Social Studies 201
Fall 2004
Answers to Problem Set No. 2
September 29, 2004

1. **Annual hours spent volunteering**. The frequency distributions for annual hours spent volunteering for 15-24 and 55-64 year olds are grouped into intervals of unequal width. In order to obtain the mode and construct the histograms, it is necessary to obtain the densities associated with each interval. In order to do this and graph the histograms accurately, it is best to use real class limits. Table 1 contains this information. The histograms obtained from the frequency distributions and densities are contained in Figures 1 and 2.

Table 1: Frequency distributions and densities for annual hours spent volunteering, 15-24 and 55-64 year olds

Hours volunteering	Real class limits	Interval Width (w)	Age 15-24		Age 55-64	
			f	Density	f	Density
0-4	-0.5-4.5	5	11	2.20	2	0.40
5-9	4.5-9.5	5	7	1.40	2	0.40
10-19	9.5-19.5	10	7	0.70	3	0.30
20-29	19.5-29.5	10	5	0.50	3	0.30
30-49	29.5-49.5	20	10	0.50	5	0.25
50-69	49.5-69.5	20	7	0.35	7	0.35
70-99	69.5-99.5	30	5	0.17	5	0.17
100-129	99.5-129.5	30	6	0.20	4	0.13
130-189	129.5-189.5	60	0	0.00	4	0.07
Total			58		35	

The mode for each distribution is at the peak of the histogram, or the interval with the greatest density of occurrence. For 15-24 year olds, the mode is at 0-4 hours, since the interval from 0 to 4, or -0.5 to 4.5 has density of 2.2, by far the highest density for this distribution. The mode is thus 0-4 hours, or the midpoint of this interval, at 2 hours.

For the 55-64 year olds, the intervals with greatest density are 0-4 and 5-9, so the mode is 0-9, or the midpoint of this interval, 4.5 hours.

2. **Record-setting cold**. For the new record, the smallest value is 38.6 and the largest is 44.9 so the range is $44.9 - 38.6 = 6.3$ degrees. For the old record, the maximum is 43.3 and the minimum value is 37.2 so the range is $43.3 - 37.2 = 6.1$ degrees.

Table 2: Calculations for mean and standard deviation, temperatute records

New record			Old record	
X	$X - \bar{X}$	$(X - \bar{X})^2$	X	X^2
41.6	-0.2	0.04	38.9	1,513.21
44.9	0.1	9.61	43.3	1,874.89
41.3	-0.5	0.25	37.2	1,383.84
40.9	-0.9	0.81	37.8	1,428.84
38.6	-3.2	10.24	37.8	1,428.84
43.3	1.5	2.25	42.2	1,780.84
42.1	0.3	0.09	41.7	1,738.89
292.7	0.1	23.29	278.9	11,149.35

The calculations for the mean and standard deviation are given in Table 2, with the two methods of obtaining the standard deviation shown in this table.

For the new record, the mean is $\Sigma X/n = 292.7/7 = 41.814$. Rounded off to the nearest tenth of a degree, the mean is 41.8 degrees below zero Celsius. The variance is

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{23.29}{6} = 3.882$$

and the standard deviation is

$$s = \sqrt{s^2} = \sqrt{3.882} = 1.970$$

or 2.0 degrees.

Figure 1: Histogram of annual hours volunteering, 15-24 year olds

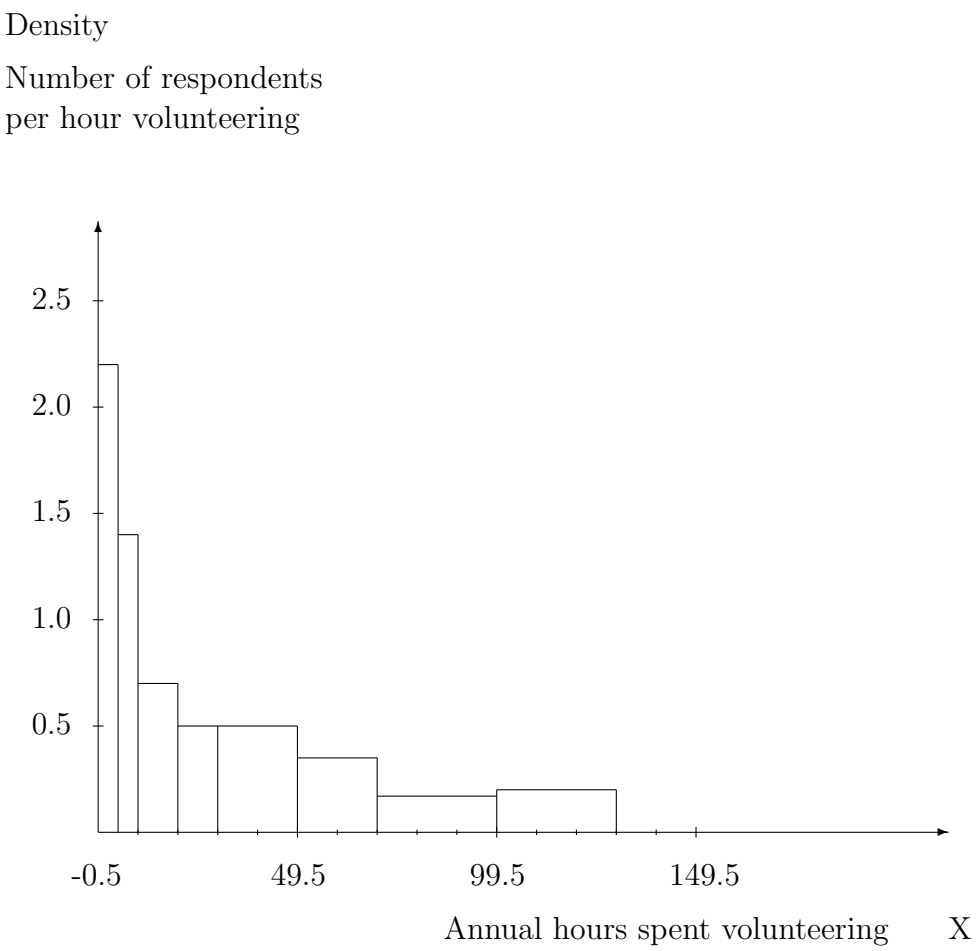
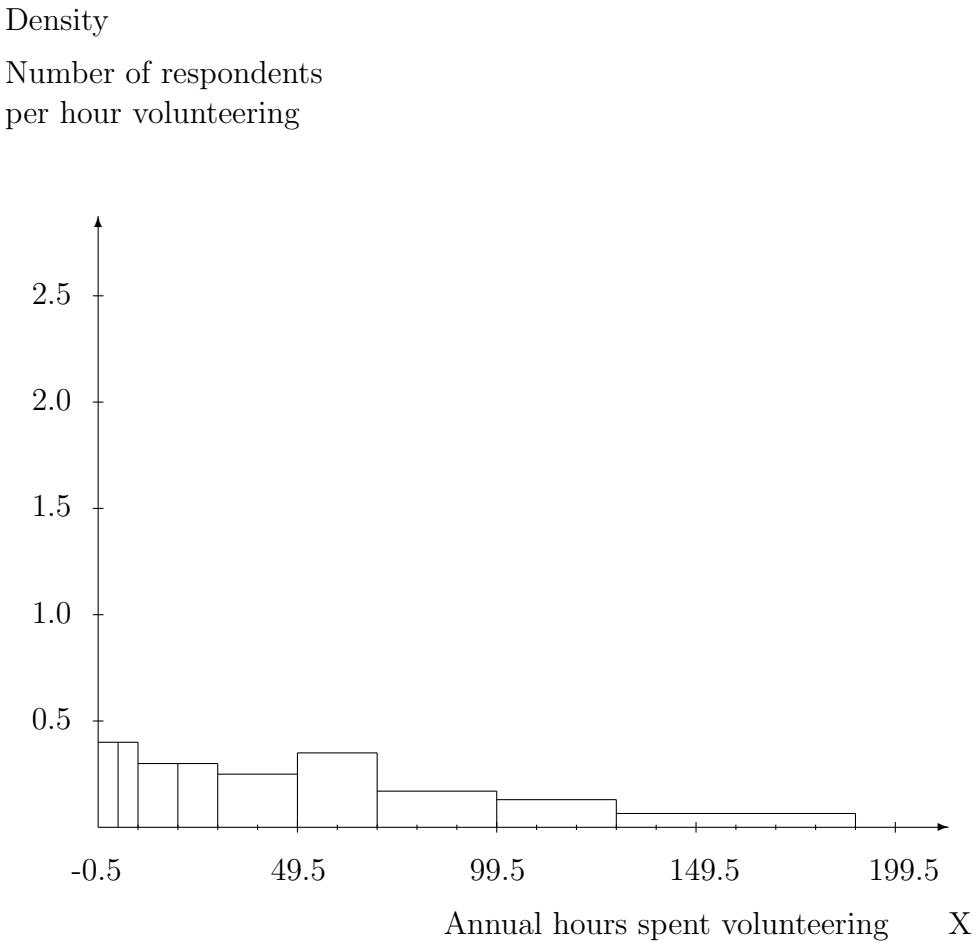


Figure 2: Histogram of annual hours volunteering, 55-64 year olds



For the old record, the mean is $\Sigma X/n = 278.9/7 = 39.843$. To the nearest tenth of a degree the old record was minus 39.8 degrees Celsius. Using the alternative formula, the variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right) \\ &= \frac{1}{6}\left(11,149.35 - \frac{278.9^2}{7}\right) \\ &= \frac{11,149.35 - 11,112.17}{6} \\ &= \frac{37.177}{6} \\ &= 6.196 \end{aligned}$$

and the standard deviation is

$$s = \sqrt{s^2} = \sqrt{6.196} = 2.489$$

or 2.5 degrees.

For the new record, the temperatures, in order from low to high, are 38.6, 40.9, 41.3, 41.6, 42.1, 43.3, 44.9. There are seven values and the middle value is the fourth, that is, 41.6 degrees, associated with Regina. For the old record, the temperatures, in order from low to high are 37.2, 37.8, 37.8, 38.9, 41.7, 42.2, 43.3. There are seven values and the middle value is the fourth, that is, 38.9 degrees, also associated with Regina.

3. From the original Table 3 in Problem Set 2, for incomes under \$20,000, the category with the largest percentage of cases is 'Medium' or 3, with 41.2% of cases. For those with income of \$60,000 plus, it is those who responded 'Good' or 4, who occur most frequently, with 47.6% of cases, a larger percentage than any other response. The mode for the lower income group is thus 'Medium' or 3, and the mode for the higher income group is 'Good' or 4.

   The tabular calculations for the means are contained in Table 3.

   For respondents with income of less than \$20,000, the mean health status is

$$\bar{X} = \frac{\Sigma(PX)}{100} = \frac{332.8}{100} = 3.328$$

Table 3: Calculations for mean health status, low and high income households

Health	<$20,000		$60,000 plus	
Status (X)	P	PX	P	PX
1	6.1	6.1	0.4	0.4
2	9.4	18.8	1.6	3.2
3	41.2	123.6	20.6	61.8
4	32.2	128.8	47.6	190.4
5	11.1	55.5	29.8	149.0
Total	100.0	332.8	100.0	404.8

and for those with incomes of $60,000 plus, the mean health status is

$$\bar{X} = \frac{\Sigma(fX)}{n} = \frac{404.8}{100} = 4.048$$

Rounded to the nearest tenth of a point, the mean health status for individuals from lower income households is 3.3 and for individuals from higher income households is 4.0.

The percentages and cumulative percentages required for obtaining the median and interquartile range are contained in Table 3.

The variable, health status, has a discrete, ordinal scale so the percentiles are the values of health status where the cumulative percentages first reach the required level. From the cumulative percentages in Table 4, the median health status for the lower income group is at 3, since there are only 15.5% of respondents with a health status of 2 or less but over fifty per cent (56.7%) with a health status of 3 or less. For the respondents from higher income households, the fifty per cent point is first reached at a health status of 4 (only 22.6% with status of 3 or less but 70.2% with status of 4 or less).

The interquartile range is the 75th percentile (P_{75}) minus the 25th percentile (P_{25}). For the lower income respondents, P_{25} occurs at health status 3 and P_{75} is at health status 4, so the interquartile range is 1.

$$IQR = P_{75} - P_{25} = 4 - 3 = 1$$

Table 4: Percentages and cumulative percentages for obtaining median health status, low and high income respondents

Health	<$20,000		$60,000 plus	
status	P	Cum. P	P	Cum. P
1	6.1	6.1	0.4	0.4
2	9.4	15.5	1.6	2.0
3	41.2	56.7	20.6	22.6
4	32.2	88.9	47.6	70.2
5	11.1	100.0	29.8	100.0
Total		100.0		100.0

For the higher income respondents, P_{25} does not occur until health status 4 and P_{75} is at health status 5, so the interquartile range is 1.

$$IQR = P_{75} - P_{25} = 5 - 4 = 1$$

Summary comparison. Table 5 provides a summary of the statistics calculated.

Table 5: Summary statistics for health status, low and high income respondents

	Household income	
Statistic	<$20,000	$60,000 plus
Mode	3	4
Mean	3.3	4.0
Median	3	4
P_{75}	4	5
P_{25}	3	4
IQR	1	1

From these summary statistics, the health status reported by those respondents from households with incomes of less than $20,000 averages approximately one point lower than those from households with

incomes of $60,000 plus. The mode and median are each 3 for the lower income individuals and 4 for those with higher incomes. The means differ by a little less than one unit on the health status scale (3.3 for low income and 4.0 for high income) but this appears to be a considerable difference, given that the health status scale has a range of only five points. Variation, as measured by the IQR, is identical for the two distributions, one unit of health status. From this it can be concluded that individuals from lower income households report lower health status than do individuals from higher income households.

4. **Household income of volunteers and non-volunteers**. The calculations for obtaining the means are contained in Table 6. Since income is continuous, the midpoints of the intervals are used for the X values. For the lower open-ended interval, I assumed that incomes are not lower than zero, so the lower interval is from 0 to 20, with a midpoint of 10. For the upper open-ended interval, I could have selected a value of 90 for the midpoint. However, I selected 100 since incomes might be much larger than 100 thousand dollars, so the value of 100 might be a reasonable approximation to the mean incomes of respondents in this upper interval. For volunteers, the mean income is

Table 6: Calculations for mean income of volunteers and non-volunteers

	Midpoint	Volunteer		Non-volunteer	
Income	X	f	fX	f	fX
Under 20	10	230	2,300	132	1,320
20-40	30	278	8,340	119	3,570
40-60	50	247	12,350	61	3,050
60-80	70	185	12,950	51	3,570
80 plus	100	56	5,600	8	800
Total		996	41,540	371	12,310

$$\bar{X} = \frac{\Sigma(fX)}{n} = \frac{41,540}{996} = 41.707$$

or, rounded to the nearest hundred dollars, \$41,700. For non-volunteers, the mean income is

$$\bar{X} = \frac{\Sigma(fX)}{n} = \frac{12,310}{371} = 33.181$$

or, rounded to the nearest hundred dollars, \$33,200.

The percentages for obtaining the medians and seventy-fifth percentiles are contained in Table 7.

Table 7: Percentages for obtaining median income, volunteers and nono-volunteers

	Width	Volunteers		Non-volunteers	
Income		P	Cum. P	P	Cum. P
Under 20	20	23.1	23.1	35.6	35.6
20-40	20	27.9	51.0	32.1	67.7
40-60	20	24.8	75.8	16.4	84.1
60-80	20	18.6	94.4	13.7	97.8
80 plus	–	5.6	100.0	2.2	100.0
Total		100.0		100.0	

Using the percentage distributions in Table 7, the median of income for volunteers is in the interval from 20 to 40 thousand dollars. Interpolating in this interval, the median is

$$\begin{aligned} P_{50} &= 20 + \left(\frac{50 - 23.1}{27.9} \times 20\right) \\ &= 20 + (26.9/27.9 \times 20) \\ &= 20 + (0.964 \times 20) \\ &= 20 + 19.28 \\ &= 39.28 \end{aligned}$$

or \$39,300.

$$\begin{aligned} P_{50} &= 20 + \left(\frac{50 - 35.6}{32.1} \times 20\right) \\ &= 20 + (14.4/32.1 \times 20) \\ &= 20 + (0.449 \times 20) \\ &= 20 + 8.97 \\ &= 28.97 \end{aligned}$$

or \$29,000.

For the volunteers, the seventy-fifth percentile is in the interval from 40 to 60 thousand dollars, since there are only 51.0% of cases with incomes of 40 thousand or less but when all those with incomes of 40-60 thousand are included, there are 75.8% of cases.

$$\begin{aligned} P_{75} &= 40 + \left(\frac{75 - 51.0}{24.8} \times 20\right) \\ &= 40 + (24.0/24.8 \times 20) \\ &= 40 + (0.968 \times 20) \\ &= 40 + 19.355 \\ &= 59.355 \end{aligned}$$

or \$59,400.

For non-volunteers, the seventy-fifth percentile is in the same 40 to 60 thousand dollar interval, since there are only 67.7% of cases with incomes of 40 thousand or less but when all those with incomes of 40-60 thousand are included, there are 84.1% of cases.

$$\begin{aligned} P_{75} &= 40 + \left(\frac{75 - 67.7}{16.4} \times 20\right) \\ &= 40 + (7.3/16.4 \times 20) \\ &= 40 + (0.445 \times 20) \\ &= 40 + 8.902 \\ &= 48.902 \end{aligned}$$

or \$48,900.

Comparison of distributions. In order to assist in comparing the two distributions, the statistics calculated in the question are collected together in Table 8.

Table 8: Summary statistics of income for volunteers and non-volunteers, in thousands of dollars

Statistic	Volunteers	Non-volunteers
Mean	41.7	33.2
Median	39.3	29.0
P_{75}	59.4	48.9

Probably the main point to be made about the two distributions is that non-volunteers generally report lower income than do volunteers. In the original distributions in Table 4 of Problem Set 2, the non-volunteers tended to be clustered at incomes under forty thousand dollars. About one-half of volunteers have incomes below this but about two-thirds of non-volunteers are at these lower incomes. As a result, it is no surprise that the mean, median, and seventy-fifth percentile of income of non-volunteers are each well below the level of the corresponding statistic for volunteers. For the percentile measures, the difference is about ten thousand dollars (39.3-29.0=10.3 and 59.4-48.9=10.5), while the difference is eight and nine thousand dollars for the mean (41.7-33.2=8.5).

The major similarity between the two distributions is that both groups cluster more at the low to middle incomes than at the upper incomes. There is only a small proportion of respondents in either group who report incomes of eighty thousand dollars or more. For each group, the mode is under forty thousand dollars.

In summary, each distribution peaks at a low to middle income, with fewer respondents at the upper income levels. But the distribution of income for non-volunteers is more concentrated at te very lowest incomes, while the distribution for volunteers is more concentrated around the middle incomes.

5. **Averages**

(a) The 'average' student could be the mode. Since the description is fairly general, with little indication of what characteristics belong to this average student, it may be that the author means that more students fit this category of not being athletically inclined than any other category. If the author surveyed students and found that the mean hours wasted were 10 hours each week, then the average student might represent this mean of ten. As for whether socializing is wasting time, that is a value judgment that may not be justified. I'd think this average is either the mode or median.

(b) Again, the 'average Joe' is presumably some sort of conception of what an average person might be like. Since this is not very well defined, this is most likely to be the mode. But in he quote concerning technical skills in the Crowns, the average might be either the median or mean. Whether the author is correct or not, the claim appears to be that the median or mean skill level of many ordinary workers is well below that of those with high technical skills. Again, this is a highly judgmental statement, with no proof provided and not much indication of what skill types and levels are being considered. Ordinary Joes might be just as skilled as those with more technical skills, it is just that the skill set is different.

(c) Marriage as the only acceptable social institution would be the mode. That is, the variable would be type of social institution, a variable that is no more than a classification. The mode would thus have to be used. The claim here is that all or most couple would have been married at some time in the past.

(d) The later union formation could be the modal age of forming unions, but is more likely the median or mean. People form unions at many different ages, so the mode is not so distinct, and would not usually be used to measure this. The author could be referring to the median, arguing that the age by which one-half of couples are formed is later than in earlier years. Or the mean age of forming unions could be greater than in earlier years.