Social Studies 201
Winter 2004
Answers to Problem Set No. 2
February 3, 2004

1. **Distribution of age last worked**.

   (a) The variable X is "age last did paid work" and there are samples from two groups, male respondents and female respondents. Since the data are categorized into intervals of different width, in order to construct the histogram for each group, it is necessary to calculate the densities. These calculations are in Table 1, with the histograms in Figures 1 and 2. The densities for the large group of respondents who retired at age 65 could have been graphed with the densities shown in Table 1, assuming all retired at age 65, with an interval width of one year. Alternatively, given these as open-ended intervals, I drew the bars for this group at a height that seemed reasonable, given the shape of the histogram, and the open-ended aspect of the 65 and over age group.

   One other issue is whether the real class limits should be 14.5, 24.5, 34.5, etc. or 15, 25, 35, etc. Either is acceptable, although in this presentation I used the former.

   (b) The tabular calculations for the means are contained in Table 2. For males, the mean age last worked is

$$\bar{X} = \frac{\Sigma(fX)}{n} = \frac{73,166.5}{1,385} = 52.82$$

   and for females, the mean age last worked is

$$\bar{X} = \frac{\Sigma(fX)}{n} = \frac{118,875}{2,912} = 40.82$$

   .

   Using the percentage distributions in Table 3, the median of age last worked for males is in the interval from 54.5 to 59.5. Interpolating in this interval, the median is

$$P_{50} = 54.5 + \left(\frac{50 - 41.6}{59.9 - 41.6} \times 5 \right)$$

Figure 1: Histogram of age last worked for males

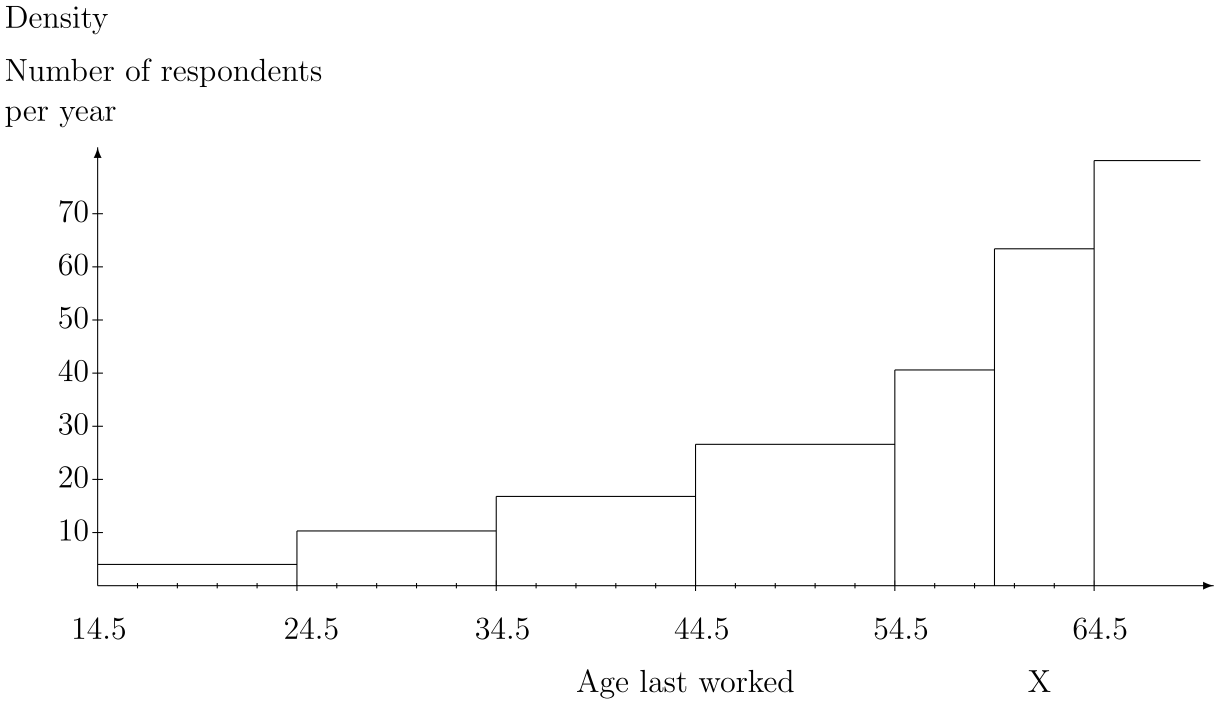

Figure 2: Histogram of age last worked for females

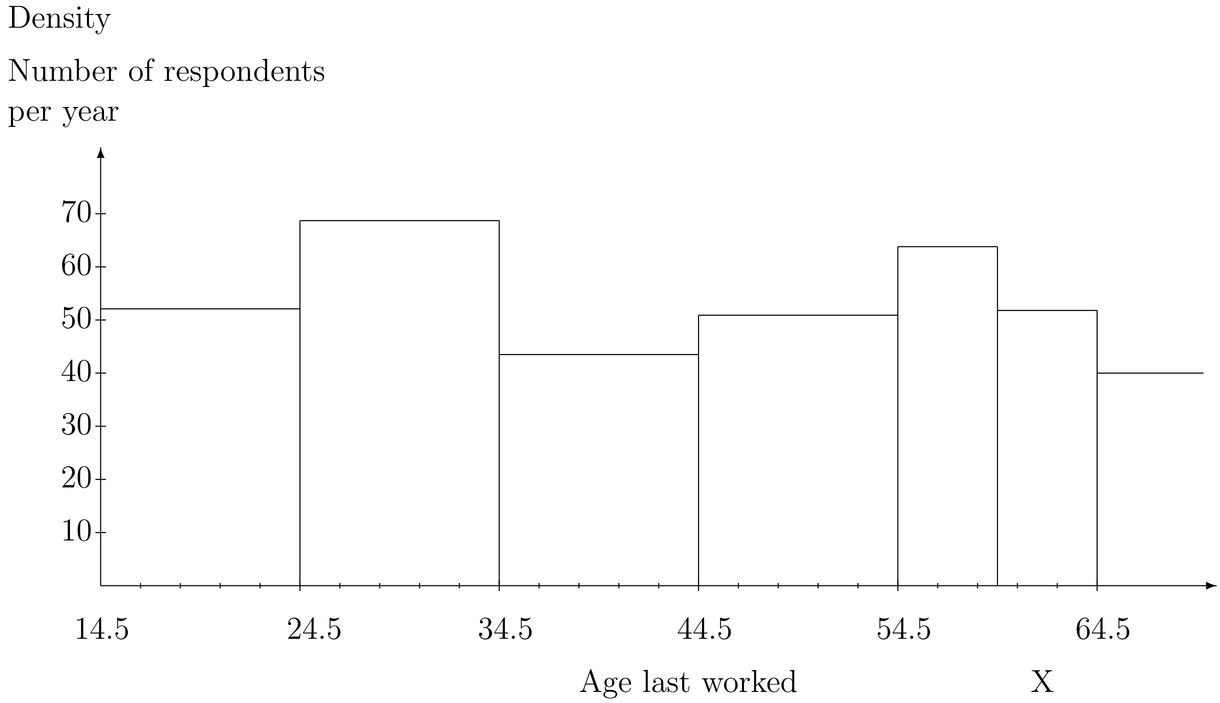

Table 1: Frequency distributions and densities for age last did paid work, male and female respondents

Age in	Interval	Male		Females	
years	Width	Number	Density	Number	Density
14.5-24.5	10	40	4.0	521	52.1
24.5-34.5	10	103	10.3	687	68.7
34.5-44.5	10	168	16.8	435	43.5
44.5-54.5	10	266	26.6	509	50.9
54.5-59.5	5	253	50.6	319	63.8
59.5-64.5	5	317	63.4	259	51.8
64.5-65.5	1	238	238.0	182	182.0
Total		1,385		2,912	

$$\begin{aligned} &= 54.5 + (8.4/18.3 \times 5) \\ &= 54.5 + (0.459 \times 5) \\ &= 54.5 + 2.3 \\ &= 56.8 \end{aligned}$$

years.

$$\begin{aligned} P_{50} &= 34.5 + \left(\frac{50 - 41.4}{56.3 - 41.4} \times 10 \right) \\ &= 34.5 + (8.6/14.9 \times 10) \\ &= 34.5 + (0.577 \times 10) \\ &= 34.5 + 5.8 \\ &= 40.1 \end{aligned}$$

(c) Comparison of distributions. One similarity of the two distributions is the large number of respondents who report last working at age 65. This is consistent with the situation in many workplaces, where the normal retirement age is 65. While not required for this question, the mode (or peak) of each distribution would be at 65. For females, the rest of the distribution is relatively flat

Table 2: Calculations for mean age last worked, male and female respondents

Age in	Midpoint	Males		Females	
years	X	f	fX	f	fX
14.5-24.5	19.5	40	780.0	521	10,159.5
24.5-34.5	29.5	103	3,038.5	687	20,266.5
34.5-44.5	39.5	168	6,636.0	435	17,182.5
44.5-54.5	49.5	266	13,167.0	509	25,195.5
54.5-59.5	57.0	253	14,421.0	319	18,183.0
59.5-64.5	62.0	317	19,654.0	259	16,058.0
64.5-65.5	65.0	238	15,470.0	182	11,830.0
Total		1,385	73,166.5	2,912	118,875.0

or uniform, with similar numbers and percentages reporting each age group as the age when last worked. For females, there is a definite secondary peak at ages 25-34, consistent with the fact that some women leave the workplace after birth of a child or children. For males, the distribution rises as age increases, at least up to age 65. That is, at each successive age group, there are relatively more males who report that age as the age last worked.

These differences in distribution produce the summary statistics of Table 4. Given the relatively uniform distribution for females, the distribution is not all that asymmetrical, so the mean and median are very similar to each other, at age 40 or 41. This is a much lower age for the centre of the distribution that for males, where the mean and median are close to the age of the mid-50s. The greater concentration of males at ages 45 and higher produces a larger mean and median for males, as compared with females.

2. **Length of time using internet**. From the bar charts, the data have been reorganized into Tables 5 and 6. In order to determine the mode for each group of users, it is necessary to obtain the density of occurrence – Figures 1 and 2 of the question sheet are bar charts but they are not histograms, since interval widths differ. The calculations

Table 3: Percentages for obtaining median age last worked, male and female respondents

Age in	Width	Males		Females	
years	w	P	Cum. P	P	Cum. P
14.5-24.5	10	2.9	2.9	17.9	17.9
24.5-34.5	10	7.4	10.3	23.5	41.4
34.5-44.5	10	12.1	22.4	14.9	56.3
44.5-54.5	10	19.2	41.6	17.5	73.8
54.5-59.5	5	18.3	59.9	11.0	84.8
59.5-64.5	5	22.9	82.8	8.9	93.7
64.5-65.5	1	17.2	100.0	6.3	100.0
Total		100.0		100.0	

Note. The percentage for females in the second row was reduced by 0.1 in order for the sum of percentages to total one hundred per cent.

Table 4: Summary statistics for aged last worked, males and females

Statistic	Male	Female
Mode	65.0	65.0
Mean	52.8	40.8
Median	56.8	40.1

of density are in Table 5. In order to use a consistent unit of time across all categories, the times reported in months have been reorganized into years, so that 0-6 months is 0 to 0.5 years and 6-12 months is 0.5 to 1 year.

The mode for infrequent users is less than 6 months, or 3 months, since this is the category with the greatest density. For regular users the mode is 6-12 months or, as a single value, the mode is 9 months, since the density for this interval is greater than at 1-4 years or at any other category.

From the percentages and cumulative percentages of Table 6 the per-

Table 5: Percentage distributions and densities for length of time used internet, infrequent and regular users

Length of time	Interval	Infrequent users		Regular users	
in years	Width	Per cent	Density	Per cent	Density
0-0.5	0.5	22	44.0	6	12
0.5-1	0.5	18	36.0	8	16
1-4	3	49	16.3	46	15.3
4-7	3	10	3.3	31	10.3
7-15	8	1	0.1	9	1.1
Total		100		100	

centiles and interquartile ranges are as follows.

The 75th percentile for infrequent users is

$$P_{75} = 1 + \left(\frac{75 - 40}{89 - 40} \times 3 \right) = 1 + (0.714 \times 3) = 1 + 2.1 = 3.1$$

and the 25th percentile is

$$P_{25} = 0.5 + \left(\frac{25 - 22}{40 - 22} \times 0.5 \right) = 0.5 + (0.167 \times 0.5) = 0.5 + 0.08 = 0.6.$$

For infrequent users, the interquartile range is $3.1 - 0.6 = 2.5$.

The 75th percentile for regular users is

$$P_{75} = 4 + \left(\frac{75 - 60}{91 - 60} \times 3 \right) = 4 + (0.517 \times 3) = 4 + 1.6 = 5.6$$

and the 25th percentile is

$$P_{25} = 1 + \left(\frac{25 - 14}{60 - 14} \times 3 \right) = 1 + (0.239 \times 3) = 1 + 0.7 = 1.7.$$

For regular users, the interquartile range is $5.6 - 1.7 = 3.9$.

Table 6: Percentage and cumulative percentage distributions for length of time used internet, infrequent and regular users

Length of time	Interval	Infrequent		Regular	
in years	Width	Per cent	Cum. per cent	Per cent	Cum. per cent
0-0.5	0.5	22	22	6	6
0.5-1	0.5	18	40	8	14
1-4	3	49	89	46	60
4-7	3	10	99	31	91
7-15	8	1	100	9	100
Total		100		100	

3. **Extent of use of internet**.

For number of times used internet in the last month, the values ordered from lowest to highest are 3, 6, 10, 12, 15, 27, 27, 28, 30, 30. There are ten values so the middle values are the 5th and 6th values, that is, 15 and 27. The median is these two values or, more likely the average of these two values, so the median is $(15 + 27)/2 = 42/2 = 21$ times used the internet in the past month.

For the number of hours used the internet, the hours ordered from low to high are 1, 1, 2, 2, 3, 4, 7, 10, 10, 22. Again, the middle value is the 5th and 6th or 3 and 4. More commonly the median would be reported as the average of these two values, or $(3 + 4)/2 = 7/2 = 3.5$ hours used the internet in the past month.

The range is $30 - 3 = 27$ times, or 3 and 30 times, for number of times used the internet in the past month. The number of hours used the internet in the past month ranges from 1 to 22 hours, so the range is $22 - 1 = 21$ hours.

The calculations for the mean and standard deviation are given in Table 7, with the two methods of obtaining the standard deviation provided in the table. For times used the internet, the mean is $\Sigma X/n = 188/10 = 18.8$.

Times			Hours	
X	$X - \bar{X}$	$(X - \bar{X})^2$	X	X^2
30	11.2	125.44	3	9
12	-6.8	46.24	4	16
28	9.2	84.64	10	100
30	11.2	125.44	22	484
27	8.2	67.24	7	49
3	-15.8	249.64	1	1
6	-12.8	163.84	1	1
15	-3.8	14.44	2	4
10	-8.8	77.44	2	4
27	8.2	67.24	10	100
188	0.0	1,021.60	62	768

Table 7: Calculations for Mean and Standard Deviation, Use of Internet

The variance is

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n-1} = \frac{1,021.60}{9} = 113.51$$

and the standard deviation is

$$s = \sqrt{s^2} = \sqrt{113.51} = 10.654$$

or 10.7 times.

For number of hours used internet in the past month, the mean is $\Sigma X/n = 62/10 = 6.2$. Using the alternative formula, the variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right) \\ &= \frac{1}{9}\left(768 - \frac{62^2}{10}\right) \\ &= \frac{(768 - 384.40)}{9} = 42.62 \end{aligned}$$

and the standard deviation is

$$s = \sqrt{s^2} = \sqrt{42.62} = 6.53$$

or 6.5 hours.

4. **Averages and percentiles**

   (a) The mean is relevant since this suggests a total value of wealth of \$3.9 trillion across all Canadians and this amounts to \$121,900 for each Canadian. This appears to be the mean wealth per Canadian, if the total wealth were to be divided equally across all Canadians.

   Mode, since this implies that Thatcher is no risk for more people than to those for whom he might be a risk. The categories implied are risk and no risk, with no risk being more people than for whom he might be a risk.

   (b) This implies the median, or close to the median, since Regina is near the middle over a set of cities ordered by level of property tax. Presumably cities have been ranked from cheapest in terms of property taxes to most expensive in terms of property taxes. Regina is actually at the 64th percentile (16/25 is 64/100), although the newspaper considers this close to the middle, or near the median.

   (c) If cities are ranked from least dry (or wettest) to most dry, then Saskatoon is in the 94th percentile, since there are only six of 100 ranking greater on the dryness index. Regina is at the 89th percentile, with eleven cities higher on the dryness scale, or eighty-nine cities less dry.