Social Studies 201

Answers to Problem Set 4

November 9, 2004

1. **Religiosity and volunteer work**

   (a) The probability of selecting someone who performs less than 30 hours of volunteer work is 0.302.

   $$P(\text{less than 30 hours}) = \frac{\text{no. less than 30 hours}}{\text{total no. of cases}} = \frac{294}{975} = 0.302$$

   The probability of selecting someone who is very religious is 0.192.

   $$P(\text{very religious}) = \frac{\text{no. very religious}}{\text{total no. of cases}} = \frac{187}{975} = 0.192$$

   (b) The chance of selecting an individual who is somewhat religious and performs 30-99 hours of volunteer work is 0.195.

   $$\frac{\text{no. somewhat religious and no. 30-99}}{\text{total no. of cases}} = \frac{190}{975} = 0.195$$

   (c) The chance of selecting someone who is not very religious (NVR) or not at all religious (NR) is 0.240.

   $$P(\text{NVR}) + P(\text{NR}) - P(\text{NVR and NR}) = \frac{158}{975} + \frac{76}{975} - \frac{0}{975} = \frac{234}{975} = 0.240$$

   (d) The chance of selecting someone who is not very religious (NVR) or volunteers for 30 to 99 hours per year is 0.423.

   $$P(\text{NVR}) + P(\text{30-99}) - P(\text{NVR and 30-99}) = \frac{158}{975} + \frac{301}{975} - \frac{47}{975} = \frac{412}{975} = 0.423$$

   (e) The probability of volunteering less than 30 hours, given very religious, is 0.257.

   $$P(\text{less than 30 hours/very religious}) = \frac{48}{187} = 0.257$$

The probability that the individual volunteers less than 30 hours, given not at all religious is 0.408.

$$P(\text{less than 30/not at all religious}) = \frac{31}{76} = 0.408$$

(f) If A is the event of volunteering for 100 plus hours and B is the event of being very religious, one way of checking to see whether the two events are independent or not is to see whether $P(A/B)$ and $P(A)$ are equal. These probabilities are

$$P(A/B) = \frac{96}{187} = 0.513$$

$$P(A) = \frac{380}{975} = 0.390$$

These two probabilities differ by 0.123, so are not real close to each other. As a result, these two events can be considered dependent events.

Alternatively, a check to see whether $P(B/A)$ and $P(B)$ are equal is another way to examine this. These probabilities are

$$P(B/A) = \frac{96}{380} = 0.253$$

$$P(B) = \frac{187}{975} = 0.192$$

Again there is a considerable difference between these two probabilities so the events of being very religious and volunteering for 100 plus hours are dependent events.

(g) It appears that the very religious volunteer more than those who are less religious. From the first part of (f), the event of volunteering for 100 plus hours, given very religious, is 0.513, considerably greater than 0.390, the overall probability of volunteering for 100 plus hours. Thus those who consider themselves very religious state that they volunteer more than average.

Similarly, from (e), the probility of volunteering for less than 30 hours, the lowest category of volunteering, is greater if the person

is not religious (0.408). In contrast, the very religious are less likely to be in this lowest category of volunteering.

From Table 1, suppose the conditional probabilities of volunteering for 100 plus hours, given each of the categories of religiosity, are calculated. The conditional probability is greatest for the very religious ($96/187 = 0.513$) and less for each of the other categories of religiosity ($212/554 = 0.383$, $48/158 = 0.304$, and $24/76 = 0.316$).

From these considerations, a conclusion that those who regard themselves as very religious tend to volunteer more than those who are in the other categories of religiosity.

2. **Probability statements**.

There are several pairs of dependent events. Some of the dependent pairs and independent pairs are as follows. are as follows.

- The event of poor interpersonal relations and the event of stress could be considered to be dependent events for the respondents who say these are related. Apparently not all respondents, however, considered these to be related, so they would be independent for those respondents who said there is no relation between them.
- The event of poor relations being a source of stress and the event of type of industry or occupation of work of employees could be dependent events. That is, fewer primary industry workers felt this was a cause of stress than for workers in health industries. The event of being in a health industry appears to raise the probability of stress and the event of being in a primary industry appears to lower the cause of stress.
- The last sentence in the first quote implies independence of occupation and stress for most occupations. That is, the likelihood does not vary much from the average for most occupations, so for these unstated occupations, poor relations causing stress appears to be independent of the occupations.
- The event of excellent health appears to have lower chance of occuring given the event of being in Canada a long time. This means that these two events are dependent on each other.

- The final quote implies an independence of health status and immigrant/non-immigrant status (IM), when the immigrants have been in Canada a long time. That is, for immigrants who have been in Canada for a long time the probability of a specific health status given the immigrant status is equal to the overall probability of this health status.

$$P(\text{specific health status/IM}) = P(\text{specific health status})$$

3. Standardized normal distribution.

   (a) For $Z = 1.75$, the A area if 0.4599 and this is the area between $Z = 0$ and $Z = 1.75$.

   (b) The area between 0.7 and 1.8 is the area between the centre and $Z = 1.8$ (an A area), minus the area between the centre of the normal distribution and $Z = 0.70$ (also an A area). The area between the centre of the normal distribution and $Z = 1.80$ is 0.4641. The area between centre and $Z = 0.70$ is 0.2580. The area between $Z = 0.70$ and $Z = 1.80$ is thus $0.4641 - 0.2580 = 0.2061$.

   (c) The proportion of cases between $Z = -0.80$ and $Z = 1.50$ is the sum of the areas between the centre of the distribution and each of these Z values. Between $Z = 0$ and $Z = -0.80$, the area is 0.2881 while the area between $Z = 0$ and $Z = 1570$ is 0.4332. The sum of these two areas is $0.2881 + 0.4332 = 0.7203$. As a percentage of the total area, this is $0.7203 \times 100 = 72.03\%$.

   (d) The area to the right of $Z = -0.87$ is the A area associated with $Z = 0.87$ (from -0.87 to centre) plus the one-half of the area to the right of centre. For $Z = 0.87$, the A area is 0.3079. The required proportion is $0.3079 + 0.5000 = 0.8079$.

   (e) The area under the normal curve below $Z = 1.27$ is the one-half of the area to the left of centre, plus the area between centre and $Z = 1.27$. For $Z = 1.27$, the A area is 0.3980 and thus the required area is $0.5000 + 0.3980 = 0.8980$.

   (f) One-half standard deviation is associated with a Z of 0.5. Required is the area under the curve below $Z = -0.50$ and above $Z = 0.50$. For $Z = 0.50$, the B area is 0.3085, and this is the area

above $Z = 0.50$. By symmetry, the area below $Z = -0.50$ is also 0.3085. As a result, the area beyond one-half standard deviation from the mean is $0.3085 + 0.3085 = 0.6170$, or $0.6170 \times 100 = 61.7\%$.

(g) For an area of 0.40 in the left tail of the distribution, it is necessary to find a B area as close as possible to 0.4000. This corresponds to $Z = -0.25$, with an area of 0.3013 below this. The fortieth percentile thus occurs at $Z = -0.25$.

(h) If there is 0.0425 of the area in one tail of the distribution (B area), the appropriate Z value associated with this is $Z = 1.72$. For this Z, the B area is 0.0427, just over 0.0424. The required Z values are thus -1.72 and +1.72, with $2 \times 0.427 = 0.0854$ in the two tails, just a little more than the 0.085 specified.

(i) The seventieth percentile occurs at the Z-value such that 0.7000 of the area is below this and the other 0.3000 is above this. The easiest way to find this is to look for a B area of 0.3000, and this occurs at $Z = 0.52$ or $Z = 0.53$. Below this Z there is 0.5000 of the the area to the left of centre plus another 0.2000 between centre and this Z.

(j) If 10%, or 0.1000 of the area, is to be deleted from each end of the distribution, look for a B area of 0.1000. The closest Z is at 1.28. The trimming points for this trimmed mean are at $Z = -1.28$ and $Z = +1.28$.

4. **Annual hours worked**. For this question $\mu = 1,420$ and $\sigma = 730$.

(a) If the distribution is assumed to be normally distributed with mean $\mu = 1,420$ and standard deviation $\sigma = 730$, the results are as follows.

i. For $X = 2,250$, $Z = (X - \mu)/\sigma = (2,250 - 1,420)/730 = 830/730 = 1.14$. The proportion who work more than 2,250 hours annually is the B area for $Z = 1.14$, and this is 0.1271.

ii. For $X = 750$ hours, $Z = (X - \mu)/\sigma = (750 - 1,420)/730 = -670/730 = -0.92$. The proportion who work more than 750 hours annually is the B area for $Z = -0.92$, and this is 0.1788. As a percentage, this is $0.1788 \times 100 = 17.88\%$.

iii. The area between 750 and 1,750 is the area from 750 to centre plus the area from centre to 1,750. For $X = 750$, $Z = -0.92$ from (ii), and the area between this and centre is 0.3212. For $X = 1,750$ hours, $Z = (X - \mu)/\sigma = (1,750 - 1,420)/730 = 330/730 = 0.45$. The area between this and centre is 0.1736. The total area is the sum of these two areas, or $0.3212 + 0.1736 = 0.4948$. If there are 502 individuals, this means there are $502 \times 0.4948 = 248.3$ or 248 individuals who work between 750 and 1,750 hours annually.

(b) Comparison of the actual and normal distribution. From the diagram of the normal distribution superimposed on the histogram of hours worked by 15-24 year olds, it is apparent that the actual distribution is not very close to normally distributed. As compared with a normal distribution, there are more 15-24 year olds at a small number of hours worked, fewer at the middle number of hours worked, and more at around 2000 hours. It may be that the actual distribution is fairly close to the normal distribution at the upper end.

The following table, using the data from part a., demonstrates this as well. In order to make the two distributions comparable, all areas, proportions, and numbers of cases have been converted into percentages.

Annual hours	Normal %	Actual %
<750	17.8%	25.7%
750-1,750	49.5%	36.7%
2,250 plus	12.7%	11.4%

At the lower end, less than 750 hours worked at jobs annually, there would be 17.8% of cases if the distribution were exactly normally distributed. In fact, there are 25.7% who work this few hours, approximately eight percentage points more that indicated by a normal distribution. As seen in the diagram, at the middle levels, there would be more cases (almost 50%) from 750 to 1,750 if the distribution were normal – in fact, there are only 36.7%

in this interval. Finally, the actual distribution and the normal distribution are fairly similar at 2,250 hours plus.

In conclusion, while the actual distribution of hours worked at jobs annually may be similar to a normal distribution at the upper end, the rest of the actual distribution is very different than a normal distribution.

5. Computer problems

a. Is pay normally distributed?

Statistics

	N		Mean	Std. Deviation
	Valid	Missing		
PAY Hourly pay in dollars	384	323	9.7200	5.4286

Only the relevant section of the frequency distribution is included here – see explanation below.

PAY Hourly pay in dollars

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	1	.1	.3	.3
	3.00	2	.3	.5	.8
	3.75	1	.1	.3	1.0
	5.00	4	.6	1.0	2.1
	5.50	1	.1	.3	2.3
	5.55	1	.1	.3	2.6
	5.60	30	4.2	7.8	10.4
	14.84	1	.1	.3	85.7
	15.00	15	2.1	3.9	89.6
	15.50	1	.1	.3	89.8
	16.00	2	.3	.5	90.4
	16.50	1	.1	.3	90.6
	17.00	6	.8	1.6	92.2
	18.00	4	.6	1.0	93.2
	18.50	1	.1	.3	93.5
	18.83	1	.1	.3	93.8
	19.00	2	.3	.5	94.3
	19.75	2	.3	.5	94.8
	20.00	6	.8	1.6	96.4
	22.00	3	.4	.8	97.1
	22.50	2	.3	.5	97.7
	25.00	3	.4	.8	98.4
	28.00	2	.3	.5	99.0
	30.00	1	.1	.3	99.2
	40.00	1	.1	.3	99.5
	45.00	2	.3	.5	100.0
	Total	384	54.3	100.0	

For the variable PAY, the mean is $9.72 and the standard deviation is $5.43. The intervals within one and two standard deviations of the mean are:

1 s.d. From 9.72 - 5.43 = 4.29 to 9.72 + 5.43 = 15.15 or (4.29 , 15.15).

2 s.d. is $5.4286 x 2 = 10.86. 9.72 - 10.86 is less than zero, so 0 is the lower end point of this interval, since pay cannot be less than 0. 9.72 + 10.86 = 20.58. Thus the two standard deviations from the mean interval is the interval (0 , 20.58).

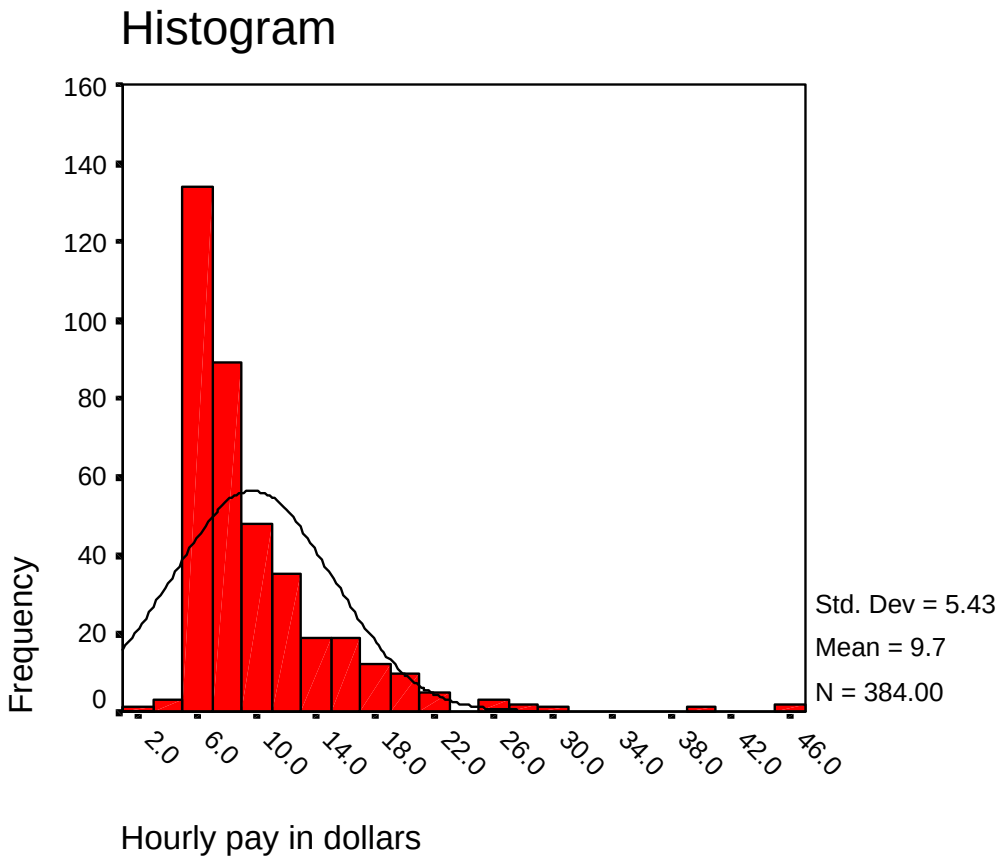
Obtaining the percentage of cases within each of these intervals involves examining the frequency distribution, counting the number of cases within each of these intervals, and multplying these by 100 to obtain the percentage. Alternatively, the cumulative percentages can be used as follows.

For the one standard deviation interval, there is a cumulative percentage of 89.6% of cases with PAY less than $15.15. At the lower end of the interval, there is only 1% of the cases with PAY less than $4.29. As a result, there is 89.6% - 1% = 88.6% of cases within the interval from $4.29 to $15.15. From the table of the normal curve, there are 0.3413 of the cases between Z = 0 and Z = 1, so all together there are 0.3413 + 0.3413 = 0.6826 or 68.26% of the cases within one standard deviation of the mean for the normal distribution. There is a considerably larger percentage of cases within one standard deviation of the mean in the actual frequency distribution of pay. As a result, the actual distribution of pay is not normally distributed over the interval within one standard deviation of the mean. The actual distribution of pay is much more concentrated around the centre than in the case of a normal distribution. This can be seen in the histogram, where the bars are more concentrated close to the mean than in the case of the normal distribution curve.

For two standard deviations, the interval is from 0 to $20.58. The cumulative percentage of cases up to $20.00, just below $20.58, is 96.4% of cases. From the normal distribution, there are 0.4772 of cases between Z = 0 and Z = 2. By symmetry, there are 0.4772 + 0.4772 = 0.9544, or 95.44% of the cases within two standard deviations of the mean in a normal distribution. This is similar to the 96.4% of cases within two standard deviations of the mean in the actual frequency distribution. While the distribution of cases is not spread out like a normal distribution over this two standard deviation interval, the percentage of cases within two standard deviations is very similar for the normal and actual distribution.

From the histogram, it can be seen that there are a lot of cases close to the mean (tall bars) but fewer cases between about $8 and $20 (short bars), so the greater concentration close to the mean and the lesser concentration at

$8 to $20 balance out to produce the above result. But from the normal curve superimposed on the actual histogram, it is clear that pay is not really normally distributed.



5. b. Probabilities from a crosstabulation.

Using the crosstabulation table below, the probabilities are as follows.

i. P(somewhat important or very important)
= P(somewhat important) + P(very important). Since these two events do not overlap, they are mutually exclusive, so the AND probability for the two events is 0.

P (somewhat important) = 291/682
P(very important) = 200/682.

The required probability is 291/682 + 200/682 = 491/682 = 0.720.

ii. P(4 or 5) = P(4) + P(5) – P(4 and 5), but the latter part is again 0.

P(4) = 158/682
P(5) = 122/682
P(4 or 5) = 158/682 + 122/682 = 280/682 = 0.411.

iii. P(5 / very important) = 27/200 = 0.135.
P(5 / not at all important) = 18/62=0.290.

iv. P(strongly disagree) = 139/682 = 0.204.
P(strongly disagree/very important) = 74/200 = 0.370.
Since these two probabilities are quite different, these two events are dependent on each other.

v. It is sometimes possible to find two events that are quite close to being independent of each other. When the variables have at least an ordinal scale of measurement, such a pair of events is sometimes found near the centre of the table.

Try the events of having a neutral response (3) on V4 and regarding religious and spiritual values as somewhat important.

P(3) = 177/682 = 0.260.
P(3/somewhat important) = 80/291 = 0.275.

These two probabilities are fairly close to each other, so the events of being neutral on the gay-lesbian question and regarding religious or spiritual values as somewhat important are very close to being independent of each other.

'4 Gay and Lesbians Married * VALUES Importance of Spiritual Values Crosstabulation

Count

		VALUES Importance of Spiritual Values				
		1 Not at all important	2 Of little importance	3 Somewhat important	4 Very important	Total
V4 Gay and Lesbians Married	1 Strongly Disagree	3	19	43	74	139
	2	8	12	38	28	86
	3	19	36	80	42	177
	4	14	38	77	29	158
	5 Strongly Agree	18	24	53	27	122
Total		62	129	291	200	682