Social Studies 201

October 4, 2006

Positional Measures – Percentiles.

See text, section 5.6, pp. 208-213.

Note: The examples in these notes may be different than used in class on October 4, 2006. However, the examples are similar and the methods used are identical to what was presented in class.

Introduction

In this section of the notes positional measures are discussed. Positional measures are termed percentiles and indicate the per cent of cases located either below or above a specific value of the variable X. These are useful measures in describing how a distribution unfolds, from lowest to highest values. Rather than being restricted to the centre of the distribution, as are measures of centrality, percentiles indicate the position associated with a particular percentage of cases – any one of the possible one hundred percentages can be indicated.

The median, at the fifty per cent point, is a percentile, the fiftieth percentile. That is, the median is the value of the variable such that one-half of the cases are less than the median and the other half are greater than or equal to the median.

But there are ninety-nine other perctiles. "Per cent" means per one hundred, and in total, there are one hundred percentiles, corresponding to each possible per cent of cases. In practice, not all one hundred percentiles are obtained, but the method illustrated in this section is general in that it can be used to determine any of the percentiles.

> **Percentiles – definition and notation**. The rth percentile of a distribution is the value of the variable X, such that r per cent of cases are less than or equal to this value, and the other $(100-r)$ per cent of cases are greater than or equal to this value.
>
> In these notes, the rth percentile is given the symbol P_r, so that r per cent of cases are less than or equal to P_r and the remaining $(100-r)$ per cent of cases are greater than or equal to P_r.

Percentiles are most readily obtained from the cumulative percentage distribution and using a formula similar to that used for determining the median.

Before examining methods of calculating percentiles, these notes discuss various aspects of percentiles and some examples of how they are used.

1. Recall that there are one hundred per cent of cases in a distribution. Cumulative percentages begin with zero per cent of cases at the smallest value of the variable and proceed to one hundred per cent of cases at the largest value of the variable. Percentiles are particular values of the variable between these lower and upper limits. They indicate the value of the variable such that a specific percentage of cases is less than or equal to this value.

2. In order to determine percentiles, a variable must have at least an ordinal level of measurement. Interval and ratio level measurements are ordinal, so percentiles can be obtained for variables with these higher levels of measurement. But percentiles cannot be obtained for variables that have no more than a nominal scale of measurement, variables such as sex, ethnicity, or political party supported.

3. Various positional measures are used in statistical analysis. All of these are essentially percentiles, but there are several variants. The positional measures most commonly used are the following.

   - **Percentiles**. Values of a variable X dividing the cases into one hundred equal parts. The values of X are $P_0, P_1, P_2, ..., P_{99}, P_{100}$. One per cent of the the cases are between each of these values.
   - **Deciles**. Values of a variable X dividing the cases into ten equal parts. The values of X are $P_0, P_{10}, P_{20}, ..., P_{90}, P_{100}$. Ten per cent of the the cases are between each of these values.
   - **Quintiles**. Values of a variable X dividing the cases into five equal parts. The values of X are $P_0, P_{20}, P_{40}, P_{60}, P_{80}, P_{100}$. Twenty per cent of the the cases are between each of these values.
   - **Quartiles**. Values of a variable X dividing the cases into four equal parts. The values of X are $P_0, P_{25}, P_{50}, P_{75}, P_{100}$. Twenty-five per cent of the the cases are between each of these values.

- **Median**. Value of a variable X dividing the cases into two equal parts. The median value of X is P_{50} and fifty per cent of cases are less than or equal to P_{50} and the other fifty per cent of cases are greater than or equal to P_{50}.

Example of percentiles in a hypothetical distribution

A simple, hypothetical distribution is illustrated in Table 1, along with its corresponding histogram in Figure 1. The position of several percentiles is indicated in the figure.

Table 1: Hypothetical percentage and cumulative percentage distribution to illustrate percentiles

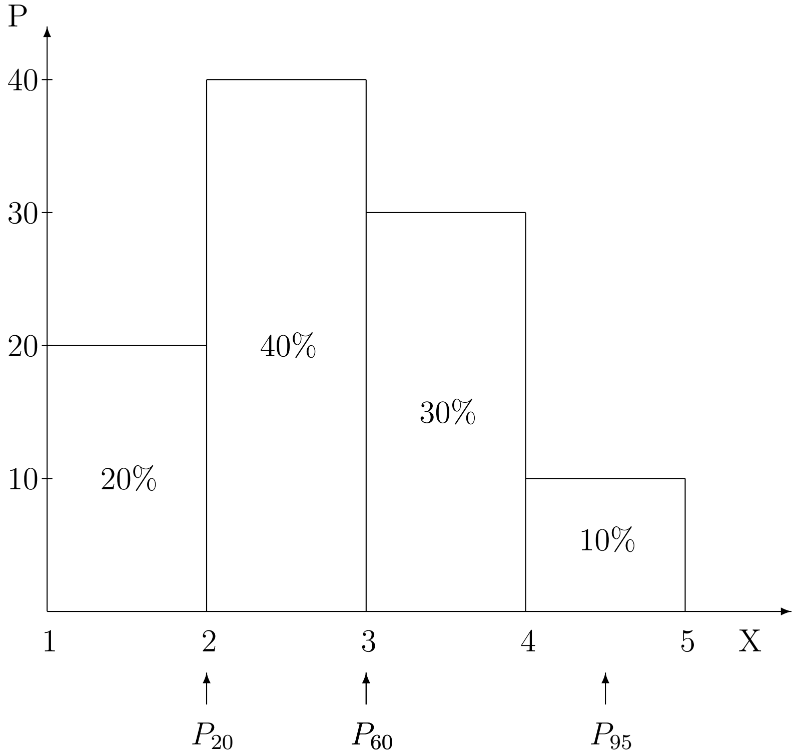
	Per	Cumulative
X	cent	per cent
1-2	20	20
2-3	40	60
3-4	30	90
4-5	10	100
Total	100	

This distribution has a variable X with 1 as the smallest value and 5 as the largest value. Since the intervals are each of equal width, one unit of the variable, the percentage distribution in Figure 1 is a histogram. The percentage of cases in each bar of the histogram is indicated in the middle of each bar.

The twentieth percentile is at $X = 2$, that is, $P_{20} = 2$. From the table or the figure, it can be seen that the first twenty per cent of cases in this distribution are encountered when $X = 2$, since there are twenty per cent of cases with a value less than or equal to 2. The other $100 - 20 = 80$ per cent of cases have values greater than or equal to 2.

Similarly, it can be seen that the sixtieth percentile is at $X = 3$, since that is where the cumulative percentage distribution first reaches sixty per cent. That is, there are twenty per cent of cases less than or equal to 2, and another forty per cent between 2 and 3, so the sixty per cent point occurs at $X = 3$. Thus $P_{60} = 3$.

Figure 1: Histogram and percentiles for hypothetical distribution



It may be a little more difficult to determine P_{95}, the ninety-fifth percentile. At $X = 4$, ninety per cent of cases have been accounted for and by $X = 5$, one hundred per cent of cases have been encountered. Assuming the cases are uniformly spread across the interval from 4 to 5, the ninety-fifth percentile is half way along this interval since ninety-five is exactly half way between ninety and one hundred per cent. As a result, the ninety-fifth percentile is at $X = 4.5$. That is, $P_{95} = 4.5$, as indicated in Figure 1.

Other percentiles can be determined in a similar manner. The notes in this section present procedures for obtaining any percentile, using the method of interpolation.

Example of how percentiles are used

- **Measuring literacy skills**. Percentiles are often used to indicate where an individual stands, relative to others, on a test or how an individual or family compares with others on a social science scale such as status or prestige. Those with a higher percentile score more highly than those with a lower percentile. An example concerning literacy skills illustrates this.

  A July 2004 report from Statistics Canada concerning literacy skills in Canada, and across other industrialized countries (OECD countries), found that western and central Canadian provinces rated well but literacy skills were lower in Atlantic Canada. The test was conducted among a sample of 15-year olds. There was considerable variation in literacy skills both among and within provinces. Skills were measured using the Programme for International Student Assessment (PISA), a standardized measure of literacy skills. Scores on the PISA test range from approximately 300 to 700, with the test structured to have a mean score of 500 across the countries participating.

  According to the report (p. 7),

  > The scores on the reading, mathematics, and science assessment for PISA were scaled to have a mean of 500 and a standard deviation of 100 for students from all OECD countries participating in the study. As a way to interpret the magnitude of one-point on the PISA in substantive terms, consider two hypothetical scores on the reading test of 500 and 515, respectively. The student with a score of 500 would be near the 50th percentile among all OECD students, while the student with a 15-point advantage would be at about the 56th percentile.

  That is, about one-half of all those taking the test had a score of 500 or below and fifty-six per cent of those taking the test scored about 515 or below. From other information provided in the report, the mean score for all Canadian fifteen-year olds was 529, twelve points below the score for Finland, the country with the highest rating. Saskatchewan respondents had a mean of 529, about the same as respondents in Australia.

One finding of the study is that literacy skills increase with socioeconomic status (SES). That is, respondents coming from disadvantaged socioeconomic background (low SES) score lower on PISA than do respondents from more advantaged socioeconomic background (high SES). In particular, it is noted in the report (p. 18),

> despite Canada's relative success with less advantaged students, there is a large performance gap between students from low and high SES backgrounds. A typical student at the 5th percentile scored approximately 479, while a typical student at the 95th percentile scored approximately 580 – a difference of about 100 points.

In this finding, percentiles are used to indicate relative socioeconomic status. The 5th percentile represents students from the lower end of the SES spectrum, a disadvantaged level – they score an average of only 479. In contrast, students from the top five per cent of SES (95th percentile) score one hundred points higher.

> All information in this example taken from J.Douglas Willms, *Variation in Literacy Skills Among Canadian Provinces: Finding from the OECD PISA*, Statistics Canada catalogue number 81-595-MIE, No. 0012, Ottawa, 2004. Available at the Statistics Canada web site:
>
> http://www.statcan.ca/Daily/English/040714/d040714b.htm

Example of how percentiles are used

- **Income Distributions**. Another situation where percentiles are commonly used is when examining distributions of income. Rather than listing all percentiles, reports of income distributions typically list income quintiles (the 20th, 40th, 60th, and 80th percentiles). Table 2 gives such a list for individual incomes of Canadian adults in 2002. The data for this table is adapted from Table 202-0604, Statistics Canada, *Income Trends in Canada 1980-2002*, catalogue number 13F0022XCB, 2004.

Table 2: Quintiles and percentiles for distribution of after-tax income of economic families with two persons or more Canada, 2002; and shares, 1996 and 2002

	Quintile value	Share of total income (%)	
Quintile (percentile)	in 2002 dollars	1996	2002
Lowest quintile (P_{20})	32,000	7.4	7.4
Second quintile (P_{40})	45,500	13.1	12.9
Third quintile (P_{60})	60,800	18.1	17.7
Fourth quintile (P_{80})	81,900	23.8	23.5
Highest quintile		37.6	38.5

From Table 2, the first quintile is at \$32,000. This means there are one-fifth, or twenty per cent, of Canadian families with after-tax incomes at or below \$32,000 in 2002. That is, $P_{20} = \$32,000$. The second quintile, or fortieth percentile, is $P_{40} = \$45,500$. That is, there are forty per cent of families with after-tax incomes less than or equal to this. Also, there are twenty per cent of families with after-tax incomes between \$32,000 and \$45,500. Similarly, the sixtieth percentile occurs at \$60,800 and the eightieth at \$81,900. This latter percentile is the income such that only twenty per cent of families have after-tax incomes higher than this. That is, the twenty per cent of families with the highest incomes have incomes from \$81,900 to the very highest income. The one-hundredth percentile does not occur until the very highest income, probably several million dollars, a value that Statistics Canada does not report.

Statistics Canada, in May 2004, reported on changes in the distribution of family income for the period ending in 2002. Part of this report stated

> By expressing the income of each quintile as a share of the income of all families, we concentrate on relative changes

> among quintiles. Any increase to a particular quintile is necessarily a decrease for some other quintiles. How did the shares of income received by lower-, middle-, and higher-income families evolve in the last several years?
>
> There was a very small and gradual shift in favour of the highest quintile families from 1996 to 1998, as their share of after-tax income rose from 38% to 39%. Their share did not fluctuate between 1998 and 2002, at an average of 39%. Any changes in the shares of market income were even less evident over the period from 1996 to 2002.
>
> From Statistics Canada " Analysis of income in Canada," Chapter VII, "Income inequality in relative terms" section of chapter, available at web site:
>
> http://www.statcan.ca/Daily/English/040520/d040520b.htm

In this Statistics Canada report, the share of total income received by each quintile permits an analysis of changes in income inequality. For 1996 and 2002, these shares are given in the last two columns of Table 2. As reported in the above quote and Table 2, the percentage of total income received by the top quintile, the best-off twenty per cent of families, was 39 per cent in 2002, up slighlty from 38 per cent in 1996. The share of income for the poorest one-fifth, that is the bottom quintile, remained unchanged. However, from Table 2, it was the middle quintiles that suffered a decline in their share of income. In summary, over the 1996-2002 period, there was a shift of after-tax income from middle to higher income families.

Percentiles for grouped data using cumulative percentages

In order to determine percentiles, it is generally best to begin by constructing the cumulative percentage distribution. From the cumulative percentage distribution, the category containing the desired percentile can be readily determined. Where the data are numerical and organized into intervals, the method of interpolation is then used to determine an exact value for the percentile.

Percentiles for grouped data – discrete variable

See text, section 5.6.2, pp. 209-210.

The rth percentile of the variable is the value of the variable such that r per cent of cases are less than this value and the other $(100 - r)$ per cent of cases are greater than or equal to this value. The rth percentile is usually given the notation P_r.

Using the cumulative percentage distribution, the rth percentile, P_r, is in the category or interval where a cumulative percentage of r per cent or more is first reached. This category or interval then has r% of cases less than or equal to P_r, and the other $(100 - r)$% of cases greater than or equal to P_r.

> **Example – attitudes to treating gays and lesbians as married**. A sample of just under seven hundred undergraduate students was asked to state their view about the statement, "Tax laws and job benefits should recognize gay and lesbian couples as married." Respondents stated their degree of agreement or disagreement about this statement on a five-point scale, from 1, indicating strongly disagree with the statement, to 5, indicating strongly agree with the statement. Responses are reported in Table 3. The frequency, percentage, and cumulative percentages are all given in this table.
>
> Obtain the fifteenth, fiftieth, sixty-eighth, and ninety-second percentiles.
>
> **Answer**. The variable is attitude or opinion concerning the gay and lesbian issue, with responses measured on a 1-5 scale. Responses are ordered from strongly disagree to strongly agree, so this is an ordinal scale and percentiles can be meaningfully obtained.

Table 3: Frequency, percentage, and cumulative percentage distributions of views of undergraduate student respondents to "Gays and lesbians married"

Response (X)	Number	Per cent	Cumulative per cent
1 (strongly disagree)	141	20.3	20.3
2 (somewhat disagree)	86	12.4	32.7
3 (neutral)	183	26.3	59.0
4 (somewhat agree)	161	23.2	82.2
5 (strongly agree)	124	17.8	100.0
Total	695	100.0	

The fifteenth percentile, P_{15}, is at category $X = 1$, or strongly disagree. This first category where $X = 1$ contains just over twenty per cent of cases (20.3%), and this is more than the required fifteen per cent. That is, the fifteen per cent of cases with the smallest values of the opinion variable are response 1. Thus $P_{15} = 1$, or strongly disagree.

The fiftieth percentile, P_{50}, is also the median value, and this occurs at $X = 3$, or a neutral response. At $X = 1$ and $X = 2$, there are only 32.7% of responses. But when all the respondents who answer 1, 2, or 3 are included, the cumulative per cent of cases is 59.0%. As a result, the fiftieth percentile, or median, is a response 3, or neutral.

The sixty-eighth percentile is at response 4, or somewhat agree. This is the response at which the cumulative percentage first reaches 68% of cases, so $P_{68} = 4$ or somewhat agree.

$P_{92} = 5$, or strongly agree. At $X = 4$, the cumulative percentage reaches only 82.2 per cent. Ninety-two per cent of responses are not encountered until those with a response of $X = 5$ are included. The ninety-second percentile is thus 5 or strongly agree.

Interpolation to obtain percentiles. See text, section 5.6.3, pp. 211-214

For data where the variable is numeric and grouped into intervals, the exact value of the percentile is obtained by interpolating across the interval containing the percentile. The cumulative percentage is used to identify the interval containing the percentile. Then the following formula is used to determine a more exact value of the variable for the percentile (text, p. 211). The rth percentile is

$$P_r = \begin{array}{l}\text{Value of variable}\\ \text{at lower end of}\\ \text{interval}\end{array} + \left[\frac{r - \begin{array}{c}\text{Cumulative per cent}\\ \text{at lower end of interval}\end{array}}{\text{Per cent of cases in interval}}\right]\begin{bmatrix}\text{Interval}\\ \text{width}\end{bmatrix}$$

This is the same interpolation procedure and formula used when obtaining the median. The only difference is that in the numerator of the fraction in the large bracket, the particular percentile r replaces the value of 50 used in the formula for the median.

The formula presented here is more general than the formula for the median and applyies to all one hundred percentiles. Since the median is a particular percentile, the fiftieth percentile, the formula here can be used to obtain the median, by merely making $r = 50$.

Examples of how to obtain the median follow.

> **Example – grade point averages**. Use the distribution of grade point averages of five hundred and seventy three students in Table 4 to obtain the twentieth, and eighty-fifth percentiles.

From Table 4, the twentieth percentile occurs in the interval 65-70, where the cumulative percentages first reach twenty per cent. That is, there are only 9.4% of students with grades of 65 or less, and another 16.2% of students with grades between 65 and 70. As a result, the cumulative percentage reaches 25.6% when all students with grades of 70 or less are included.

Interpolating within the interval from 65 to 70, the twentieth percentile is

$$P_{20} = 65 + \left(\frac{20 - 9.4}{16.2} \times 5\right)$$

Table 4: Frequency, percentage, and cumulative percentage distributions of grade point averages

Grade	Number	Per cent	Cumulative per cent
Less than 60	14	2.4	2.4
60-65	40	7.0	9.4
65-70	93	16.2	25.6
70-75	146	25.5	51.1
75-80	138	24.1	75.2
80-85	119	20.8	96.0
85 plus	23	4.0	100.0
Total	573	100.0	

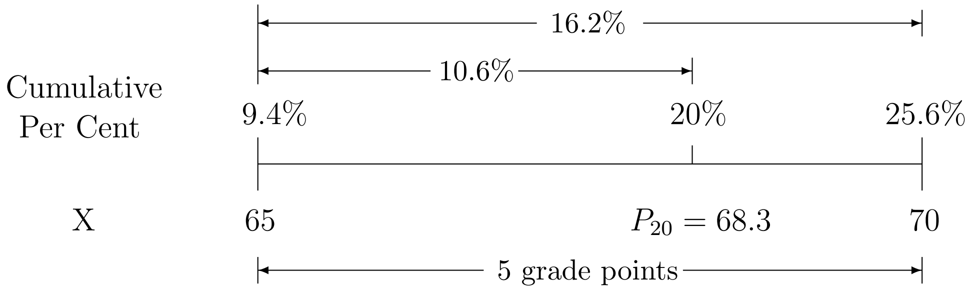
$$\begin{aligned}
&= 65 + \left(\frac{10.6}{16.2} \times 5\right) \\
&= 65 + (0.654 \times 5) \\
&= 65 + 3.27 \\
&= 68.27.
\end{aligned}$$

That is, P_{20} is to the right of 65, by the fraction of the distance along the interval required to reach the twenty per cent point. This fraction is the per cent from 9.4 to 20, as a proportion of the 16.2 per cent of cases in the interval. This fraction or proportion is converted into units of the variable, by multiplying the proportion (0.654) by the interval width of five grade points. This occurs at 3.27 grade points above the lower end of the interval. Adding this 3.27 grade points to 65, the grade point at the lower end of the interval, results in a total of 68.27. Rounded to the nearest tenth of a grade point, the twentieth percentile is 68.3.

This process is illustrated diagramatically in Figure 2, using a diagram similar to that for obtaining the median. The difference from the diagram used to obtain the median is that the aim is to obtain a particular percentile, in this case, P_{20}. That is, the diagram is constructed to obtain the position at which the variable X reaches the twenty per cent point.

The eighty-fifth percentile of grades occurs in the interval from 80 to 85.

Figure 2: Calculation of P_{20} of grade point averages



There are only 75.2% of students with grades less than or equal to 80, but 96.0% of students with grades less than or equal to 85. Thus P_{85} occurs within the 80-85 grade point interval. Using the same formula as earlier, 80 is the lower end point of the interval and the interval is 5 grade points wide (from 80 to 85). There are 20.8% of students in this interval (denominator of fraction in large brackets), $r = 85$, and the cumulative per cent at the lower end of the interval is 75.2%. From these values,

$$\begin{aligned}
P_{85} &= 80 + \left(\frac{85 - 75.2}{20.8} \times 5 \right) \\
&= 80 + \left(\frac{9.8}{20.8} \times 5 \right) \\
&= 80 + (0.471 \times 5) \\
&= 80 + 2.36 \\
&= 82.36.
\end{aligned}$$

From the above, there are twenty per cent of students with grades of 68.3% or less, and eighty-five per cent of students with grades of 82.4% or less.

Example – Alcohol consumption by income. The *Canadian Community Health Survey*, cycle 1.2, conducted by Statistics Canada gives the data in Table 5.

Use these data to compute the median and seventieth percentiles of number of alcoholic drinks consumed per week for adults in each of the two income groups.

Table 5: Distribution of Saskatchewan adults by number of alcoholic drinks consumed per week, by personal income

Number of drinks per week	Number of respondents with income of:	
	< \$30,000	\$30,000 plus
None	370	188
1-4	214	185
5-9	94	106
10-19	54	74
20-39	30	29
Total	762	582

The percentages for obtaining the medians and seventieth percentiles are contained in Table 6. Since there is a gap between the end points of the apparent class limits, the real class limits are used to obtain more accurate estimates of the percentiles.

Using the percentage distributions in Table 6, the median consumption for those of low income is in the interval from 1 to 4 or 0.5 to 4.5 drinks per week, where the cumulative percentage first reaches more than fifty per cent. Interpolating in this interval, the median is

$$\begin{aligned} P_{50} &= 0.5 + \left(\frac{50 - 48.6}{28.1} \times 4 \right) \\ &= 0.5 + \left(\frac{1.4}{28.1} \times 4 \right) \\ &= 0.5 + (0.0498 \times 4) \\ &= 0.5 + 0.199 \\ &= 0.699 \end{aligned}$$

or 0.7 drinks per week.

For the high income individuals, the median is also in the same interval. Interpolating in this inteval gives

$$P_{50} = 0.5 + \left(\frac{50 - 32.3}{31.8} \times 4 \right)$$

Table 6: Percentages for obtaining percentiles of alcohol consumption, low and high income Saskatchewan respondents

Alcohol	Interval	Low income		High income	
consumption	width	P	Cum. P	P	Cum. P
-0.5 - 0.5	1	48.6	48.6	32.3	32.3
0.5 - 4.5	4	28.1	76.7	31.8	64.1
4.5 - 9.5	5	12.3	89.0	18.2	82.3
9.5 - 19.5	10	7.1	96.1	12.7	95.0
19.5 - 39.5	20	3.9	100.0	5.0	100.0
Total		100.0		100.0	

$$\begin{aligned} &= 0.5 + \left(\frac{17.7}{31.8} \times 4\right) \\ &= 0.5 + (0.5566 \times 4) \\ &= 0.5 + 2.226 \\ &= 2.726 \end{aligned}$$

or 2.7 drinks per week.

For low income individuals, the seventieth percentile is in the interval 1-4, or from 0.5 to 4.5 drinks, since there are only 48.6% of cases with zero drinks per week and another 28.1% of cases in this interval, making a cumulative total of 76.7% once all cases in this interval are included. Interpolating in this interval, the seventieth percentile is

$$\begin{aligned} P_{70} &= 0.5 + \left(\frac{70 - 48.6}{28.1} \times 4\right) \\ &= 0.5 + \left(\frac{21.4}{28.1} \times 4\right) \\ &= 0.5 + (0.7616 \times 4) \\ &= 0.5 + 3.046 \\ &= 3.546 \end{aligned}$$

or 3.5 drinks per week.

For the high income individuals, the median is in the interval from 5 to 9, or 4.5 to 9.5, where the seventy per cent point of the cumulative percentage distribution is first reached. Interpolating in this inteval gives

$$\begin{aligned} P_{70} &= 4.5 + \left(\frac{70 - 64.1}{18.2} \times 5 \right) \\ &= 4.5 + \left(\frac{5.9}{18.2} \times 5 \right) \\ &= 4.5 + (0.3242 \times 5) \\ &= 4.5 + 1.621 \\ &= 6.121 \end{aligned}$$

or 6.1 drinks per week.

Last edited October 5, 2006.