Statistics 1

Revision Notes

June 2012

Contents

1 Statistical modelling

Statistical modelling

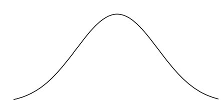
Example: When a die is rolled, we say that the probability of each number is $\frac{1}{6}$. This is a statistical model, but the assumption that each face is equally likely might not be true.

Suppose the die is weighted to increase the chance of a six. We might then find, after experimenting, that the probability of a *six* is $\frac{1}{4}$ and the probability of a *one* is $\frac{1}{12}$, with the probability of other faces remaining at $\frac{1}{6}$. In this case we have *refined*, or improved, the model to give a truer picture.

Example: The heights of a large group of adults are measured. The mean is 172·3 cm and the standard deviation is 12·4 cm.

It is thought that the general shape of the histogram can be modelled by the curve

$$f(x) = \frac{1}{12{\cdot}4\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-172{\cdot}3)^2}$$



This might not give a true picture, in which case we would have to change the equation, or *refine the model*.

Definition

A *statistical model* is a *simplification* of a real world situation. It can be used to make *predictions* about a real world problem. By analysing and *refining* the model an *improved understanding* may be obtained.

Advantages

- the model is *quick* and *easy* to produce
- the model helps our *understanding* of the real world problem
- the model helps us to make *predictions*
- the model helps us to *control* a situation – e.g. railway timetables, air traffic control etc.

Disadvantages

- the model *simplifies* the situation and *only describes a part* of the real world problem.
- the model may *only work in certain situations*, or for a *particular range of values*.

2 Representation of sample data

Variables

Qualitative variables

Non-numerical - e.g. red, blue or long, short etc.

Quantitative variables

Numerical - e.g. length, age, time, number of coins in pocket, etc

Continuous variables

Can take **any** value within a given range - e.g. height, time, age etc.

Discrete variables

Can only take certain values - e.g. shoe size, cost in £ and p, number of coins.

Frequency distributions

Frequency tables

A list of discrete values and their frequencies.

Example: The number of *M&M s* is counted in several bags, and recorded in the frequency table below:

number of M&M s	37	38	39	40	41	42	43
frequency	3	8	11	19	13	7	2

Cumulative frequency

Add up the frequencies as you go down the list

number of M&M s	37	38	39	40	41	42	43
frequency	3	8	11	19	13	7	2
cumulative frequency	3	11	22	41	54	61	63

Stem and leaf & back-to-back stem and leaf diagrams

Line up the digits on the leaves so that it looks like a bar chart.
Add a key; e.g. 5|2 means 52, or 4|3 means 4·3 etc.

Comparing two distributions from a back to back stem and leaf diagram.

8/1 = 81

A		B
	8	1 2 2 4 5 6 7 8
3 1 0	7	2 3 5 6 7 7 8 8 9 9 9
3 2 2 1 1	6	1 3 4 5 5 6 6 7 8 8 8 9 9
7 6 5 5 3 3	5	0 0 1 2 3 4 5 6 8
9 8 7 4 3 0 0	4	
8 7 5 3 1 1	3	
9 9 7 4 2	2	
5 4 3 2 1	1	
8 7	0	

1. The values in **A** are on average smaller than those in **B**
2. The values in **A** are more spread out than those in **B.**

Grouped frequency distributions

Class boundaries and widths

When deciding class *boundaries* you **must not leave a gap** between one class and another, whether dealing with *continuous* or *discrete* distributions.

For *discrete* distributions avoid leaving gaps between classes by using class boundaries as shown below:
X 0, 1, 2, 3, 4, 5, 6, 7, …

Class interval – as given	Class contains	**Class boundaries without gaps**
0 – 4	0, 1, 2, 3, 4	$\mathbf{0 - 4\frac{1}{2}}$
5 – 9	5, 6, 7, 8, 9	$\mathbf{4\frac{1}{2} - 9\frac{1}{2}}$
10 – 12	10, 11, 12	$\mathbf{9\frac{1}{2} - 12\frac{1}{2}}$
etc		

For *continuous* distributions the class boundaries can be anywhere.

Cumulative frequency curves for grouped data

class interval	class boundaries	frequency	**class**	**cumulative frequency**
0 - 4	0 to 4 ½	27	**≤ 4 ½**	**27**
5 – 9	4 ½ to 9 ½	36	**≤ 9 ½**	**63**
10 – 19	9 ½ to 19 ½	54	**≤ 19 ½**	**117**
20 – 29	19 ½ to 29 ½	49	**≤ 29 ½**	**166**
30 – 59	29 ½ to 59 ½	24	**≤ 59 ½**	**190**
60 –99	59 ½ to 99 ½	10	**≤ 99 ½**	**200**

Plot points at ends of intervals, **$(4\frac{1}{2}, 27)$, $(9\frac{1}{2}, 63)$, $(19\frac{1}{2}, 117)$** etc. and join points with a smooth curve.

Histograms

Plot the axes with a *continuous scale* as normal graphs.

There are *no gaps* between the bars of a histogram.

Area equals frequency.

Note that the total area under a *frequency* histogram is N, the total number and the area from a to b is the number of items between a and b.

To draw a histogram, first draw up a table showing the class intervals, class boundaries, class widths, frequencies and then heights = $\frac{frequency}{width}$ – as shown below:

class interval	**class boundaries**	class width	frequency	**height**
0 - 4	**0 to $4\frac{1}{2}$**	$4\frac{1}{2}$	27	**6**
5 – 9	**$4\frac{1}{2}$ to $9\frac{1}{2}$**	5	36	**7·2**
10 – 19	**$9\frac{1}{2}$ to $19\frac{1}{2}$**	10	54	**5·4**
20 – 29	**$19\frac{1}{2}$ to $29\frac{1}{2}$**	10	49	**4·9**
30 – 59	**$29\frac{1}{2}$ to $59\frac{1}{2}$**	30	24	**0·8**

60 –99	**59 $\frac{1}{2}$ to 99 $\frac{1}{2}$**	40	10	**0·25**

Example: A grouped frequency table for the weights of adults has the following entries:

weight kg	…	…	50 – 60	…	70 – 85	…
frequency	…	…	60	…	20	…

In a histogram, the bar for the class 50 – 60 kg is 2 cm wide and 9 cm high.

Find the width and height of the bar for the 70 – 85 kg class.

Solution: 50 – 60 is usually taken to mean $50 \le$ weight < 60

The width of the 50 – 60 class is 10 kg $\equiv$ 2 cm

$\Rightarrow$ width of the 70 – 85 class is 15 kg $\equiv \frac{15}{10} \times 2 = 3$ cm

The area of the 50 – 60 bar is $2 \times 9 = 18$ cm^2 $\equiv$ frequency 60

$\Rightarrow$ the frequency of the 70 – 85 bar is 20 $\equiv$ an area of $\frac{20}{60} \times 18 = 6$ cm^2

$\Rightarrow$ the height of the 70 – 85 bar is *area* $\div$ *width* = $6 \div 3 = 2$ cm.

Answer width of 70 – 85 kg bar is 3 cm, and height is 2 cm.

3 Mode, mean (and median)

Mode

The mode is the value, or class interval, which occurs most often.

Mean

The mean of the values $x_1, x_2, \dots, x_n$ with frequencies $f_1, f_2, \dots, f_n$ the mean is

$$m = \bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i\,, \qquad \text{where } N = \sum_{i=1}^{n} f_i$$

Example: Find the mean for the following table showing the number of children per family.

Solution:

Number of children x	Frequency f	xf
0	5	0
1	8	8
2	12	24
3	18	54
4	9	36
5	4	20
	56	142

$\Sigma\, x_i f_i = 142$, and $N = \Sigma\, f_i = 56$

$\Rightarrow \quad \bar{x} = \frac{142}{56} = 2{\cdot}54$ to 3 S.F.

In a grouped frequency table you must use the mid-interval value.

Example: The table shows the numbers of children in prep school classes in a town.

Solution:

Number of children	Mid-interval value x	Frequency f	xf
1 - 10	5·5	5	27·5
11 - 15	13	8	104
16 – 20	18	12	216
21 - 30	25·5	18	459
31 - 40	35·5	11	390·5
		54	1197

$\Sigma x_i f_i = 1197$, and $N = \Sigma f_i = 54$

$\Rightarrow \quad \bar{x} = \frac{1197}{54} = 22{\cdot}2$ to 3 S.F.

Coding

The weights of a group of people are given as $x_1, x_2, \ldots x_n$ in *kilograms*. These weights are now changed to *grammes* and given as $t_1, t_2, \ldots t_n$.

In this case $t_i = 1000 \times x_i$ – this is an example of *coding*.

Another example of coding could be $t_i = \frac{x_i - 20}{5}$.

Coding and calculating the mean

With the coding, $t_i = \frac{x_i - 20}{5}$, we are subtracting 20 from each x-value and then dividing the result by 5.

We first find the mean for t_i, and then we reverse the process to find the mean for x_i

$\Rightarrow$ we find the mean for t_i, multiply by 5 and add 20, giving $\bar{x} = 5\bar{t} + 20$

Proof: $t_i = \frac{x_i - 20}{5} \quad \Rightarrow \quad x_i = 5t_i + 20$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i = \frac{1}{N}\sum_{i=1}^{n}(5t_i + 20)f_i$$

$$\Rightarrow \quad \bar{x} = \frac{5}{N}\sum_{i=1}^{n} t_i f_i + \frac{20}{N}\sum_{i=1}^{n} f_i$$

$\Rightarrow \quad \bar{x} = 5\bar{t} + 20$ since $\bar{t} = \frac{1}{N}\sum_{i=1}^{n} t_i f_i$ and $N = \sum_{i=1}^{n} f_i$

Example: Use the coding $t_i = \frac{x_i - 165}{10}$ to find the mean weight for the following distribution.

Weight, kg	Mid-interval x_i	Coded value $t_i = \frac{x_i-165}{10}$	Frequency f_i	$t_i f_i$
140 - 150	145	–2	9	–18
150 - 160	155	–1	21	–21
160 - 170	165	0	37	0
170 - 180	175	1	28	28
180 - 190	185	2	11	22
			106	11

$\Rightarrow \quad \bar{t} = \frac{11}{106}$

and $\quad t_i = \frac{x_i - 165}{10} \quad \Rightarrow \quad \bar{x} = 10\,\bar{t} + 165 = 10 \times \frac{11}{106} + 165 = 166{\cdot}0377358$

$\Rightarrow$ mean weight is $166{\cdot}04$ kg to 2 D.P.

Here the coding simplified the arithmetic for those who like to work without a calculator!

Median

The median is the middle number in an ordered list. Finding the median is explained in the next section.

When to use mode, median and mean

Mode

You should use the mode if the data is qualitative (colour etc.) or if quantitative (numbers) with a clearly defined mode (or bi-modal). It is not much use if the distribution is fairly even.

Median

You should use this for quantitative data (numbers), particularly when there are extreme values (outliers).

Mean

This is for quantitative data (numbers), and uses all pieces of data. It gives a true measure, but is affected by extreme values (outliers).

4 Median (Q_2), quartiles (Q_1, Q_3) and percentiles

Discrete lists and discrete frequency tables

To find medians and quartiles

1. Find $k = \frac{n}{2}$ (for Q_2), $\frac{n}{4}$ (for Q_1), $\frac{3n}{4}$ (for Q_3).
2. If k is an integer, use the mean of the k^{th} and $(k + 1)^{\text{th}}$ numbers in the list.
3. If k is not an integer, use the next integer **up**, and find the number with that position in the list.

Interquartile range

The interquartile range, I.Q.R., is $Q_3 - Q_1$.

Discrete lists

A discrete list of 10 numbers is shown below:

x	11	13	17	25	33	34	42	49	51	52

$n = 10$ for Q_1, $\frac{n}{4} = 2\frac{1}{2}$ so use 3^{rd} number, $\Rightarrow$ $\boldsymbol{Q_1 = 17}$

for Q_2, $\frac{n}{2} = 5$ so use mean of 5^{th} and 6^{th}, $\Rightarrow$ $\boldsymbol{Q_2 = 33\frac{1}{2}}$ median

for Q_3, $\frac{3n}{4} = 7\frac{1}{2}$ so use 8^{th} number, $\Rightarrow$ $\boldsymbol{Q_3 = 49}$

The interquartile range, I.Q.R., is $Q_3 - Q_1 = 49 - 17 = 32$.

Discrete frequency tables

x	5	6	7	8	9	10	11	12
f	3	6	8	10	9	8	6	4
cum freq	3	9	17	27	36	44	50	54

$n = 54$ for Q_1, $\frac{n}{4} = 13\frac{1}{2}$ so use 14^{th} number, $\Rightarrow$ $\boldsymbol{Q_1 = 7}$

for Q_2, $\frac{n}{2} = 27$ so use mean of 27^{th} and 28^{th}, $\Rightarrow$ $\boldsymbol{Q_2 = 8\frac{1}{2}}$ median

for Q_3, $\frac{3n}{4} = 40\frac{1}{2}$ so use 41^{st} number, $\Rightarrow$ $\boldsymbol{Q_3 = 10}$

The interquartile range, I.Q.R., is $Q_3 - Q_1 = 10 - 7 = 3$.

Grouped frequency tables, continuous and discrete data

To find medians and quartiles

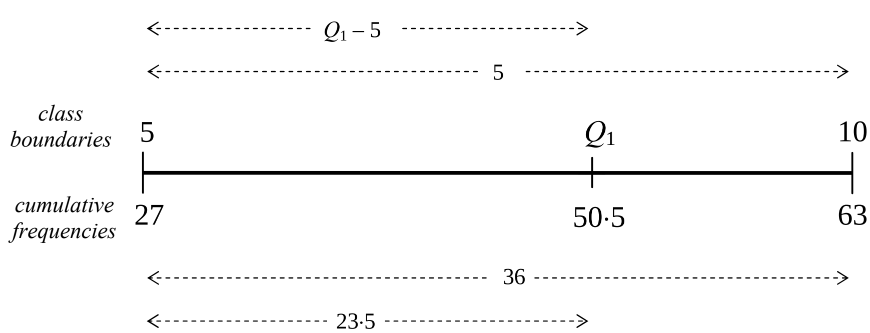
1. Find $k = \frac{n}{2}$ (for Q_2), $\frac{n}{4}$ (for Q_1), $\frac{3n}{4}$ (for Q_3).
2. **Do not round *k* up or change it in any way.**
3. Use linear interpolation to find median and quartiles – **note** that you must use the correct intervals for discrete data (start at the $\frac{1}{2}$ s).

Grouped frequency tables, continuous data

class boundaries	frequency	cumulative frequency
$0 \le x < 5$	27	27
5 to 10	36	63
10 to 20	54	117
20 to 30	49	166
30 to 60	24	190
60 to 100	12	202

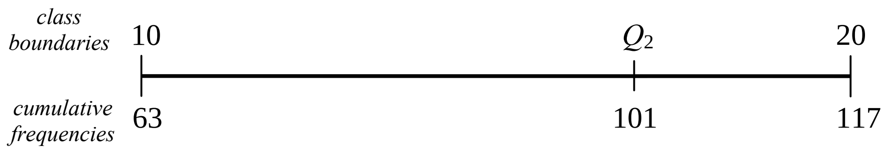
With *continuous* data, the end of one interval is the same as the start of the next – no gaps.

To find Q_1, $n = 202$ $\Rightarrow$ $\frac{n}{4} = 50\frac{1}{2}$ **do not change it**



From the diagram $\frac{Q_1 - 5}{5} = \frac{23\cdot5}{36} \Rightarrow Q_1 = 5 + 5 \times \frac{23\cdot5}{36} = 8\cdot263888889 = 8\cdot26$ to 3 S.F.

To find Q_2, $n = 202$ $\Rightarrow$ $\frac{n}{2} = 101$ **do not change it**



From the diagram $\frac{Q_2 - 10}{20 - 10} = \frac{101 - 63}{117 - 63}$ $\Rightarrow$ $Q_2 = 10 + 10 \times \frac{38}{54} = 17{\cdot}037\ldots = 17{\cdot}0$ to 3 S.F.

Similarly for Q_3, $\frac{3n}{4} = 151{\cdot}5$, so Q_3 lies in the interval (20, 30)

$\Rightarrow$ $\frac{Q_3 - 20}{30 - 20} = \frac{151.5 - 117}{166 - 117}$ $\Rightarrow$ $Q_3 = 20 + 10 \times \frac{34.5}{49} = 27{\cdot}0408\ldots = 27{\cdot}0$ to 3 S.F.

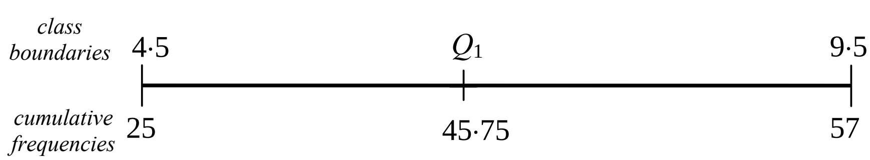
Grouped frequency tables, discrete data

The *discrete* data in grouped frequency tables is treated as *continuous*.

1. Change the class boundaries to the $4\frac{1}{2}$, $9\frac{1}{2}$ etc.
2. Proceed as for grouped frequency tables for continuous data.

class interval	class boundaries	frequency	cumulative frequency
0 – 4	0 to 4 ½	25	25
5 – 9	4 ½ to 9 ½	32	57
10 – 19	9 ½ to 19 ½	51	108
20 – 29	19 ½ to 29 ½	47	155
30 – 59	29 ½ to 59 ½	20	175
60 –99	59 ½ to 99 ½	8	183

To find Q_1, $n = 183$ $\Rightarrow$ $\frac{n}{4} = 45{\cdot}75$



From the diagram $\frac{Q_1 - 4{\cdot}5}{9{\cdot}5 - 4{\cdot}5} = \frac{45{\cdot}75 - 25}{57 - 25}$

$\Rightarrow \qquad Q_1 = 4{\cdot}5 + 5 \times \frac{20{\cdot}75}{32} = 7{\cdot}7421875\ldots = 7{\cdot}74$ to 3 S.F.

Q_2 and Q_3 can be found in a similar way.

Percentiles

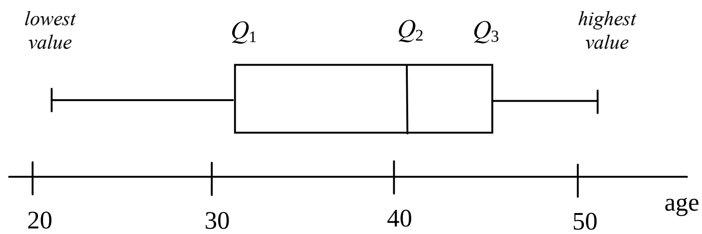
Percentiles are calculated in exactly the same way as quartiles.

Example: For the 90th percentile, find $\frac{90n}{100}$ and proceed as above.

Box Plots

In a group of people the youngest is 21 and the oldest is 52. The quartiles are 32 and 45, and the median age is 41.

We can illustrate this information with a box plot as below – remember to include a scale.



Outliers

An outlier is an extreme value. You are not required to remember how to find an outlier – you will always be given a rule.

Example: The ages of 11 children are given below.

age	3	6	12	12	13	14	14	15	17	21	26

$Q_1 = 12$, $Q_2 = 14$ and $Q_3 = 17$.

Outliers are values outside the range $Q_1 - 1{\cdot}5 \times (Q_3 - Q_1)$ to $Q_3 + 1{\cdot}5 \times (Q_3 - Q_1)$.

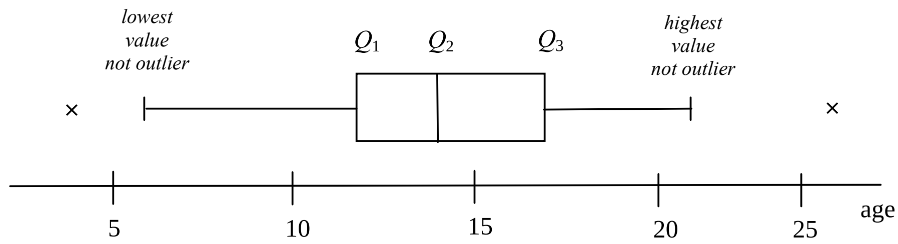
Find any outliers, and draw a box plot.

Solution: Lower boundary for outliers is $12 - 1{\cdot}5 \times (17 - 12) = 4{\cdot}5$

Upper boundary for outliers is $17 + 1{\cdot}5 \times (17 - 12) = 24{\cdot}5$

$\Rightarrow$ 3 and 26 are the only outliers.

To draw a box plot, put crosses at 3 and 26, and draw the lines to 6 (the lowest value which is *not* an outlier), and to 21 (the highest value which is *not* an outlier).



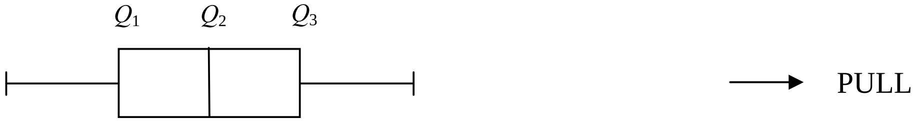
Note that there are other ways of drawing box plots with outliers, but this is the safest and will never be wrong – so why not use it.

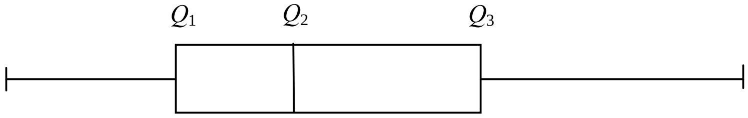
Skewness

A distribution which is symmetrical is **not** skewed

Positive skew

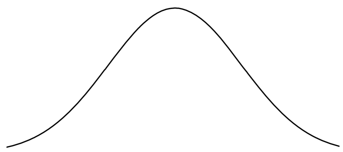
If a symmetrical box plot is stretched in the direction of the positive x-axis, then the resulting distribution has *positive* skew.





For positive skew the diagram shows that $Q_3 - Q_2 > Q_2 - Q_1$

The same ideas apply for a continuous distribution, and a little bit of thought should show that for *positive* skew *mean* > *median* > *mode*.



PULL

mode *mean*

median

Negative skew

If a symmetrical box plot is stretched in the direction of the negative x-axis, then the resulting distribution has *negative* skew.

Q_1 Q_2 Q_3

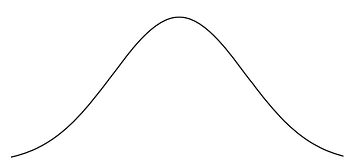
PULL

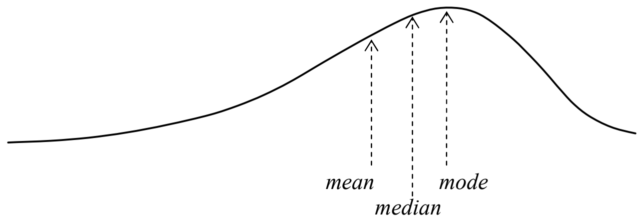
Q_1 Q_2 Q_3

For negative skew the diagram shows that $Q_3 - Q_2 < Q_2 - Q_1$

The same ideas apply for a continuous distribution, and a little bit of thought should show that for *negative* skew *mean* < *median* < *mode*.

PULL





5 Measures of spread

Range & interquartile range

Range

The *range* is found by subtracting the smallest value from the largest value.

Interquartile range

The *interquartile range* is found by subtracting the lower quartile from the upper quartile, so I.Q.R. = $Q_3 - Q_1$.

Variance and standard deviation

Variance is the square of the standard deviation.

$$s_x{}^2 = \frac{1}{N}\sum(x_i - \bar{x})^2 f_i, \quad \text{or}$$

$$s_x{}^2 = \frac{1}{N}\sum x_i{}^2 f_i - \bar{x}^2.$$

When finding the variance, it is nearly always easier to use the second formula.

Variance and standard deviation measure the *spread* of the distribution.

Proof of the alternative formula for variance

$$s_x{}^2 = \frac{1}{N}\sum(x_i - \bar{x})^2 f_i = \frac{1}{N}\sum(x_i{}^2 - 2x_i\bar{x} + \bar{x}^2) f_i$$

$$= \frac{1}{N}\sum x_i{}^2 f_i - \frac{1}{N}\sum 2x_i\bar{x}f_i + \frac{1}{N}\sum \bar{x}^2 f_i$$

$$= \frac{1}{N}\sum x_i{}^2 f_i - \frac{2\bar{x}}{N}\sum x_i f_i + \frac{\bar{x}^2}{N}\sum f_i$$

since $\bar{x} = \frac{1}{N}\sum xf$ and $N = \sum f$

$$= \frac{1}{N}\sum x_i{}^2 f_i - 2\bar{x}^2 + \bar{x}^2 \;\; = \frac{1}{N}\sum x_i{}^2 f_i \; - \; \bar{x}^2$$

Rough checks, $m \pm s$, $m \pm 2s$

When calculating a standard deviation, you should check that there is approximately 65 - 70% of the population within 1 s.d. of the mean and

approximately 95% within 2 s.d. of the mean.

These approximations are best for a fairly symmetrical distribution.

Coding and variance

Using the coding $t_i = \frac{x_i - k}{a}$ we see that

$x_i = at_i + k \quad \Rightarrow \quad \bar{x} = a\bar{t} + k$

$$\Rightarrow \; s_x{}^2 = \frac{1}{N}\sum (x_i - \bar{x})^2 f_i \quad = \frac{1}{N}\sum \big((at_i + k) - (a\bar{t} + k)\big)^2 f_i$$

$$= \frac{1}{N}\sum (at_i - a\bar{t})^2 f_i \quad = \frac{a^2}{N}\sum (t_i - \bar{t})^2 f_i$$

$$\Rightarrow \; s_x{}^2 = a^2 s_t{}^2 \quad \Rightarrow \quad s_x = a s_t$$

Notice that subtracting k has no effect, since this is equivalent to translating the graph, and therefore does not change the *spread*, and if all the x-values are divided by a, then we need to multiply s_t by a to find s_x.

Example: Find the mean and standard deviation for the following distribution.

Here the x-values are nasty, but if we change them to form $t_i = \frac{x_i - 210}{5}$ then the arithmetic in the last two columns becomes much easier.

x	$t_i = \frac{x_i - 210}{5}$	f	$t_i f_i$	$t_i{}^2 f_i$
200	–2	12	–24	48
205	–1	23	–23	23
210	0	42	0	0
215	1	30	30	30
220	2	10	20	40
		117	3	141

the mean of t is $\bar{t} = \frac{1}{N}\sum t_i f_i = \frac{3}{117} = \frac{1}{39}$

and the variance of t is $s_t{}^2 = \frac{1}{N}\sum t_i{}^2 f_i - \bar{t}^2 = \frac{141}{117} - \left(\frac{1}{39}\right)^2 = 1{\cdot}204470743$

$s_t = \sqrt{1 \cdot 204470743} = 1 \cdot 097483824 = 1.10$ to 3 S.F.

To find $\bar{x}$, using $t_i = \frac{x_i - 210}{5}$, $\Rightarrow$ $\bar{x} = 5\bar{t} + 210 = 215.5$ to 1 D.P.

To find the standard deviation of x

$s_x = 5s_t = 5 \times 1{\cdot}0974838\ldots = 5{\cdot}49$ to 3 S.F.

We would need to multiply the variance by $5^2 = 25$

$\Rightarrow \quad s_x{}^2 = 25s_t{}^2 = 25 \times 1{\cdot}204470743 = 30{\cdot}1$ to 3 S.F.

---

6 Probability

Relative frequency

After tossing a drawing pin a large number of times
the *relative frequency* of it landing point up is $\frac{\textit{number of times with point up}}{\textit{total number of tosses}}$;
this can be thought of as the experimental probability.

Sample spaces, events and equally likely outcomes

A *sample space* is the set of all possible outcomes, all *equally likely*.
An *event* is a set of possible outcomes.

$$\text{P}(A) = \frac{\text{number of ways } A \text{ can happen}}{\text{total number in sample space}} = \frac{n(A)}{N},$$ where N is number in sample space.

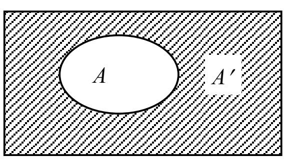
Probability rules and Venn diagrams

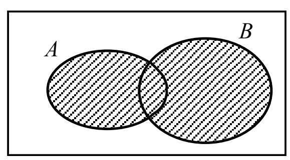
All outcomes must be equally likely to happen.

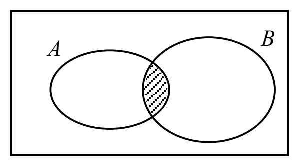
$P(A) = \dfrac{n(A)}{N}$

$P(A') = P(\text{not } A) = 1 - P(A)$

A' is the *complement* of A. $A \cup B$ means A *or* B *or both*. $A \cap B$ means *both* A *and* B,







$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \mid B)$ means the probability that A has occurred *given that* we know that B has already occurred, and should always be re-written as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

If we know that B has already happened, we can think of B as the new sample space with $n(B)$ elements.
Then the number of ways that A can now occur is $n(A \cap B)$

$$\Rightarrow \quad P(A \mid B) = \frac{n(A\cap B)}{n(B)} = \frac{\frac{n(A\cap B)}{N}}{\frac{n(B)}{N}} = \frac{P(A\cap B)}{P(B)}$$

Diagrams for two dice etc.

When considering two dice, two spinners or a coin and a die, the following types of diagram are often useful – they ensure that all outcomes are *equally likely* to happen.

Two dice

green						
6	×	×	×	×	×	×
5	×	×	×	×	×	×
4	×	×	×	×	×	×
3	×	×	×	×	×	×
2	×	×	×	×	×	×
1	×	×	×	×	×	×
	1	2	3	4	5	6 *red*

Coin and die

6	×	×
5	×	×
4	×	×
3	×	×
2	×	×
1	×	×
	H	T

Two coins

H	H
H	T
T	H
T	T

Three coins

H	H	H
H	H	T
H	T	H
T	H	H
H	T	T
T	H	T
T	T	H
T	T	T

From these diagrams it should be easy to see that

For two dice: $P(\text{total } 10) = \frac{3}{36}$, $P(red > green) = \frac{15}{36}$, $P(\text{total } 10 \mid 4 \text{ on green}) = \frac{\frac{3}{36}}{\frac{6}{36}} = \frac{1}{2}$.

For coin and die: $P(\text{Head and an even number}) = \frac{3}{12}$.

For three dice: $P(\text{exactly two Heads}) = \frac{3}{8}$.

Tree diagrams

The rules for tree diagrams are

Select which branches you need

Multiply along each branch

Add the **results of each branch** needed.

Make sure that you include enough working to show which branches you are using (method).

Be careful to allow for selection *with and without replacement*.

Example: In the launch of a rocket, the probability of an electrical fault is 0·2. If there is an electrical fault the probability that the rocket crashes is 0·4, and if there is no electrical fault the probability that the rocket crashes is 0·3.

Draw a tree diagram. The rocket takes off, and is seen to crash. What is the probability that there was an electrical fault?

Solution:

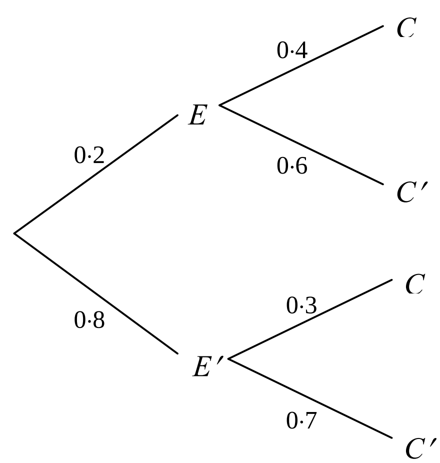
We want to find $P(E \mid C)$.

$P(E \mid C) = \frac{P(E \cap C)}{P(C)}$

$P(E \cap C) = 0{\cdot}2 \times 0{\cdot}4 = 0{\cdot}08$

and $P(C) = 0{\cdot}2 \times 0{\cdot}4 + 0{\cdot}8 \times 0{\cdot}3 = 0{\cdot}32$

$\Rightarrow \quad P(E \mid C) = \frac{0{\cdot}08}{0{\cdot}32} = 0{\cdot}25$



Independent events

Definition. A and B are independent $\Leftrightarrow$ $P(A \cap B) = P(A) \times P(B)$

It is also true that $P(A \mid B) = P(A \mid B') = P(A)$.

A and B are not linked, they have no effect on each other.

To prove that *A* and *B* are independent

first find $P(A)$, $P(B)$ and $P(A \cap B)$ **without** assuming that $P(A \cap B) = P(A) \times P(B)$,

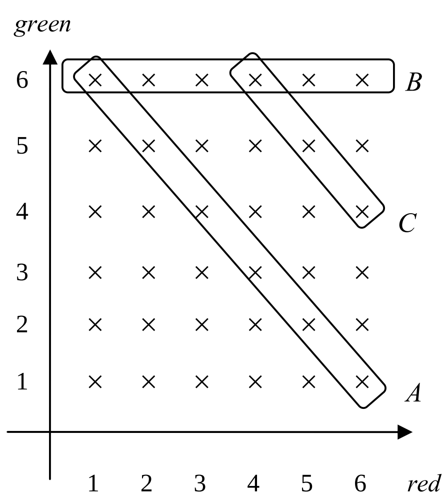
second show that $P(A \cap B) = P(A) \times P(B)$.

Note: If A and B are **not** independent then $P(A \cap B) \neq P(A) \times P(B)$, and must be found in another way, usually considering sample spaces and/or Venn diagrams.

Example: A red die and a green die are rolled and the total score recorded.

A is the event 'total score is 7', B is the event 'green score is 6' and C is the event 'total score is 10'.
Show that A and B are independent, but B and C are not independent.

Solution: The events A, B and C are shown on the diagram.



$P(A) = \frac{6}{36} = \frac{1}{6}$, $P(B) = \frac{6}{36} = \frac{1}{6}$

and $P(A \cap B) = \frac{1}{36}$ from diagram

$P(A) \times P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} = P(A \cap B)$

$\Rightarrow$ A and B are independent.

$P(B) = \frac{6}{36} = \frac{1}{6}$, $P(C) = \frac{3}{36} = \frac{1}{12}$

and $P(B \cap C) = \frac{1}{36}$ from diagram

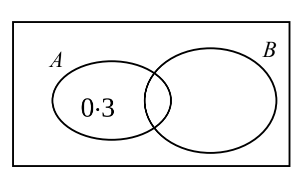
$P(B) \times P(C) = \frac{1}{6} \times \frac{1}{12} = \frac{1}{72} \neq P(B \cap C)$

$\Rightarrow$ B and C are **not** independent

Example: A and B are independent events. $P(A) = 0{\cdot}5$ and $P(A \cap B') = 0{\cdot}3$. Find $P(B)$.

Solution: $P(A) = 0{\cdot}5$ and $P(A \cap B') = 0{\cdot}3$

$\Rightarrow$ $P(A \cap B) = 0{\cdot}5 - 0{\cdot}3 = 0{\cdot}2$



But $\quad P(A \cap B) = P(A) \times P(B)$

$\Rightarrow \quad 0{\cdot}2 = 0{\cdot}5 \times P(B)$

$\Rightarrow \quad P(B) = \dfrac{0{\cdot}2}{0{\cdot}5} = 0{\cdot}4$

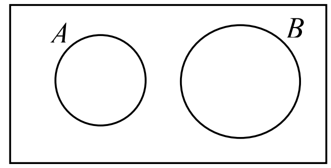
Exclusive events

Definition. A and B are mutually exclusive

$\Leftrightarrow P(A \cap B) = 0$

i.e. they cannot both occur at the same time

$\Rightarrow \quad P(A \cup B) = P(A) + \mathrm{P}(B)$

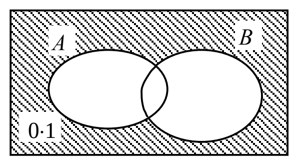


Note: If A and B are **not** exclusive then $P(A \cup B) \neq P(A) + P(B)$, and must be found in another way, usually considering sample spaces and/or Venn diagrams.

Example: $P(A) = 0{\cdot}3$, $P(B) = 0{\cdot}9$ and $P(A' \cap B') = 0{\cdot}1$.

Prove that A and B are mutually exclusive.

Solution: $A' \cap B'$ is shaded in the diagram



$\Rightarrow \quad P(A' \cap B') = 1 - P(A \cup B)$

$\Rightarrow \quad P(A \cup B) = 1 - 0{\cdot}1 = 0{\cdot}9$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$\Rightarrow \quad 0.9 = 0{\cdot}3 + 0{\cdot}6 - P(A \cap B)$

$\Rightarrow \quad P(A \cap B) = 0$

$\Rightarrow \quad A$ and B are mutually exclusive.

Number of arrangements

Example: A bag contains 5 Red beads, 7 Yellow beads, and 6 White beads. Three beads are drawn *without replacement* from the bag. Find the probability that there are 2 Red beads and 1 Yellow bead.

Solution: These beads can be drawn in any order, *RRY, RYR, YRR*

$\Rightarrow \quad P(RRY \text{ or } RYR \text{ or } YRR)$

$= \quad P(RRY) + P(RYR) + P(YRR)$

$= \quad \frac{5}{18} \times \frac{4}{17} \times \frac{7}{16} + \frac{5}{18} \times \frac{7}{17} \times \frac{4}{16} + \frac{7}{18} \times \frac{5}{17} \times \frac{4}{16} = \frac{35}{408} = 0.0858$ to 3 S.F.

You must always remember the possibility of more than one order. In rolling four DICE, exactly TWO SIXES can occur in **six** ways:

SSNN, SNSN, SNNS, NSSN, NSNS, NNSS, each of which would have the same probability $\left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2 = \frac{25}{1296}$

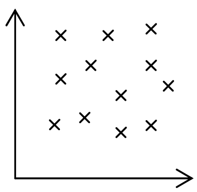
and so the probability of exactly two sixes with four dice is $6 \times \frac{25}{1296} = \frac{25}{216}$.

---

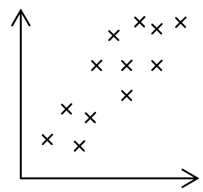
7 Correlation

Scatter diagrams

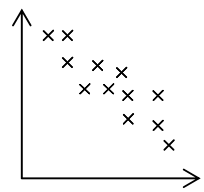
Positive, negative, no correlation & line of best fit.



no correlation



positive correlation



negative correlation

The pattern of a scatter diagram shows **linear** correlation in a general manner.
A line of best fit can be draw by eye, **but only when the points nearly lie on a straight line**.

Product moment correlation coefficient, PMCC

Formulae

The S_{**} are all similar to each other and make other formulae simpler to learn and use:

$$S_{xy} = \sum (x_i - \bar{x})(y - \bar{y}) = \sum x_i y_i - \frac{1}{N}\left(\sum x_i\right)\left(\sum y_i\right)$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i{}^2 - \frac{1}{N}\left(\sum x_i\right)^2$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i{}^2 - \frac{1}{N}\left(\sum y_i\right)^2$$

The **PMCC** $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \quad -1 \le r \le +1$ for proof, see appendix

To calculate the PMCC first calculate S_{xx}, S_{yy} and S_{xy} using the second formula on each line.
N.B. These formulae are all in the formula booklet.

Coding and the PMCC

To see the effect of coding on the PMCC, it is better to use the first formula on each line.

Example: Investigate the effect of the coding $t = \frac{x-k}{a}$ on the PMCC.

Solution:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$$t_i = \frac{x_i - k}{a} \quad \Rightarrow \quad x_i = at_i + k \quad \text{and} \quad \bar{x} = a\,\bar{t} + k$$

$$\Rightarrow \quad r_{xy} = \frac{\sum\left(at_i + k - (a\bar{t} + k)\right)(y - \bar{y})}{\sqrt{\sum\left(at_i + k - (a\bar{t} + k)\right)^2 \sum(y_i - \bar{y})^2}}$$

$$= \frac{\sum a(t_i - \bar{t})(y - \bar{y})}{\sqrt{\sum a^2(t_i - \bar{t})^2 \sum(y_i - \bar{y})^2}}$$

$$= \frac{a\sum(t_i - \bar{t})(y - \bar{y})}{\sqrt{a^2}\sqrt{\sum(t_i - \bar{t})^2 \sum(y_i - \bar{y})^2}}$$

$$= \frac{\sum(t_i - \bar{t})(y - \bar{y})}{\sqrt{\sum(t_i - \bar{t})^2 \sum(y_i - \bar{y})^2}}$$

$$= r_{ty}$$

In other words, the coding on x has had no effect on the PMCC. Similarly, coding on y has no effect on the PMCC.

$\Rightarrow$ **Coding has no effect on the PMCC.**

Interpretation of the PMCC

It can be shown that $-1 \le r \le +1$ see appendix

if $r = +1$ there is perfect *positive linear correlation*,
if $r = -1$ there is perfect *negative linear correlation*,
if $r = 0$ (or close to zero) there is *no linear correlation* .

PMCC tests to see if there is a **linear connection** between the variables.
For strong correlation, the points on a scatter graph will lie very close to a straight line, and r will be close to 1 or –1.

Example: Bleep tests are used to measure people's fitness. A higher score means a higher level of fitness. The heart rate, p beats per minute, and bleep score, s, for 12 people were recorded and coded, using $x = p - 60$ and $y = 10s - 50$.

x	0	–6	9	–1	5	8	30	19	28	20	36	23
y	55	62	38	–7	50	44	8	8	3	20	–14	3

$\Sigma x = 171$, $\Sigma y = 270$, $\Sigma x^2 = 4477$, $\Sigma y^2 = 13540$, $\Sigma xy = 1020$.

(*a*) Find the PMCC between x and y.

(*b*) Write down the PMCC between p and s.

(*c*) Explain why your answer to (*b*) might suggest that there is a linear relationship between p and s.

(*d*) Interpret the significance of the PMCC.

Solution: (*a*) $$S_{xy} = \sum x_i y_i - \frac{1}{N}\left(\sum x_i\right)\left(\sum y_i\right) = 1020 - \frac{171 \times 270}{12} = -2827{\cdot}5$$

$$S_{xx} = \sum x_i^2 - \frac{1}{N}\left(\sum x_i\right)^2 = 4477 - \frac{171^2}{12} = 2040{\cdot}25$$

$$S_{yy} = \sum y^2 - \frac{1}{N}\left(\sum y_i\right)^2 = 13540 - \frac{270^2}{12} = 7465$$

$$\Rightarrow \quad r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-2827{\cdot}5}{\sqrt{2040{\cdot}25 \times 7465}} = -0{\cdot}7245127195 = -0{\cdot}725 \text{ to 3 S.F.}$$

(*b*) As coding has no effect on the PMCC, the product moment correlation coefficient for p and s is also $-0{\cdot}725$, to 3 S.F.

(*c*) $r = -0{\cdot}725$ is 'quite close' to -1, and therefore the points on a scatter diagram would lie close to a straight line

$\Rightarrow$ there is evidence of a linear relation between p and s.

(*d*) There is **negative correlation** between p and s, which means that
as heart rate increases, the bleep score decreases, or
people with higher heart rate tend to have lower bleep scores.

---

8 Regression

Explanatory and response variables

h

d

In an experiment a toy car is released from rest on a ramp from a height of h. The horizontal distance, d, is then measured. The experimenter can control the height, h, and the distance, d, depends on the height chosen.

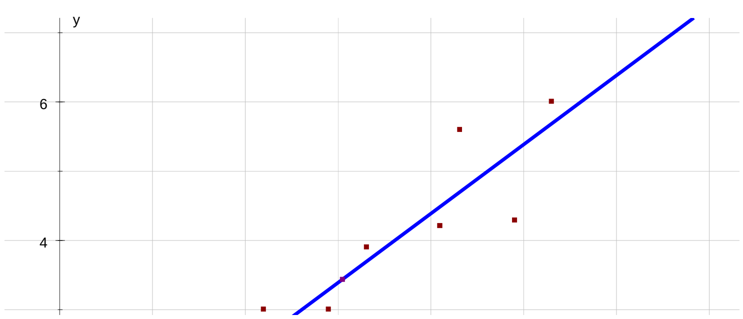
h is called the **explanatory** variable and is plotted on the horizontal axis. d is called the **response** variable and is plotted on the vertical axis.

In some cases it may not be possible to control the explanatory variable. For example the temperature at a given time may affect the sales of ice cream; the researcher cannot control the temperature, but it is the temperature which affects the ice cream sales.

Therefore the temperature is the *explanatory* variable, and the ice cream sales is the *response* variable.

Regression line

Least squares regression line



The scatter diagram shows the regression line of y on x. The *regression* line is drawn to minimise the sum of the squares of the vertical distances between the line and the points.

It can be shown that the *regression* line has equation $y = a + bx$, where $b = \dfrac{S_{xy}}{S_{xx}}$,

also that the *regression* line passes through the 'mean point', $(\bar{x}, \bar{y})$,

and so we can find a from the equation $\bar{y} = a + b\bar{x} \Rightarrow a = \bar{y} - b\bar{x}$

Interpretation

In the equation $y = a + bx$

a is the value of y when x is zero (or when x is not present)
b is the amount by which y increases for an increase of 1 in x.
You must write your interpretation in the context of the question.

Example: A local authority is investigating the cost of reconditioning its incinerators. Data from 10 randomly chosen incinerators were collected. The variables monitored were the operating time x (in thousands of hours) since last reconditioning and the reconditioning cost y (in £1000). None of the incinerators had been used for more than 3000 hours since last reconditioning.

The data are summarised below,

$$\Sigma x = 25.0,\ \Sigma x^2 = 65.68,\ \Sigma y = 50.0,\ \Sigma y^2 = 260.48,\ \Sigma xy = 130.64.$$

(*a*) Find the equation of the regression line of y on x.

(*b*) Give interpretations of a and b.

Solution:

(*a*) $\bar{x} = \dfrac{25{\cdot}0}{10} = 2{\cdot}50$, $\quad \bar{y} = \dfrac{50{\cdot}0}{10} = 5{\cdot}00$,

$S_{xy} = 130{\cdot}64 - \dfrac{25{\cdot}0 \times 50{\cdot}0}{10} = 5{\cdot}64$, $\quad S_{xx} = 65.68 - \dfrac{(25{\cdot}0)^2}{10} = 3{\cdot}18$

$\Rightarrow \quad b = \dfrac{S_{xy}}{S_{xx}} = \dfrac{5.64}{3.18} = 1{\cdot}773584906$

$\Rightarrow \quad a = \bar{y} - b\bar{x} = 5{\cdot}00 - 1{\cdot}773584906 \times 2{\cdot}50 = 0{\cdot}5660377358$

$\Rightarrow$ regression line equation is $y = 0{\cdot}566 + 1{\cdot}77x$ to 3 S.F.

(*b*) *a* is the cost in £1000 of reconditioning an incinerator which has not been used, so the cost of reconditioning an incinerator which has not been used is £566.
(*a* is the value of *y* when *x* is zero)

b is the increase in cost (in £1000) of reconditioning for every extra 1000 hours of use, so it costs an extra £1774 to recondition an incinerator for every 1000 hours of use. (*b* is the gradient of the line)

---

9 Discrete Random Variables

Random Variables

A random variable must take a numerical value:

Examples: the number on a single throw of a die
the height of a person
the number of cars travelling past a fixed point in a certain time

But **not** the *colour* of hair as this is not a number

Continuous and discrete random variables

Continuous random variables

A continuous random variable is one which can take **any** value in a certain interval;
Examples: height, time, weight.

Discrete random variables

A discrete random variable can only take certain values in an interval
Examples: Score on die (1, 2, 3, 4, 5, 6)
Number of coins in pocket (0, 1, 2, ...)

Probability distributions

A probability distribution is the set of possible outcomes together with their probabilities, similar to a frequency distribution or frequency table.

Example:

score on two dice, X	2	3	4	5	6	7	8	9	10	11	12
probability, $f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

is the probability distribution for the random variable, X, the total score on two dice.

Note that the sum of the probabilities **must** be 1, i.e. $\sum_{x=2}^{12} P(X = x) = 1$.

Cumulative probability distribution

Just like cumulative frequencies, the cumulative probability, F, that the total score on two dice is less than or equal to 4 is $F(4) = P(X \le 4) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36} = \frac{1}{6}$.

Note that $F(4.3)$ means $P(X \le 4.3)$ and seeing as there are no scores between 4 and 4.3 this is the same as $P(X \le 4) = F(4)$.

Expectation or expected values

Expected mean or expected value of *X*.

For a discrete probability distribution the expected mean of X, or the expected value of X is

$$\mu = \mathrm{E}[X] = \sum x_i p_i$$

Expected value of a function

The expected value of any function, $f(X)$, is defined as

$$\mathrm{E}[X] = \sum f(x_i) p_i$$

Note that for any constant, k, $\mathrm{E}[k] = k$, since $\sum k p_i = k \sum p_i = k \times 1 = k$

Expected variance

The expected variance of X is

$$\sigma^2 = \mathrm{Var}[X] = \sum (x_i - \mu)^2 p_i = \sum x_i{}^2 p_i - \mu^2, \quad \text{or}$$

$$\sigma^2 = \mathrm{Var}[X] = \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - \mu^2 = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

Expectation algebra

$\text{E}[aX + b] = \sum (ax_i + b)p_i = \sum ax_i p_i + bp_i = a \sum x_i p_i + b$

$= a\text{E}[X] + b$ since $\sum x_i p_i = \mu$, and $\sum p_i = 1$

$\text{Var}[aX + b] = \text{E}[(aX + b)^2] - (\text{E}[(aX + b)])^2$

$= \text{E}[(a^2X^2 + 2abX + b^2)]) - (a\text{E}[X] + b)^2$

$= \{a^2\, \text{E}[X^2] + 2ab\, \text{E}[X] + \text{E}[b^2]\} - \{a^2\, (\text{E}[X])^2 + 2ab\, \text{E}[X] + b^2\}$

$= a^2\, \text{E}[X^2] - a^2\, (\text{E}[X])^2 = a^2\{\text{E}[X^2] - (\text{E}[X])^2\} = a^2\, \text{Var}[X]$

Thus we have two important results:

$\text{E}[aX + b] = a\text{E}[X] + b$

$\text{Var}[aX + b] = a^2\, \text{Var}[X]$

which are equivalent to the results for coding done earlier.

Example: A fair die is rolled and the score recorded.

(a) Find the expected mean and variance for the score, X.

(b) A ‘prize’ is awarded which depends on the score on the die. The value of the prize is \$$Z = 3X - 6$. Find the expected mean and variance of Z.

Solution: (a)

score	probability	$x_i p_i$	$x_i^2 p_i$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{4}{6}$
3	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{9}{6}$
4	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{16}{6}$
5	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{25}{6}$
6	$\frac{1}{6}$	$\frac{6}{6}$	$\frac{36}{6}$
		$\frac{21}{6}$	$\frac{91}{6}$

$\Rightarrow \quad \mu = \text{E}[X] = \sum x_i p_i = \frac{21}{6} = 3\frac{1}{2}$

and $\sigma^2 = \text{E}[X^2] - (\text{E}[X])^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{35}{12} = 2\frac{11}{12}$

$\Rightarrow$ The expected mean and variance for the score, X, are $\mu = 3\frac{1}{2}$ and $\sigma^2 = 2\frac{11}{12}$

(*b*) $Z = 3X - 6$

$\Rightarrow$ $E[Z] = E[3X - 6] = 3E[X] - 6 = 10\frac{1}{2} - 6 = 4\frac{1}{2}$

and $Var[Z] = Var[3X - 6] = 3^2 Var[X] = 9 \times \frac{35}{12} = 26\frac{1}{4}$

$\Rightarrow$ The expected mean and variance for the prize, $\$Z$, are $\mu = 4\frac{1}{2}$ and $\sigma^2 = 26\frac{1}{4}$

The discrete uniform distribution

Conditions for a discrete uniform distribution

- The discrete random variable X is defined over a set of n distinct values
- Each value is equally likely, with probability $^1/_n$.

Example: The random variable X is defined as the score on a single die. X is a discrete uniform distribution on the set {1, 2, 3, 4, 5, 6}

The probability distribution is

Score	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Expected mean and variance

For a discrete uniform random variable, X defined on the set $\{1, 2, 3, 4, ..., n\}$,

X	1	2	3	4	...	...	n
Probability	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$			$\frac{1}{n}$

By symmetry we can see that the Expected mean $= \mu = E[X] = \frac{1}{2}(n+1)$,

or $\mu = E[X] = \sum x_i p_i = 1 \times \frac{1}{n} + 2 \times \frac{1}{n} + 3 \times \frac{1}{n} + \ldots + n \times \frac{1}{n}$

$= (1 + 2 + 3 + \ldots + n) \times \frac{1}{n} = \frac{1}{2}n(n+1) \times \frac{1}{n} = \frac{1}{2}(n+1)$

The expected variance,

$$\text{Var}[X] = \sigma^2 = \text{E}[X^2] - (\text{E}[X])^2 = \sum x_i^2 p_i - \mu^2$$

$$= (1^2 + 2^2 + 3^2 + \ldots + n^2) \times \frac{1}{n} - \left(\frac{1}{2}(n+1)\right)^2$$

$$= \frac{1}{6} n(n+1)(2n+1) \times \frac{1}{n} - \frac{1}{4}(n+1)^2 \qquad \text{since } \sum i^2 = \frac{1}{6} n(n+1)(2n+1)$$

$$= \frac{1}{24}(n+1)\{(8n+4) - (6n+6)\}$$

$$= \frac{1}{24}(n+1)(2n - 2)$$

$$\Rightarrow \qquad \text{Var}[X] = \sigma^2 = \frac{1}{12}(n^2 - 1)$$

These formulae can be quoted in an exam (if you learn them!).

Non-standard uniform distribution

The formulae can sometimes be used for non-standard uniform distributions.

Example: X is the score on a fair 10 sided spinner. Define $Y = 5X + 3$.

Find the mean and variance of Y.

Y is the distribution $\{8, 13, 18, \ldots 53\}$, all with the same probability $\frac{1}{10}$.

Solution: X is a discrete uniform distribution on the set $\{1, 2, 3, \ldots, 10\}$

$\Rightarrow \quad \text{E}[X] = \frac{1}{2}(n+1) = 5\frac{1}{2}$

and $\quad \text{Var}[X] = \frac{1}{12}(n^2 - 1) = \frac{99}{12} = 8\frac{1}{4}$

$\Rightarrow \quad \text{E}[Y] = \text{E}[5X + 3] = 5\text{E}[X] + 3 = 30\frac{1}{2}$

and $\quad \text{Var}[X] = \text{Var}[5X + 3] = 5^2\,\text{Var}[X] = 25 \times \frac{99}{12} = 206\frac{1}{4}$

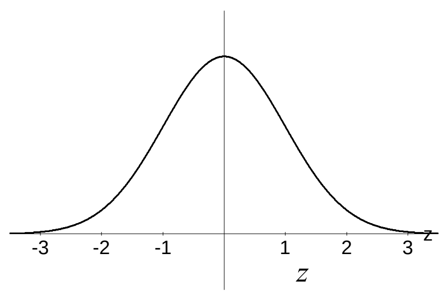
$\Rightarrow \quad$ mean and variance of Y are $27\frac{1}{2}$ and $206\frac{1}{4}$.

---

10 The Normal Distribution N(μ, σ^2)

The standard normal distribution N(0, 1^2)

The diagram shows the standard normal distribution



Mean, μ, = 0

Standard deviation, σ, = 1

The tables give the area, $\Phi(z)$, from $-\infty$ upto z;

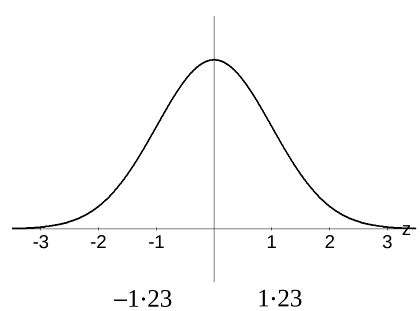
To find other probabilities, sketch the curve and use your head

Example: $P(Z < -1{\cdot}23)$

= area upto $-1{\cdot}23$ = $\Phi(-1{\cdot}23)$

= area beyond $+1{\cdot}23$ = $1 - \Phi(+1{\cdot}23)$

= $1 - 0{\cdot}8907$ = $0{\cdot}1093$ to 4 D.P.



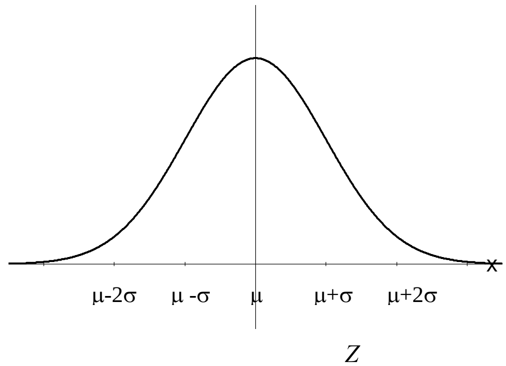
The general normal distribution N(μ, σ^2)

Use of tables

To use the tables for a Normal distribution with

mean μ and standard deviation σ

We use $Z = \dfrac{X - \mu}{\sigma}$ (see appendix) and look in the tables under this value of Z



Example: The length of life (in months) of Blowdri's hair driers is approximately Normally distributed with mean 90 months and standard deviation 15 months.

(*a*) Each drier is sold with a 5 year guarantee. What proportion of driers fail before the guarantee expires?

(*b*) The manufacturer decides to change the length of the guarantee so that no more than 1% of driers fail during the guarantee period. How long should he make the guarantee?

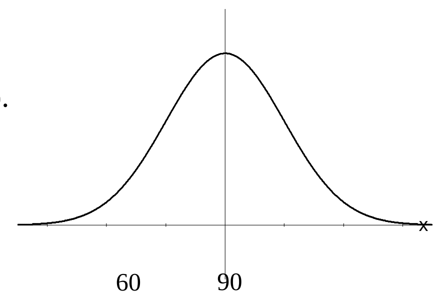
Solution:

(*a*) X is the length of life of drier $\Rightarrow$ $X \sim$ N(90, 15^2).

5 years = 60 months

$\Rightarrow$ we want $P(X < 60)$ = area upto 60

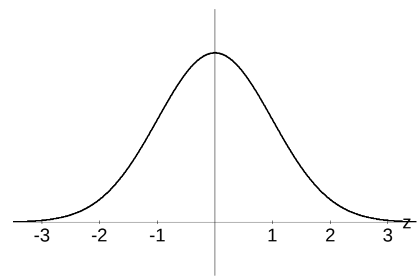
$$\Rightarrow \; Z = \frac{X - \mu}{\sigma} = \frac{60 - 90}{15} = -2.0$$



so we want area to left of $Z = -2$

$= \Phi(-2) = 1 - \Phi(2)$

$= 1 - 0{\cdot}9772 = 0{\cdot}0228$ to 4 D.P. from tables.



$\Rightarrow$ the proportion of hair driers failing during the guarantee period is 0.0288 to 4 D.P.

(*b*) Let the length of the guarantee be t years

$\Rightarrow$ we need $P(X < t) = 0{\cdot}01$.

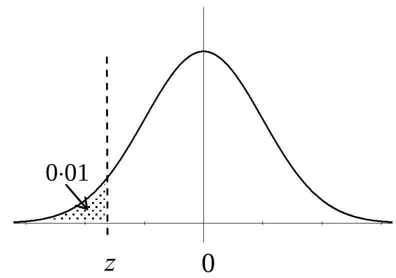
0·01

t 90

We need the value of Z such that $\Phi(Z) = 0{\cdot}01$

From the tables $Z = -2{\cdot}3263$ to 4 D.P. from tables (remember to look in the small table after the Normal tables)

Standardising the variable



$\Rightarrow \quad Z = \dfrac{X-\mu}{\sigma} = \dfrac{t-90}{15}$

$\Rightarrow \quad \dfrac{t-90}{15} = -2{\cdot}3263$ to 4 D.P. from tables

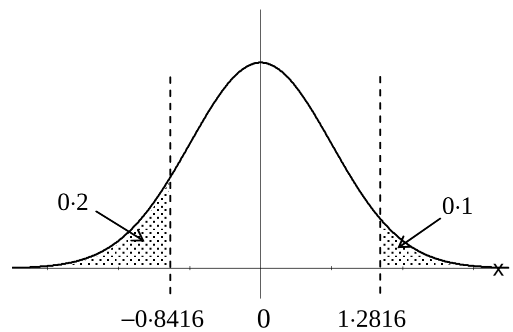
$\Rightarrow \quad t = 55{\cdot}1$ to 3 S.F.

so the manufacturer should give a guarantee period of 55 months (4 years 7 months)

Example: The results of an examination were Normally distributed. 10% of the candidates had more than 70 marks and 20% had fewer than 35 marks.

Find the mean and standard deviation of the marks.

Solution:

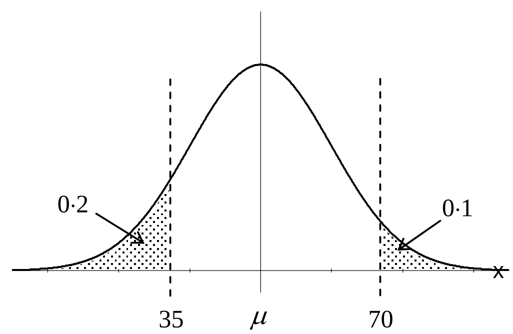


First we need the values from the tables

$\Rightarrow \quad \Phi(-0{\cdot}8416) = 0{\cdot}2,$

and $1 - \Phi(1{\cdot}2816) = 0{\cdot}1$

Using $Z = \dfrac{X-\mu}{\sigma}$ we have



$-0{\cdot}8416 = \dfrac{35-\mu}{\sigma}$

$\Rightarrow \quad \mu = 35 + 0{\cdot}8416\sigma$

and $1{\cdot}2816 = \dfrac{70-\mu}{\sigma}$

$\Rightarrow \quad \mu = 70 - 1{\cdot}2816\sigma$

$\Rightarrow \quad \sigma = 16{\cdot}5$ and $\mu = 48{\cdot}9$ to 3 S.F. simultaneous equations

Example: The weights of chocolate bars are normally distributed with mean 205 g and standard deviation 2·6 g. The stated weight of each bar is 200 g.

(*a*) Find the probability that a single bar is underweight.

(*b*) Four bars are chosen at random. Find the probability that fewer than two bars are underweight.

Solution:

(a) Let W be the weight of a chocolate bar, $W \sim \mathrm{N}(205, 2{\cdot}6^2)$.

$Z = \frac{W-\mu}{\sigma} = \frac{200-205}{2{\cdot}6} = -\,1{\cdot}9230769\ldots$

$\mathrm{P}(W < 200) = \mathrm{P}(Z < -\,1{\cdot}92) = 1 - \Phi(1{\cdot}92) = 1 - 0{\cdot}9726$

$\Rightarrow$ probability of an underweight bar is $0{\cdot}0274$.

(*b*) We want the probability that 0 or 1 bars chosen from 4 are underweight.

Let U be underweight and C be correct weight.

P(1 underweight) = P($CCCU$) + P($CCUC$) + P(CUCC) + P(UCCC)

$= 4 \times 0{\cdot}0274 \times 0{\cdot}9726^3 = 0{\cdot}1008354753$

P(0 underweight) = $0{\cdot}9276^4 = 0{\cdot}7403600224$

$\Rightarrow$ the probability that fewer than two bars are underweight $= 0{\cdot}841$ to 3 S.F.

---

11 Context questions and answers

Accuracy

You are required to *give your answers to an appropriate degree of accuracy*.

There is no hard and fast rule for this, but the following guidelines should never let you down.

1. If stated in the question give the required degree of accuracy.
2. When using a calculator, give 3 S.F.
   unless finding $S_{xx,}$, S_{xy} etc. in which case you can give more figures – you should use *all* figures when finding the PMCC or the regression line coefficients.
3. Sometimes it is appropriate to give a mean to 1 or 2 D.P. rather than 3 S.F.
4. When using the tables and doing simple calculations (which do not *need* a calculator), you should give 4 D.P.

Statistical models

Question 1

(*a*) Explain briefly what you understand by

(i) a statistical experiment,

(ii) an event.

(*b*) State one advantage and one disadvantage of a statistical model.

Answer

(*a*) a test/investigation/process for collecting data to provide evidence to test a hypothesis.

A subset of possible outcomes of an experiment

(*b*) Quick, cheap can vary the parameters and predict

Does not replicate real world situation in every detail.

Question 2

Statistical models can be used to describe real world problems. Explain the process involved in the formulation of a statistical model.

Answer

Observe real world problem

Devise a statistical model and collect data

Compare observed against expected outcomes and test the model

Refine model if necessary

Question 3

(*a*) Write down two reasons for using statistical models.

(*b*) Give an example of a random variable that could be modelled by

(i) a normal distribution,

(ii) a discrete uniform distribution.

Answer

(*a*) To simplify a real world problem

To improve understanding / describe / analyse a real world problem

Quicker and cheaper than using real thing

To predict possible future outcomes

Refine model / change parameters possible *Any* 2

(*b*) (i) height, weight, etc. (ii) score on a face after rolling a fair die

---

Histograms

Question 1

Give a reason to justify the use of a histogram to represent these data.

Answer

The variable (minutes delayed) is continuous.

Averages

Question 1

Write down which of these averages, mean or median, you would recommend the company to use. Give a reason for your answer.

Answer

The median, because the data is skewed.

Question 2

State whether the newsagent should use the median and the inter-quartile range or the mean and the standard deviation to compare daily sales. Give a reason for your answer.

Answer

Median & IQR as the data is likely to be skewed

Question 3

Compare and contrast the attendance of these 2 groups of students.

Answer

Median 2nd group < Median 1st group;

Mode 1st group > Mode 2nd group;

2nd group had larger spread/IQR than 1st group

Only 1 student attends all classes in 2nd group

Question 4

Compare and contrast these two box plots.

Answer

Median of Northcliffe is greater than median of Seaview.

Upper quartiles are the same

IQR of Northcliffe is less than IQR of Seaview

Northcliffe positive skew, Seaview negative skew

Northcliffe symmetrical, Seaview positive skew (quartiles)

Range of Seaview greater than range of Northcliffe

any 3 acceptable comments

Skewness

Question 1

Comment on the skewness of the distribution of bags of crisps sold per day.
Justify your answer.

Answer

$Q_2 - Q_1 = 7$; $Q_3 - Q_2 = 11$; $Q_3 - Q_2 > Q_2 - Q_1$ so positive skew.

Question 2

Give two other reasons why these data are negatively skewed.

Answer

For negative skew; Mean < median < mode: 49·4 < 52 < 56

$Q_3 - Q_2 < Q_2 - Q_1$: 8 < 17

Question 3

Describe the skewness of the distribution. Give a reason for your answer.

Answer

No skew *or* slight negative skew.

$0.22 = Q_3 - Q_2 \approx Q_2 - Q_1 = 0.23$ *or* $0.22 = Q_3 - Q_2 < Q_2 - Q_1 = 0.23$

or mean (3.23) ≈ median (3.25), *or* mean (3.23) < median (3.25)

Correlation

Question 1

Give an interpretation of your PMCC (–0·976)

Answer

As height increases, temperature decreases **(must be in context).**

Question 2

Give an interpretation of this value, PMCC = –0·862.

Answer

As sales at one petrol station increases, sales at the other decrease **(must be in context)**.

Question 3

Give an interpretation of your correlation coefficient, 0·874.

Answer

Taller people tend to be more confident (**must be in context**).

Question 4

Comment on the assumption that height and weight are independent.

Answer

Evidence (in question) suggests height and weight are positively correlated / linked, therefore the assumption of independence is not sensible (**must be in context**).

Regression

Question 1

Suggest why the authority might be cautious about making a prediction of the reconditioning cost of an incinerator which had been operating for 4500 hours since its last reconditioning.

Answer 4500 is well outside the range of observed values, and there is no evidence that the model will apply.

Question 2

Give an interpretation of the slope, 0·9368, and the intercept, 19, of your regression line.

Answer

The slope, b – for every extra hour of practice on average 0·9368 fewer errors will be made

The intercept, a – without practice 19 errors will be made.

Question 3

Interpret the value of b (coefficient of x in regression line).

Answer

3 extra ice-creams are sold for every 1°C increase in temperature

Question 4

At 1 p.m. on a particular day, the highest temperature for 50 years was recorded. Give a reason why you should not use the regression equation to predict ice cream sales on that day.

Answer

Temperature is likely to be outside range of observed values.

Question 5

Interpret the value of a, (regression line)

Answer

Number of ……… sold if no money spent on advertising

Question 6

Give a reason to support fitting a regression model of the form $y = a + bx$ to these data.

Answer

Points on the scatter graph lie close to a straight line.

Question 7

Give an interpretation of the value of b.

Answer

A flight costs **£2.03 (or about £2)** for every extra **100km** or about **2p** per extra **km**.

Discrete uniform distribution

Question 1

A discrete random variable is such that each of its values is assumed to be equally likely.

(a) Write down the name of the distribution that could be used to model this random variable.

(b) Give an example of such a distribution.

(c) Comment on the assumption that each value is equally likely.

(d) Suggest how you might refine the model in part (a).

Answer

(a) Discrete uniform

(b) Example – Tossing a fair die /coin, drawing a card from a pack

(c) Useful in theory – allows problems to be modelled, but the assumption might not be true in practice

(d) Carry out an experiment to find the probabilities – which might not fit the model.

Normal distribution

Question 1

The random variable X is normally distributed with mean 177·0 and standard deviation 6·4.

It is suggested that X might be a suitable random variable to model the height, in cm, of adult males.

(*a*) Give two reasons why this is a sensible suggestion.

(*b*) Explain briefly why mathematical models can help to improve our understanding of real-world problems.

Answer

(*a*) Male heights cluster round a central height of approx 177/178 cm

Height is a continuous random variable.

Most male heights lie within $177 \pm 3\times 6.4$

(*b*) Simplifies real world problems

Enable us to gain some understanding of real world problems more quickly/cheaply.

Question 2

Explain why the normal distribution may not be suitable to model the number of minutes that motorists are delayed by these roadworks.

Answer For this data skewness is 3.9, whereas a normal distribution is symmetrical and has no skew.

Question 3

Describe two features of the Normal distribution

Answer

Bell shaped curve; symmetrical about the mean; 95% of data lies within 2 s.d. of mean; etc. (any 2).

Question 4

Give a reason to support the use of a normal distribution in this case.

Answer

Since **mean and median are similar** (or **equal** or **very close**), the distribution is (nearly) **symmetrical** and a **normal distribution may be suitable**.

Allow mean or median close to mode/modal class $\Rightarrow$), the distribution is (nearly) **symmetrical** and a **normal distribution may be suitable**.

---

12 Appendix

$-1 \leq$ P.M.C.C. ≤ 1

Cauchy-Schwartz inequality

Consider $(a_1^2 + a_2^2)(b_1^2 + b_2^2) - (a_1b_1 + a_2b_2)^2$

$= a_1^2 b_1^2 + a_1^2 b_2^2 + a_2^2 b_1^2 + a_2^2 b_2^2 - a_1^2 b_1^2 - 2a_1b_1a_2b_2 - a_2^2 b_2^2$

$= a_1^2 b_2^2 - 2a_1b_1a_2b_2 + a_2^2 b_1^2$

$= (a_1b_2 - a_2b_1)^2 \geq 0$

$\Rightarrow (a_1^2 + a_2^2)(b_1^2 + b_2^2) - (a_1b_1 + a_2b_2)^2 \geq 0$

$\Rightarrow (a_1b_1 + a_2b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2)$

This proof can be generalised to show that

$$(a_1b_1 + a_2b_2 + \ldots + a_nb_n)^2 \leq (a_1^2 + a_2^2 + \ldots + a_n^2)(b_1^2 + b_2^2 + \ldots + b_n^2)$$

or $$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

P.M.C.C. between –1 and +1

In the above proof, take $a_i = (x_i - \bar{x})$, and $b_i = (y_i - \bar{y})$

$\Rightarrow S_{xy} = \Sigma\,(x_i - \bar{x})(y_i - \bar{y}) = \Sigma\, a_ib_i$

$S_{xx} = \Sigma\,(x_i - \bar{x})^2 = \Sigma\, a_i^2$ and $S_{yy} = \Sigma\,(y_i - \bar{y})^2 = \Sigma\, b_i^2$

P.M.C.C. $= r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

$\Rightarrow r^2 = \dfrac{S_{xy}^2}{S_{xx}S_{yy}} = \dfrac{(\sum a_ib_i)^2}{(\sum a_i^2)(\sum b_i^2)} \leq 1$ using the Cauchy-Schwartz inequality

$\Rightarrow \quad -1 \le r \le +1$

Regression line and coding

The regression line of y on x has equation $y = a + bx$, where $b = \frac{S_{xy}}{S_{xx}}$, and $a = \bar{y} - b\bar{x}$.

Using the coding $x = hX + m$, $y = gY + n$, the regression line for Y on X is found by writing $gY + n$ instead of y, and $hX + m$ instead of x in the equation of the regression line of y on x,

$\Rightarrow \quad gY + n = a + b(hX + m)$

$\Leftrightarrow \quad Y = \frac{a+bm-n}{g} + \frac{h}{g} bX \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad$ equation **I**.

Proof

$\bar{x} = h\bar{X} + m$, and $\bar{y} = g\bar{Y} + n$

$\Rightarrow \quad (x - \bar{x}) = (hX + m) - (h\bar{X} + m) = (hX - h\bar{X})$, and similarly $(y - \bar{y}) = (gY - g\bar{Y})$.

Let the regression line of y on x be $y = a + bx$, and
let the regression line of Y on X be $Y = \alpha + \beta X$.

Then $b = \frac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\bar{x}$

also $\beta = \frac{S_{XY}}{S_{XX}}$ and $\alpha = \bar{Y} - \beta\bar{X}$.

$$b = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum(hX-h\bar{X})(gY-g\bar{Y})}{\sum(hX-h\bar{X})^2} = \frac{hg\sum(X-\bar{X})(Y-\bar{Y})}{h^2\sum(X-\bar{X})^2} = \frac{g\sum(X-\bar{X})(Y-\bar{Y})}{h\sum(X-\bar{X})^2}$$

$$\Rightarrow \quad b = \frac{g}{h}\beta$$

$$\Rightarrow \quad \beta = \frac{hb}{g}$$

$$\alpha = \bar{Y} - \beta\bar{X} = \frac{\bar{y}-n}{g} - \frac{hb}{g}\left(\frac{\bar{x}-m}{h}\right) = \frac{\bar{y}-b\bar{x}+bm-n}{g} = \frac{a+bm-n}{g} \qquad \text{since } a = \bar{y} - b\bar{x}$$

and so

$$Y = \alpha + \beta X \Leftrightarrow Y = \frac{a+bm-n}{g} + \frac{h}{g} bX$$

which is the same as equation **I**.

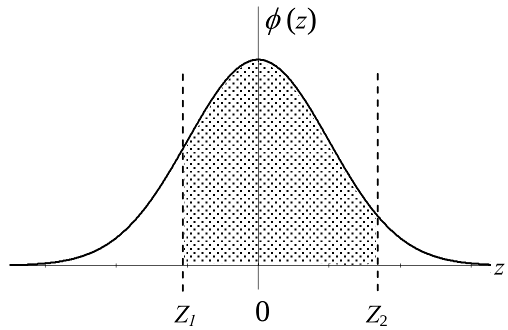
Normal Distribution, $Z = \frac{X-\mu}{\sigma}$

The standard normal distribution with mean 0 and standard deviation 1 has equation

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

The normal distribution tables allow us to find the area between Z_1 and Z_2.

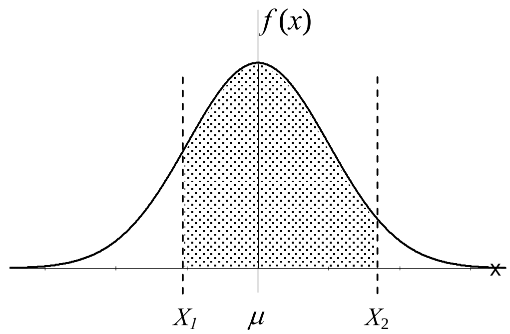
$$= \int_{Z_1}^{Z_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$



The normal distribution with mean μ and standard deviation σ has equation

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

$$P(X_1 \le X \le X_2) = \int_{X_1}^{X_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$



Using the substitution $z = \frac{x-\mu}{\sigma}$

$dz = \frac{1}{\sigma}\, dx$, $Z_1 = \frac{X_1-\mu}{\sigma}$ and $Z_2 = \frac{X_2-\mu}{\sigma}$

$$\Rightarrow P(X_1 \le X \le X_2) = \int_{Z_1}^{Z_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$

= the area under the standard normal curve, which we can find from the tables using

$Z_1 = \frac{X_1-\mu}{\sigma}$ and $Z_2 = \frac{X_2-\mu}{\sigma}$.

Thus $P(X_1 \le X \le X_2) = P(Z_1 \le Z \le Z_2) = \Phi(Z_2) - \Phi(Z_1)$

Index