$$\text{Mean for grouped data} = \frac{\sum f_i * x_i}{\sum f_i}$$

INTERPRETATION OF THE MEAN WHEN THE VARIABLE IS CODED

We often have data files where a quantitative variable is not given in its original form, but coded into a small number of categories. For instance, the variable Respondent's Income could be given in the form shown in Table 3.2.

Table 3.2 **Example of a quantitative variable that is coded into 21 categories, with a 22nd category for those who refused to answer**

Category	Code
Less than $1000	1
$1000–2999	2
$3000–3999	3
$4000–4999	4
$5000–5999	5
$6000–6999	6
$7000–7999	7
$8000–9999	8
$10,000–12,499	9
$12,500–14,999	10
$15,000–17,499	11
$17,500–19,999	12
$20,000–22,499	13
$22,500–24,999	14
$25,000–29,999	15
$30,000–34,999	16
$35,000–39,999	17
$40,000–49,999	18
$50,000–59,999	19
$60,000–74,999	20
$75,000 and more	21
Refused to answer	22

Thus, we would not know the exact income of a respondent. We would only know the category he or she falls into.

This kind of measuring scale poses a challenge. If we compute the mean with SPSS, we will not get the mean income. We will get the mean code, because it is the codes that are used to perform the computations. There is a data file that comes with SPSS where the income is coded in this way. This data file contains information about 1500 respondents, including information on the income bracket they fall into, coded as shown in Table 3.2. When we exclude the 22nd category, which consists of the people who refused to answer this question, the computation of the mean with SPSS produces the following result:

Mean = 12.35

What is the use of this number? It is not a dollar amount! If we look at Table 3.2, we see that the code 12 stands for an income of between $17,500 a year and $20,000 a year (with that last number excluded from the category). To interpret this number, we should first translate it into a dollar amount (it can be done with a simple rule). But even without transforming it into the dollar amount it corresponds to, we could use the mean code for comparisons. For instance, we will see in Lab 3 that if we compute the mean income separately for men and women, we get

Mean income for men: 13.9

Mean income for women: 10.9

(excluding the category of people who refused to answer).

Although the mean code does not tell us exactly the mean income for men and women, it still tells us that there is a big difference between men and women for that variable. Table 3.2 tells us that the code 13 corresponds to the income bracket $22,500–25,000, while the code 10 represents the income bracket $12,500–15,000. We can conclude that the difference in income between men and women, for that sample, is roughly around $10,000 a year.

We see that that when the variables are coded, the interpretation of the mean requires us to translate the value obtained into what it stands for. For quantitative variables coded this way, it may also be useful to find the frequencies of the various categories, as we did for nominal variables. For the example at hand, we would get Table 3.3 as shown.

The conclusion of the preceding discussion is that when we have an ordinal variable with few categories, or even a quantitative variable that has been recoded into a small number of categories, it may be useful to compute the frequency table of the various categories, in addition to the mean and other descriptive measures.

Weighted Means

Consider the following situation: you want to find the average grade in an exam for two classes of students. The first class averaged 40 out of 50 in the exam, and the second class averaged 46 out of 50. If you put the two classes together, you *cannot* conclude that the average is 43. This is so because the classes may have different numbers of students. Suppose the first class has 20 students, and the second one 40 students. In other words, we have the data shown in Table 3.4.

To compute the average grade for the two classes taken together, we do not need to know the individual scores of each student. Indeed, we have seen before that a sum of n scores is equal to its average times n. We will use this to obtain the formula shown below for weighted means.

The mean for the two classes taken together can be written as

Table 3.3 **Frequencies of the various income categories for the variable Income**

Respondent's income	Frequency	Valid Percent
LT $1000	26	2.6
$1000–2999	36	3.6
$3000–3999	30	3.0
$4000–4999	24	2.4
$5000–5999	23	2.3
$6000–6999	23	2.3
$7000–7999	15	1.5
$8000–9999	31	3.1
$10,000–12,499	55	5.5
$12,500–14,999	54	5.4
$15,000–17,499	64	6.4
$17,500–19,999	58	5.8
$20,000–22,499	55	5.5
$22,500–24,999	61	6.1
$25,000–29,999	84	8.5
$30,000–34,999	83	8.4
$35,000–39,999	54	5.4
$40,000–49,999	66	6.6
$50,000–59,999	38	3.8
$60,000–74,999	23	2.3
$75,000+	44	4.4
Refused to answer	47	4.7
Total	994	100.0
Missing	506	
Grand Total	1500	

Table 3.4 **Two classes of different size and the mean grade in each**

	Average Grade out of 50	Number of Students
Class A	40	20
Class B	46	40

$$\frac{\text{Sum of all scores in class A} + \text{Sum of all scores in class B}}{60}$$

The sum of all scores in class A can be replaced by the average score (40) times 20, since there are 20 students in this class. And the sum of all scores in class B can be replaced also by its average score (46) times 40, since this class includes 40 students. The equation for the mean becomes:

$$\frac{(40 \times 20)}{60} + \frac{(46 \times 40)}{60}$$

This can now be written as:

$$\text{mean of the two classes combined} = 40 \times (\mathbf{20/60}) + 46 \times (\mathbf{40/60})$$

or again as:

$$\text{mean of the two classes combined} = 40 \times (\mathbf{1/3}) + 46 \times (\mathbf{2/3})$$

The last formula is important: we see that the average grade of class A is multiplied by the **weight** of class A, which is its relative importance in the total population. Class A forms 1/3 of the total population (20 students out of 60) and class B 2/3 of the total (40 students out of 60). The underlying formula is:

$$\text{Average grade for the two classes: } 40 \times w_1 + 46 \times w_2$$

The $\mathbf{w_i}$'s are called the **weights** of the various classes. In this case, the weight is an expression of the number of people in each class compared to the total population of the two classes.

The general formula is as follows.

If you have n values	$x_1, x_2, x_3, \ldots$ etc.,
each having the corresponding weights:	$w_1, w_2, w_3, \ldots$ etc.,
the **weighted mean** is given by	$x_1 w_1 + x_2 w_2 + x_3 w_3 + \cdots + x_n w_n$

The weights are positive numbers and must add up to 1. That is:

$$w_1 + w_2 + w_3 + \cdots + w_n = 1.$$

The weights are not always a reflection of the size of the various groups involved. If you are computing the weighted average of your grades during your college studies, the weights could be proportional to the credits given to each course, or they could be an expression of the importance of the course in a given program of studies. A Faculty of Medicine may weight the grades of its candidates by giving a bigger weight to Chemistry and Biology than Art History, for instance.

Example

A buyer wants to evaluate several houses she has seen. She attributes a score out of ten to each house on each of the following items: size, location, internal design, and quality of construction. Any house having a score less than 5 on any item would not be acceptable. The resulting scores for three houses that are seen as acceptable on all grounds are recorded in Table 3.5. The buyer does not

attribute the same importance to each item. The size of the house is the most important quality. The quality of the construction is also very important, but not as important. The buyer attributes a weight to each item, which reflects the importance of that item for her. The weights are given in the last column.

Table 3.5 **Scores given to three houses on four items, and their weights**

Item	House A	House B	House C	Weight of item
Size	9	7	6	0.4
Location	5	9	10	0.1
Internal design	6	5	8	0.2
Quality of construction	7	9	7	0.3

We can now calculate the weighted average score for each house, using the formula for weighted means given above.

For house A: weighted mean score: $10 \times 0.4 + 5 \times 0.1 + 6 \times 0.2 + 7 \times 0.3 = 7.8$
For house B: weighted mean score: $7 \times 0.4 + 9 \times 0.1 + 5 \times 0.2 + 9 \times 0.3 = 7.4$
For house C: weighted mean score: $6 \times 0.4 + 10 \times 0.1 + 8 \times 0.2 + 7 \times 0.3 = 7.1$

We see that house A obtained the highest weighted score. The total, unweighted score of house C is higher than that of house A. But because the items do not all have the same importance, house A ended up having a higher weighted score.

THE MEDIAN AND THE MODE

The **median** is another measure of central tendency for quantitative variables. It is defined as the value that sits right in the middle of all data entries when they are listed in ascending order. If the number of entries is odd, there will be one data entry right in the middle. If the number of entries is even, we will have *two* data entries in the middle, and the median in this case will be their average. Here are two examples.

Case 1: variable X 2, 3, 4, 4, 5, 5, **5**, 6, 7, 8, 11, 13, 13
Case 2: variable Y 2, 3, 4, 4, 5, **5**, **6**, 7, 8, 11, 13, 13

For the variable X we have 13 entries. The value 5 sits in the middle, with six entries equal or smaller than it, and six entries equal or larger. The median for X is thus 5. But for variable Y, we have 12 entries. There are therefore two entries in the middle of the ordered list, not just one. The median will be the average of the two, that is $(5 + 6) \div 2 = \mathbf{5.5}$.

The median is not sensitive to extreme values. Suppose, for instance, that the entries for variable X were: 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 11, 13, 60. Although the last

entry is very large compared to the others, it does not affect the median, which is still 5. The mean, however, would have been affected (compute it yourself for the two situations and see how different it would be). For this reason, the median is a better representative of the center when there are extremely large values on one side of it. But the mean is more useful for statistical computations, as we will see in the coming sections.

Half the population has a score that is lower than or equal to the median, and the other half has a score larger than the median or equal to it. This way of formulating the median is very useful in situations where the distribution is skewed (such as the distribution of income) or in situations where time is involved, especially when processes have not been completed by everybody, as illustrated below.

Examples of the use of the median

- We are told that the average age at first marriage for a population is 22 years for women, and 25 for men. The median for women is 21, and for men it is 24. This means that by the time they reached 21 years of age, half the women in this population were married. For men, half of them were married by the age of 24.
- In a research on the time taken by immigrants to find a job, 500 new immigrants who arrived at least three years ago are interviewed. The mean can not be found because some of them have not found a regular or full-time job yet. But it is found that the median time taken for them to find a regular, full-time job was 18 months for men, and 5 months for women. This means that by the 18th month after arrival, 50% of the men had found a job. Women were faster in finding regular full-time jobs: 50% had a job within 5 months of their date of arrival.

Because the median involves only the *ordered* list of data entries, it can be used if the quantitative variable is measured at the ordinal level. But if the number of categories is small, the median is not very useful.

The **mode** can also be used for quantitative variables. When the values are grouped into classes, the mode is defined as it is for qualitative variables: it is the class that has the highest frequency. But the mean and median remain the best descriptive measures for quantitative variables. If the variable is continuous and the values have not been grouped into classes, the mode is the value at which a *peak* occurs in the graph representing the distribution.

COMPARISON OF THE MEAN AND THE MEDIAN

Both the mean and the median are measures of central tendency of a distribution, that is, they give us a central value around which the other values are found. They are therefore very useful for comparing different samples, or different populations,

or samples with a population, or a given population at different moments in time to see how it has evolved. However, each of the mean and the median has its advantages and its drawbacks.

The mean takes into account every single value that occurs in the data. Therefore, it is sensitive to every value. A single very large value can boost the mean up if the number of entries is not very large. For instance, if one worker in a group of 20 workers won a $1 million lottery ticket, the average wealth of those 20 would look artificially high. The median is not sensitive to every single value. In a distribution where the largest value is changed from 60 to 600, the median would not change. The mean would.

It follows from these remarks that the mean is a more sophisticated measure, because it takes every value into account. Indeed, it is the mean that is used to compute the standard deviation, which a measure of dispersion that will be seen below. However, in situations where the distribution is not very symmetric, and where there are some extreme values on only one side of the distribution, the mean will tend to be shifted towards the extreme values, whereas the median will stay close to the bulk of the data. Therefore, whenever the distribution is highly skewed, the median is a better representative of the center of the distribution than the mean. This is true for variables such as income or wealth, where the distribution among individuals in a country, and also worldwide, is highly skewed. For such a variable, the median is a more accurate representative of the central tendency of the distribution.

Measures of Dispersion

For Qualitative Variables

There are not many measures of dispersion for qualitative variables. One of the measures we can compute is the **variation ratio**. It tells us whether a large proportion of data is concentrated in the modal category, or whether it is spread out over the other categories. The variation ratio is defined as

$$\text{variation ratio} = \frac{\text{number of entries not in the modal class}}{\text{total number of entries}}$$

It is a positive number smaller than one. If this ratio is close to zero, it indicates a great homogeneity, almost every unit being in the modal class. The farther it is from zero, the greater the dispersion of the data over the other categories. Like many other measures, this one is easy to interpret when doing comparisons. For instance, if we compare the sizes of the various linguistic groups in two cities where several languages are spoken, we can use the variation ratio to assess the degree of heterogeneity in each city. Here is an example.

City	Linguistic groups	Percentage
City A	French speaking	30%
	English speaking	34%
	Chinese speaking	20%
	Other	16%
	Total	**100%**
City B	French speaking	28%
	English speaking	40%
	Chinese speaking	20%
	Other	12%
	Total	**100%**

The variation ratio for city A would be $(30 + 20 + 16)/100 = 0.66$, and for city B it would be $(28 + 20 + 12)/100 = 0.60$, showing that city A is a little more heterogeneous than city B.

For Quantitative Variables

There are many ways of measuring the dispersion for quantitative variables. The simplest is the range, but we also have various forms of restricted range, we have the deviation from the mean, the standard deviation, the variance and finally the coefficient of variation. Let us go through these measures one at a time.

RANGE

The **range** is the simplest way of measuring how spread out the data is. You simply subtract the smaller entry from the larger one and add 1, and this tells you the size of the interval over which the data is spread out. For example, you would describe a range of values for the variable **Age** as follows:

> In this sample, the youngest person is 16 years old and the oldest 89, spanning a range of 74 years (89 – 16 + 1).

But we may have extreme values that give a misleading impression about the dispersion of the data. For instance, suppose that a retired person decided to enroll in one of our classes. We could then say that the ages of the students in this class range from 16 years up to 69 years, but that would be misleading, as the great majority of students are somewhere between 17 years old and maybe 23 or 24 years old. For this reason, we can introduce variants of the notion of range.

The $\mathbf{C_{10\text{-}90}}$ **range**, for instance, computes the range of values after we have dropped 10% of the data at each end: the 10% largest entries and the 10% smallest