Social Studies 201

October 4, 2004

Measures of Variation

Overview

Measures of variation (range, interquartile range, standard deviation, variance, and coefficient of relative variation) are presented in the notes in this module. These measures are less commonly used than are averages, and are not as familiar as are measures of centrality (mode, median, or mean). Measures of variation indicate the extent of dispersion or variation of a distribution.

In these notes, each measure is defined and examples are presented. The following quick examples demonstrate ways that measures of variation form an essential aspect of social science analysis.

Quick examples

- The range of grades in Social Studies 201 last semester was 42 percentage points, from a minimum of 50% to a maximum of 92%.
- Literacy skills vary both across and within provinces. A Statistics Canada study demonstrated that while "a student's socio-economic background is a key factor, it accounted for less than half the variation in provincial literacy scores."

  From http://www.statcan.ca/Daily/English/040714/d040714b.htm
- Class grades vary more in an introductory than in upper level classes – the standard deviation of grades in a 100-level class was 8 percentage points but was only 3 percentage points in a 300-level class.

- "Most countries today are culturally diverse. ... In very few countries can the citizens be said to share the same language, or belong to the same ethnonational group."

  Will Kymlicka, *Multicultural Citizenship*, Oxford, Clarendon Press, 1995, p. 1.

- " Income inequality among families remained stable."

  From http://www.statcan.ca/Daily/English/040520/d040520b.htm

- "Between 1981 and 1996, the second, third, and fourth quintiles lost 2.8 per cent of their before-tax income, a total of $14 billion, to the upper quintile .... These figures support the idea that there is increasing polarization of income in Canada."

  From L. Tepperman and J. Curtis, *Sociology: A Canadian Perspective*, Don Mills, Oxford University Press, 2004, p. 371.

Each of these quick examples refers to how members of a sample or population differ from each other – variation in grades, literacy skills, culture, and income. While the measures of centrality for each of these variables describes the average or typical member, just as important is how members differ from each other. It is this variation that forms the topic for this module.

Topics

- Introduction
- Range
- Interquartile range
- Variance and standard deviation
  - Ungrouped data
  - Grouped data
- Coefficient of relative variation
- Conclusion

Readings

Chapter 5, Sections 5.7 - 5.11

Learning objectives

By the end of this module, you should be familiar with measures of variation: range, interquartile range, standard deviation, variance, and coefficient of relative variation. In particular, you should:

- Understand the definition of each measure.
- Know how to calculate each measure from data presented in either ungrouped or grouped format.
- Be able to interpret the meaning of each of the measures in written statements and day-to-day usage.

Introduction

Measures of variation are statistics describing the extent of variation among members of a sample or population. These measures are not as well understood or as commonly used as are measures of centrality (mode, median, and mean). But measures of variation are just as important as are averages to understanding distributions and populations. Not everyone in a population is at the average, and measures of variation describe how members of a population differ.

Measures of variation are single numbers, or statistics, that summarize how varied or how dispersed members of a population are. The measures of variation that we discuss in this class are as follows. Formal definitions and examples of each are contained in later parts of Module 4.

- **Range**. The range is the maximum value of a variable minus its minimum value.
- **Interquartile range**. The interquartile range is the value of the 75th percentile minus the value of the 25th percentile. This is the range for the "middle one-half" of a distribution.
- **Standard deviation**. The standard deviation is a measure of the "average" or "standard" amount by which individual cases differ from

the mean of a distribution. The exact form of this average, or standard, is provided later in these notes.

- **Variance**. The variance is the square of the standard deviation. Alternatively stated, the variance is an "average" of squares of differences of individual values from the mean. This will become apparent when you study the formulae and examples later in this module.

- **Coefficient of relative variation**. The coefficient of relative variation is the value of the standard deviation divided by the mean. This measure of variation provides an indication of how much variation there is in a distribution, relative to the mean of the distribution.

Most of these measures of variation are not used in ordinary conversations or in newspapers and magazines – or they are much less commonly used than averages. One key to understanding measures of variation is to use them to compare two distributions. In the examples and exercises in this module, two distributions are often compared. When you work through these examples and exercises, carefully examine the two distributions and the associated measures of variation. This will help you develop an understanding of how measures of variation can be used to compare variability.

Consider the hypothetical distributions A and B in Figures 1 and 2. Measures of variation for distribution A have smaller values than the corresponding measures of variation for distribution B. These indicate that distribution A is less varied and distribution B is more varied. Each distribution is symmetrical and has the same mean, at the centre of the distribution. But distribution B in Figure 2 is twice as spread out as distribution A in Figure 1. Most measures of variation for the distribution in Figure 2 will be double the size of the corresponding measure for the distribution in Figure 1.

Differences between variation of distributions A and B are summarized in Table 1. Each of the words **variation**, **dispersion**, and **concentration** can be used to contrast differences in variation, or spread of values, in the two distributions.

Figure 1: Less varied distribution A

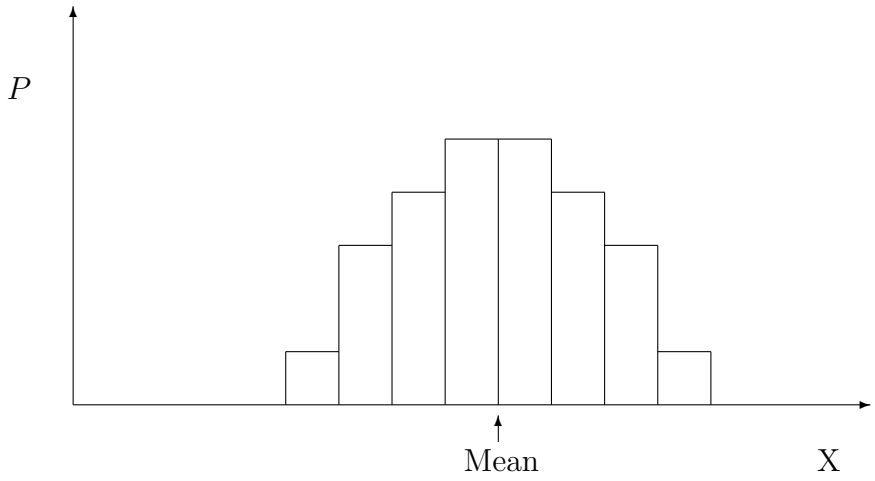


Figure 2: More varied distribution B

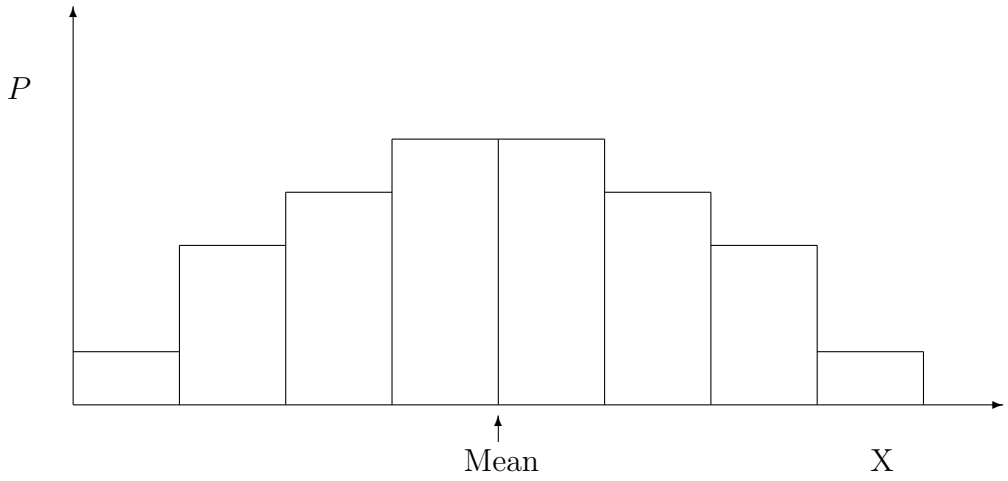


Table 1: Variation of distributions of A and B – Figures 4.1 and 4.2

Figure 1	Figure 2
less varied	more varied
less dispersed	more dispersed
more concentrated	less concentrated

Range

See text, section 5.8.1, pp. 216-218.

> "Provinces clearly vary in their reading performance, ranging from 501 in New Brunswick to 550 in Alberta."
>
> 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, p. 14. Available at the Statistics Canada web site:
> http://www.statcan.ca/Daily/English/040714/d040714b.htm

The first measure of variation is the range, the set of values over which the variable varies.

> **Definition**. The range is the maximum value of the variable minus the minimum value of the variable.
>
> $$\text{Range} = \text{maximum} - \text{minimum}.$$
>
> That is, the range is the largest minus the smallest value of the cases in a data set.
>
> Alternatively, the range is sometimes defined as a listing of the minimum and maximum values.

Since the range requires determining the smallest and largest values, the a variable must have at least an ordinal level of measurement in order to determine the range.

The unit of measure for the range is the same as the unit of measure for the variable X. For example, if X represents income of individuals in dollars, and the minimum and maximum incomes are measured in dollars, then the range is measured in dollars.

Example 4.1 – Range of grades. A student takes five classes during her first semester at the University of Regina. The grades obtained are 64, 74, 68, 79, and 85.

What is the range of grades of the student?

Answer. The variable is grade, presumably in the unit of one percentage point, so this variable has at least an interval level of measurement and the range can be meaningfully determined.

The lowest grade was 64 and the highest was 85, so the range is 21.

$$\text{Range} = \text{maximum} - \text{minimum} = 85 - 64 = 21$$

The range of grades for this student is 21 percentage points.

Example 4.2 – Range of ages. For a sample of eleven University of Regina undergraduates, the ages are 27, 19, 18, 18, 18, 29, 20, 36, 24, 18, and 19 years.

What is the range of ages for these eleven students?

Answer. The variable is age in years, a variable with a ratio scale of measurement so the range can be meaningfully determined.

The lowest age is 18 and the highest age is 36, so the range is 18.

$$\text{Range} = \text{maximum} - \text{minimum} = 36 - 18 = 18$$

The range of ages for this sample of students is 18 years.

Example 4.3 – Attitudes to treating gay and lesbian couples 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 with this statement on a five-point scale,

Table 2: Frequency, percentage, and cumulative percentage distributions of views of undergraduate student responses to treating gay and lesbian couples as married

Response (X)	Frequency	%	Cumulative %
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	

from 1, indicating strongly disagree, to 5, indicating strongly agree. The distribution of responses is reported in Table 2.

What is the range of views on this topic?

Answer. For this statement, responses are measured on a scale from 1 (strongly disagree) to 5 (strongly agree). This measurement is at the ordinal level, so the range is meaningful.

The range of opinions is $5-1=4$. Alternatively stated, in words, the range is from 1 to 5, or from strongly disagree to strongly agree.

Note. The distribution of cases between these lower and upper limits has no effect on the determination of the range. All that matters for the range is what the smallest and largest values are.

Recap – some characteristics of the range

The range provides a useful first indication of the dispersion or concentration of values in a distribution. At the same time, as demonstrated by Example 4.3, the range is limited in that it does not indicate how varied the cases are between the minimum and maximum values.

Imagine two rooms with ten people in each room – in room A the ages of the people have a range from 18 to 22, a range of 4 years; in room B the

ages of the people have a range from 15 to 55, a range of 40 years. A glance around each room will quickly indicate that the people in room A have a similar set of ages (more concentrated or less dispersed), while the people in room B are more varied in their ages (less concentrated or more dispersed).

One of the first questions a researcher asks about a variable is what is the range of the variable across the data set. This does not tell the researcher a great deal about the values of the variable, but provides a useful first indication of the spread or variation of the values. The researcher is likely to follow up by calculating some of the other measures of variation in the following notes.

Interquartile Range (IQR)

See text, section 5.8.2, pp. 218-224.

The interquartile range (IQR) is the range of the values of a variable over the middle part of a distribution. Specifically it is the range from the 25th to the 75th percentile of a variable. The IQR is a range, once the upper one-quarter and lower one-quarter of cases of a distribution are eliminated from consideration.

> **Definition**. The interquartile range (IQR) is the seventy-fifth percentile minus the twenty-fifth percentile. In symbols,
>
> $$\text{IQR} = P_{75} - P_{25}$$
>
> Alternatively, the interquartile range is the third quartile minus the first quartile, since the third quartile is defined as the seventy-fifth percentile and the first quartile is the twenty-fifth percentile.

The unit of measure for the IQR is the same as the unit of measure for the variable X. For example, if X represents income of individuals in dollars, each percentile is also in dollars and the IQR is in dollars.

Compared with the range, one advantage of the interquartile range is that it indicates the spread or concentration for the middle one-half of the distribution, ignoring the extremes of the distribution. This is worthwhile for statistical analysis when the extremes are of less interest and more consideration of the middle part of the distribution is required. That is, the difference between the seventy-fifth and twenty-fifth percentiles provides a good indication of the range of the values for the middle, or more typical cases, of the distribution.

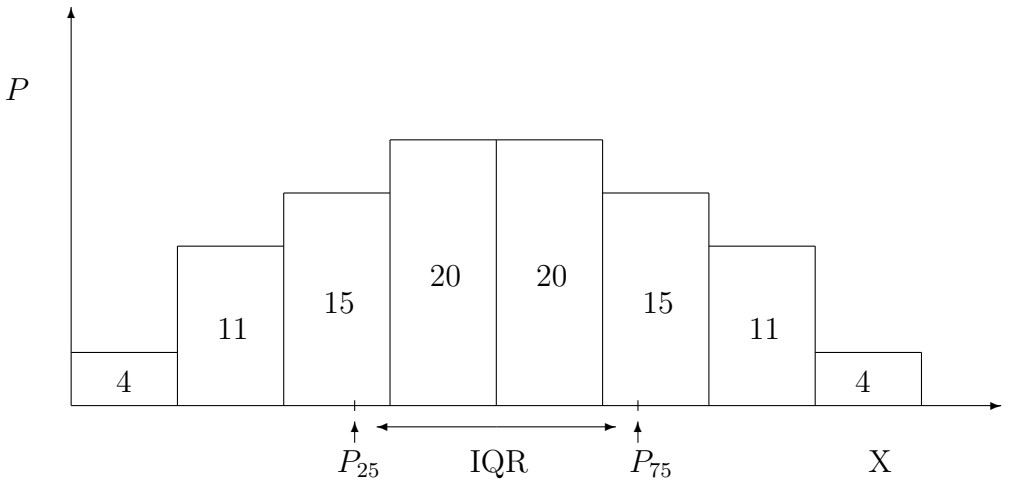
The IQR eliminates the effect of extreme values. For example, a distribution such as an income distribution may be primarily composed of people with low and middle income. Adding a few individuals with very high incomes (say millionaires) to this distribution could dramatically increase the range of the distribution. In contrast, the interquartile range would change little, since it indicates the range of the middle part of the income distribution.

One defect of the IQR as a measure of variation is that it is based on only two specific percentiles, and does not take other values of the variable into account. This occurs because the IQR is a positional measure, indicating only the difference between two other positional measures, P_{75} and P_{25}. In general though, this defect is outweighed by the advantages noted earlier.

Diagrammatic illustration of the interquartile range – Figure 3

The distribution in Figure 3 is a histogram, with the approximate percentage of cases in each bar listed in the middle of the bar; the sum of these percentages is one hundred per cent.

Figure 3: Diagrammatic illustration of interquartile range



The approximate position of the twenty-fifth percentile is indicated by P_{25} – about two-thirds of the way across the bar labelled 15. That is, there are 4% of cases in the first interval and another 11% in the second interval, for a total of 15%. The location of P_{25} is the value of X such that 25% of cases are less than or equal to P_{25}. Since there are 15% of cases in the first two categories, and another 15% in the third category, the twenty-fifth per cent point is reached at about two-thirds of the way across the third category.

By similar reasoning, the seventy-fifth percentile occur about one-third of the way across the sixth category – there are seventy per cent of cases in the first five categories and about one-third of the way across the sixth category

is where P_{75} is located. This is indicated by P_{75} in Figure 3, the value of X such that 75% of cases are less than or equal to P_{75}.

The IQR is the distance between P_{25} and P_{75}. There are 25% of cases less than P_{25}, another 25% of cases greater than P_{75}, and the middle 50% of cases occur within the interval represented by the IQR.

Example 4.3 – Variation in attitudes to social issues. A sample of just under seven hundred undergraduate students was asked to state their view about two statements, "Tax laws and job benefits should recognize gay and lesbian couples as married" and "More provincial tax dollars should be devoted to universal health care." 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.

Table 3: Frequency distributions of views of undergraduate students to statements concerning gay and lesbian couples and expenditures for health care

Response	Frequency	
X	Gay/lesbian couples	Health care
1 (strongly disagree)	141	35
2 (somewhat disagree)	86	71
3 (neutral)	183	230
4 (somewhat agree)	161	222
5 (strongly agree)	124	126
Total	695	686

Question. Obtain the interquartile range for responses to the two statements. In words, briefly compare the variability of the two distributions.

Answer. For each variable, the scale is ordinal, since these are attitudes or opinions concerning the respective issue, with responses

measured on a 1-5 scale. Responses are ordered from strongly disagree to strongly agree, an ordinal scale, so percentiles and the IQR can be meaningfully obtained.

Since the IQR is the difference between the seventy-fifth and twenty-fifth percentiles, it makes most sense to reconstruct the frequency distributions into percentage and cumulative percentage distributions, so percentiles can be easily determined. This is done in Table 4, with the two distributions separated with a vertical line.

Table 4: Percentage (P) and cumulative percentage (Cum. P) distributions of views about gay and lesbian couples and expenditures for health care

Response	Gay/lesbian couples		Health care	
X	P	Cum. P	P	Cum. P
1 (strongly disagree)	20.3	20.3	5.1	5.1
2 (somewhat disagree)	12.4	32.7	10.3	15.4
3 (neutral)	26.3	59.0	33.5	48.9
4 (somewhat agree)	23.2	82.2	32.4	81.3
5 (strongly agree)	17.8	100.0	18.7	100.0
Total	100.0		100.0	

From Table 4, percentiles can be readily determined. Attitude responses are measured on a discrete scale (1, 2, 3, 4, or 5). Interpolation is not necessary; the percentiles are the values of the response X where the appropriate percentage of cases is first obtained.

Gay and lesbian couples. For the statement concerning gay and lesbian couples, the twenty-fifth percentile is at $X = 2$. There are 20.3% of cases at $X = 1$ and another 12.4% of cases at $X = 2$, totalling 32.7% of cases with values of 2 or less. Thus $P_{25} = 2$.

The seventy-fifth percentile is at $X = 4$, since there are fewer than 75% of cases at $X = 3$ or less but more than 75% at 4 or less. Thus $P_{75} = 4$.

The IQR is thus two.

$$IQR = P_{75} - P_{25} = 4 - 2 = 2$$

The interquartile range is 2 for responses to the statement about treating gay and lesbian couples as married.

Health care expenditures. For the health care issue, the twenty-fifth percentile is not obtained until $X = 3$, since there are only 15.4% of cases at values 1 and 2 of responses, and the 25 per cent point is first reached at a response of 3. Also $P_{75} = 4$, since seventy-five per cent of cases is first reached when $X = 4$. Thus the IQR is 1.

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

The interquartile range is 1 for responses to the statement about health care expenditures.

Comparison of variability. Comparing these two distributions, views are more varied on the gay and lesbian couples issue ($IQR = 2$) and less varied on the health care expenditure issue ($IQR = 1$). The range for the two distributions is identical, from an attitude response of 1 to and attitude response of 5. But across these possible values for the variable, responses are more concentrated on the health care issue, with few respondents disagreeing and most respondents having a response of 3 or more. This produces a relatively small IQR of only 1 point on the five-point attitude scale. In contrast, responses are more varied on the gay and lesbian couples issue, with considerable percentages of responses at each of the five possible responses.

In summary, there is great variation of student views concerning treating gay and lesbian couples as married, with many respondents in support of it but with many also opposed to it. In contrast, views concerning tax dollars for health care are more uniform or concentrated, with respondents generally in agreement that this would be a good policy.

Example 4.4 – Variation in earnings in Saskatchewan and Alberta The distributions of earnings for respondents in Saskatchewan and Alberta were examined in Exercise 3.xxx. These distributions are provided again in Table 5.

Table 5: Distribution of earnings, Saskatchewan and Alberta respondents, 2002

Income in thousands of $	Percentage of respondents Saskatchewan	Alberta
Less than 5	16.2	11.7
5-10	12.9	11.4
10-15	8.8	8.8
15-20	8.2	7.9
20-30	15.6	14.0
30-40	12.3	13.3
40-50	9.4	9.9
50-60	6.6	6.8
60 plus	10.0	16.2
Total	100.0	100.0

Adapted from Table 202-0101, Statistics Canada, *Income Trends in Canada 1980-2002*, catalogue number 13F0022XCB, 2004.

Question. Obtain the interquartile range for the distribution of earnings among respondents in each of Saskatchewan and Alberta. Briefly comment on differences in the variation of earnings in the two provinces.

Answer. In Exercise 3.xxx, the twenty-fifth and seventy-fifth percentiles (P_{25} and P_{75}, respectively) were calculated. The value of these percentiles, in dollars, is given in Table 6. If you do not recall how to obtain these percentiles, review Exercise 3.xxxx.

The interquartile range for Saskatchewan is

$$IQR = P_{75} - P_{25} = 41,100 - 8,400 = 32,700$$

Table 6: Summary of statistics for Saskatchewan and Alberta earnings

	Value of statistic ($)	
Measure	Saskatchewan	Alberta
P_{75}	41,100	48,000
P_{25}	8,400	11,100
IQR	32,700	36,900

and for Alberta is

$$IQR = P_{75} - P_{25} = 48,000 - 11,1400 = 36,900$$

The interquartile range is $32,700 for Saskatchewan earners and $36,900 for Alberta earners.

From these two measures, and examining the table, the variation in earnings is greater in Alberta than in Saskatchewan. There is a difference of just over $4,000 in the interquartile ranges ($36,900 minus $32,700) for the two provinces, with Alberta having greater variation than Saskatchewan. The range of earnings is the same in the two provinces, from less than $5,000 to $60,000 plus. But in Saskatchewan there is a greater concentration of respondents at lower and middle earnings levels. For Alberta, there is a greater percentage of respondents at the high end of earnings, resulting in a greater variation for Alberta. The respective interquartile ranges are consistent with the larger variation for Alberta than Saskatchewan, with the range over the middle one-half of earnings being $36,900 for Alberta and $32,700 for Saskatchewan.

Recap of Interquartile range (IQR)

The interquartile range is a useful measure of variation in that it describes the extent of variation over the middle part of a distribution. The advantage of the IQR over the range is that the influence of the extremes of a distribution are eliminated – only the difference between the seventy-fifth and twenty-fifth percentiles is considered. For variables such as income or earnings, the IQR is especially useful. These variables may have a few extreme values, with either very large or very small incomes or earnings. By eliminating the lower one-quarter and the upper one-quarter of cases, the IQR provides a good summary of how varied the more typical incomes are.

vspace0.1cm

The IQR is also an easy measure to calculate, at least once the percentiles are obtained. It is merely the difference between the seventy-fifth and twenty-fifth percentiles.

The disadvantage of the interquartile range is that it is a positional measure, based on only the twenty-fifth and seventy-fifth percentiles. The next measures of variation to be examined in these notes, the standard deviation and variance, remedy this defect. For these next measures, the value associated with each case is taken into account.

In summary, the IQR is used when the researcher wishes to eliminate the influence influence of extreme values and consider the variation for the more typical cases in a distribution.