Chapter 01: The nature of econometrics and economic data

Solutions to Problems

1 (i) Ideally, we could randomly assign students to classes of different sizes. That is, each student is assigned a different class size without regard to any student characteristics such as ability and family background. We would like substantial variation in class sizes (subject, of course, to ethical considerations and resource constraints).

(ii) A negative correlation means that larger class size is associated with lower performance. We might find a negative correlation because larger class size actually hurts performance. However, with observational data, there are other reasons we might find a negative relationship. For example, children from more affluent families in Australia might be more likely to attend schools with smaller class sizes, and affluent children generally score better on standardized tests. Another possibility is that, within a school, a principal might assign the better students to smaller classes. Or, some parents might insist their children are in the smaller classes, and these same parents tend to be more involved in their children's education.

(iii) Given the potential for confounding factors – some of which are listed in (ii) – finding a negative correlation would not be strong evidence that smaller class sizes actually lead to better performance. Some way of controlling for the confounding factors is needed, and this is the subject of multiple regression analysis.

2 (i) Here is one way to pose the question: If two firms, say *A* and *B*, are identical in all respects except that firm *A* supplies job training one hour per worker more than firm *B*, by how much would firm *A*'s output differ from firm *B*'s?

(ii) Manufacturing firms in Victoria are likely to choose job training depending on the characteristics of workers. Some observed characteristics are years of schooling, years in the workforce, and experience in a particular job. Firms might even discriminate based on age, gender, or race. Perhaps firms choose to offer training to more or less able workers, where 'ability' might be difficult to quantify but where a manager has some idea about the relative abilities of different employees. Moreover, different kinds of workers might be attracted to firms that offer more job training on average, and this might not be evident to employers.

(iii) The amount of capital and technology available to workers would also affect output. So, two firms with exactly the same kinds of employees would generally have different outputs if they use different amounts of capital or technology. The quality of managers would also have an effect.

(iv) No, unless the amount of training is randomly assigned. The many factors listed in parts (ii) and (iii) can contribute to finding a positive correlation between *output* and *training* even if job training does not improve worker productivity.

3 It does not make sense to pose the question in terms of causality. Economists would assume that students choose a mix of studying and working (and other activities, such as attending class, leisure, and sleeping) based on rational behaviour, such as maximizing utility subject to the constraint that there are only 168 hours in a week. We can then use statistical methods to measure the association between studying and working, including regression analysis. But we would not be claiming that one variable 'causes' the other. They are both choice variables of the student.

Multiple Choice Questions

1. c
2. d
3. d
4. b
5. c
6. c
7. a

Computer Exercises

C1 (i) The average of *educ* is about 12.6 years. There are two people reporting zero years of education, and 19 people reporting 18 years of education.

(ii) The average of *wage* is about $5.90, which seems low in the year 2008.

(iii) Using Table B-60 in the 2004 *Economic Report of the President*, the CPI was 56.9 in 1976 and 184.0 in 2003.

(iv) The sample contains 252 women (the number of observations with *female* = 1) and 274 men.

C2 (i) There are 1,388 observations in the sample. Tabulating the variable *cigs* shows that 212 women have *cigs* > 0.

(ii) The average of *cigs* is about 2.09, but this includes the 1,176 women who did not smoke. Reporting just the average masks the fact that almost 85 percent of the women did not smoke. It makes more sense to say that the 'typical' woman does not smoke during pregnancy; indeed, the median number of cigarettes smoked is zero.

(iii) The average of *cigs* over the women with *cigs* > 0 is about 13.7. Of course this is much higher than the average over the entire sample because we are excluding 1,176 non-smoker women.

(iv) The average of *fatheduc* is about 13.2. There are 196 observations with a missing value for *fatheduc*, and those observations are necessarily excluded in computing the average.

C3 (i) 185/445 ≈ .416 is the fraction of men receiving job training, or about 41.6%.

(ii) For men receiving job training, the average of *re78* is about 6.35, or $6,350. For men not receiving job training, the average of *re78* is about 4.55, or $4,550. The difference is $1,800, which is very large. On average, the men receiving the job training had earnings about 40% higher than those not receiving training.

(iii) About 24.3% of the men who received training were unemployed in 1978; the figure is 35.4% for men not receiving training. This, too, is a big difference.

(iv) The differences in earnings and unemployment rates suggest the training program had strong, positive effects. Our conclusions about economic significance would be stronger if we could also establish statistical significance.

C4 (i) The smallest and largest values of *children* are 0 and 13, respectively. The average is about 2.27.

(ii) Out of 4,358 women, only 611 have electricity in the home, or about 14.02 percent.

(iii) The average of *children* for women without electricity is about 2.33, and for those with electricity it is about 1.90. So, on average, women with electricity have .43 fewer children than those who do not.

(iv) We cannot infer causality here. There are many confounding factors that may be related to the number of children and the presence of electricity in the home; household income and level of education are two possibilities. For example, it could be that women with more education have fewer children and are more likely to have electricity in the home (the latter due to an income effect).

C5

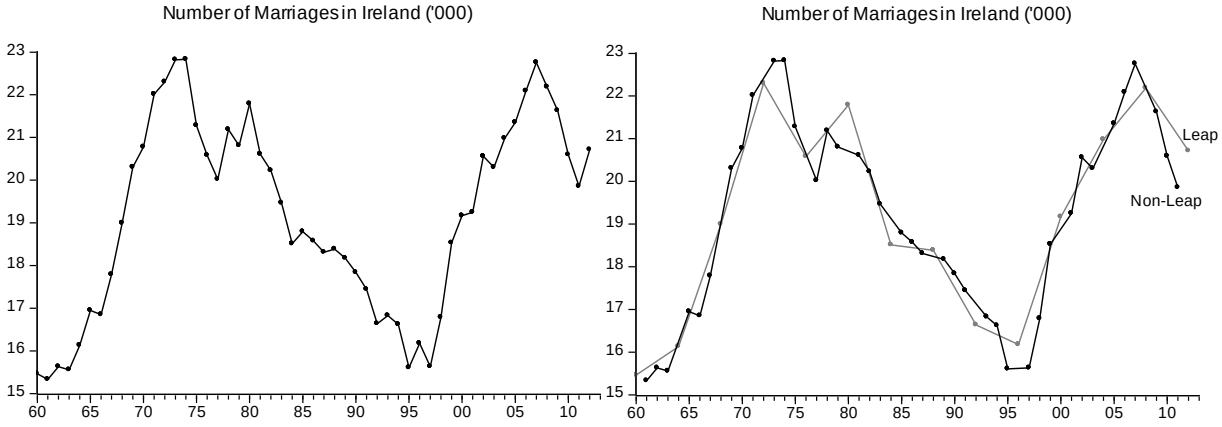
(i)

There appears to be a downward trend in the data.

(ii)

The number of marriages is lower in the leap years although the differential more recently appears to be getting smaller.

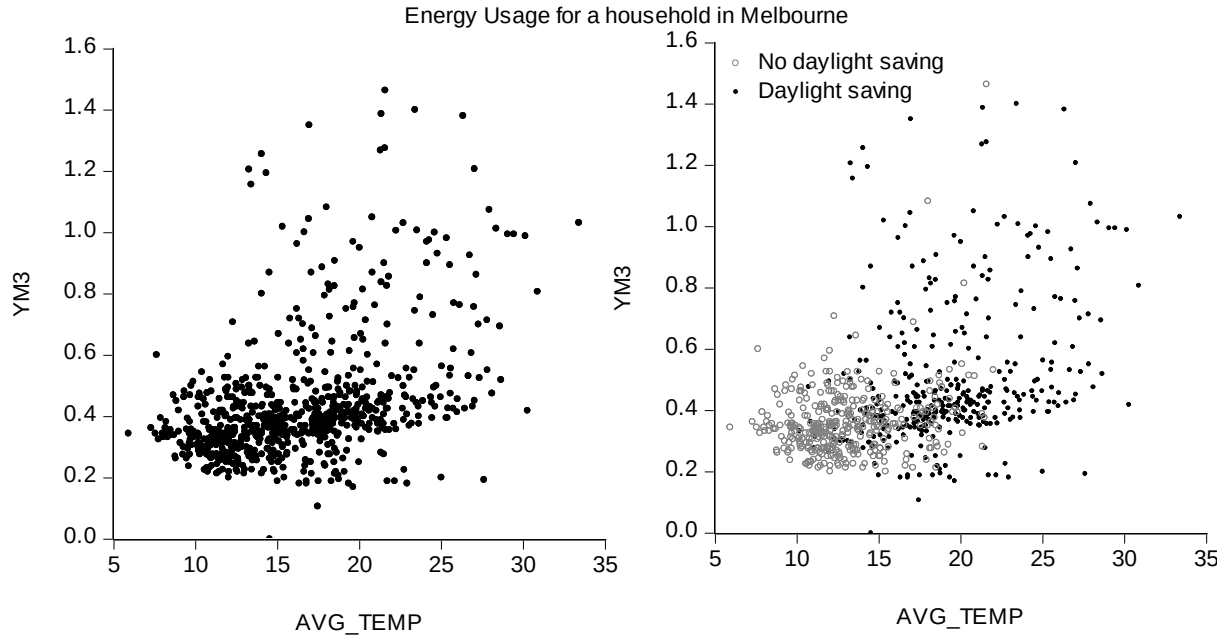
(iii)



The number of marriages has fluctuated over time with some periods of downturns and other periods where there has been an increase. There is no obvious difference when we plot the data separately for leap and non-leap years.

C6

(i)



In the first plot we observe that energy usage increases with increases in average temperature. The relationship doesn't appear to be linear.

(ii) In the second plot we divide the data by whether the observation is in the daylight savings period or not. When the observations are not in the day light savings period these mainly correspond to winter and in the graph we can see these observations are generally clustered in the area of lower average temperatures and lower energy usage. However, the nonlinearity in the data can still be observed.

Chapter 02: Basic mathematical tools

Solutions to Problems

1 The table is extended with required calculatyions as follows,

Observation	X_i	Y_i	X_i^2	$X_i Y_i$	$\left(X_i-\bar{X}\right)^2$
1	2	1	4	2	$(2-2.6)^2=.36$
2	0	3	0	0	$(0-2.6)^2=6.76$
3	–1	–2	1	2	$(-1-2.6)^2=12.96$
4	5	4	25	20	$(5-2.6)^2=5.76$
5	7	3	49	21	$(7-2.6)^2=19.36$
	$\sum_{i=1}^{5} X_i = 13$	$\sum_{i=1}^{5} Y_i = 9$	$\sum_{i=1}^{5} X_i^2 = 79$	$\sum_{i=1}^{5} X_i Y_i = 45$	$\sum_{i=1}^{5}\left(X_i-\bar{X}\right)^2$ =45.2

Note that $\bar{X} = \sum_{i=1}^{5} X_i \Big/ 5 = 2.6$ and $\bar{Y} = \sum_{i=1}^{5} Y_i \Big/ 5 = 1.8$

(i) As the tabole shows, $\sum_{i=1}^{5} X_i$ =13 and $\sum_{i=1}^{5} Y_i$ = 9

(ii) From the table, $\sum_{i=1}^{5} X_i^2$ = 79 and $\left(\sum_{i=1}^{5} X_i\right)^2$ =$(13)^2$ =169

(iii) $\sum_{i=1}^{5} X_i Y_i$ =45 and $\left(\sum_{i=1}^{5} X_i\right)\left(\sum_{i=1}^{5} Y_i\right)$=13*9= 117

(iv) $\sum_{i=1}^{5}\left(X_i+Y_i\right)$=13+9 = 22 and $\left(\sum_{i=1}^{5} X_i+\sum_{i=1}^{5} Y_i\right)$=13+9=22. Actually, these are same quantities.

(v) $\sum_{i=1}^{5}\left(X_i-Y_i\right) = \sum_{i=1}^{5} X_i - \sum_{i=1}^{5} Y_i = 13-9 = 4$

(vi) $\sum_{i=1}^{5} 2X_i = 2\sum_{i=1}^{5} X_i = 2*13 = 26$

(vii) $\sum_{i=1}^{5} 2 = 2*5=10$

(viii) $\sum_{i=1}^{5}\left(X_i - \bar{X}\right)=0$

2.

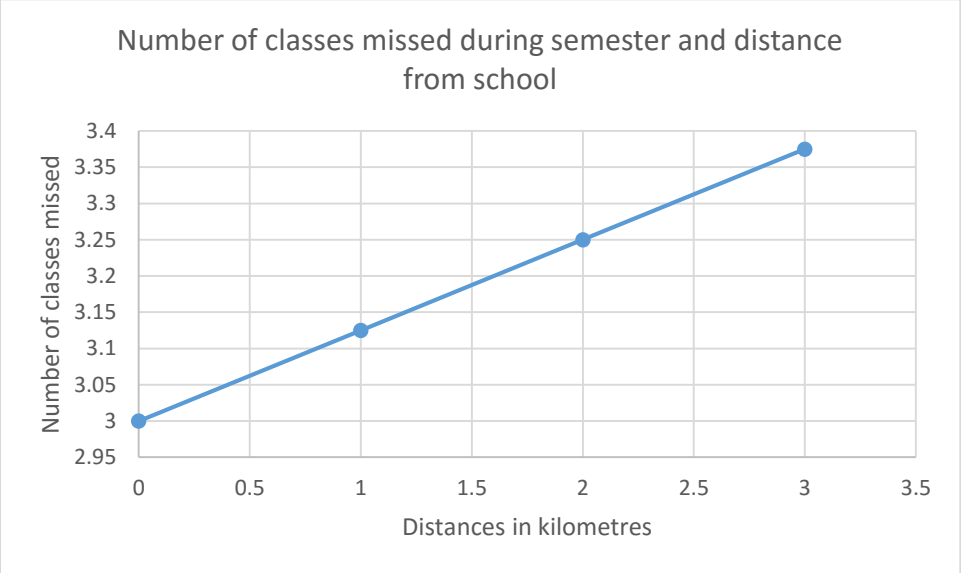
(i) The average monthly housing expenditure is $566.

(ii) The average monthly expenditure would be 5.66, respectively, measured in hundreds of dollars.

(iii) The average monthly housing expenditure increases to $586.

3.

(i) This is just a standard linear equation with intercept equal to 3 and slope equal to .125 The intercept is the number of missed classes for a student who lives on campus.



(ii) The average number of classes missed by students who live 8 kilometres away is, missed = 3 + .125(8) = 4.0 or approximately 4 classes.

(iii) The difference between the average number of classes missed by student living 16 kilometres and 32 kilometres away is = [3 + .125(32)]–[3 + .125(16))] =7–5 = 2 class.

4. If *price* = 15 and *income* = 200, *quantity* = 20 – 1.8(15) + .03(200) = –1, which is nonsense. This shows that linear demand functions generally cannot describe demand over a wide range of prices and income.

5.

(i) The percentage point change is 6 – 4 = 2, or a two percentage point increase in the unemployment rate.

(ii) The percentage change in the unemployment rate is 100[(6 – 4)/4] = 50%. i.e., unemployment increased by 50%.

6. The majority shareholder is referring to the percentage point increase in the stock return, while the CEO is referring to the change relative to the initial return of 15%. To be precise, the shareholder should specifically refer to a 3 percentage *point* increase.

7.

(i) The person b's salary exceeds that of person B by 100[42,000 – 35,000)/35,000] = 20%.

(ii) The approximate proportionate change is log(42,000) – log(35,000) ≈ .182, so the approximate percentage change is %18.2. [Note: log(·) denotes the natural log.]

8.

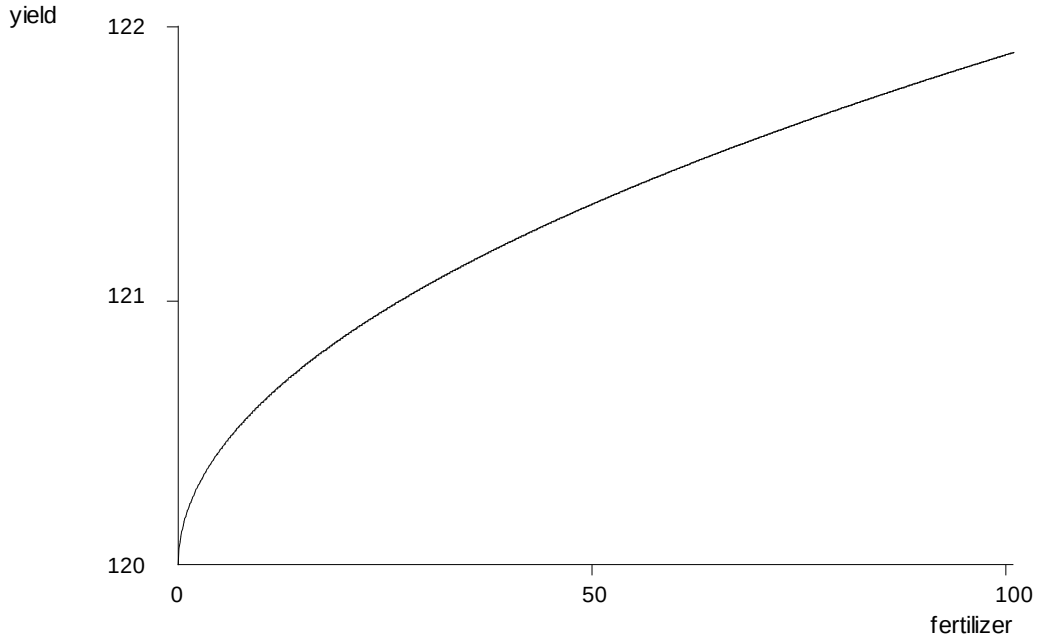
(i) When *exper* = 0, log(*salary*) = 10.6; therefore, *salary* = exp(10.6) ≈ $40,134.84. When *exper* = 5, *salary* = exp[10.6 + .027(5)] ≈ $45,935.80.

(ii) The approximate proportionate increase is .027(5) = .135, so the approximate percentage change is 13.5%.

(iii) 100[(45,935.80 – 40,134.84)/40,134.84) ≈ 14.5%, so the exact percentage increase is about one percentage point higher.

9.

(i) The relationship between *yield* and *fertilizer* is graphed (see over page).



(ii) Compared with a linear function, the function

$$yield = 120 + .13\sqrt{fertilizer}$$

has a diminishing marginal effect, and the slope approaches zero as *fertilizer* gets large. The initial kilogram of fertilizer has the largest effect, and each additional kilogram has an marginal effect smaller than the previous kilogram.

10.

(i) The value 20.5 is the intercept in the equation, so it literally means that if *age* = 0 then the BMI is 20.5. Of course, *age* = 0, measured in years would indicate the BMI of 20.5 as shown by that the intercept is the BMI of new born babies or precisely, babies less than a year old. The intercept by itself is not much of interest since body fat 1-of babies less than a year old is not usually a concern. Also, the intercept should ideally reflect a dataset on age and BMI of the adult population, so, by itself, 20.5 is not of much interest.

(ii) We use calculus to obtain the maximum BMI:

$$\frac{dBMI}{dAge} = .2 - .004Age \qquad \text{and} \quad \frac{d^2BMI}{dAge^2} = -.004 < 0.$$

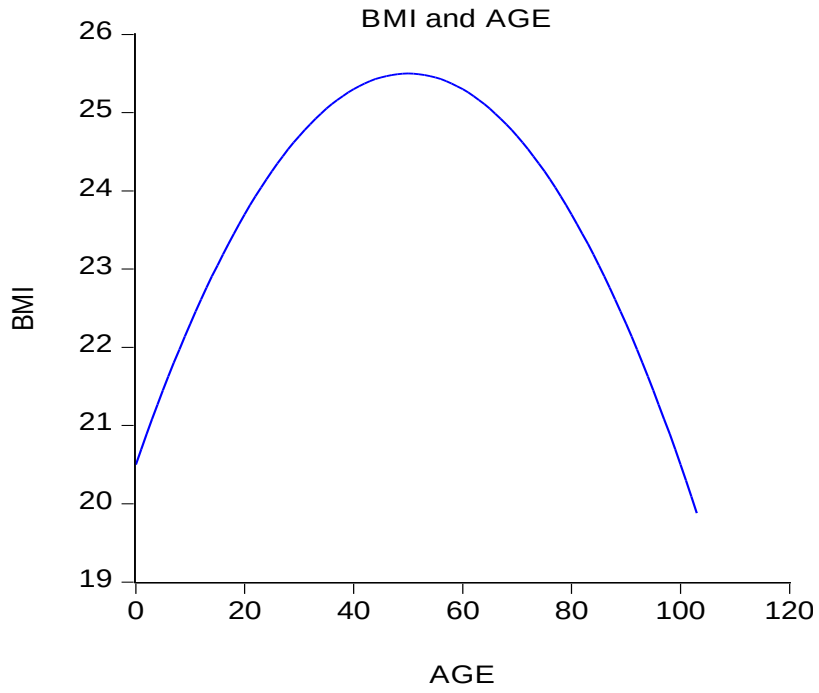
Hence. The BMI function has a maximum. Letting, the first derivative equal to 0,

$$\frac{dBMI}{dAge} = .2 - .004Age = 0$$

$$Age = \frac{.2}{.004} = 50$$

Therefore, BMI is maximum at the age of 50 years.

(iii) The following graph shows the solution rounded to the nearest integer:



(iv) It is not at all realistic to think that BMI and age will have a determinictic relationship. BMI is also scientifically measured in accordance with the height of a person. Besides there are many other factors that affect BMI of a person, such as a person's general lifestyle, eating habits, health awareness, income. Multiple regression analysis allows for many observed factors to affect a variable like BMI, and also recognizes that there are unobserved factors that are important and that we can never directly account for.

Multiple Choice Questions:

1. d
2. b
3. d
4. a
5. c
6. b
7. c

8. c
9. a
10. a

Chapter 03: Fundamentals of statistics: a review

Solutions to Problems

1.

(i) P(0< Z < 1) = .8413–.5=.3413

(ii) P(–1< Z < 1) = .8413–.1587 = .6826

(iii) P(Z > 2.55) = 1–.9946 =.0054

(iv) P(Z < –1.60) = .0548

2.

(i) P($X \leq 6$) = P[($X - 5$)/2 $\leq$ (6 – 5)/2] = P($Z \leq .5$) $\approx$ .692, where Z denotes a Normal (0,1) random variable. [We obtain P($Z \leq .5$) from Table G.1.]

(ii) P($X > 4$) = P[($X - 5$)/2 > (4 – 5)/2] = P($Z > -.5$) = P($Z \leq .5$) $\approx$ .692.

(iii) P($|X - 5| > 1$) = P($X - 5 > 1$) + P($X - 5 < -1$) = P($X > 6$) + P($X < 4$) $\approx$ (1 – .692) + (1 – .692) = .616, where we have used answers from parts (i) and (ii).

3. Let X denote family income. Then given the information we find the required probabilities as shown below,

(i) $P(X < 30000) = P(Z < \frac{30000 - 50000}{10000}) = P(Z < -2) = .0228$

(ii) $P(X > 70000) = P(Z > \frac{70000 - 50000}{10000}) = P(Z > 2) = 1 - .9772 = .0228$

4. Let X representds the marks obtained by the students and X* denote the lowest mark that will be awarded an A grade. Given that $X \sim N(70, 6)$ we first find out the value of standard normal variable Z, such that the probability of Z exceeding this value is 10% or .10. That is, we need to find the value of Z that leaves out 10% area under the right tail of the Z distribution.

From the appendix on areas under the standard normal distribution, we find that the relevant value of Z is 1.28 (approximately). Hence we get,

$$\frac{X^*-70}{6}=1.28$$
$$\Rightarrow X^*=(1.28)\times 6+70=77.68$$

Hence, the lowest mark that will be awarded an A grade is 77.68 or 78 (apporximately).

5. Let Y_{it} be the binary variable equal to one if fund *i* outperforms the market in year *t*. By assumption, $P(Y_{it} = 1) = .5$ (a 50-50 chance of outperforming the market for each fund in each year). Now, for any fund, we are also assuming that performance relative to the market is independent across years. But then the probability that fund *i* outperforms the market in all 10 years, $P(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{i,10} = 1)$, is just the product of the probabilities: $P(Y_{i1} = 1) \cdot P(Y_{i2} = 1) \ldots P(Y_{i,10} = 1) = (.5)^{10} = 1/1024$ (which is slightly less than .001). In fact, if we define a binary random variable Y_i such that $Y_i = 1$ if and only if fund *i* outperformed the market in all 10 years, then $P(Y_i = 1) = 1/1024$.

6. In eight attempts the expected number of free throws is 8(.74) = 5.92, or about six free throws.

7.

Tossing three coins give the following sample space or the possible combinations of events,

HHH, HHT, HTH, HTT, THH, THT,TTH, TTT

Since P(H)= .5 and P(T) =.5 hence probability of each event, say of the event HTH is P(HTH) = .5*.5*.5=.125

Given the X represents the number of tails, we can construct the probability distribution of X, that takes the values of 0 (no tail), 1 (one tail), 2(two tails) and 3 (three tails).

X	0	1	2	3
Prob.	.125	.375	.375	.125

E(X) = 0*.125 + 1*.375 + 2*.375 + 3*.125 = 1.5

$E(X^2) = 0^2*.125 + 1^2*.375 + 2^2*.375 + 3^2*.125 = 3$

Profit = $(X^2+X) – 5$

E(Profit) = $E(X^2+X) –5 = E(X^2)+E(X) – 5 = 3 + 1.5 –5 = –.5$ hence there is a loss of 50 cents.

8. If *Y* is salary in dollars then $Y = 1000 \cdot X$, and so the expected value of *Y* is 1,000 times the expected value of *X*, and the standard deviation of *Y* is 1,000 times the standard deviation of *X*. Therefore, the expected value and standard deviation of salary, measured in dollars, are $57,000 and $14,600, respectively.

9.

(i) P(male wins) = 40/60 = .667 apx

(ii) P(married/ male) = $\frac{P(married\ \&\ male)}{P(male)} = \frac{10/60}{40/60} = \frac{10}{60} * \frac{60}{40} = \frac{10}{40} = .25$

10. E(*GRADE*|*ATAR=65*) = 10.5 + .85 (65) = 65.75. Similarly, E(*GRADE*|*ATAR=95*) = 10.5 + .85 (95) = 91.25. The difference in expected grade obtained in the subject is substantial, but the difference in ATAR scores is also rather large.

11.

(i) This is just a special case of what we covered in the text, with $n = 4$: $E(\bar{Y}) = \mu$ and $Var(\bar{Y}) = \sigma^2/4$.

(ii) $E(W) = E(Y_1)/8 + E(Y_2)/8 + E(Y_3)/4 + E(Y_4)/2 = \mu[(1/8) + (1/8) + (1/4) + (1/2)] = \mu(1 + 1 + 2 + 4)/8 = \mu$, which shows that W is unbiased. Because the Y_i are independent,

$$Var(W) = Var(Y_1)/64 + Var(Y_2)/64 + Var(Y_3)/16 + Var(Y_4)/4$$

$$= \sigma^2[(1/64) + (1/64) + (4/64) + (16/64)] = \sigma^2(22/64) = \sigma^2(11/32).$$

(iii) Because 11/32 > 8/32 = 1/4, $Var(W) > Var(\bar{Y})$ for any $\sigma^2 > 0$, so $\bar{Y}$ is preferred to W because each is unbiased.

12.

(i) $E(W_a) = a_1E(Y_1) + a_2E(Y_2) + \ldots + a_nE(Y_n) = (a_1 + a_2 + \ldots + a_n)\mu$. Therefore, we must have $a_1 + a_2 + \ldots + a_n = 1$ for unbiasedness.

(ii) $Var(W_a) = a_1^2 Var(Y_1) + a_2^2 Var(Y_2) + \ldots + a_n^2 Var(Y_n) = (a_1^2 + a_2^2 + \ldots + a_n^2)\sigma^2$.

(iii) From the hint, when $a_1 + a_2 + \ldots + a_n = 1$ – the condition needed for unbiasedness of W_a – we have $1/n \leq a_1^2 + a_2^2 + \ldots + a_n^2$. But then $Var(\bar{Y}) = \sigma^2/n \leq \sigma^2(a_1^2 + a_2^2 + \ldots + a_n^2) = Var(W_a)$.