## Suggested Exercises from M&M Chapter 1 *[Homegrown exercises begin on next page]*

*This first page was updated on September 3*

To start with, do some of the odd-numbered exercises. answers to all odd-numbered exercises are given on textbook pages S-1 onwards.

Do some or all of the following even-numbered exercises. You are asked to hand in answers to designated ones.. see the list, and the deadline, on the main course page. Some of these will be discussed in tutorials or answers to them posted on the course web page

Section 1.1
(beginning p 23)

1.10
1.14
1.16,
1.22
1.24
1.31
1.33 [see NB1, NB2]
1.35
1.36

Section 1.2
(beginning p 58)

1.42 1.46
1.50 1.52
1.54 1.56
1.58 1.60
1.62 1.66

Section 1.2
(beginning p 85)

1.70 1.72
1.74 1.76
1.78 1.80
1.82 1.84
1.86 1.88
1.96

[NB1, 1.33] This exercise, based on the overall distribution shown at the top of page 36, is helpful, since it deals with **class intervals of unequal width**. A similar situation is found in the blood lead levels example (homegrown exercise -1- on next page) where the authors use a *wider first interval*. Another *important for epidemiology* example of this is the use of the 1-year wide age category 0-1, the 4-year wide category 1-5, and the 5-year categories 5-9, 10-14, etc. in tabulating age distributions of entire populations. It is not easy to get a spreadsheet to make a correct histogram of such a table: one must get it to draw rectangles with bases of different widths, and adjust the heights so that the *area* is proportional to frequency.

[NB2, 1.33] In exercise 1.33, the authors updated the $$ salary categories from the 2nd edition, but forgot to match the table with the corresponding text: the text says a salary of $20,000 belongs in the second class; if my understanding of their convention of including the leftmost endpoint but not the rightmost is correct, then their example salary of $20,000 belongs in the *third* class i.e. 20K-25K, not in the *second*.

These exercises, from the overall "**Chapter 1 Exercises**" that begin on page 93, are quite helpful, since their placement does not indicate exactly which section of the chapter they relate to

1.106 1.107 1.108 1.110 1.112
1.113 1.114 1.120 1.122

### Note: calculating SD (or variance) from grouped data

The mean and SD functions in some calculators can handle such data. The variance/SD functions in spreadsheets can not: they assume each value in the specified range occurs with a frequency of 1. One way around this is to implement the long (definitional) formula for variance (see M&M p 51)

Grouped data consist of a column of x values (or midpoints of categories, if the data are continuous rather than discrete) and a column of frequencies]

1. Calculate the mean

* Make a new column, containing the product of each x value multiplied by the number of times it occurs (the *frequency*)

* Divide the sum of these products by the sum of the *frequencies* ("*n*")

2. Calculate the sum of the squared deviations from the mean

* Make a new column, containing the deviation of each value [or midpoint] from the mean. Make another with the squares of these deviations.

* Make yet another new column, containing the product of each squared deviation multiplied by the number of times it occurs (the *frequency*)

* Obtain the sum of these products

3. Calculate the variance, i.e., the "average" squared deviation

* Divide the sum of the products by *n* - 1. For SD, take square root.

Programming these extra steps in a spreadsheet helps show what the variance is. There are shorter ways to *calculate* it, that avoid having to first calculate the mean, but they do not really show the *concept* of variance.

## "Homegrown" Exercises around M&M Chapter 1

### -1- Maternal Lead Levels after Alterations to Water Supply"

p203-204 The Lancet July 25, 1981

Sir,—From 1970 to 1980, studies have shown that soft, plumbosolvent water supplies in older houses with lead-soldered water pipes and lead-lined storage tanks will result in a significant uptake of lead by people in these areas. The Lawther report has drawn attention to this, but the idea was first put forward in 1844 by Professor Christison in Edinburgh. Plumbosolvency can be greatly reduced by increasing water pH and hardness, and we have confirmed this in Glasgow. The greatest effect is achieved by a change in pH, which may be accomplished by addition of calcium hydroxide, as lime or other alkaline materials, to the water. We report here on the effects of such a change in plumbosolvency in Glasgow upon the blood lead concentrations of a specific population.

We have noticed that the pH of tap-water in Glasgow was about 6.3 and that the median blood lead concentrations for a number of different populations in Glasgow were consistently around 0.8-0.9 μmol/l. In a comparative study we examined 236 mothers in the postnatal wards of Stobhill Hospital in the autumn and winter of 1977. The women were aged from 17-37 and none had a history of industrial lead exposure. Blood and domestic tap-water samples were collected and analysed as previously described. Blood lead analysis, by flameless atomic absorption spectrophotometry, was checked by quality control and by a 10% interchange of samples with another laboratory. The distribution of blood lead values was skewed. The median value for the maternal blood lead distribution was 0.8 μmol/l and 3% of the maternal bloodlead levels were over 2 μmol/l. This gave a geometric mean maternal blood lead of 0.7 μmol/l (see table). We concurrently studied mothers with 6 week-old children as a confirmatory test, and found that median blood lead was also 0.8 μmol/l and that the geometric mean blood lead was 0.79 μmol/l. In both studies we found a curvilinear relationship between blood lead and water lead, with the log of the maternal blood lead level varying as the cube root of both first flush and running water lead concentrations.

DISTRIBUTION OF BLOOD LEAD CONCENTRATIONS

| Blood lead | **1977** study | **1980** study |
|---|---|---|
| range μmol/l | No.(%) of | No.(%) |
| | 236 mothers | of 475 mothers |
| 0–0.25 | 30(12.7) | 90(18.9) |
| 0.26–0.45 | 53(22.5) | 225(47.5) |
| 0.46–0.65 | 29(12.3) | 90(18.9) |
| 0.66–0.85 | 39(16.8) | 44( 9.3) |
| 0.86–1.05 | 31(13.1) | 19( 4.0) |
| 1.06–1.25 | 17( 7.2) | 3( 0.6) |
| 1.26–1.45 | 16( 6.8) | 2( 0.4) |
| 1.46–1.65 | 7( 3.0) | 2( 0.4) |
| 1.66 or more | 14( 5.9) | 0 |

In April, 1978, after the introduction of an automatic, closed-loop lime-dosing system, the pH of the Loch Katrine supply to 0.92 million consumers in Glasgow was raised from 6.3 to 7.8 and maintained at this level. Before lime-dosing more than 50% of random daytime water samples taken within the city had lead concentrations in excess of 0.48 μmol/l (100μg/l), the WHO standard. After lime addition, 80% of random daytime samples contained less than 0.48 μmol/l. We found, however, that the pH was not maintained in the water distribution network since many samples had a pH of 7.0 or less and previous experience and studies at the Water Research Centre indicated an optimum pH of 8.5 for minimum plumbosolvency. Accordingly, in August, 1980, the pH of Glasgow's water supply was raised to 9, which maintained a pH of over 8 at the tap. We then estimated that more than 95% of random daytime samples taken from the Loch Katrine supply area would give water lead concentrations of below 0.48 μmol/l. These figures are consistent with the calculated relative plumbosolvencies of the water and with the changes in individual plumbing systems over this time.

...

## -1- Maternal Lead Levels after Alterations to Water Supply"

p203-204 The Lancet July 25, 1981

We later examined another population of 475 mothers in the postnatal wards of Stobhill Hospital in the autumn and winter of 1980. Their median blood level was 0.32 and their geometric mean blood level was 0.39 µmol/l (see table). These two studies, in similar populations from the same catchment area, show that there has been a dramatic drop in mean blood lead concentration as a result of water treatment since we can identify no other exposure source which could have altered in this time, although change in housing stock either by demolition or by rehabilitation will have helped this downward trend. 7.5% of mothers in 1977 had blood lead levels in excess of 1.5 µmol/l but only 0.4% of the equivalent group of women had blood lead levels in excess of 1.5 µmol/l in 1980. These studies were unlikely to have been affected by variables such as fluctuation in maternal blood lead during pregnancy or annual variations in population blood lead since all of this work was carried out in women at the same stage of pregnancy and at the same time of year. Our findings are of importance since much emphasis has been put on limiting lead exposure in pregnant mothers because of the deleterious effects of high maternal blood lead levels on the mental development of the fetus and neonate.

Exposure to lead in water results not only from drinking the water but also from the use of such water in food preparation and the resultant absorption of lead onto food during cooking. Emphasis has been laid upon the removal from petrol of alkyl leads used as antiknocking agents. Yet, in Germany where lead concentrations in petrol were reduced in 1976 from 0.4 to 0.15g/l, producing an 80% reduction in vehicle lead emission, it was found that the mean reduction in blood lead levels in inhabitants of three cities ranged between 0.01 and 0.13 µmol/l, a drop of about 10%. Despite the drop in air lead, blood lead levels increased or remained unchanged in 40% of men and 50% of women, figures consistent with those theoretically calculated in the Lawther Report (16%), but much less than those given by the Conservation Society(32-69%) as the likely contribution of inhaled lead to total body lead. In the U.S.A. where a significant hazard exists from lead-containing paints, similar findings have been made. Although a link may be found between faecal lead excretion in children and their degree of exposure to lead paint, no association between faecal lead and traffic density could be found. This means that lead uptake from food stuffs, including water, is the primary contributor to total lead intake and implies that the importance of petrol lead emission in terms of over-exposure to lead must be secondary.

The relationship of blood lead level to IQ is of great importance in the neonate. One view on the subject is that IQ is affected continuously from low levels of lead exposure, and that a measurable deficit in intelligence will be reached by the time blood lead concentrations are as high as 1.5 µmol/l. The alternative view is that no IQ deficit will be seen until some breakpoint (e.g., 1.5 µmol/l) is reached, at which time the decreases in IQ will be progressive.

Our study shows that lead in water, as an obvious source of lead exposure, may be diminished by water treatment which can reduce the percentage of over-exposed fetuses (maternal blood lead concentration > 1.5 µmol/l) by 94%.

### EXERCISE

a What blood lead concentrations fall into the different tabulated intervals used to construct the two frequency distributions? (if you need to, see Colton pp 19-20* on the question of "true" intervals) What are the midpoints of these true intervals ?

*More modern books, that assume use of computers rather than manual methods, do not seem to fuss about such issues -- there is just a passing comment -- in an exercise -- at bottom of page 35 of M&M.*

** Excerpts from Colton are under Resources on the web page.*

b Sketch histograms for the 1977 and 1980 distributions, putting them both on the same graph in a way that allows them to be easily compared. *Use plain ruled paper (or, make your own -- don't go out and buy graph paper!) . See note on previous page re grouped data with unequal intervals. And, assume, arbitrarily, that the last category ends at 2.00. You may find this part easier to do by hand than by computer!!*

c Paraphrase the statement "the median value was 0.8µmol/l" (paragraph 2).

d Calculate the (arithmetic) mean and standard deviation of the grouped 1977 data (use a semi-arbitrary "midpoint" for the 14 values in the last interval). Is the SD very useful? What might be a better summary?

## -2- Premature Death in Jazz Musicians: Fact or Fiction?

*A letter to Am J Public Health 1991 Jun;81(6):804-5*

"Jazz musicians tend to be more liable than other professions to die early deaths from drink, drugs, women, or overwork."[1]

"The career of the ODJB (Original Dixieland Jazz Band) was both as fantastic and as typical as any that jazz has had to offer. Its story features... the petty jealousies, alcoholism, premature deaths, and all the rest."[2]

"Catlett's career was a singularly queer one, even for jazz, whose history is filled with the wreckage of poverty, sudden obscurity, and premature death."[3]

Statistical study of 86 jazz musicians listed in a university syllabus refutes these tenets,[4] the second and third of which were made by two of America's most respected critics, and all of which foster the commonly held view that jazz players die prematurely.

Dates of birth, and of death when it had occurred, were tabulated, and longevity matched with that expected in the United States by year of birth, race, and sex.[5-7] One musician who had not reached the age of his life expectancy was excluded from the list; the musicians were born in the US.

Birth years ranged from 1862 to 1938; 16 births occurred before 1900, 23 between 1900 and 1909, 19 between 1910 and 1919, 22 between 1920 and 1929, and five between 1930 and 1939. Comparison with national values showed that 70 (82%) of the musicians exceeded their life expectacy; four-fifths of the Black men, three fourths of the White men, and all the women lived longer than expected as shown in this frequency distribution.

| | Male | | | Female | | |
|---|---|---|---|---|---|---|
| | Total | n | % | Total | n | % |
| White | 19 | 14 | 74 | - | - | - |
| Black | 59 | 49 | 83 | 7 | 7 | 100 |

Jazz was born in the "sporting houses" of New Orleans and nurtured in the speakeasies and night clubs of Chicago, Kansas City, and New York. Its association with vice and crime in its early days has led to the assumption that to play jazz is to court depravity and death. Although the size and sex distribution of the sample limits the inferences to be drawn, the data suggest that jazz musicians do not die young. Most of the 85 musicians in this study have survived the potential hazards of irregular hours of work and meals, the ready temptation of drugs and alcohol, and the perils of racial prejudice, and to have overcome "the problem of the artist who is creative within a socially and racially discrnniminatory world."[8]

References (1) Lindsay M: Teach Yourself Jazz. London: English Universities Press, 1958:(2) Schuller G: Early jazz. Its Roots and Musical Development. New York: Oxford University Press, 1968:176. (3) Balliett W: The Sound of Surprise. New York: Da Capo Press, 1918;144. (4) Norton P, Schumacher HJ: Topics in American Culture: Jazz Styles. Ann Arbor, MI: University of Michigan Extension Service, 1978;xiv-Xvi. (5) Chilton J: Who's Who of Jazz. Philadelphia:Chilton Book Co. 1972. (6) Feather L: The Encyclopedia of Jazz. New Ed. Bonanza Books, 1960. (7) US Department of Commerce: Historical Statistics of the United States. Colonial Times to 1957. Washington, DC: Govt Print ing Olfice, 1461;24-25. (8) Berendt J: The Jazz Book. New York: Lawrcnce HilI, 1915:256.

### EXERCISE

The design & analysis contain several epidemiologic and statistical defects. Identify and comment on the most important ones.

## -3- Clinical Research in General Medical Journals: A 30-Year Perspective

Fletcher et al., NEJM 301:180-183, 1979

Fletcher et al. studied the characteristics of 612 randomly selected articles published in the NEJM, JAMA and Lancet since 1946. Two of the attributes they examined was the number of authors per article and the number of subjects studied in each article; they found:

| | No. articles | No. authors | | No. subjects |
|---|---|---|---|---|
| Year | examined | Mean | (SD) | Median |
| 1946 | 151 | 2.0 | (1.4) | 25 |
| 1956 | 149 | 2.3 | (1.6) | 36 |
| 1966 | 157 | 2.8 | (1.2) | 16 |
| 1976 | 155 | **4.9** | (**7.3**) | 30 |

**EXERCISE (see also exercise -12- )**

a Sketch a suitable graphic representation of the trends.

b From the mean and SD, roughly reconstruct the actual frequency distribution of the number of authors per article for 1946 [it may save you time if you use spreadsheet on course web page under Resources]

c Why report median (rather than mean) no. of subjects per study ?

d Can '76 SD=**7.3** really be larger than the mean=**4.9**? Explain.

## -4- from the U.K. Media...

*The Independent*

The usually wonderful Jeremy Paxman, introducing a Newsnight discussion last Friday on the teaching of reading skills, expressed dismay that 'a third of our primary schoolchildren have below-average reading ability'. Had he paid more attention in his 'rithmetic lessons, perhaps Paxman would have realised that half our schoolchildren are below average in everything. As, indeed. are half our Newsnight presenters.

**EXERCISE** Set *The Independent* straight

## -5- Statistics in the U.K. Parliament

*Hansard 29 November 1991*

*Mr Arbuthnot*: the Labour party's suggestion of a minimum wage is in itself rather obscure and bizarre. As I understand it, it is tied to the average and would therefore not only be relatively high at £3.40 but would increase as the average wage itself increased. With each increase in the average rate of pay, the minimum wage itself would have to go up and it would be forever chasing its own tail.

*Mr Tony Lloyd:* Perhaps I can help the hon. Gentleman. It will be tied to the median, which is not the same as the average. It is simply a mid-point on the range and would not be affected by changes in the minimum wage.

*Mr Arbuthnot*: From what I understand, even an amount tied to the median would be affected because if the lowest wage were increased to £3.40 per hour, the median would have to rise.

*Mr Tony Lloyd*: I shall put the matter in simple terms. The median, the mid-point in a series of numbers such as 2.2, 5.6 and 7, is defined as being the difference between 2 and 7, which is 3.5. If we alter the figures 2 and 2 to 3.5, the middle figure of 5 would remain unaltered because it is independent of the bottom figures.

*Mr Arbuthnot*: I do not understand the hon. Gentleman's mathematics and I slightly doubt whether he does.

*Mr Matthew Carrington (Fulham)*: I am extremely confused. I studied mathematics for some years at school and I have not totally forgotten all of them. The median is not the mid-point between the first number and the last. It is where the largest number of items in a sample comes to, whereas the average is obviously the sample multiplied by the number of items. The hon. Member for Stretford (Mr Lloyd) is obviously extremely confused. The median has a precise mathematical definition which is absolutely right, and my hon. Friend is correct in saying that the median is bound to alter if the number at the bottom on the scale is changed. That will alter the average as well in a different way, but it is bound to alter the median. Perhaps the hon. Member for Stretford wishes to define median in a non mathematical sense.

*Mr Arbuthnot*. I am grateful to my hon. Friend for sorting out at least the hon. Gentleman's mathematics with obvious skill and knowledge.

**EXERCISE** Correct the honourable Gentlemen

### -6- various representations of temporal changes in rate of deaths from cancer

from 1st edition of Moore and McCabe

The rate of deaths from cancer in the United States has increased since 1930 as follows:

Cancer deaths per 100,000 population

| Year | 1930 | '35 | '40 | '45 | '50 | '55 | '60 | '65 | '70' | '75 | '80 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rate | 97 | 108 | 120 | 134 | 140 | 147 | 149 | 154 | 163 | 170 | 184 |

a Draw a line graph of these data designed to emphasize the rise in death rates. (Imagine you are trying to persuade Congress to appropriate more money to fight cancer.)

b Draw another line graph of the same data designed to show only a moderate increase in the death rate.

### -7- The variability of young children's energy intake

Birch LL. et al. NEJM 324(4):232-5, 1991 Jan 24

a Who is more variable? a younger child who consumes 887, 672, 757, 867, 899 and 872 calories on 6 observed days or an older child who consumes 1155, 1193, 1167, 1315, 1401 and 1133?

b What is the mean and SD for the younger child if energy is recorded in kJ? (92 Cal = 380 kJ, according to a table on a food package)

### -8- knowing your way around the Gaussian distribution...

[q. from early edition of Armitage]

The iodine level of a tin of salt is stated to be between 433µg and 753 µg. Assuming that the iodine content is a normally distributed random variable and that it lies within the given limits with probability 0.94 and below the lower limit with probability 0.01, find the probability that the iodine content exceeds:

(a) 500 µg (b) 700 µg

### -9- Heights and Weights of Alberta children

from "Alberta Study" by Spitzer et al. in mid 1980-s

The following summaries were derived from random samples of males and females in the "Alberta Study"

| Variable | n | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| males | | | | | |
| Height (cm) | 102 | 176.9 | 7.1 | 150 | 197 |
| Weight (Kg) | 101 | 83.0 | 15.1 | 35 | 125 |
| females | | | | | |
| Height (cm) | 107 | 164.3 | 6.1 | 142 | 182 |
| Weight (Kg) | 107 | 72.3 | 14.8 | 49 | 115 |

(a) Is height more variable than weight (among males)?

(b) Are women's heights more variable than men's?

(c) Fit a Gaussian ("Normal") distribution to the observed distribution of weights in the 101 males

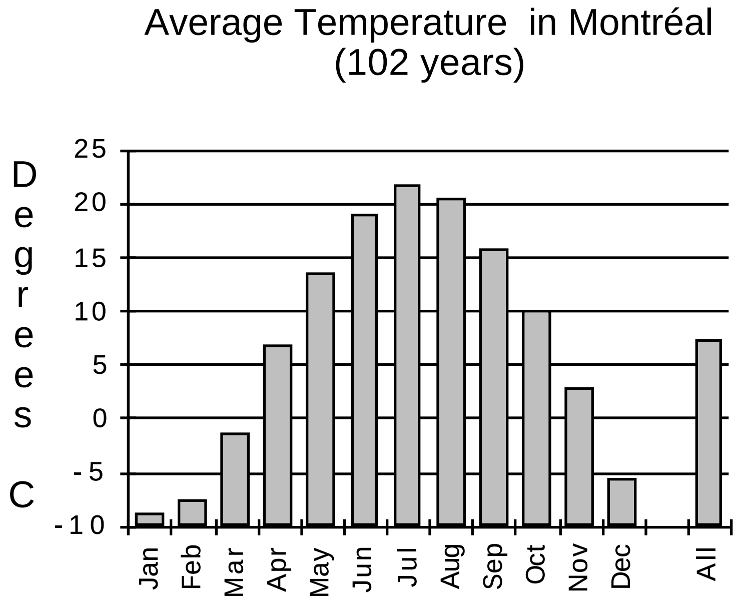
```
 3|5
 4|6
 5|9
 6|1223455678
 7|0112233344455666677777778999
 8|000000000001223333444445555567779
 9|0011111445789
10|11333579
11|278
12|115
```

by calculating how many you would "expect" in each weight category if the weights followed a Normal distribution with mean 83 Kg and standard deviation 15.1 Kg .

*Hint: In Excel, make 2 columns, one for the beginnings and one for the endings of the class intervals; for each interval, subtract the value of cumulative Normal function at the beginning of the interval from the value at the end; and multiply this difference by 101.*

### -10- Montreal Temperatures

data from Environment Canada



Is this a histogram? Justify your answer.

### -11- The Average Adolescent

Excerpt from article in the Montreal Gazette, September 30, 1993

"In Canada, the average adolescent has their first full sexual experience at 15. By Grade 9, 31 per cent of boys and 21 per cent or girls have had full sexual activity. By Grade 11, 49 per cent of boys and 46 per cent of girls. High-school dropouts, 90 per cent of boys, 81 per cent of girls. By college/university, 77 per cent of young men, 73 per cent of young women. This is a 1988 study. A 1990 update showed that the amount of sexual activity had increased in all age groups."

### Get Number Right!

Follow-up letter (from a student in Med-1 at McGill)

"In the Sept. 30 column, the journalist mentions that the average teen has a full sexual experience by Grade 9. More precisely, her statistics are 31 per cent of boys and 21 per cent of girls in this age group. Assuming equal numbers of boys and girls, this comes out to 26 per cent of Grade 9 students - a far cry from the 50 per cent "average" that the journalist claims. Perhaps some Gazette columnists would benefit from arithmetic courses".

**EXERCISE:**

a This med studedent hasn't taken 607! Straighten out the med-1 student -- and the Montreal Gazette journalist -- by rewriting the first sentence of the article, and by explaining to both of them why your version is more accurate. [use same "age = grade + 6" that journalist uses]

b Sketch (i) the histogram (ii) the cumulative distribution for what your guess as to the distribution of "age at first sexual experience" amomg Canadian males and females of your age. Use separate curves for males and females, but put them in the same graph for comparison.

### -12- How much can one tell about a distribution from just the mean and SD?

[Data from Ftetcher et al. NEJM 301, 180-183, 1979]

In a sample of n=151 articles from 1946, the average number of authors per article was 2.0 and the SD was 1.4.

a Try to reconstruct the frequency distribution. To make the calculations easier, a spreadsheet is available on the web page under the resources for Chapter 1.

b For this type of variable, do you think there are other very different shaped distributions with the same mean and SD, that would be compatible with these summaries (For example, could you take the mirrow image of the distribution (so you would have the same SD) and just slide it left or right until you got the correct mean?)