TOPIC 3
FREQUENCY DISTRIBUTIONS AND TABLES

Getting information from a table is like extracting sunlight from a cucumber.

Farquhar and Farquhar (1891)

Purpose of frequency distributions

need to summarise the observations to give them meaning

When we undertake statistical investigations we end up with a series of observations of some variables from a number of statistical units. Usually we will observe both quantitative and qualitative variables for each unit. Given a series of observations, we usually want to summarise this information so that we can make some sense out of it. We may want to make an estimate for the whole population of units of which our group may only be a sample; or we may just want to have some convenient way to summarise the basic data to reduce the amount of information to a manageable size.

frequency distributions summarise data

One form of summarising data which is often used is a frequency distribution. A frequency distribution is obtained by dividing the range of observed values into a number of classes (that is, groups or ranges of values to which observations can be assigned or allocated), allocating observations to these classes and determining the class frequency for each class (that is, the number of observations in each class).

A frequency distribution helps in understanding the nature of the data as it tells us how the data is spread out across the range of possible values, (that is, how the data is 'distributed') and readily brings out the classes having large and small frequencies. For some purposes, the frequency distribution is completely adequate and provides all the information we need from the original set of observations.

A frequency distribution is a grouping of data into categories showing the number of observations in each category.

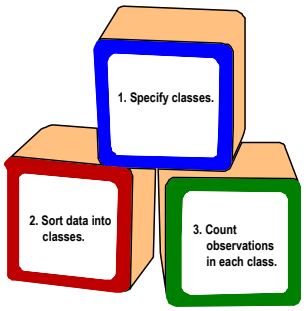
qualitative variables

In the case of qualitative variables, the frequency distribution consists of the frequencies of the units falling into the categories of the variable because data will have been measured using nominal or ordinal scales. Examples of such frequency distributions are provided by the frequency distributions of persons by marital status, persons by economic activity status, households by type of accommodation and businesses by industry group.

continuous data

Frequency distributions for continuous data use class intervals as the categories. These class intervals contain a range of values rather than a single value. This is to simplify and summarise the data.

Steps in the construction of a frequency distribution



3 broad steps

The three broad steps in constructing a frequency distribution are:

1. Specify the classes into which the data will be grouped.
2. Sort the data into the classes.
3. Count the number of observations in each class.

The most complicated part is specifying the classes to group the data into.
There are a number of guidelines used which help specify the classes. However, it should be remembered that these are only guidelines and common sense should always be applied.

Specifying Classes

qualitative data

If you are working with qualitative data, the data will probably be sorted into classes already, based on the response categories. Sometimes, when dealing with a large number of response categories, the observations can be grouped or combined. An example where qualitative data is grouped is when making tables on the number of employees in different occupations. It is often useful to group the occupation categories, so that the number of occupation classes is manageable.

quantitative data

When working with quantitative data it can take a number of steps to specify the classes into which the data will be grouped. Two things need to be considered:

1 the number of classes needed; and

2 the size of each class (i.e. the range of values in each class).

Number of classes

no less than 5 and no more than 20

When deciding on the number of classes the general principle used is that *no fewer than 5 and no more than 20 classes* should be used in the construction of a frequency distribution. Too few, or too many, classes give little information about the distribution of the data.

large number of observations

The number of classes is also determined by the total number of observations and their concentration and spread. A large number of classes may be desirable when there is a large number of observations. If there are some isolated high and low values, try not to let these dictate the number of classes. It needs to be remembered that the purpose of frequency distributions is to summarise data, not to provide details such as extreme values. If you have too many classes you may as well be looking at the original observations and the frequency distribution will not provide an effective summary.

small number of classes

If you have too few classes you hide genuine differences in the observations and cause a substantial loss of information. Quite often we decide on the number of classes and the size of the classes at the same time.

spread of the data

The first thing we need to know is the spread of the data – how many different observations does our frequency distribution have to include? The spread of the data is known as the *range*[2], and is the difference between the highest observation and the lowest observation. The range is then used to specify the classes so that they divide up the range (usually evenly) across the data.

range

Range = highest observation - lowest observation

fish example

It is probably easier to understand how to define the number of classes by looking at an example. The example will also show the relationship between number of classes and class size – which we will cover next. The data is illustrative, but for our purpose it is landed fish from one area in a fish survey.

Table 3.1 Weight of 63 fish (weights are given in kgs)

4.6	3.9	2.8	6.6	4.2	3.7	3.7	5.9
3.2	2.2	3.2	4.1	3.1	3.0	4.8	4.1
2.1	4.2	5.0	4.6	5.4	2.4	6.3	2.9
5.3	4.0	4.7	3.6	3.3	6.9	4.5	2.5
5.4	5.7	3.8	4.1	7.9	6.2	3.0	3.3
5.0	5.4	3.4	4.4	4.0	3.6	5.0	4.1
4.8	7.2	6.4	3.0	3.5	5.3	7.7	3.9
2.6	5.6	3.3	5.5	4.3	3.9	6.3	

Source: Illustrative data only.

class size

The size of a class is the difference between the lowest value in the class and the lowest value in the next class. It is common practice to use intervals that are multiples of 5 or 10, such as 5, 10, 20, 100 or 1,000.

We also try to keep the classes of equal size, because this makes understanding and interpreting the data easier. Sometimes it is not desirable to have equal sized classes because it could mean that you have too many or too few classes. In practice, we determine the number of classes and class size together. Firstly, we determine the number of classes, and this then acts as a guide or indicates the size of the classes.

number of classes from Table 3.1

From Table 3.1, the highest value is 7.9 (underlined), the lowest value is 2.1 (underlined). Therefore the range is:

Range = 7.9 - 2.1

= 5.8 kg

The range tells us that we have a spread of 5.8 kg in weight from the smallest to the largest observation. In this case, the range of 5.8 kg suggests six classes because we prefer to have equal class intervals.

[2] *See Topic 5.*

width of class interval

The width of the classes are approximated by dividing the range by the number of classes. That is:

$$\text{Approximate width of class interval} = \frac{\text{range}}{\text{number of classes}}$$

consider the number of classes

This does not mean that the width of the class interval has to be this number, but this number gives an indication of the suitable width of the interval. Remember that in practice we determine the number of classes and the width of the class interval together.

For the data in Table 3.1, six classes were thought to be suitable. This would give an approximate class interval width of $^{5.8}/_{6}$ = 0.97 kg. This suggests a class interval width of 1 kg would be suitable, and the observations can be covered by six classes.

Class limits

class limits

The next thing to consider is the upper and lower boundaries of the classes. The smallest and largest values used to specify the class are called its **class limits**. We have defined a class interval of 1 kg for the data in Table 3.1. This would suggest that the classes be specified as:

2.0 – 3.0 kg

3.0 – 4.0 kg

4.0 – 5.0 kg

etc.

However, the problem with such classes is that they are not **mutually exclusive**. To see this, in what class would you place a fish weighing 3.0 kg? In the 2.0 – 3.0 class or in the 3.0 – 4.0 kg class? In order to overcome this problem, we change the classes to be:

Table 3.2 Class limits of fish weight data (kg)

Classes	Class limits	Lower class limit	Upper class limit
2.0–2.9	2.0 & 2.9	2.0	2.9
3.0–3.9	3.0 & 3.9	3.0	3.9
4.0–4.9	4.0 & 4.9	4.0	4.9
5.0–5.9	5.0 & 5.9	5.0	5.9
6.0–6.9	6.0 & 6.9	6.0	6.9
7.0–7.9	7.0 & 7.9	7.0	7.9

Source: Table 3.1

The class limits define the range of values that are included in each class after the observations have been recorded. The lower and upper class limit define the smallest and largest recorded observation which can be included in the class. The class 2.0 - 2.9 from Table 3.2 includes ten values: 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, and 2.9. The other five classes similarly include ten values.

Frequency distribution

frequency distribution

So now we can construct a frequency distribution for the data from Table 3.1 with six classes of 1 kg in size:

Table 3.3 Frequency distribution of fish weight data (kg)

Weight (kg)	Tally Mark	Frequency
2.0–2.9	𝍸 //	7
3.0–3.9	𝍸 𝍸 𝍸 ////	19
4.0–4.9	𝍸 𝍸 𝍸 /	16
5.0–5.9	𝍸 𝍸 //	12
6.0–6.9	𝍸 /	6
7.0–7.9	///	3
Total		63

Source: Table 3.1

other criteria for frequency distribution

In addition to these basic criteria, there are a number of other principles used when constructing frequency distributions. These concern the occurrence of observations in classes.

Occurrence of observations

once and only once

There are two other main criteria to consider when specifying classes:

1 No possible observation is to be able to be included in more than one class; and

2 Every possible observation must be able to be included in one of the classes.

example

It is particularly important that the range of each class should be defined so that each observation can go into one, and only one, class interval. We saw in the fish example that what initially seemed to be reasonable classes (2.0 – 3.0, 3.0 – 4.0, etc) in fact lead to confusion. Another example could be constructing a frequency distribution for the height of a group of students. Our classes might be:

incorrect classes

1.00 - 1.10 metres

1.10 - 1.20 metres

1.20 - 1.30 metres

1.30 - 1.40 metres

and so on.

double counting

In this case a student with a height of exactly 1.20 metres could go into one of two classes. A much better set of classes would be:

correct classes

1.00 to less than 1.10 metres

1.10 to less than 1.20 metres

1.20 to less than 1.30 metres

1.30 to less than 1.40 metres

and so on.

In this case, there is no ambiguity and an observation of 1.20 metres can only fit in one class.

True class limits

continuous data

The class limits define the range of values that are included in each class after the observations have been recorded. For **continuous data** there may be gaps between the upper limit of one class and the lower limit of the next class.

Recall the classes and class limits set out in Table 3.2. Recall also that we were trying to construct classes with a size of 1 kg. But the difference between the upper and lower class limit is 0.9 kg. Did we fail?

Consider a fish that actually weighed 2.94 kg. In what class would such a fish be included? Would it fail our criteria that “every observation must be able to be included in one of the classes”? You will notice from Table 3.1 that all the observations have been recorded in tenths of kilograms. It is reasonable to assume that all the fish weights have been recorded to the nearest tenth of a kilogram (after rounding), so a fish that actually weighed 2.94 kg would be recorded as weighing 2.9 kg and therefore included in the first class. In fact all fish weighing between 2.90 kg and 2.95 kg will be rounded and recorded as 2.9 kg and be included in the first class. Similarly, all fish weighing between 2.95 kg and 3.00 kg will be rounded and recorded as 3.0 kg and be included in the second class.

true class limits

This leads to the idea of **true class limits**. The true class limits specify the actual range of values included in the class before rounding. For the fish data in Table 3.1, the class limits are shown over:

class limits

2.0–2.9 kg

3.0–3.9 kg

4.0–4.9 kg

5.0–5.9 kg

6.0–6.9 kg

7.0–7.9 kg

'true' limits

The true class limits would be:

Table 3.4 Class limits and true class limits for fish data (kg)

Class	Class limits	True lower class limit	True upper class limit
2.0–2.9	2.0 - 2.9	1.95	2.95
3.0–3.9	3.0 - 3.9	2.95	3.95
4.0–4.9	4.0 - 4.9	3.95	4.95
5.0–5.9	5.0 - 5.9	4.95	5.95
6.0–6.9	6.0 - 6.9	5.95	6.95
7.0–7.9	7.0 - 7.9	6.95	7.95

Source: Table 3.3

rounded data

Notice that the distinction between class limits and true class limits is only important for rounded data, which only occurs for continuous variables.

discrete data

For discrete variables the class limits and the true class limits will be the same.

Class intervals

use true class limits to find the class interval

We have defined the class limits and true class limits in order to avoid any ambiguity in the way we assign values to classes. The **class interval** (that is, the range of values in a class) is the difference between the true class limits of the class. You must note it is not necessarily the difference between the upper and lower limit of the class.

equal sized classes

Whenever possible, the class intervals should be the same. Class intervals of equal lengths make it much easier to comprehend the distribution and to draw suitable diagrams. If unequal intervals are used, it is often difficult to compare one class frequency with another. Sometimes, however, it is impossible to avoid unequal class intervals; the variability of the data or the need for confidentiality sometimes require unequal intervals.

common sense should prevail

Don't forget, although the guidelines for constructing frequency distributions presented in this topic are very useful, common sense should always prevail. For example, what would you do if the observations in Table 3.1 ranged from 0.1 kg to 7.9 kg? Would you construct classes of equal size?

No – it would not be practical. Take some time to think about the classes you would have with this range of data.

open-ended class intervals

In preparing a frequency distribution, it is better to avoid having **open-ended class intervals** (that is, no upper [or lower] class limits). However, in many situations the very small or very large values of observations falling in the end-classes are so different to the other observations in the class that there is considerable loss of information in putting them in one class at the end with a closed class interval. In such situations the best procedure would be to specify the open end class interval and all the values of the observations in the class or at least indicate their average value and the range, as this would avoid loss of relevant information and would allow further calculations on the data. If, for example, no information other than the frequency is available for an end class which is open, the midpoint and the

upper-class limit may have to be estimated by taking into account the frequencies in the previous class or classes.

Class midpoint

definition

The **class midpoint** is the midpoint of the class, and is obtained by taking the average of the upper and lower true class limits.

formula

$$\text{Class midpoint} = \frac{\text{True upper class limit} + \text{True lower class limit}}{2}$$

other terms for midpoints

In the example in Table 3.5 below, the class midpoints are 2.45, 3.45, 4.45, 5.45, 6.45 and 7.45. The class midpoints are sometimes referred to as class marks, class mid-marks or class mid-values. Once a frequency distribution is formed, the observations in a class are assumed to have the same value and that value is taken as the class midpoint unless the average value of that class is available. Class midpoints are necessary for summarising grouped data as we will see in later topics.

Table 3.5 Frequency distribution of fish data of a sample of 63 fish

Class	True class limits	Class midpoint
2.0–2.9	1.95 & 2.95	2.45
3.0–3.9	2.95 & 3.95	3.45
4.0–4.9	3.95 & 4.95	4.45
5.0–5.9	4.95 & 5.95	5.45
6.0–6.9	5.95 & 6.95	6.45
7.0–7.9	6.95 & 7.95	7.45

Source: Table 3.2

Quantitative frequency distributions – grouping data

loss of information for quantitative data

Note that once a frequency distribution is prepared, the identities of the units are lost and they just become counts in their respective classes. Further, for a **quantitative variable,** the actual value of a unit gets merged with those of the other units in a class. Thus, there is a loss of information in reducing the primary data into a frequency distribution, but this loss of information is often not crucial for the purposes for which the data are used.

frequency distributions summarise data

A frequency distribution of a quantitative variable immediately tells us how many large, medium and small values there are in the population. It also gives us an idea of the most often occurring values at a glance. It throws considerable light on the nature or shape of the population under study, bringing out quite clearly whether the frequencies are evenly spread or are concentrated at particular values.

To study quantitative frequency distributions, we need to look at the two different types of variables we covered in Topic 2, because the problems of constructing a frequency distribution and drawing diagrams of the distributions are somewhat different for each type.

continuous and discrete variables

In general, we distinguish between two types of quantitative variables; the first where the variable is allowed to take any value within a specified range, and the second where the variable can only take certain values. The first type is what we call a continuous variable and the second type we call a discrete variable. In most cases discrete variables take whole number values.

approximations

In practice, continuous data is not measured or recorded continuously. We could, for example, measure the height of a person to the nearest millimetre if we had sufficiently accurate equipment. With an ordinary tape measure, however, we could probably measure it correctly to the nearest centimetre. It was similarly the case for the fish data in Table 3.1 where the fish weights were measured to the nearest tenth of a kilogram.

discrete data

With discrete data there are two possibilities; we can define one class to be either just one single value of the variable, or it can include a range of values. Which one is used will depend on the data. Two examples of frequency distributions of discrete data are given in Tables 3.6 and 3.7.

Table 3.6 Number of Destinations on Holiday by Number of Persons (Cook Islands Visitor Survey 1991)

Number of Destinations	Persons
1	1,811
2	683
3	342
4	273
5	137
6	103
7	68
Total	**3,417**

Source: Cook Islands Visitor Survey 1991, Survey Report No. 13, TCSP, Table 23, p. 31.

Table 3.7 Net Monthly Income by Number of Households, Solomon Islands (1993 Income and Expenditure Survey)

Net monthly income	Households
0 – 50	42,872
51 – 150	1,213
151 – 250	1,591
251 – 350	1,861
351 – 500	1,383
501 – 700	827
701 – 900	607
900 and more	737
Total	**51,091**

Source: Solomon Islands Statistical Bulletin No. 18/95, Table 3.1.2, p8.

comparing different populations

Table 3.6 shows the distribution of the number of destinations in the holiday by the number of persons visiting the Cook Islands. In this case the range of data is quite small, i.e. from 1 to 7 destinations, so we use the actual number of destinations to define each class. In Table 3.7, we have the distribution of net monthly income for households in the Solomon Islands. In this case, the range of values observed is high, from 0 to more than 900, so we use classes that contain a range of values. In this survey income was collected as a discrete variable, i.e. in whole dollars. So for the 1,213 households with between $51 and $150 net monthly income, they could have income of $51, $52, $53, $54 etc through to $150. Hence values like $51.29999 or indeed any other values which are not integers between 51 to 150 cannot occur.

unequal class intervals

As a general rule we would prefer to choose classes that have equal class intervals; this certainly makes interpretation and the drawing of diagrams much easier. However, if we have a large number of small values and a small number of large values, then equal class intervals are not really suitable. In Table 3.7, for instance, there are many small incomes, but only a few large ones. If we used equal intervals for these data we would either lose a lot of information or fail in our attempt to summarise the data. In other words, we would either have just one or two classes with almost all the observations or have so many classes that the data would not be summarised. So in this case we need unequal class intervals to summarise the data while retaining the essential information.

Not reported or not stated cases

always include 'not stated' in the frequency distribution

In collecting data, it is always possible that there would be some observations for which the required information is not stated. It is extremely desirable to show the number of such cases as a separate category in a frequency distribution as this would give an indication of the extent to which the data is incomplete. Treatment of such cases in subsequent analysis present many problems as there is an obvious loss of information in this case. A real life example of such a situation is provided in Table 3.8.

Table 3.8 Age and sex distribution of employees (in completed years) in Fiji, 1993

Age group	Number of Persons Employed	
	Male	Female
15–19	2,170	1,209
20–29	17,075	9,659
30–39	17,257	9,454
40–49	11,669	4,819
50–59	4,774	1,249
Over 60	498	104
Age not stated	17,472	4,632
Total	**70,915**	**31,126**

Source: Annual Employment Survey 1993, Bureau of Statistics, Fiji, Table 8, p. 19.

'not reported' can be very significant

Apart from the simple loss of information, the not reported cases can alert analysts to possible false conclusions being drawn by only considering the reported cases. In the above example it may be suspected that the reason for employer not stating the age of employees was embarrassment at ages being either over 60 years old, or younger than 15 years. If these suspicion is correct, the frequency distribution would look very different if those cases not stating their age were accurately recorded.

For a discussion of other important aspects of frequency distributions, refer to the end of this topic "More on frequency distributions".

Tables

statistical tables

In practice, when we are preparing statistical reports or publications we apply a number of guidelines to the layout and format of frequency distributions. We call these formatted frequency distributions statistical tables. Tables are used to present univariate (for example, a table of age distribution) or bivariate (for example, a table with age categories represented in the rows and sex categories represented in the columns) data. They are also used to show more than two variables – such as tables with age, sex and region.

definition

A table is an arrangement of data in a number of rows and columns. The simplest form of a table is a column or row of numbers representing the number of units falling in the categories of a single variable and is called a *one-way classification table*.

guidelines

The guidelines for published or released statistical tables are:

- ☆ Have a reference to the table (such as a table number);
- ☆ Have a clear title;
- ☆ Have rows and columns clearly labelled;
- ☆ Specify the units of the data in the table (for example, kg);
- ☆ Include the source of the data;
- ☆ Use vertical and horizontal lines to separate the labels from the data themselves;
- ☆ Usually do not have the columns separated by vertical lines or rows by horizontal lines – this splits the table up too much;
- ☆ Space the table entries so that the table is easy to read;
- ☆ Use summary statistics (eg. sub-totals, means) to provide additional summary information;
- ☆ Include footnotes to explain any strange features in the data;
- ☆ Use appropriate rounding (usually to one or two decimal places); and
- ☆ Make sure that you have not breached confidentiality by disclosing personal or commercially sensitive information.

formatting tables

The following principles will help you format your table:

1. Put numbers most likely to be compared with each other in columns.
2. Where practical, put columns with larger values at the left of the table, and columns with smaller values at the right of the table.

A table set out following these guidelines will be much easier to read and understand.
When in doubt keep tables SIMPLE.

Parts of a table

INFORMATION BOX 3: Parts of a Table

(a) Number *(b) Title*

Table 11 Foreign Aid by Major Donors, 1995

(c) Headnote ***'000 Australian Dollars***

(d) Headings *(e) Captions* *(e) Captions*

Donor	Country: Fiji	PNG	Samoa	Tonga	Vanuatu	Total
Australia	14,151	266,667	5,862	8,600	12,173	**307,453**
New Zealand	5,094	-	4,943	4,600	4,506	**19,143**
France	472	-	-	400	10,494	**11,366**
EC	19,245	18,841	1,667	-	4,593	**44,345**
United Kingdom	377	-	-	-	3,333	**3,711**
USA	-	-	-	-	778	**778**
Canada	-	-	-	400	370	**770**
Japan	12,736	-	-	10,000	4,926	**27,662**
UNDP	660	-	977	700	2,519	**4,856**
ADB	-	-	-	-	4,099	**4,099**
Other (1)	1,415	-	18,620	600	22,975	**43,611**
Total	**54,151**	**285,507**	**32,069**	**25,300**	**70,765**	**467,793**

(f) Stubb *(g) Body*

(h) Footnote

(1) Includes both other countries and other organisations.

(i) Source

Source: SPESS, South Pacific Commission, 1998.

(a) table number — This identifies the table and precedes the title. If any report or publication contains more than one table they should all be numbered.

(b) title — The **title** is placed above the main body of the table. It should be brief and concise but fully self-explanatory. In some cases several lines of title are necessary. If a title is too long, then an **abbreviated title** may be used above the full title.

(c) headnote — In some tables it may be necessary to include a **headnote**. This is usually printed in smaller types than the title and provides supplementary information about the table or a substantial section of it. **Headnotes** are often used to specify the units of the data in the table, or the survey the data was collected in.

(d) headings — The variables in the rows and columns of the table should be defined by a **heading**.

(e) captions — The **caption** is the designation at the top of each column and it explains what each column represents.

(f) stub — This is the left hand column and its caption. It indicates the description of each row in the table.

(g) body — The **body** of the table includes the numerical information that is placed in appropriate cells governed by row and column headings. A **cell** is the intersection of one row and one column.

(h) footnotes — These provide explanations concerning individual numbers or column or rows of numbers. They are placed at the bottom of the table and are usually in smaller type. They are denoted by either letters of the alphabet or numbers and should run left to right down the page. A **new set of footnotes** should be provided with each table. Only in cases where lengthy repetition will be avoided should the words “See footnote .. to table ..” be used instead of repeating a footnote.

(i) source notes — If statistics are collected from a secondary source then this should be acknowledged below the title or more usually below the footnotes.

Rounding

reasons for rounding

Rounding is often the first step in simplifying and summarising statistical data. Good rounding is essential if a table is going to be easy to understand.

There is often a fear that "accuracy is being lost" when rounding is done. There are two arguments against this which you should consider:

accuracy versus understanding

1. Just because a computer can produce a number with lots of decimal places in it does not mean the number is really that accurate. *A number is no more accurate than the instrument which measured it.*
2. Even if the number is super–accurate and merits lots of digits, and the alternative is a rounded number which can be understood at a glance, you would normally opt for the rounded option.

rules

The following are the general guidelines for performing rounding:

> Numbers **less than 5** are **rounded down**
>
> Numbers **greater than 5** are **rounded up**
>
> If the number is **5** then it is rounded **up or down at random**

example

9**3** is rounded to 90

9**6** is rounded to 100

9**5** is rounded to 90 or 100 at random

rounding of the number 5

If the method you use for rounding the digit 5 is not the same as this, continue to use your method. The main thing is that the method used is applied *consistently* throughout tables.

make sure totals are correct

In Table 3.9 you will see that if you add up the rounded numbers of females you get 51,500 but if you round the raw number total (51,583) you get 51,600. The general guideline here is that the rounded totals should be consistent with the unrounded totals – if you were presenting only the rounded numbers the total should be 51,600.

Table 3.9 Population by State, Federated States of Micronesia, 1994

State	Un-rounded		Rounded to the nearest '00	
	Males	Females	Males	Females
Yap	5,565	5,613	5,600	5,600
Pohnpei	17,253	16,439	17,300	16,400
Kosrae	3,806	3,511	3,800	3,500

Chuuk	27,299	26,020	27,300	26,000
Total	**53,923**	**51,583**	**54,000**	**51,500**

Source: 1994 FSM Census of Population and Housing, Detailed Social and Economic Characteristics Report, 1996.

Calculating percentages

calculating percentages

Often it is helpful to present your data as percentages. To change an amount to a percentage divide it by the total and multiply by 100.

percentage

$$\text{Percent} = \frac{\text{amount}}{\text{total}} \times 100$$

guidelines

Generally we do not have more than two decimal places with percentages. The total of the percentages should add up to 100. You should also take care to say, either in the column title or as a footnote, what number was used for the total – especially if the overall total was not used. For example:

Table 3.10 Population by State, Federated States of Micronesia, 1994

State	Number		Percent of total population	
	Males	Females	Males	Females
Yap	5,565	5,613	5.3	5.3
Pohnpei	17,253	16,439	16.4	15.6
Kosrae	3,806	3,511	3.6	3.3
Chuuk	27,299	26,020	25.9	24.7
Total	**53,923**	**51,583**	**51.1**	**48.9**

Source: 1994 FSM Census of Population and Housing, Detailed Social and Economic Characteristics Report, 1996.

% total

In table 3.10, the column heading clearly states *percent of the total population* – and the total for males (51.1%) and females (48.9%) adds up to 100%.

… Exercises …

1. The local airline employs 30 people. The length of service, in completed years, for each employee is as follows:

4	1	4	12	10	5	8	14	4	1
8	8	1	4	1	8	8	0	7	3
3	2	7	12	13	11	10	6	2	3

Construct a frequency distribution. Use 0 as the lower limit of the first class and a class interval of 3 years.

2. Police files reveal the ages of persons arrested for purse snatching: 16, 41, 25, 21, 30, 17, 29, 50, 30 and 39.

(a) using 15 years as the lower limit of the first class, and an interval of 10 years, organise the age data into a frequency distribution.

(b) What are the numbers 16, 41, 25, … , called?

(c) Based on the data contained in the frequency distribution, describe the age distribution of the purse snatchers.

(b) The numbers are called _____

(c) The age distribution of purse snatchers

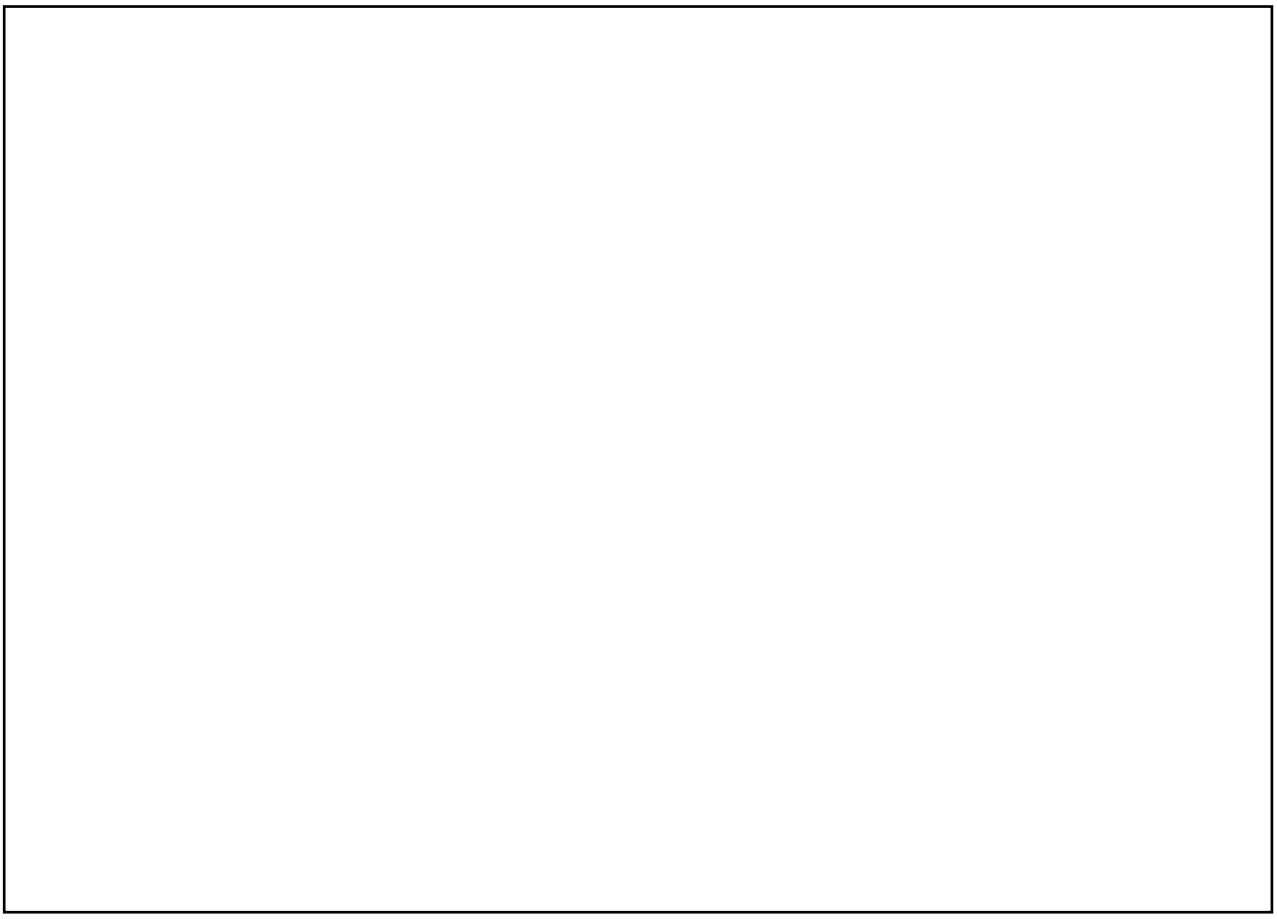
3. The local market reported the following number of people buying vegetables for the past 30 days:

85	81	65	58	47	30	51	92	85	42
55	37	31	82	63	33	44	93	77	57
44	74	63	67	46	73	52	53	47	35

Construct a table for this data. Determine the class intervals and sizes yourself.

Working:

TABLE:



… Self-Review …

1. The following data represent the amount spent (in dollars) by a random sample of 35 households on basic food items for one month:

57	34	27	41	25	18	39
33	37	39	38	47	31	42
60	58	31	47	37	16	64
34	41	43	50	46	63	51
41	30	42	37	48	28	34

(a) Starting with \$15 as the beginning number for the first class, and using an interval of 10, organise the data into a frequency distribution.

(b) Briefly describe the distribution.

2. The following is a sample of data from the Visitors Survey in Port Vila, Vanuatu conducted from June to November 1997. Construct a table containing the number of visitors by country of residence and sex. Also include the percentages for the % of the total population for each sex.

	Country	Sex		Country	Sex		Country	Sex
1.	Australia	Female	12.	Australia	Male	23.	New Zealand	Female
2.	Australia	Female	13.	Australia	Male	24.	New Zealand	Female
3.	Australia	Female	14.	Australia	Male	25.	New Zealand	Female
4.	Australia	Female	15.	Japan	Female	26.	New Zealand	Male
5.	Australia	Female	16.	Japan	Male	27.	New Zealand	Male
6.	Australia	Male	17.	Japan	Male	28.	New Zealand	Male
7.	Australia	Male	18.	New Caledonia	Female	29.	USA	Female
8.	Australia	Male	19.	New Caledonia	Female	30.	USA	Male
9.	Australia	Male	20.	New Caledonia	Female	31.	USA	Male
10.	Australia	Male	21.	New Caledonia	Male	32.	USA	Male
11.	Australia	Male	22.	New Caledonia	Male	33.	USA	Male

Working and Table:

More on Frequency Distributions

other types of frequency distributions

Apart from the frequency distributions covered in this topic, there are three other types of frequency distributions which are of significant practical use in analysing data:

- ☆ relative frequency distributions,
- ☆ frequency densities, and
- ☆ cumulative frequency distributions.

Relative frequency distributions

uses ratios

The **relative class frequency** is the ratio of the class frequency to the total frequency. The class relative frequencies given in Table 3.11 are based on the data from Table 3.3.

relative class frequency

$$\text{Relative class frequency} = \frac{\text{class frequency}}{\sum \text{class frequencies}}$$

notation

The Σ or sigma sign means sum or total. It is one of the most commonly used symbols in statistics.

Table 3.11 Class relative frequency for fish weight data (kg)

Class Group	Class Frequency	Class Relative Frequency
2.0–2.9	7	0.11
3.0 - 3.9	19	0.30
4.0 - 4.9	16	0.25
5.0 - 5.9	12	0.19
6.0 - 6.9	6	0.10
7.0 - 7.9	3	0.05
Total (Σ)	**63**	**1.00**

Source: Table 3.3

Frequency density

unequal class intervals can be compared using frequency densities

As stated earlier in this topic, unequal class intervals should be avoided. However, sometimes this is impossible and when they are used, we need to be very careful in comparing the class frequencies to determine the nature of the distribution. Class frequencies of unequal class intervals are not directly comparable. However, they can be made comparable by dividing the frequency in a class by the class interval length. This ratio is called the **frequency density**. An alternative method is to determine a suitable class interval and divide the class frequencies by the ratio of their class interval to the class

interval considered suitable.

example

For example, if a frequency distribution has ten classes, the first eight with class intervals of 5 years and the last two with class intervals of 10 years, a suitable class interval to determine the frequency density would be 5 years. For the first eight classes the frequency density would equal the class frequency since the division would be by one ($^5/_5 = 1$). For the last two classes the frequency density would equal half the class frequency since the division would be by two ($^{10}/_5 = 2$). The frequency density would then be 'per 5 years'.

An example of how to calculate frequency densities is given in Tables 3.12. It is useful to know how to calculate frequency densities when creating charts for unequal class intervals.

Table 3.12 Age at marriage of the brides and grooms married in Guam, 1994

Age at Marriage	Class Interval	Brides		Grooms	
		Class Frequency	Frequency Density (per 5 years)	Class Frequency	Frequency Density (per 5 years)
15–19	5	124	124	37	37
20–24	5	536	536	518	518
25–29	5	391	391	359	359
30–34	5	248	248	256	256
35–44	10	218	109	272	136
45 and over[1]	30	75	13	150	25
Total		**1,592**		**1,592**	

[1] Maximum age group was 70 – 74 years, so this class interval: 74 – 45 =30 (inclusive).
Source: Office of Planning and Evaluation, 1995, Guam.

Cumulative frequency distribution

more or less than a specified value

The frequency distribution gives information on the number of units in the different class intervals, thus bringing out the number of small-sized, medium-sized and large-sized units. However, there are many situations where the interest is mainly in finding out the number or percentage of units having values less than or greater than a specified value.

applications

For instance, in studying the distribution of land holdings for purposes of formulating legislation dealing with land, it is probably important to know the number or percentage of holdings having sizes less than or more than specified values. Similarly, in studying the income distributions of households or individuals, it is important to know the number or percentage of households or persons having incomes less than or greater than certain values. This type of information can be readily obtained from a frequency distribution by computing the cumulative frequencies or cumulative relative frequencies.

progressive totals of each class interval

The **cumulative frequency distribution** is obtained by calculating the progressive totals of the frequencies against each class interval. You can start at either the first class interval and proceeding down, or start from the last class interval and proceed upwards, depending on which type of cumulative frequencies are desired. The former frequency distribution is called the **'less than' cumulative frequency distribution** and the latter is called the **'greater than' cumulative frequency distribution**. If the relative frequencies in Table 3.12 are cumulated instead of the frequencies, then we get the less

than or greater than **cumulative relative frequency distribution**.

examples

Two examples of cumulative frequency distributions are provided in Tables 3.13 and 3.14.

Table 3.13 Age distribution of employees, Fiji, 1993

Age group	Frequency	Cumulative frequency (less than)	Cumulative frequency (greater than)
15–19	3,379	3,379	102,041
20–29	26,734	30,113	98,662
30–39	26,711	56,824	71,928
40–49	16,488	73,312	45,217
50–59	6,023	79,335	28,729
Over 60	602	79,937	22,706
Age not stated	22,104	102,041	22,104
Total	**102,041**		

Source: Annual Employment Survey 1993, Bureau of Statistics, Suva, Fiji

Table 3.14 Age at marriage of brides married in Guam, 1994

Age group	Frequency	Cumulative frequency (less than)	Cumulative frequency (greater than)
15–19	124	124	1,592
20–24	536	660	1,468
25–29	391	1,051	932
30–34	248	1,299	541
35–44	218	1,517	293
45 and over	75	1,592	75
Total	**1,592**		

Source: Table 3.12

cumulative frequency – less than

The cumulative frequency (less than) allows us to easily find out information such as the number of people aged under 30 years. In the example in Table 3.13 we can see 30,113 persons under 30 years were employed in Fiji in 1993 and 56,284 people below the age of 40 were employed. In Table 3.14 we can see that during 1994 in Guam in 1979, 124 brides married at an age of less than 20 years, and 660 brides were married under 25 years of age. It is very informative and at the same time very simple to read.

cumulative frequency – greater than

The cumulative frequency (greater than) is just as simple. From Table 3.13, we can see that there were 22,706 employed people who are over the age of 60 or whose ages have not been stated. From Table 3.14 we can see that there were 293 brides who were at least 35 years of age and 541 brides who were 30 years of age or more.

not stated generally excluded

These kinds of observations are very useful in a number of situations and hence the need for calculating cumulative frequencies. Care needs to be taken with the treatment of ‘not stated’. For most purposes, it is generally better to calculate the cumulative frequencies excluding this category.

Excel – PivotTable Reports

What is a PivotTable?

A PivotTable is an interactive table that quickly summarises or cross tabulates large amounts of data. You can rotate its rows and columns to see different summaries of the source data, filter the data by displaying different pages, or display the details for areas of interest.

You can create a PivotTable from a Microsoft Excel worksheet, an external database, multiple Microsoft Excel worksheets, or another PivotTable.

Creating a PivotTable to summarise data for easy analysis

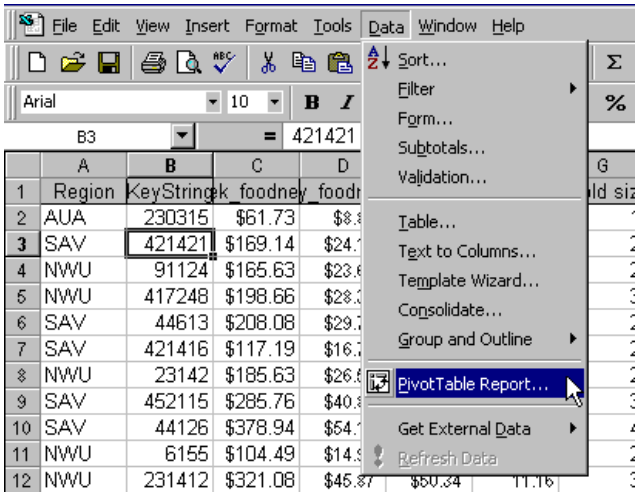
Sometimes you have to make a large number of statistical tables, especially if you are working with a dataset for the first time or looking for new trends or changes. It is sometimes easier to prepare such tables in Excel, where you can see individual data cells, sort, filter and quickly move around your data.

TIP

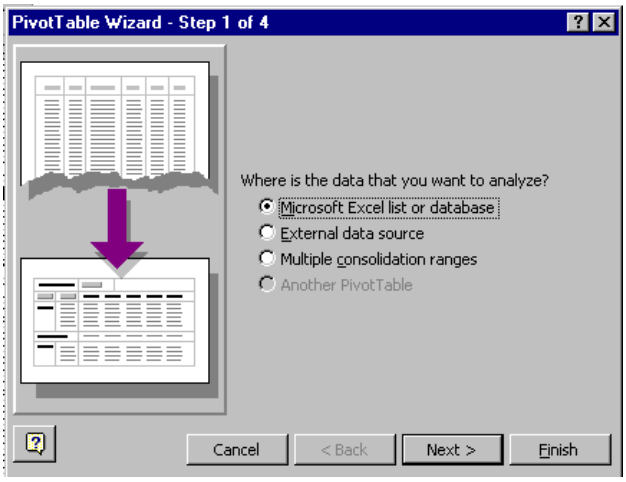
When you enter data in Excel (or import it from other applications), you can have a maximum of 65,536 ROWS (i.e. records in your dataset) and 256 COLUMNS (i.e. variables in your dataset). If you have more records in your dataset you have to 'divide' it up before you import it to Excel, and import or enter the data in separate sheets.

Steps to create a PivotTable:

1. Click any cell in your data, and then, on the Data menu, click PivotTable Report.

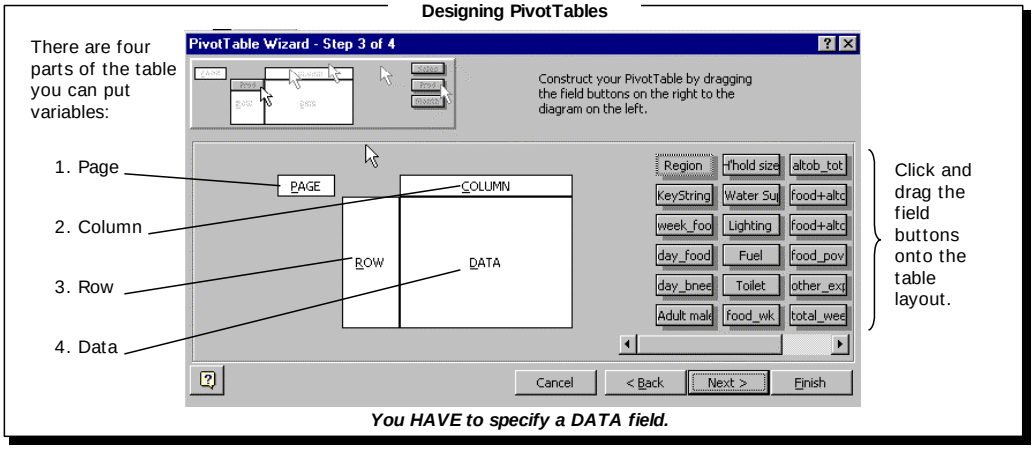


2. The PivotTable Wizard appears. In the first step (Step 1 of 4) Excel asks you where the data is you want to analyse. Make sure the first option (Microsoft Excel List Or Database) is selected and click next.

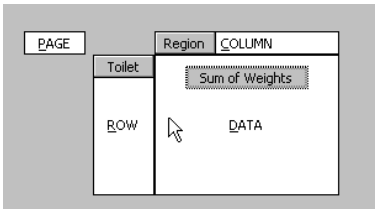


3. The Step 2 dialog box shows the range Excel has selected as the data for the PivotTable. Because you did not specify a cell range for the PivotTable, Excel by default selects all the data in the worksheet. If you have a large dataset, you could specify the columns you wanted to use to avoid memory problems. Click next.

4. The Step 3 dialog box is where you design your PivotTable. The white boxes in the middle of the dialog box form the PivotTable layout diagram. You drag the buttons located on the right side of the dialog box, which are labelled with the field names from the data, to create your PivotTable.

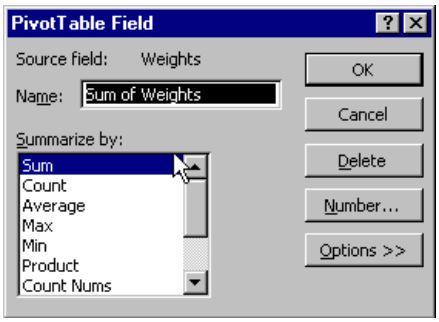


5. Drag the field buttons the COLUMN, ROW and DATA areas in the layout diagram. You can have more than one ROW or COLUMN variable, but be careful because too many variables can make the table very confusing. Keep the table design simple if you can. Usually you will use two COLUMN variables, one ROW variable and a data variable. Typical 'second' COLUMN variables are Geographic (region or province) and Gender. In the below example, Region is the COLUMN, Toilet is the ROW, and the Weights field was dragged into the DATA area.



6. Note that when you drag some fields into the data field the button changes to Sum of Weights. The PivotTable wizard knows that the Weights variable contains NUMERIC data, and by default will sum it. Since this is a sample survey, the data field should always be Sum of Weights. It is easy to change the sum option.

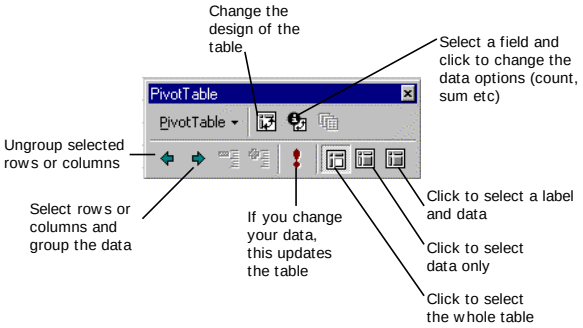
7. You can change the 'sum' option in any field (DATA, ROW or COLUMN) by double clicking on the field button. Usually the DATA field options are changed. Double click on the Sum of Weights button and the PivotTable Field dialog box opens. The two most common options are COUNT (a frequency count) or SUM (for quantitative data only). Change the Summarize by option to Count and click OK. Notice how the data field has changed to Count of Weights. Click the next button to get to the next step in the Wizard.



8. In the final step you define where you want the PivotTable stored – the new worksheet option is the default. Click on the Finish button and Excel will make the PivotTable. A PivotTable looks like this:

	A	B	C	D	E	F	G	H	I	J	K
1	Count of Weights	Region									
2	Toilet	AUA	NWU	ROU	SAV	Grand Total					
3	1	21	23	15	26	85					
4	2	14	37	47	78	176					
5	3	6	29	23	60	118					
6	9			5	2	7					
7	Grand Total	41	89	90	166	386					
8											

9. To modify your PivotTable, make sure you are in a cell in the table (or have selected rows or columns) and use the PivotTable toolbar. If the PivotTable toolbar is not displayed, first check that you have clicked inside your table to make it active. If the menu is still not displayed, from the View menu select Toolbars and make sure the PivotTable option is selected.



TIP

Usually when you have finished a PivotTable you store it as a data table rather than a PivotTable because PivotTables can take up a lot of memory. When you complete a PivotTable you select it, Copy it and use 'Paste special' to overwrite the original PivotTable. In the Paste special dialog box, select the Values option. This deletes the original PivotTable, so be sure that the table is finished!

Grouping data in PivotTables – Method 1

Pivot tables can be used to group numeric data used in columns or rows. You could group data like age into five-year groups, commodity items into a higher level of the classification, or individual incomes into ranges.

Grouping data in Pivot tables is straight forward, the only 'complication' comes with uneven class intervals or open-ended groups such as '65 +' for age, or '$60,000 +' for income. Grouping data using 'Method 2' (next) can be easier if you have uneven class intervals, providing that you don't have a large number of rows in your table.

1. Usually you group the ROW field. Click on the ROW field. Notice how the whole ROW of the PivotTable is highlighted.

2. Click on the Group button  on the PivotTable toolbar to open the Grouping dialog box. The dialog box displays the smallest and largest numbers in your table, and suggests the class interval for the table. You can change the class interval. Click on OK when you have finished.

TROUBLE?

If you are unable to group data in a PivotTable, it is most likely because the data is not formatted properly and Excel doesn't recognise it as a number. To fix this, change the format of the field to a Number format, and then refresh the PivotTable using the  button.

3. To ungroup the class intervals, click in the row and then click on the Ungroup button  on the PivotTable toolbar.

4. Open ended classes can be a problem – there are two ways of creating groups like '65 years and over' or '$100,000+'. The first is to create them manually when you have finished your pivot table and use the Sum function in Excel. The second is to use the second method of grouping data in the table – with this method you specify the observations to include in each class interval.

Grouping data in pivot tables – Method 2

In the first method of grouping data you grouped all of the row data at once. It is possible to specify the rows to include in each group.

1. Highlight the rows in the PivotTable you want to group together into one class interval. Click on the group button. Your PivotTable should look something like this:

	A	B	C	D	E	F
1	Count of age		sex			Count
2	age2	age	1	2	Grand Total	age
3	Group1	0	4	1	5	0-4
4		1	3	7	10	5-9
5		2	4	6	10	10-14
6		3	2	2	4	15-19
7		4	3	1	4	20-24
8	5	5	4	4	8	25-29

2. The group has been created, but the data for the individual observations (in this example ages) is still being displayed. Click on the Hide detail button to summarise the data by the group you have created.
3. It is easy to change the labels of the groups you have created. By default, they are called 'Group1, Group2 etc. Click on the Group1 cell and type in the correct name for the row, such as 0-4.
4. Using this method you can easily create different sized class intervals and open ended first or last classes.

The choice of method you use for defining the groups in your pivot table depends on the data you are grouping.

Showing details

Sometimes you create a PivotTable and see something in the table that needs investigation – it might be an inconsistency in the data, or a distribution that you didn't expect. This is useful when you are doing 'output editing' and creating output tables and correcting inconsistencies.

The PivotTable has a 'drilldown' capability that is used to see the data included in each cell. You can look at the specific details used to calculate a cell quickly by doing a drilldown. A drilldown creates a new worksheet that lists all the records used to calculate a PivotTable figure.

1. Go to the cell in the data field which contains the value you want to investigate (for example, it could be a higher or lower frequency count than expected, or an average which doesn't seem correct).
2. Double click on the cell. A new worksheet is added to the workbook, containing a list of all the observations in the data which went into the frequency count (or average or sum).

Creating page reports

It is simple to create a series of Pivot tables for different PAGE views using PivotTables. You would do this for different states, major groups in a classification (such as occupation or commodity items) or different businesses or people (administrative records or business activity).

1. Go into the PivotTable Wizard and add a field to the PAGE of the PivotTable.
2. Use the right mouse button to click any cell in the PivotTable.
3. On the shortcut menu, click Show Pages.
4. In the Show Pages dialog box, click OK.

A new worksheet for each observation of the PAGE field is added to the workbook. Each new worksheet contains a PivotTable that summarises the average age information for each State.

WARNING!

Be careful when selecting the field for the PAGE field – the number of observations of the PAGE variable is the number of worksheets that will be created in your workbook.

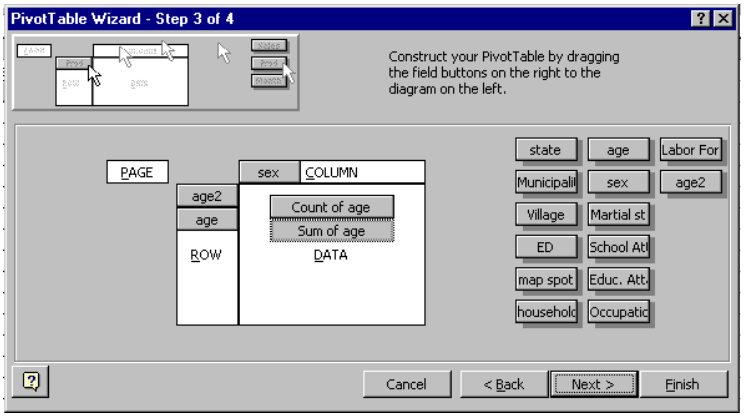
One step further ...

Changing the data field in a pivot table

Statistical tables often have both class frequency and percentages included in the table. Here a table is created with both the frequency and the percent.

1. Click on the  button on the PivotTable toolbar to change the design of the PivotTable.

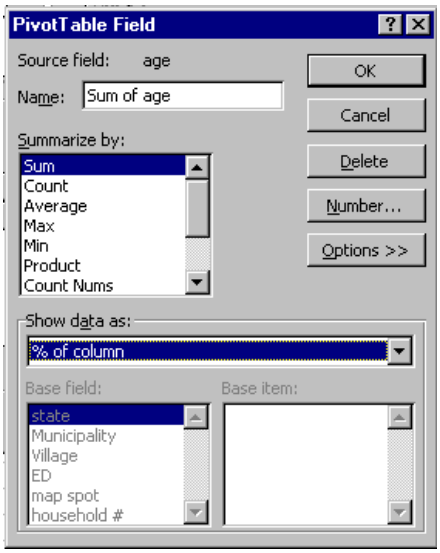
2. Click and drag a second Age button onto the DATA field. The PivotTable should look something like this:



3. Double click on the 

button to open the Pivot Table Field dialog box. Click on the 

button to open the Show data as panel.



4. Change the Show Data As option from Normal to % of column. Note the other different ways the DATA field can be formatted. Click on the OK button.

5. You will be asked to confirm that its OK to replace the contents of the destination cells. Click on OK. Your completed pivot table should look like this:

	A	B	C	D	E	F
1				sex		
2	age2	age	Data	1	2	Grand Total
3	0-4		Count of age	16	17	33
4			Sum of age2	1.18%	1.30%	1.24%
5	5-9		Count of age	15	14	29
6			Sum of age2	4.14%	4.21%	4.18%
7	10-14		Count of age	13	14	27
8			Sum of age2	5.93%	7.84%	6.84%
9	15-19		Count of age	9	8	17
10			Sum of age2	6.17%	6.14%	6.16%
11	20-24		Count of age	9	7	16
12			Sum of age2	7.96%	6.50%	7.26%
13	25-29		Count of age	7	5	12
14			Sum of age2	7.63%	5.69%	6.71%
15	30-34		Count of age	8	5	13
16			Sum of age2	10.39%	7.31%	8.93%
17	35-39		Count of age	8	5	13
18			Sum of age2	12.18%	8.47%	10.42%

6. Note how the percentage is calculated – each column's total is 100% (which might not be appropriate for all situations, the % of Row option might be more appropriate). Click on the Undo button () to return the pivot table to number format if you wish.

Summary

To	Do this	Button
Create a PivotTable	Select any cell in a data list. On the Data menu, click Pivot Table Report. Follow the steps in the Wizard, and arrange the data by dragging the field buttons and placing them where you want them in the layout diagram.	
Group data # 1	Use the right mouse button to click a ROW or COLUMN data label, and then, on the Shortcut menu, point to Group and Outline and click Group. Change the grouping if you want and click OK.	
Group data # 2	Highlight the ROW labels you want to combine into one group. On the PivotTable menu click on the Group button. Double click on the group to hide the details, or use the Hide Details button.	 
Create page fields	Use the right mouse button to click on any part of the table (or click on the PivotTable Wizard button on the PivotTable toolbar). Drag the field button for the data you want in the PAGE field into the PAGE box in the layout diagram, and then click Finish.	
Format PivotTable numbers	Click on the PivotTable Wizard button on the PivotTable toolbar. On the Shortcut menu, click Field. In the PivotTable Field dialog box, click Number, and then select a number format for the data.	
Add a new field	Click on the PivotTable Wizard button on the PivotTable toolbar. Change the design of the PivotTable by dragging the field button for the data you want to add into the layout diagram. Click Finish.	
Remove or change data orientation	In the PivotTable, drag a field button off the pivot table to remove the data from the table. Or drag the field button to a new orientation position.	
Change the summary function	Click on the PivotTable Wizard button on the PivotTable toolbar. Double click on the DATA field button (or drag another button into the DATA area). In the PivotTable Field dialog box, select a new summary function in the Summarize By list, click OK, and then click Finish.	
Refresh data	Click on the Refresh Data button on the PivotTable toolbar.	
Show underlying details	In the PivotTable, double-click the data cell for which you want to find additional information.	