Name: Rajesh Swaminathan
Block: D
Date: 7-Jun-2004

Math Methods 11

Portfolio Assignment 2

TRANSFORMING DATA

1. Mean and Standard Deviation of Raw Data

177	175	137	155	150	166
132	146	179	140	169	177
141	148	130	176	135	130
157	172	178	143	143	136
132	166	130	151	145	178
131	171	160	140	179	166
145	142	177	176	132	135
164	179	161	145	134	179
139	149	135	142	172	148
159	160	137	130	130	164

Mean	=AVERAGE(A1:F10)	152.92
STD	=STDEVP(A1:F10)	17.09

2. (a)

182	180	142	160	155	171
137	151	184	145	174	182
146	153	135	181	140	135
162	177	183	148	148	141
137	171	135	156	150	183
136	176	165	145	184	171
150	147	182	181	137	140
169	184	166	150	139	184
144	154	140	147	177	153
164	165	142	135	135	169

Mean	=AVERAGE(A1:F10)	157.92
STD	=STDEVP(A1:F10)	17.09

5cm Added to Each Height

(b)

165	163	125	143	138	154
120	134	167	128	157	165
129	136	118	164	123	118
145	160	166	131	131	124
120	154	118	139	133	166
119	159	148	128	167	154
133	130	165	164	120	123
152	167	149	133	122	167
127	137	123	130	160	136
147	148	125	118	118	152

Mean	=AVERAGE(A1:F10)	140.92
STD	=STDEVP(A1:F10)	17.09

12cm Subtracted from Each Height

In 2 (a), 5 cm was added to each of the heights in the data. In doing so, the mean was increased by 5 cm, but the standard deviation remained the same. The same thing

happened in 2 (b), where 12 cm was subtracted from each of the heights. The mean was reduced by 12 cm, and the standard deviation stayed the same.

This makes perfect sense because when a constant value is added to each of the elements, the entire data-set is moved up a bit. This is easy to visualize if each height is positioned on a number line of its own. Adding a constant value causes the mean to move too, as the mean can be thought as the mid-point of the data on the number line. This can be show mathematically:

$\mu_1 = \frac{x_1 + x_2 + x_3 + ...}{n}$, where *n* is the number of students.

When adding *a* to each height:

$$\mu_2 = \frac{(x_1 + a) + (x_2 + a) + (x_3 + a) + ...}{n}$$
$$\mu_2 = \frac{(na) + x_1 + x_2 + x_3 + ...}{n}$$
$$\mu_2 = \frac{na}{n} + \frac{x_1 + x_2 + x_3 + ...}{n}$$
$$\mu_2 = \mu_1 + a$$

However, the relative distance of any value from the mean, or $(x - \mu)$, does not change, and therefore, the standard deviation doesn't change either.

3. (a)

88.5	87.5	68.5	77.5	75	83
66	73	89.5	70	84.5	88.5
70.5	74	65	88	67.5	65
78.5	86	89	71.5	71.5	68
66	83	65	75.5	72.5	89
65.5	85.5	80	70	89.5	83
72.5	71	88.5	88	66	67.5
82	89.5	80.5	72.5	67	89.5
69.5	74.5	67.5	71	86	74
79.5	80	68.5	65	65	82

Mean	=AVERAGE(A1:F10)	76.46
STD	=STDEVP(A1:F10)	8.54

Each Height Multiplied by 5

(b)

35.4	35	27.4	31	30	33.2
26.4	29.2	35.8	28	33.8	35.4
28.2	29.6	26	35.2	27	26
31.4	34.4	35.6	28.6	28.6	27.2
26.4	33.2	26	30.2	29	35.6
26.2	34.2	32	28	35.8	33.2
29	28.4	35.4	35.2	26.4	27
32.8	35.8	32.2	29	26.8	35.8
27.8	29.8	27	28.4	34.4	29.6
31.8	32	27.4	26	26	32.8

Mean	=AVERAGE(A1:F10)	30.58
STD	=STDEVP(A1:F10)	3.42

Each Height Multiplied by 0.2

Observation: In multiplying each height by 5 and 0.2, both the mean and the standard deviation were multiplied by 5 and 0.2 respectively.

Explanation: Like the previous question, multiplying each height by a constant value shifts the mean because each height "jumps" up by a constant factor. Mathematically,

$$\mu_1 = \frac{x_1 + x_2 + x_3 + \ldots}{n}$$

When each height is multiplied by *a*:

$$\mu_2 = \frac{(ax_1) + (ax_2) + (ax_3) + \ldots}{n}$$

$\mu_2 = a \times \mu_1$, when *a* is factored out.

But this time, the relative distances of each height from the mean, or $(x - \mu)$, is also increased by the same factor, and so there is an increase in the standard deviation too.

$$\sigma_1 = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \ldots}{n}}$$

$$\sigma_2 = \sqrt{\frac{(ax_1 - a\mu)^2 + (ax_2 - a\mu)^2 + (ax_3 - a\mu)^2 + \ldots}{n}}$$, when each height is multiplied by *a*.

$$\sigma_2 = \sqrt{\frac{a^2(x_1 - \mu)^2 + a^2(x_2 - \mu)^2 + a^2(x_3 - \mu)^2 + \ldots}{n}}$$

$$\sigma_2 = \sqrt{a^2\left[\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + \ldots}{n}\right]}$$

$\sigma_2 = a \times \sigma_1$

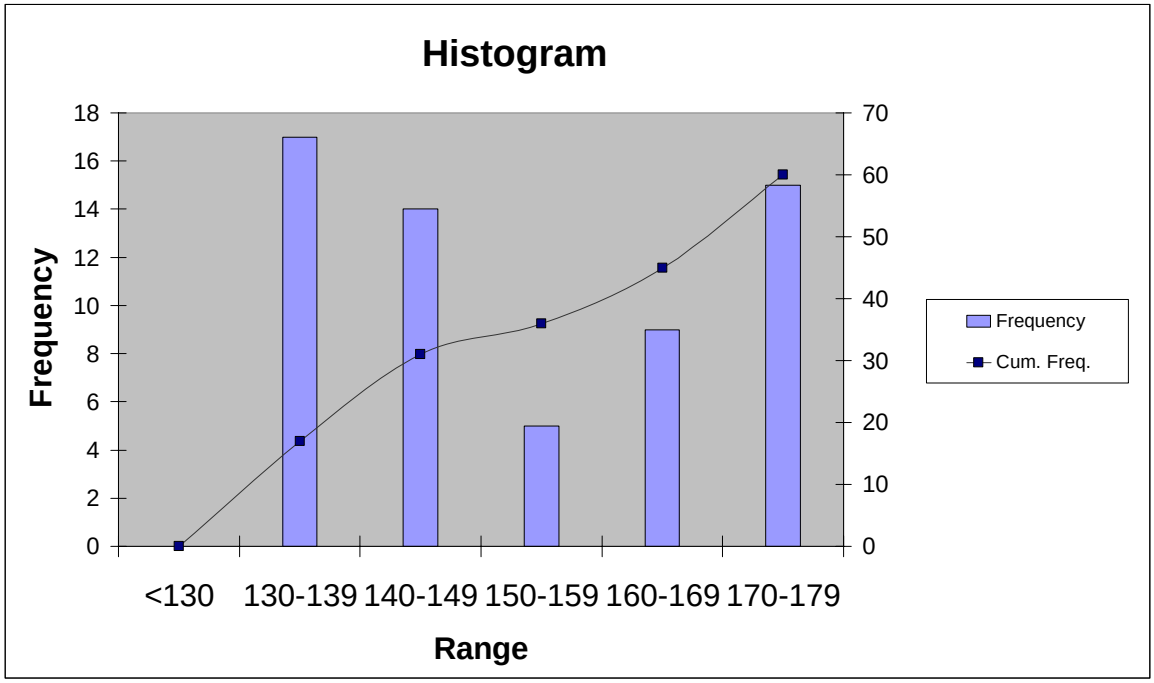
However, when $a < 0$, there is small change in that the mean is multiplied by *a*, but the standard deviation is multiplied by a factor of $|a|$. This is because we assume that the

standard deviation is always positive (deviation is thought as a distance from the mean), and therefore, when take the square root of a^2, we take only the positive root, and discard the negative root. Thus if all the heights are multiplied by -5, then the mean will be multiplied by -5, but the standard deviation will be multiplied by 5, and *not* -5.

For example, if all of the marks in a certain class are scaled by a certain factor, both the mean and the standard deviation will be scaled by the same factor.

4.

Bin	*Range*	*Frequency*	*Cum. Freq.*	*Cumulative %*
129	<130	0	0	0.00%
139	130-139	17	17	28.33%
149	140-149	14	31	51.67%
159	150-159	5	36	60.00%
169	160-169	9	45	75.00%



Frequency Distribution and Cumulative Frequency Ogive

Note: Since the use of technology is encouraged, the above table and graphs have all been generated with Microsoft Excel, and **no** statistical analysis has been done manually.

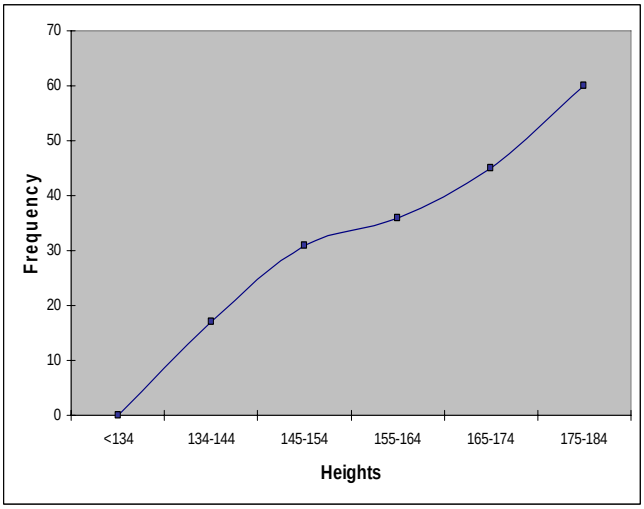
To find the median, we average out the 30th and the 31st height. From the graph,

$$\text{Median} = \frac{148+149}{2} = 148.5 \text{ cm}$$

Inter-quartile Range = Upper range – Lower Range
= 169.5 – 148.5
= 32.5

Note: The Median and the Inter-quartile ranges have been found by looking at the graph. These values have also been confirmed using Microsoft Excel's MEDIAN() and QUARTILE() functions.

5. (a) When 5 cm is added to each height, the ogive looks like this:



Ogive after 5cm is added to each height

$$\text{Median} = \frac{153+154}{2} = 153.5$$

Inter-quartile Range = 169.5 – 137
= 32.5

When we added 5cm to each height, the median was also moved up by 5cm, while the inter-quartile range remained the same.

(b) When 12cm is subtracted from each height:

$$\text{Median} = \frac{136+137}{2} = 136.5$$

Inter-quartile range = 169.5 – 137 = 32.5

When 12cm was subtracted from each height, the median was reduced by 12, but the inter-quartile range stayed the same.

We can thus generalize that adding *a* to each score in any set of data, increases the median by the same amount *a* while the inter-quartile range stays the same.

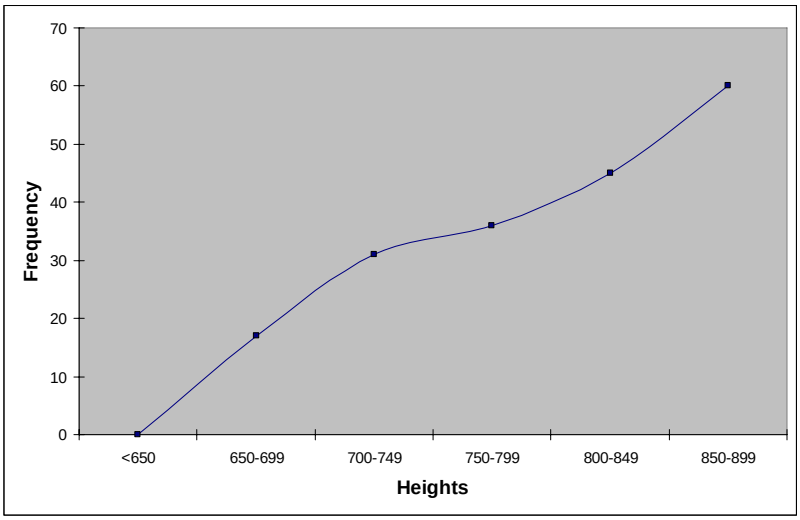
This makes sense because the median is the mid-value, and the mid-value in this case is always between the 30^{th} and the 31^{st} height. When all the heights are increased by *a*, this average will also increase by an equal amount *a*, and thus the median is also incremented by *a*. The upper and lower quartiles (the 45^{th} and the 15^{th} height respectively in this case) are also incremented by *a*, but their *difference* or the range still stays the same. Thus, there is no change to the inter-quartile range.

6. (a) Ogive obtained when each height is multiplied by 5:

$$\text{Median} = \frac{742+743}{2} = 742.5 \text{ cm}$$

Inter-quartile range = 169.5 – 137 = 32.5

Observation: The median has been increased by a factor of 5, and so has the inter-quartile range.



(b) When each of the heights is multiplied by 0.2,

Median = 29.7 cm
Inter-quartile range = 169.5 – 137 = 32.5

Observation: The median, as expected, has been multiplied by 0.2, and the inter-quartile range has also been multiplied by 0.2.

Thus when each score is multiplied by *a*, both the median and the inter-quartile range are multiplied by the same factor *a*.

However, when $a < 0$, the median is increased by *a*, but the inter-quartile range is scaled by a factor of $|a|$.

7. The results from questions 5 and 6 are summarised in the following table:

	Raw	Add 5	Subtract 12	Multiply 5	Multiply 0.2	Multiply -5
Median	148.5	153.5	136.5	742.5	29.7	-742.5
1st quartile	137	142	125	685	27.4	-847.5
3rd quartile	169.5	174.5	157.5	847.5	33.9	-685
Inter-quartile	32.5	32.5	32.5	162.5	6.5	162.5

Not surprisingly, the 1st and the 3rd quartiles also change in the same way as does the mean does, with one exception. The quartile values are different when each member of a distribution is multiplied by a negative value. In fact, they just switch in addition to being multiplied by -1, ie. the 1st quartile becomes the 3rd quartile, and vice versa. Also this is

another reason as to why the inter-quartile range is multiplied by $|a|$, when the values in the distribution are multiplied by *a*.

A point worth noting is the way changes on a distribution affect the median and the inter-quartile range exactly in the same way as the change would affect the mean and standard deviation of the distribution. When *a* is added, it affects only the mean and the median, but when *a* is multiplied, if affects all 4 – the mean, the median, the standard deviation and the inter-quartile ranges.

For example, if all of the marks in a certain class are scaled by a certain value, both the median and the inter-quartile range of the class will change.

8. (a) As we've seen in Question 2, if all the values in the data are decreased by a constant value *a*, then the mean is also decreased by the same value *a*. The original raw data had a mean of 152.95. In order to transform the data so that it has a mean of 0, all the heights need to be reduced by 152.95.

   (b) Adding or subtracting a value *a* from all the heights does not affect the standard deviation of the set. However, multiplying all of the heights, does affect the standard deviation of the set. By multiplying each score by *a*, the standard deviation can be multiplied *a*. So in order to obtain data with a standard deviation of 1, we should then divide each of the heights by the original standard deviation of 17.09.

   (c) In order to transform the given data so that it has a mean of 0 *and* a standard deviation of 1, we need to recall how adding and multiplying values affects the mean and standard deviation.

   Adding a value of a distribution affects only the mean but not the standard deviation (STD). But on the other hand, multiplying a distribution by a specific value affects both. There are thus two approaches that can be taken to perform the task:

   i. We can divide all heights by the standard deviation, and obtain the mean of this new data set (which now has a STD of 1). We then now subtract the mean of this new data from each of the values of the new data set. The final set thus obtained will have a mean of 0, and a standard deviation of 1.
   ii. Alternatively, we could subtract the mean of the data set from each of the heights (this won't affect the STD). The new data set obtained will have a mean of 0. Now we divide each of the values of this new data set by the standard deviation of the either the old data set or the new data set (they have the same values), and the final data set thus obtained will have a mean and standard deviation of 1 and 0 respectively.

   Incidentally in the second method above, we are calculating nothing but the *Z-Scores* of our distribution, which follows the formula, $Z = \frac{x - \mu}{\sigma}$. In doing so, we have converted our distribution into a *normal* distribution. For a normal distribution, the mean is always 0 and the standard deviation always 1.

Although the first method *seems* to be a different method, it can be shown to be following the same process as the second one.

In the first step, we obtained the new height x_2 by the formula

$x_2 = \frac{x}{\sigma} - \mu_2$, where μ_2 is the mean of the distribution obtained after dividing each of the heights by σ. But μ_2 is nothing but $\frac{\mu}{\sigma}$, because in dividing all values by σ, we even divided the mean of the original distribution by σ. So,

$x_2 = \frac{x}{\sigma} - \mu_2$

$x_2 = \frac{x}{\sigma} - \frac{\mu}{\sigma}$ (substitute $\mu_2 = \frac{\mu}{\sigma}$)

$\therefore x_2 = \frac{x - \mu}{\sigma}$, which is exactly what we've done in the second method. The two methods thus transform the data in the same way.