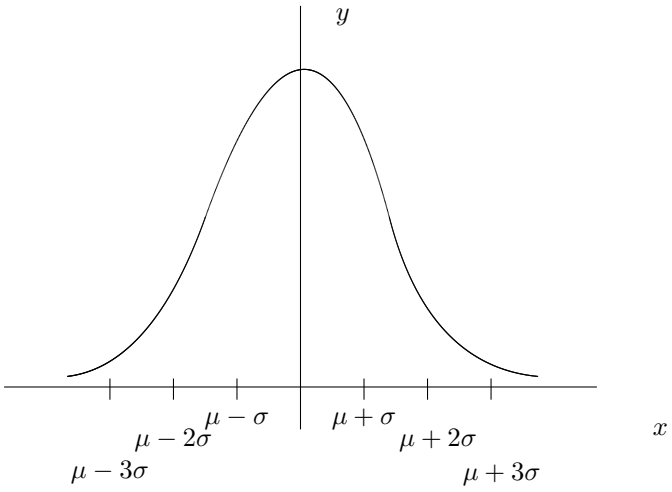
For a population of size N: $\mu = mean\ ,\ \sigma = standard\ deviation$

Sometimes it is difficult to calculate the standard deviation of the population. In such cases it is common for the standard deviation of the population σ to be estimated by examining a random sample taken from the population. A common estimator of σ is an adjusted version of the formula for the standard deviation of the sample. This formula is

$$s = \left[\frac{1}{n-1}\sum fd^2\right]^{\frac{1}{2}} \qquad where \qquad d = x - \bar{x}$$

where we use $n-1$ instead of n. This is called *Bessel's correction*. Using n instead of $n-1$ tends to underestimate the population standard deviation.

What do we mean when we say that a dataset is normally distributed?



If a dataset is *normally distributed* (i.e., conforms to the symmetric bell-shaped frequency curve shown above), the following characteristics hold:

i 67% of data lies within 1 standard deviation of the mean: i.e., $\mu \pm \sigma$

ii 97% of data lies within 2 standard deviation of the mean: i.e., $\mu \pm 2\sigma$

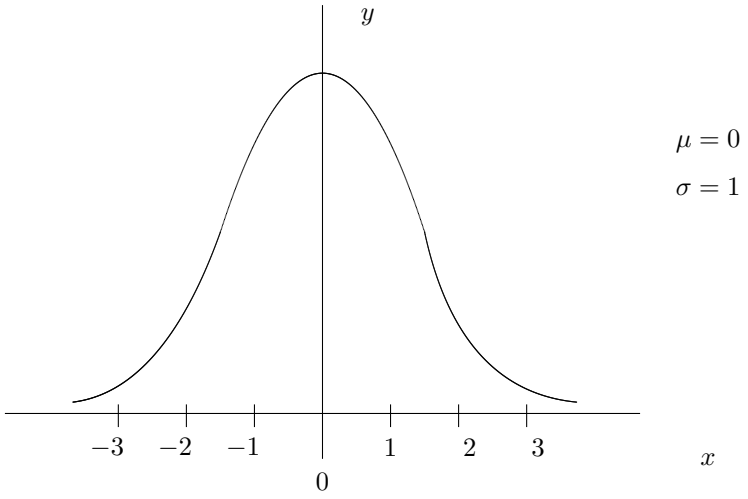
iii 99% of data lies within 3 standard deviation of the mean: i.e., $\mu \pm 3\sigma$

Different datasets will have different means and different standard deviations so we standardize the given data-set by transposing the mean to the origin 0 (i.e., subtract μ) and ensure total area under frequency curve totals 1 by dividing by the standard deviation σ.

We introduce the standard score z as

$$z = \frac{x - \mu}{\sigma}$$

This quantity is simply the number of standard deviations by which x exceeds the mean. The *standard normal curve* will have a mean $\mu = 0$ and a standard deviation $\sigma = 1$.



The equation of this curve is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

The area under this curve may be evaluated from this function to yields a value of 1. We make a link between this statistical distribution and probability by recalling the second axiom of basic probability, $P(S) = 1$. Answering probability questions based on normally distributed data simply amounts to determining areas areas under the standard normal curve.

2 Linear Regression

In computer science it is often the case that a process or an experiment produces a set of data points

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

One goal in numerical methods is to determine a formula $y = f(x)$ that relates these variables. Usually, the class of allowable formulas is chosen and then coefficients must be determined. There are many possibilities for the type of function that can be used. Often there is an underlying mathematical model based on the physical situation, that will determine the form of the function. We will consider firstly the class of *linear functions* of the form

$$y = f(x) = Ax + B$$

Our method is based on analysing the errors associated with any measurement or approximation. In taking a measurement y_k there may often be a measurement error involved so that the true value $f(x_k)$ satisfies

$$f(x_k) = y_k + e_k$$

where e_k is the measurement error. Hence

$$e_k = f(x_k) - y_k$$

for all $1 \leq k \leq n$. To determine the best fitted straight line (or curve) that goes near (not always through) the points we need to average the errors associated with every measurement. There are several such norms that can be used to measure how far the curve $y = f(x)$ lies from the data. We can choose one of the following – the maximum error $E_1(f)$, the average error $E_2(f)$ and the root-mean square error $E_3(f)$.

$$E_1(f) = max_{1 \leq k \leq n} |f(x_k) - y_k|$$

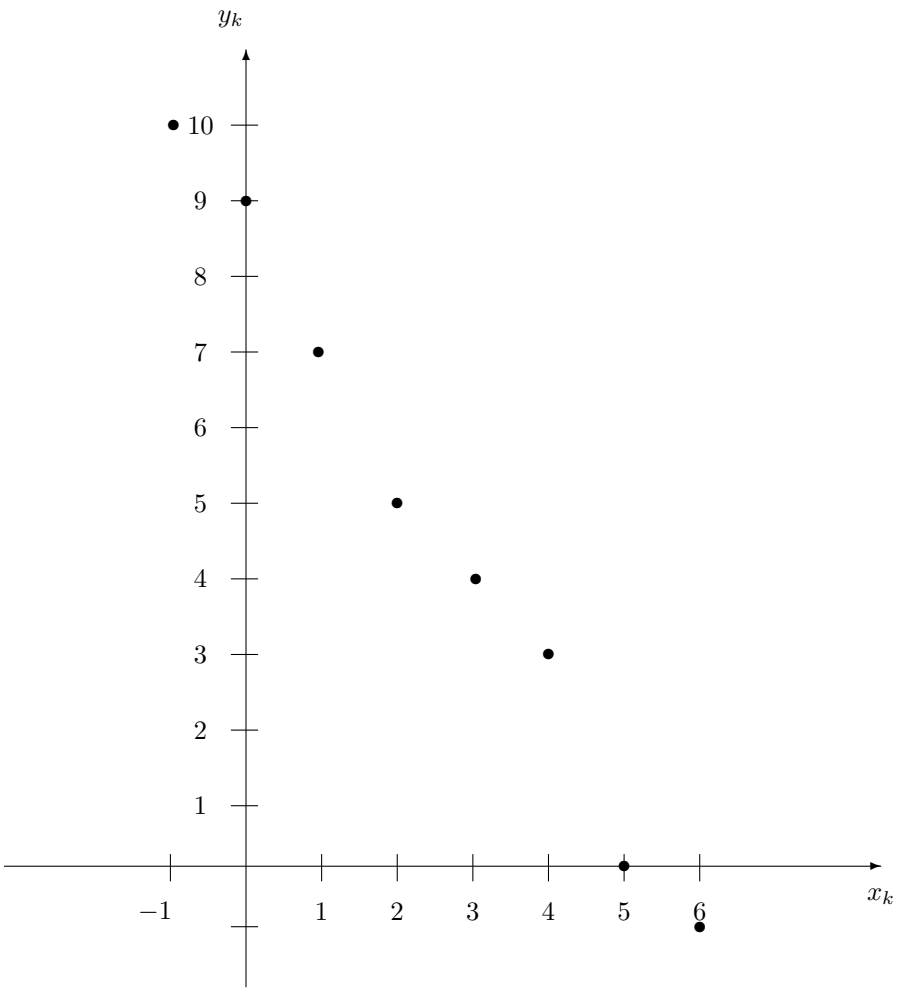
$$E_2(f) = \frac{1}{n} \sum_{k=1}^{n} |f(x_k) - y_k|$$

$$E_3(f) = \left[\frac{1}{n} \sum_{k=1}^{n} (f(x_k) - y_k)^2 \right]^{\frac{1}{2}}$$

The following example show how to apply these norms when a function and a set of points are given.

Example We can plot a *scattergraph* for the following set of data points

$$(-1, 10), (0, 9), (1, 7), (2, 5), (3, 4), (4, 3), (5, 0), (6, -1)$$



We wish to compare the maximum error, average error and root-mean square error for the linear approximation i.e., the 'best fitted' line

$$y = f(x) = 8 \cdot 6 - 1 \cdot 6x$$

x_k	y_k	$f_k = 8 \cdot 6 - 1 \cdot 6x_k$	$\lvert f(x_k) - y_k \rvert$	$(f(x_k) - y_k)^2$
-1	$10 \cdot 0$	$10 \cdot 2$	$0 \cdot 2$	$0 \cdot 04$
0	$9 \cdot 0$	$8 \cdot 6$	$0 \cdot 4$	$0 \cdot 16$
1	$7 \cdot 0$	$7 \cdot 0$	$0 \cdot 0$	$0 \cdot 00$
2	$5 \cdot 0$	$5 \cdot 4$	$0 \cdot 4$	$0 \cdot 16$
3	$4 \cdot 0$	$3 \cdot 8$	$0 \cdot 2$	$0 \cdot 04$
4	$3 \cdot 0$	$2 \cdot 2$	$0 \cdot 8$	$0 \cdot 64$
5	$0 \cdot 0$	$0 \cdot 6$	$0 \cdot 6$	$0 \cdot 36$
6	$-1 \cdot 0$	$-1 \cdot 0$	$0 \cdot 0$	$0 \cdot 00$
			$\mathbf{2 \cdot 6}$	$\mathbf{1 \cdot 40}$

Now

$$E_1(f) = max\{0 \cdot 2, 0 \cdot 4, 0 \cdot 0, 0 \cdot 4, 0 \cdot 2, 0 \cdot 8, 0 \cdot 6, 0 \cdot 0\} = 0 \cdot 8$$

$$E_2(f) = \frac{1}{8}(2 \cdot 6) = 0 \cdot 325$$

$$E_3(f) = \left[\frac{1}{8}(1 \cdot 4)\right]^{\frac{1}{2}} = 0 \cdot 41833$$

We can see that the maximum error is the largest, and if one point is badly in error, then its value determines $E_1(f)$. The average error $E_2(f)$ simply averages the absolute value of the error at the various points. It is often used because it is easy to compute. The root-mean square error $E_3(f)$ is often used when the statistical nature of the errors is considered.

The 'best-fitting' line is found by minimising one of these errors. Hence there are three best fitted lines that we can find. The root-mean square error $E_3(f)$ is the traditional choice because it is much easier to minimise.

Definition Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a set of n points. The *least squares line* $y = f(x) = Ax + B$ is the line for which the root-mean square error $E_3(f)$ is a minimum.

Given n data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, we must find parameters A and B for the line $y = f(x) = Ax + B$ which minimise the root-mean square error $E_3(f)$. So, we have

$$E_3(f) = \left[\frac{1}{n}\sum_{k=1}^{n}(Ax_k + B - y_k)^2\right]^{\frac{1}{2}}$$

Hence

$$n.E_3^2(f) = \sum_{k=1}^{n}(Ax_k + B - y_k)^2$$

It will be sufficient to minimise the following function of two variables A,B

$$E(A, B) = \sum_{k=1}^{n}(Ax_k + B - y_k)^2$$

At the point that minimises the value of $E(A, B)$, the partial derivatives of E with respect to A and then with respect to B are both zero. Take note that in this work it is x_k and y_k that are constant and A and B are variables. Hence,

$$\frac{\partial E}{\partial A} = 0 \qquad and \qquad \frac{\partial E}{\partial B} = 0$$

Holding B fixed and differentiating with respect to A yields

$$\begin{aligned}\frac{\partial E}{\partial A} &= \sum_{k=1}^{n} 2(Ax_k + B - y_k)^1.(x_k + 0 - 0) \\ &= 2\sum_{k=1}^{n}(Ax_k^2 + Bx_k - x_k y_k)\end{aligned}$$

Holding A fixed and differentiating with respect to B yields

$$\begin{aligned}\frac{\partial E}{\partial B} &= \sum_{k=1}^{n} 2(Ax_k + B - y_k)^1.(0 + 1 - 0) \\ &= 2\sum_{k=1}^{n}(Ax_k + B - y_k)\end{aligned}$$

Setting each of the partial derivatives equal to zero and using the distributive properties of the summation yields

$$0 = \sum_{k=1}^{n}(Ax_k^2 + Bx_k - x_k y_k) = A\sum_{k=1}^{n} x_k^2 + B\sum_{k=1}^{n} x_k - \sum_{k=1}^{n} x_k y_k$$

$$0 = \sum_{k=1}^{n}(Ax_k + B - y_k) = A\sum_{k=1}^{n} x_k + nB - \sum_{k=1}^{n} y_k$$

These equations can now be re-arranged to a more familiar form often referred to as the *normal equations*.

$$\left(\sum_{k=1}^{n} {x_k}^2\right)A + \left(\sum_{k=1}^{n} x_k\right)B = \sum_{k=1}^{n} x_k y_k$$

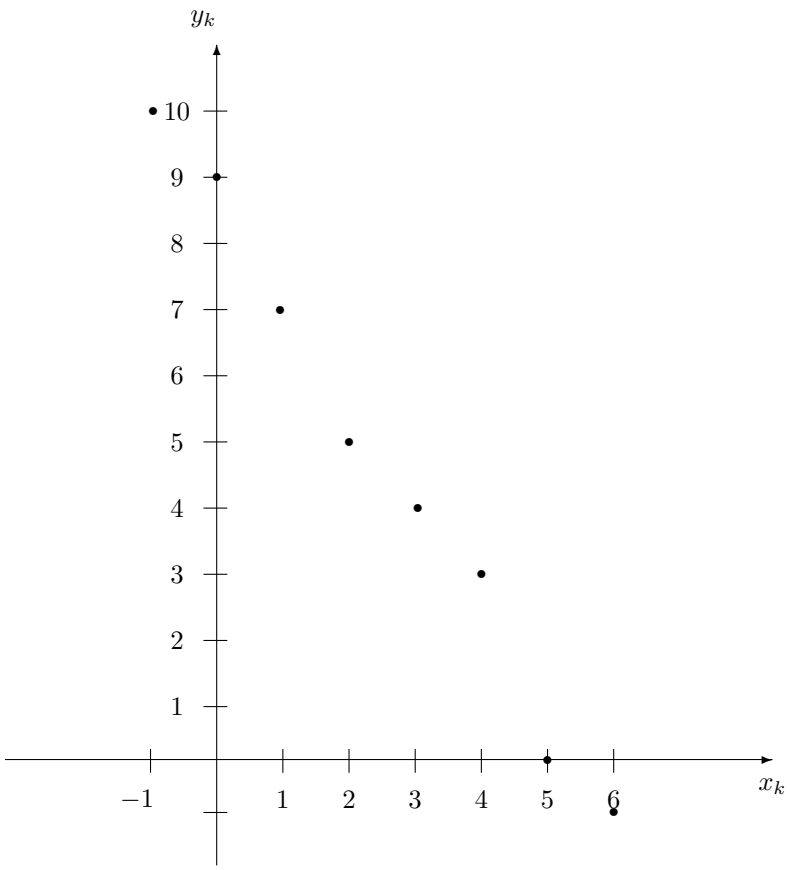
$$\left(\sum_{k=1}^{n} x_k\right)A + nB = \sum_{k=1}^{n} y_k$$

The solution of the resulting linear system can be found using matrices or algebra. The simultaneous solution of this linear system A and B again represent the coefficient of x and the constant term in the linear equation $y = f(x) = Ax + B$. This best fitted line can be used to make predictions for y for a chosen (usually future) values of x. This technique is called *forecasting*. Evaluating the root-mean square error for the same data set can be interpreted as a measure of how accurate the prediction is since it is a measure of how far the line $y = f(x)$ lies from the data i.e., the average error. The smaller the value of the root-mean square error the more confidence can be placed in any prediction that is made. This type of analysis will ensure that a person can make more informed decisions based on the numerical data collected which is a fundamental of management science.

2.1 Some Examples

Example Consider the following set of data points

$$(-1, 10), (0, 9), (1, 7), (2, 5), (3, 4), (4, 3), (5, 0), (6, -1)$$



From the scattergraph we observe a *linear relationship* between the variables x and y. We now proceed to determine the best linear function $y = f(x) = Ax + B$ using the *normal equations*. Now to obtain the required totals for the normal equations.

x_k	y_k	x_k^2	$x_k y_k$
-1	10.0	1	-10
0	9.0	0	0
1	7.0	1	7
2	5.0	4	10
3	4.0	9	12
4	3.0	16	12
5	0.0	25	0
6	-1.0	36	-6
20	**37**	**92**	**25**

The linear system involving A and B that results is

$$92A + 20B = 25$$

$$20A + 8B = 37$$

Solving yields $A = -1.6071429$ and $B = 8.6428571$. Hence the best fitted line, using the least squares method, is

$$y = -1 \cdot 6071429x + 8 \cdot 6428571$$

We now can calculate the root-mean square error $E_3(f)$. This measure of how far the straight line $y = f(x)$ lies from the data will also allow us consider how much faith we can place in any predictions made using our best fitted line. Now

$$E(f) = \left[\frac{1}{n}\sum_{k=1}^{n}(f(x_k) - y_k)^2\right]^{\frac{1}{2}}$$

To get the required total for this formula we note that

$$f(x) = -1 \cdot 6071429x + 8 \cdot 6428571$$

and proceed as follows:

x_k	y_k	$f(x_k)$	$f(x_k) - y_k$	$(f(x_k) - y_k)^2$
-1	10.0	10.25	0.25	0.0625
0	9.0	8.6428571	-0.3571429	0.127551051
1	7.0	7.0357142	0.0357142	0.001275504
2	5.0	5.4285713	0.4285713	0.183673359
3	4.0	3.8214284	-0.1785716	0.031887816
4	3.0	2.2142855	-0.7857145	0.617347275
5	0.0	0.6071426	0.6071426	0.368622136
6	-1.0	-1.0000003	0.0000003	0.0
				1.392857141

Now

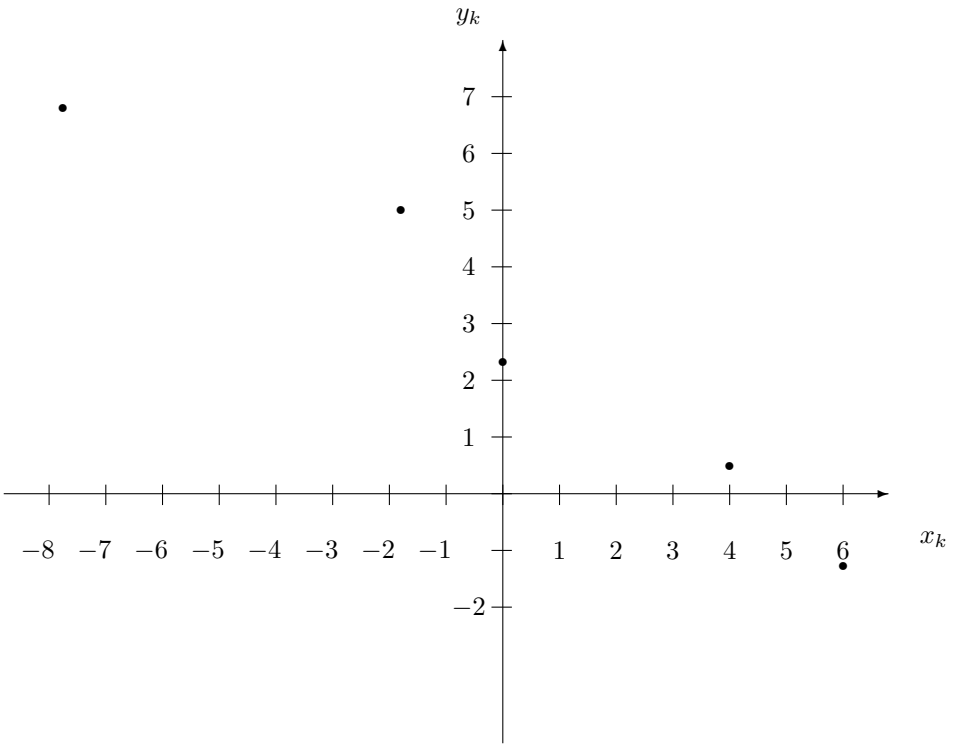
$$E(f) = \left[\frac{1.392857141}{8}\right]^{\frac{1}{2}} = 0.417261479$$

The root-mean square error for this example has been found to be 0.417261479. This value should be quoted when presenting the equation of the best fitted line to indicate how far the line lies from the data especially when making predictions using the found equation. This value is also useful when making comparisons between different data sets all generated from the same source. If the best fitted line from each data set is determined, the corresponding smallest root-mean square error would indicate the best equation to use for forecasting.

Exercise Consider the following set of data points

x	-8	-2	0	4	6
y	$6 \cdot 8$	$5 \cdot 0$	$2 \cdot 2$	$0 \cdot 5$	$-1 \cdot 3$

The *scattergraph* for the data set is as follows:



From the scattergraph we observe a *linear relationship* between the variables x and y. We now proceed to determine the best linear function $y = f(x) = Ax + B$ using the *normal equations*. Now to obtain the required totals for the normal equations.

x_k	y_k	x_k^2	$x_k y_k$
-8	$6 \cdot 8$	64	$-54 \cdot 4$
-2	$5 \cdot 0$	4	$-10 \cdot 0$
0	$2 \cdot 2$	0	0
4	$0 \cdot 5$	16	2
6	$-1 \cdot 3$	36	$-7 \cdot 8$
0	**$13 \cdot 2$**	**120**	**$-70 \cdot 2$**

The linear system involving A and B that results is

$$120A + 0B = -70 \cdot 2$$

$$0A + 5B = 13 \cdot 2$$

Solving yields $A = -0 \cdot 585$ and $B = 2 \cdot 64$. Hence the best fitted line, using the least squares method, is

$$y = -0 \cdot 585x + 2 \cdot 64$$

Now $y = f(x) = -0 \cdot 585x + 2 \cdot 64$, hence

$$f(7) = -0 \cdot 585(7) + 2 \cdot 64 = -1 \cdot 455$$

Exercise The attendance of eight computer science students at a sequence of 16 **weekly mathematics tutorials** during a single college semester was recorded. The final examination result in mathematics for each of the students was also recorded. Let x represents the number of tutorials attended and y the final examination result. The data is as follows:

Students	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
x	2	4	6	8	10	12	14	16
y	12	19	23	30	52	70	83	87

i Plot a *scattergraph*for this data and comment on the relationship you observe from your scattergraph.

ii Determine the equation of the *best-fitted line*

$$y = f(x) = Ax + B$$

where A and B are constants, using the *normal equations*.

iii If a student has attended 3 out of 16 mathematics tutorials during the semester, use the best-fitted line from part ii to predict this students final examination result.

iv Determine the *root-mean square error* $E(f)$, and comment on how it would support your prediction from part iii.

$$E(f) = \left[\frac{1}{n}\sum_{k=1}^{n}\big(f(x_k) - y_k\big)^2\right]^{\frac{1}{2}}$$

Exercise The number of sales of a particular computer game were recorded on a weekly basis and the number of sales (in hundreds) over an eight week period were presented to a meeting. The data recorded was as follows:

Week Number x	1	2	3	4	5	6	7	8
Sales in hundreds y	5	16	22	32	34	45	50	51

i Plot a scattergraph for this data. What observation can you make?

ii Find the *least-square line* $y = f(x) = Ax + B$, using the *normal equations*.

iii You are asked at this meeting what would be the expected number of sales of this computer game in week 12. What would your response be?

iv Determine the *root-mean square error* $E(f)$, and comment on how it would support your prediction from part iii.

$$E(f) = \left[\frac{1}{n}\sum_{k=1}^{n}\big(f(x_k) - y_k\big)^2\right]^{\frac{1}{2}}$$