

Descriptive Statistics and Graphs

In this chapter we look at how Stata produces descriptive, or univariate, statistics. These techniques are commonly used to explore the data before making decisions about further analysis and, when reporting analyses, to give the reader information about the nature and categories of the variables that have been used.

Deciding which statistics to use to describe a variable largely depends on the level of measurement of the variable. Here we refer to nominal, ordinal and interval levels of measurement. More details are given in Box 5.1 if you need them.

Box 5.1: A refresher on levels of measurement

Level of measurement is one of the fundamental building blocks in quantitative data analysis. Almost every statistical process starts with recognizing the level of measurement of the variable(s) to be used. It is vital that level of measurement is thoroughly understood.

Nominal

At the *nominal* level of measurement, numbers or other symbols are assigned to a set of categories for the purpose of naming, labelling or classifying the observations. Gender is an example of a nominal level variable. Using the numbers 1 and 2, for instance, we can classify our observations into the categories 'female' and 'male', with 1 representing female and 2 representing male. We could use any of a variety of words to represent the different categories of a nominal variable; however, when numbers are used to represent the different categories, we do not imply anything about the magnitude or quantitative difference between the categories.

- Other examples of nominal level variables are: ethnicity, nationality, race and case/control.

Ordinal

Ordinal level variables assign numbers to rank-ordered categories ranging from lowest to highest. The classic ordinal level measure is a Likert scale that has categories of strongly disagree, disagree, neither agree or disagree, agree, and strongly agree. We can say that a person in the category 'strongly agree' agrees with the statement more than a person in the 'agree' category, but we do not know the magnitude of the differences between the categories; that is, we don't know how much more agreement there is when 'strongly agree' is compared to 'agree'. Another example of an ordinal level variable is age groups such as young, middle-age and old or even 16–40, 41–64, 65 and over.

Interval/ratio

For *interval/ratio* (usually just referred to as interval) level of measurement the categories (or values) of a variable can be rank-ordered *and* the differences between these categories (or values) are constant. Examples of variables measured at the interval/ratio level are age, income, height and weight. With all these variables we can compare values not only in terms of which is larger or smaller, but also in terms of how much larger or smaller one is compared with another. In some discussions of levels of measurement you will see a distinction made between interval/ratio variables that have a natural zero point (where zero means the absence of the property) and those variables that have zero as an arbitrary point. For example, weight and length have a natural zero point, whereas temperature in degrees Celsius or Fahrenheit has an arbitrary zero point. Variables with a natural zero point are also called *ratio variables*. In statistical practice, however, ratio variables are subjected to operations that treat them as interval and ignore their ratio properties. Therefore, in practice there is no distinction between these two types.

Discrete or continuous?

Here we enter a maze of terminology and its uses and abuses. Let us start with the least controversial aspect. Nominal and ordinal

measures are also commonly referred to as categorical variables. They are always discrete. Provided that the categories of a nominal or ordinal variable are exhaustive (cover all potential categories) and mutually exclusive (cases can belong to only one category) then they are naturally discrete in that a case is assigned to a category with no other options. Interval level measures can either be discrete or continuous. For example, number of children is interval level but discrete in that the answers can only be 0, 1, 2, 3, ... and not 0.6 or 1.34. Age is an interval level measure that is continuous in that someone can actually be one minute or even one second older than someone else. But here is the twist – even if it is a truly continuous phenomenon, we usually measure it in a discrete way! Age is usually in whole years at the last birthday: 12, 13, 14, etc. If you want to think a bit more deeply about it, think about time. Even if we recorded the duration of some event to the nearest tenth of a second – usually not the case in social and behavioural sciences – then it is still a discrete measure because an event could be 23.1657 seconds.

So, let's accept that more often than not we measure continuous phenomena in a discrete way, and these discrete categories such as age last birthday have a standard unit of 1 year between them. How many interval level variables that we commonly use have truly standard units? Age, number of children, time and income are relatively straightforward, but how about all these scales we create from questionnaire items or standard 'instruments' such as the General Health Questionnaire and many other measures of psychological well-being and quality of life? Are the scores from these scales truly interval in that the difference in psychological well-being, for example, is the same between those scoring 4 and those scoring 5 as between those scoring 13 and those scoring 14? Or is there any real difference between those scoring 4 or 5 anyway? It is very common (we do it regularly) to make the assumption that these scales have standard units so that they can be analysed as interval level measures. In general, there is nothing wrong with making this assumption and we would be very restricted in the analysis we could do if we didn't, but it is worth revisiting the basic assumptions we often make in data analysis.

We conclude by summarizing the properties of the various levels of measurement in a table.

Properties of levels of measurement

	Exhaustive	Mutually exclusive	Ordered (rank)	Standard units	Meaningful zero
Nominal	yes	yes	no	no	no
Ordinal	yes	yes	yes	no	no
Interval	yes	yes	yes	yes	no
Ratio	yes	yes	yes	yes	yes

Adapted from Frankfort-Nachmias and Leon-Guerrero (2000: 13–14) and Bowling (2002: 144–7).

At the end of this chapter we examine some of the graphing abilities of Stata and cover some basic single-variable graphs.

FREQUENCY DISTRIBUTIONS

For frequency distributions, you use the **tabulate** command (which you can shorten to **tab** or **ta**). This command will show you the value labels associated with the variable, the frequency of each value, percentage frequency, and cumulative percentage frequency. For example:

```
tab sex
```

```
. tab sex
```

sex	Freq.	Percent	Cum.
male	4,833	47.09	47.09
female	5,431	52.91	100.00
Total	10,264	100.00	

You will notice that this command gives the value labels, but not the numerical value associated with the label. In order to get that, you must add the option **nolabel (nol)** to the **tab** command:

tab sex, nol

. tab sex,nol

sex	Freq.	Percent	Cum.
1	4,833	47.09	47.09
2	5,431	52.91	100.00
Total	10,264	100.00	

The option **missing** (or **miss** or **m**) gives you information on the missing values.

ta hlstat, miss

. ta hlstat,miss

health over last 12 months	Freq.	Percent	Cum.
excellent	2,930	28.55	28.55
good	4,613	44.94	73.49
fair	1,853	18.05	91.54
poor	641	6.25	97.79
very poor	219	2.13	99.92
.	8	0.08	100.00
Total	10,264	100.00	

The frequency table produced by Stata has three columns of figures. The left-hand one labelled **Freq.** has the actual counts in each of the categories of the variable. The middle column labelled **Percent** gives the percentage of that category so that all categories add up to 100%. Note here that when you use the **missing** option the missing-values category contributes to the 100% total so that the percentage is of all cases in your data. If the **missing** option is not used then the percentage is of the number of cases who answered that item – sometimes less than 100% of the total sample. The *hlstat* variable has only eight missing cases, but it could be much higher so that the percentages of the non-missing categories would be inaccurate. See the example below with the *jbsat1* variable. The right-hand column labelled

Cum. gives the cumulative percentage of the categories from the top of the table. The same inclusion rule applies when using the **missing** option.

```
. tab jbsat1,miss
```

job satisfaction: promotion prospects	Freq.	Percent	Cum.
does't apply	599	5.84	5.84
not satisfied	770	7.50	13.34
2	208	2.03	15.36
3	288	2.81	18.17
not satis/dissat	1,335	13.01	31.18
5	527	5.13	36.31
6	428	4.17	40.48
completely satis	985	9.60	50.08
.	5,124	49.92	100.00
Total	10,264	100.00	

```
. tab jbsat1
```

job satisfaction: promotion prospects	Freq.	Percent	Cum.
doesn't apply	599	11.65	11.65
not satisfied	770	14.98	26.63
2	208	4.05	30.68
3	288	5.60	36.28
not satis/dissat	1,335	25.97	62.26
5	527	10.25	72.51
6	428	8.33	80.84
completely satis	985	19.16	100.00
Total	5,140	100.00	

In the first table above the **missing** option is used. From this table you might (incorrectly) read that 9.6% are completely satisfied with their promotion prospects. However, when the missing cases are omitted from the second table above, 19.16% are completely satisfied with their promotion prospects. Almost half the total sample was not asked this item because they were not in employment at the time.

Two other options to use with the **tab** command are **plot** and **sort**. The **plot** option produces a basic bar chart of the relative frequencies of the variable categories. This chart is shown in the Results window and not through the graphing facility in Stata. The **sort** option rearranges the frequency table so that the categories are presented in descending order with the most frequent (mode) at the top of the table. These two options can be used together:

tab hlstat, sort plot

```
. tab hlstat, sort plot
  health over |
last 12 months |   Freq.
-----+-----+-----
      good |   4,613 | *****
  excellent |   2,930 | *****
      fair |   1,853 | *****
      poor |     641 | *****
  very poor |     219 | **
-----+-----+-----
      Total |  10,256
```

Compare this output with the frequency table given above for the variable *hlstat*. Using the **plot** option will mean that the percentages, category and cumulative, are not presented in the table but using the **sort** option by itself produces the same style of table as earlier by reordered categories.

tab hlstat, sort

```
. tab hlstat, sort
```

health over last 12 months	Freq.	Percent	Cum.
good	4,613	44.98	44.98
excellent	2,930	28.57	73.55
fair	1,853	18.07	91.61
poor	641	6.25	97.86
very poor	219	2.14	100.00
Total	10,256	100.00	

The last two options we will cover here are **nofreq** and **gen**. The **nofreq** option tells Stata not to present the frequencies, which in a single-variable table means that no table is produced. This option has more uses in two-way tables (or crosstabulations) and is covered in the next chapter. The **tab** command combined with the **gen** option produces a series of dummy (or binary) variables – one for every category of the variable in the table. After **gen** goes the name prefix (or stub) of the new series of dummy variables:

```
tab hlstat, gen(hlth)
```

The table output is the same as using just the **tab** command but Stata has generated five new variables all starting with *hlth* and called *hlth1*, *hlth2*, *hlth3*, *hlth4* and *hlth5*. These new variables are shown at the bottom of the list of variables in the Variables window. If you **tab** the new variable *hlth3* you can see that the same number of cases (1853) are given the value 1 as were in the third category (fair) on the *hlstat* variable:

```
tab hlth3
```

```
. tab hlth3
```

hlstat==fai r	Freq.	Percent	Cum.
0	8,403	81.93	81.93
1	1,853	18.07	100.00
Total	10,256	100.00	

Note that if cases are missing on the original variable (*hlstat*) then they will have missing values on these new dummy variables as well.

If you want to use the **tab** command to create a series of dummy variables but do not want to produce a frequency table, then you could use the **gen** and **nofreq** options together:

```
tab hlstat,gen(hlth) nofreq
```

This combination creates the new dummy variables but does not produce a frequency table in the Results window.

If you ask Stata to produce a frequency table of a variable that has a large number of categories/values then it will return an error message in red: too many values

The **if** qualifier works with almost all commands, including recoding and variable creation. Try:

```
tab hlstat if sex==1
```

This command gives you the frequency distribution of health status for males. Don't forget that double equals signs are required for conditional commands like this.

Alternatively, the **bysort** command can be combined with **tab**. For example, the command below will produce two frequency tables – one for the health status for males and one for health status of females. The **missing**, **nolabel**, **plot** and **sort** options can still be used in this combination.

```
bysort sex: tab mastat
```

```
. bysort sex:tab hlstat
```

```
-----
```

```
-> sex = male
```

health over	Freq.	Percent	Cum.
last 12 months			
excellent	1,536	31.79	31.79
good	2,149	44.47	76.26
fair	808	16.72	92.98
poor	246	5.09	98.08
very poor	93	1.92	100.00
Total	4,832	100.00	

-> sex = female

health over last 12 months	Freq.	Percent	Cum.
excellent	1,394	25.70	25.70
good	2,464	45.43	71.13
fair	1,045	19.27	90.39
poor	395	7.28	97.68
very poor	126	2.32	100.00
Total	5,424	100.00	

If you want to produce a series of frequency tables you do not need to type in each **tab** command separately. The command **tab1** will produce separate frequency tables for each of the variables in the list after the command. For example: **tab1 sex hlstat** will produce a frequency table for *sex* followed by one for *hlstat*. The variable list can, of course, be much longer.

. tab1 sex hlstat

-> tabulation of sex

sex	Freq.	Percent	Cum.
male	4,833	47.09	47.09
female	5,431	52.91	100.00
Total	10,264	100.00	

-> tabulation of hlstat

health over last 12 months	Freq.	Percent	Cum.
excellent	2,930	28.57	28.57
good	4,613	44.98	73.55
fair	1,853	18.07	91.61
poor	641	6.25	97.86
very poor	219	2.14	100.00
Total	10,256	100.00	

The options **nolabel**, **missing**, **sort** and **nofreq** can be used with the **tab1** command as well. The **gen** option can also be used but care needs to be taken as the category values from the original variables will be used, but if the values are duplicated then Stata will stop and send an error message as it will not create two variables with the same name. Our advice is to use the **gen** option only with single-variable tabulations.

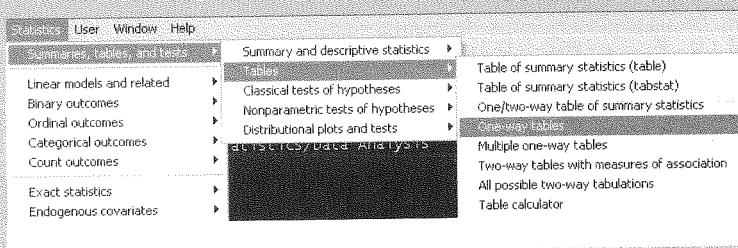
Frequency tables can also be produced by using the pull-down menus (see Box 5.2).

Box 5.2: Frequency tables using pull-down menus

Although we advocate quick progress to using do files and we use single interactive commands in the Command window in this and other chapters, Stata has the facility to use pull-down menus. You can also use these to produce descriptive statistics in frequency tables and summary statistics.

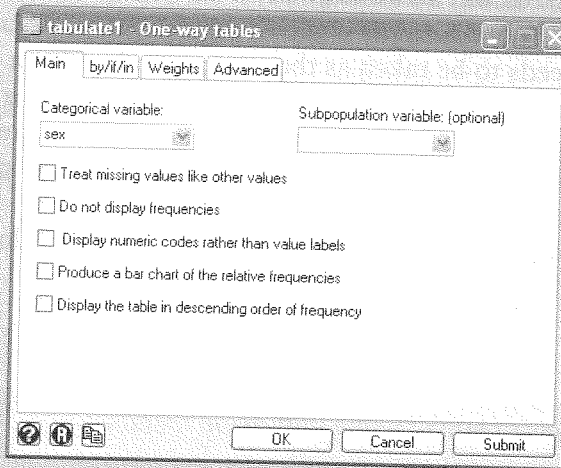
Statistics → Summaries, tables, and tests → Tables →
One-way tables

This path shows that you need to first choose the **Statistics** main pull down menu from the top toolbar then follow the blue sub menus to **One-way tables**.



This brings up the dialogue box called **tabulate 1 - one-way tables**. This dialogue box has four tabs: **Main**, **by/if/in**, **Weights** and **Advanced**.

On the **Main** tab, use the list of variables under the **Categorical variable** selector to choose the variable you want to put into a frequency table – in this case the variable **sex**.



The five option tick boxes (click in the box to choose) relate to the options:

Treat missing values like other values = **missing**

Do not display frequencies = **nofreq**

Display numeric codes rather than value labels = **no label**


Produce a bar chart of the relative frequencies = **plot**

Display the table in descending order of frequency = **sort**

In the **by/if/in** tab you can specify conditions in the same way as the **bysort** and **if** command combinations. Options in the **Advanced** tab let you create new dummy variables in the same way as the **gen** option.

SUMMARY STATISTICS

The command for common summary statistics is **summarize**, which can be shortened to **sum** or **su**. The command is followed by the variable or list of variables you wish to analyse.

Be careful, because just typing **sum** in the Command window will produce summary statistics of all the variables in the data, which can result in pages and pages of output if you have a very large data set! If you find you have done this – or made any other mistake where Stata ends up returning far too much information – you can always click on the **Break** button  to stop the command.

The **sum** command produces basic descriptive statistics such as the number of observations (Obs), mean, standard deviation (Std. Dev.), minimum and maximum values. For example:

```
sum ghqscale
```

```
. sum ghqscale
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ghqscale	9613	10.77125	4.914182	0	36

Additional statistics can be obtained by adding the **detail** option:

```
sum ghqscale, detail
```

```
. sum ghqscale, detail
```

```
subjective wellbeing (ghq) 1: likert
```

Percentiles	Smallest		
1%	3	0	
5%	5	0	
10%	6	0	Obs 9613
25%	7	0	Sum of Wgt. 9613
50%	10		Mean 10.77125
		Largest	Std. Dev. 4.914182
75%	13	36	
90%	17	36	Variance 24.14919
95%	21	36	Skewness 1.366197
99%	27	36	Kurtosis 5.713574

This expanded output includes information on the number of cases, mean and standard deviation. The minimum and maximum values can be seen with the percentile listing on the left. The 50th percentile (median) is also reported here along with the 25th and 75th percentiles for the inter-quartile range. The variance (standard deviation squared) is presented in the right-hand column along with statistics for the skewness and kurtosis of the variable (see Box 5.3).

Box 5.3: Skewness and kurtosis

The skewness statistic summarizes the degree and direction of asymmetry in the distribution of the variable. A symmetric distribution has a skewness statistic of 0. If the distribution is skewed to the left (or negatively skewed) the statistic has a negative value, and if the distribution is skewed to the right (or positively skewed) the statistic has a positive value. In the above example, the skewness has a value of 1.37 indicating that the distribution of the GHQ variable is skewed to the right. We will revisit this distribution later in the chapter when we look at graphing.

The kurtosis statistic is a summary of the shape of the distribution in relation to its peak and tails. A normal distribution has a kurtosis of 3. If the value is less than 3 then the distribution is 'flatter' than a normal distribution, with a lower peak and heavier or wider tails. If the value is greater than 3 then the distribution is 'sharper' than a normal distribution, with a higher peak and lighter or narrower tails. In this example of the *ghqscale* variable the kurtosis statistic has a value of 5.71 indicating that it is more peaked than normal, with thinner tails.

Other software calculates skewness and kurtosis statistics differently, so you need to make sure you know how the statistics are calculated so you can interpret them correctly.

In Stata you can also use the **sktest** command to formally test the normality of a variable. Two other commands – **swilk** and **sfrancia** – are also available depending on the sample size. The **sktest** command tests the skewness and kurtosis of the variable with a null hypothesis that the variable is normally distributed. Our examination of the *ghqscale* variable shows that we would reject the null hypothesis on both its skewness and kurtosis. Note that the **sktest** command does not produce the value of the statistics, and you would need to run a summary statistics command to obtain the actual values.

```
. sktest ghqscale
```

```
Skewness/Kurtosis tests for Normality
----- joint -----
Variable | Pr(Skewness) Pr(Kurtosis) adj chi2(2) Prob>chi2
-----+-----
ghqscale |          0.000          0.000
```

The **sum** command and **detail** option can be used with more than one variable. Without the **detail** option, Stata produces a list of the variables with separating lines after every fifth variable. For example:

```
sum ghqa-ghql
```

```
. sum ghqa-ghql
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ghqa	9728	2.162212	.5286127	1	4
ghqb	9728	1.7978	.778507	1	4
ghqc	9719	2.049079	.5911111	1	4
ghqd	9730	1.969476	.4843033	1	4
ghqe	9729	2.058999	.7992425	1	4
ghqf	9718	1.760136	.7077888	1	4
ghqg	9730	2.131449	.5690908	1	4
ghqh	9732	2.02949	.4746731	1	4
ghqi	9730	1.854265	.8214758	1	4
ghqj	9687	1.597399	.7404045	1	4
ghqk	9677	1.361992	.6329897	1	4
ghql	9684	2.011669	.5338565	1	4

The separating lines can be omitted by using the **sep(0)** option and changed to any other interval you wish; for example, for an interval of variables use **sep(2)**:

```
sum ghqa-ghql, sep(2)
```

```
. sum ghqa- ghql, sep(2)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ghqa	9728	2.162212	.5286127	1	4
ghqb	9728	1.7978	.778507	1	4
ghqc	9719	2.049079	.5911111	1	4
ghqd	9730	1.969476	.4843033	1	4
ghqe	9729	2.058999	.7992425	1	4
ghqf	9718	1.760136	.7077888	1	4

ghqg		9730	2.131449	.5690908	1	4
ghqh		9732	2.02949	.4746731	1	4
ghqi		9730	1.854265	.8214758	1	4
ghqj		9687	1.597399	.7404045	1	4
ghqk		9677	1.361992	.6329897	1	4
ghql		9684	2.011669	.5338565	1	4

As we have mentioned before, there are usually a few different ways of producing the statistics you want in Stata so there is some overlap with the following commands. The style of presentation differs, and for some commands this can be adjusted to suit your preferences.

If you wish to produce the standard error of the mean and confidence intervals instead of the standard deviation, minimum and maximum values you can use the **ci** command.

ci ghqscale

```
. ci ghqscale
```

Variable		Obs	Mean	Std. Err.	[95% Conf. Interval]
ghqscale		9613	10.77125	.0501212	10.673 10.8695

The default output shows the number of cases (Obs) and mean as with the **sum** command, but then the standard error of the mean (Std. Err.) and 95% confidence interval are shown. The level of the confidence intervals can be chosen by using the **level** option. For example, if you wanted the 99% confidence intervals:

ci ghqscale, level(99)

```
. ci ghqscale, level(99)
```

Variable		Obs	Mean	Std. Err.	[99% Conf. Interval]
ghqscale		9613	10.77125	.0501212	10.64212 10.90038

The **mean** command produces similar output to the default **ci** command and also has the **level** option. The other options for

mean enable you to specify how the standard error is calculated using jackknife, bootstrap or clustering adjustments.

```
. mean ghqscale
```

```
Mean estimation                Number of obs = 9613
-----+-----
      |           Mean   Std. Err.   [95% Conf. Interval]
-----+-----
ghqscale |   10.77125   .0501212   10.673  10.8695
-----+-----
```

The command **ameans** will produce the arithmetic, geometric and harmonic means with confidence intervals for the variables listed.

```
. amean ghqscale
```

```
Variable |           Type   Obs       Mean   [95% Conf. Interval]
-----+-----
ghqscale |   Arithmetic  9613   10.77125   10.673  10.8695
      |   Geometric  9602   9.802626   9.716335  9.889682
      |   Harmonic  9602   8.825263   8.724163  8.928735
-----+-----
```

An extremely useful and flexible command for producing descriptive or summary statistics is **tabstat**. This command can also be used extensively when summarizing variables by categories of another variable, and this use is covered in Chapter 6.

If you use just the **tabstat** command without any options then Stata simply returns the mean. However, the **tabstat** command can produce a large number of statistics. In the output below **tabstat** is used first on its own and then with a **statistics** option (shortened to **s**) that specifies the number of cases (**n**), mean (**me**), standard deviation (**sd**), minimum value (**min**) and maximum value (**max**). This second output is similar to that produced by the **sum** command.

```
. tabstat ghqscale
```

```
variable |           mean
-----+-----
ghqscale |   10.77125
-----+-----
```

```
. tabstat ghqscale,s(n me sd min max)
variable |          N          mean          sd    min    max
-----+-----
ghqscale |    9613    10.77125    4.914182     0    36
-----+-----
```

The **statistics** or **s** option can be used to generate other summary statistics and the output will be in the order specified inside the parentheses of the option. In this example Stata has produced the number of cases, mean and standard deviation, followed by the standard error of the mean (**sem**), skewness (**sk**) and kurtosis (**kur**). You may wish to compare this style of output with that produced by **sum** and the **detail** option.

```
. tabstat ghqscale,s(n me sd sem sk kur)
variable |          N          mean          sd se(mean) skewness kurtosis
-----+-----
ghqscale |    9613    10.77125    4.914182    .0501212    1.366197    5.713574
-----+-----
```

The default style of output of the **tabstat** command with only one variable is to put the requested statistics in one row as in the above example. However, if you are specifying a large number of statistics you may find it more convenient to have them in a column. This can be done by using a **column(variable)** option which can be shortened to **c(v)**. In the example below additional statistics have been requested. These are inter-quartile range (**iqr**) and quartiles (**q**) which includes the median (p50 in the output). The median could be requested on its own by using **med** in the statistics option brackets.

```
. tabstat ghqscale,s(n me sd sem sk kur iqr q) c(v)
stats | ghqscale
-----+-----
      N |          9613
    mean |         10.77125
      sd |         4.914182
se(mean) |         .0501212
skewness |         1.366197
kurtosis |         5.713574
      iqr |              6
      p25 |              7
      p50 |             10
      p75 |             13
-----+-----
```

One odd feature of the **tabstat** command is that when you have two or more variables the default output is to put the variables in columns. Three variables are shown in this example. So, for two or more variables, if you want the statistics to be presented in rows then you need to use the **column(statistic)** option – shortened to **c(s)** – to produce the second output below.

```
. tabstat ghqa-ghqc,s(n me sd min max)
```

stats	ghqa	ghqb	ghqc
N	9728	9728	9719
mean	2.162212	1.7978	2.049079
sd	.5286127	.778507	.5911111
min	1	1	1
max	4	4	4

```
. tabstat ghqa-ghqc,s(n me sd min max) c(s)
```

variable	N	mean	sd	min	max
ghqa	9728	2.162212	.5286127	1	4
ghqb	9728	1.7978	.778507	1	4
ghqc	9719	2.049079	.5911111	1	4

All of the summary statistics commands – **sum**, **mean**, **ci**, **ameans** and **tabstat** – can be combined with **bysort** and **if** commands. For example, if you wanted the summary statistics for these three variables, but separately for men and women, then you could use the **bysort** command.

```
. bysort sex: tabstat ghqa-ghqc,s(n me sk)
```

```
-> sex = male
```

stats	ghqa	ghqb	ghqc
N	4522	4523	4521
mean	2.125387	1.688702	2.040478
skewness	1.094095	.9064145	.8466639

```
-----
-> sex = female
```

stats	ghqa	ghqb	ghqc
N	5206	5205	5198
mean	2.194199	1.892603	2.05656
skewness	1.134487	.6264102	1.016004

However, the **tabstat** command has a **by** option that you can use to specify output split by categories of another variable. The advantage of using the **by** option is that the summary statistics for the total sample are also produced. The output below comes from the **by** option and you can see that the male and female summary statistics are given above the total sample statistics.

```
. tabstat ghqa-ghqc,s(n me sk) by(sex)
```

Summary statistics: N, mean, skewness
by categories of: sex (sex)

sex	ghqa	ghqb	ghqc
male	4522	4523	4521
	2.125387	1.688702	2.040478
	1.094095	.9064145	.8466639
female	5206	5205	5198
	2.194199	1.892603	2.05656
	1.134487	.6264102	1.016004
Total	9728	9728	9719
	2.162212	1.7978	2.049079
	1.10933	.7433614	.932651

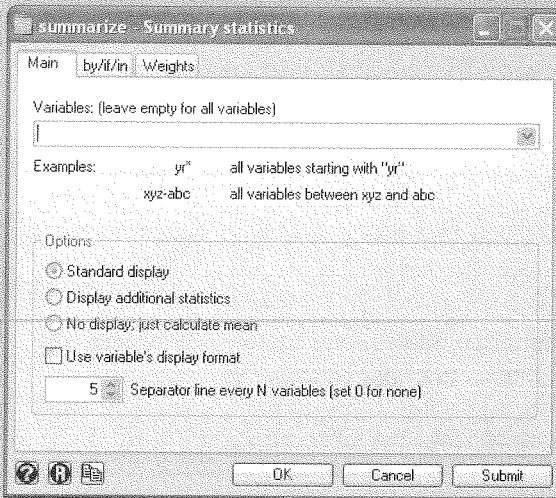
Summary statistics can also be produced using the pull-down menus (see Box 5.4).

Box 5.4: Summary statistics using pull-down menus

The path:

Statistics → **Summaries, tables, and tests** → **Summary and descriptive statistics** → **Summary statistics**

takes you to a dialogue box called **summarize - Summary statistics**. It has three tabs. In the **Main** tab you enter the variables you want statistics for by either typing them into the **Variables** box or by scrolling through the list of all variables and selecting the ones you want.



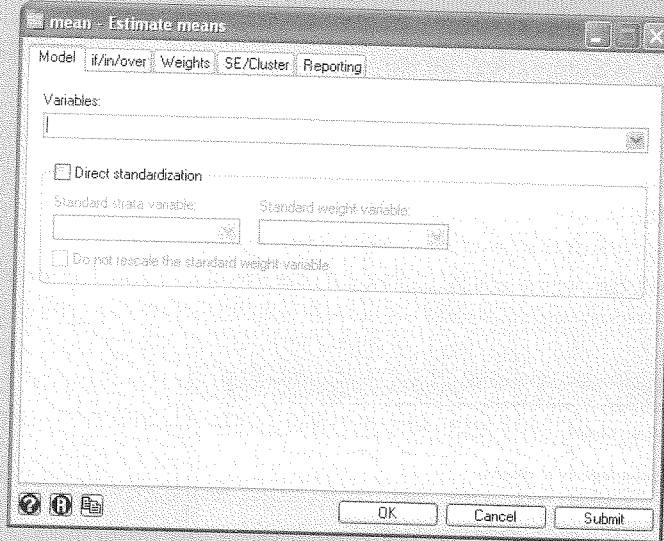
If you select the **Display additional statistics** option then Stata will present the same output as the **detail** option. In the **by/if/in** tab you can specify conditions in the same way as the **bysort** and **if** command combinations.

This path will take you to a dialogue box equivalent to the **means** command:

Statistics → **Summaries, tables, and tests** → **Summary and descriptive statistics** → **Means**

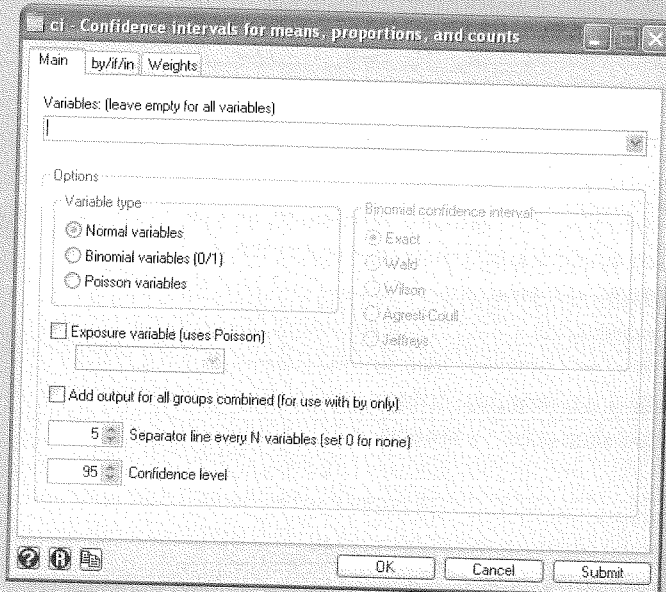
The variables are entered as before in the **Model** tab, with command combinations in the **if/in/over** tab. The **SE/Cluster** tab allows you to specify the type of standard error calculation appropriate

► for your data. The **Reporting** tab allows you to specify confidence intervals other than the default 95%.



This path takes you to a dialogue box equivalent to the **ci** command:

Statistics → **Summaries, tables, and tests** → **Summary and descriptive statistics** → **Confidence intervals**



GRAPHS

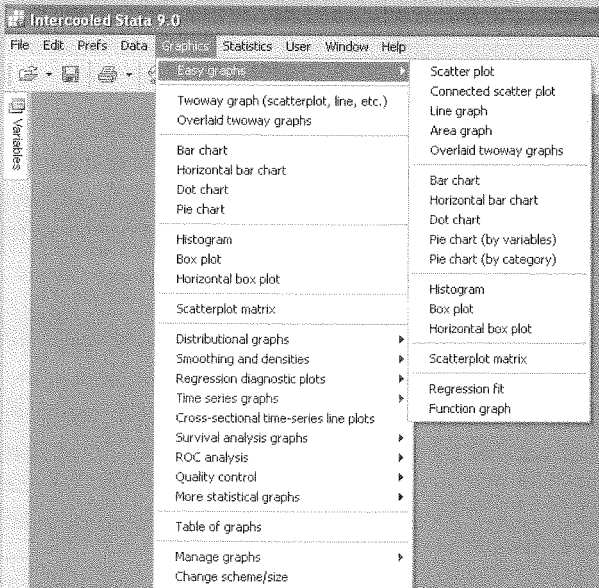
The graphing facilities in Stata have improved significantly since version 6 and now Stata is able to produce publication-quality graphics. The downside is that the commands have become more complicated and sometimes extend to two or three lines (often more), and the graphics now have their own manual. Rather than try and cover all the command options and structure, experience has taught us that it is better to use the pull-down menus to introduce graphing in Stata, much as it goes against our aim to get you using do files as soon as possible. In version 9 there is a **Graphics** pull-down menu function called **Easy graphs** which was removed in version 10, but the functions are retained in other menus. Here we cover using the main graph functions in the **Graphics** pull-down menu in version 10. If you are using version 9, then see Box 5.5 for an introduction to the use of **Easy graphs**.

Box 5.5: Easy graphs in version 9

The path through the pull-down menus is:

Graphics → **Easy graphs**

This opens a full list of graphing options:



► Pie charts

Graphics → Easy graphs → Pie chart (by category)

This path will take you to a dialogue box called **graph pie - Pie chart (by category)** where you enter the variable name for which you want to graph the categories. The pie chart produced from the default settings can be a little messy and may need tidying up before it is ready for a report.

Box plots

Graphics → Easy graphs → Box plot

Graphics → Easy graphs → Horizontal box plot

The box plot can either be vertical or horizontal depending on its use and your preference.

Histograms

Graphics → Easy graphs → Histogram

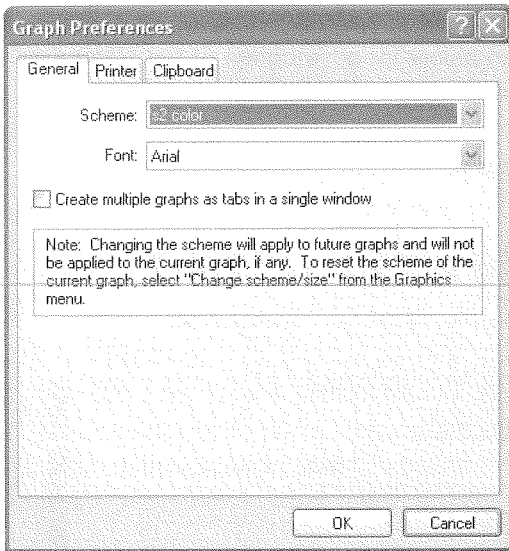
The pull-down menu takes you to a dialogue box called **histogram - Histogram for continuous and categorical variables**. In the **Main** tab you can either type in or scroll down the list of variables in the **Variable** box. On the right-hand side you select whether the variable chosen is either continuous or discrete. If you want the histogram to display percentages instead of the default density scale on the Y axis you select the **Options** tab and select **Percent** in the **Y axis** box on the lower left-hand side. The level of information given is probably enough to judge the distribution but falls well short of report quality. Note that the default is that the bars still touch even for discrete variables, which many would consider to be technically incorrect.



For continuous variables, after typing in or selecting the variable, select the **Continuous data** option in the **Main** tab. In the **Options** tab you can tick the **Add normal density plot** option in the **Density plot options** box on the right-hand side. This will show a normal distribution for the same mean and standard deviation as the variable so you can compare the actual distribution with an expected normal distribution.


The default graph format and colours are determined by the graph preferences through the pull-down menu path

Edit → Preferences → Graph Preferences

This brings up a box that shows that the default scheme is s2 color and the default font is Arial. If you prefer your graphs to be formatted differently then you can change these. Stata has numerous schemes to choose from. Once set in this box, every future graph will be formatted in this way. However, each individual graph scheme and fonts can be changed in the tabs when you are creating the graph and you should check that what you are choosing is consistent with the general scheme.




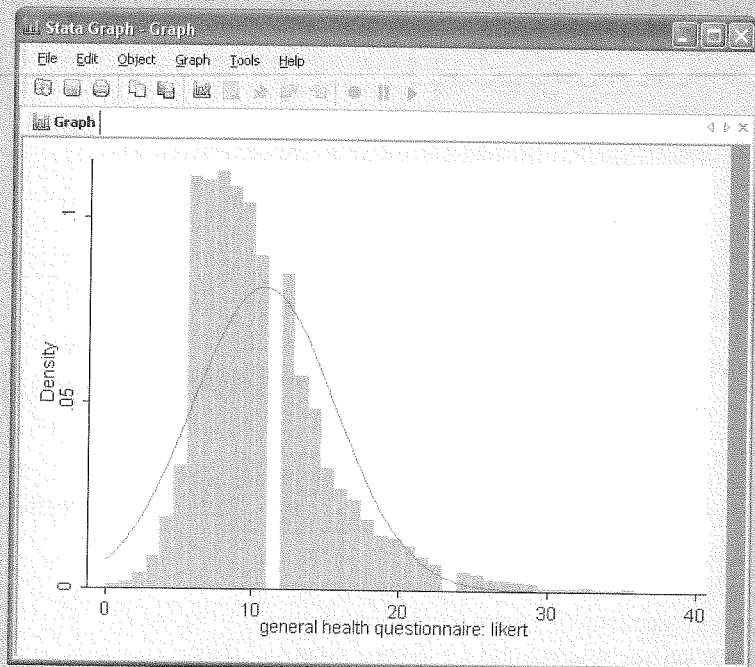
When you create a graph it opens in a new window. You can leave this window open while you return to Stata (by clicking anywhere in the main Stata windows), but then the graph is not shown. On the toolbar there is an icon –  in version 9 and  in version 10 – that will bring the graph back to the front. The icon is dimmed when there isn't a graph to show.

The biggest, and best, change in version 10 is the interactive graphics editor. To open the editor you need to right-click on the **Graph** window and select **Start Graph Editor** or click on this icon  on the toolbar above the graph when it's first created (see Box 5.6). You can do this at any time a graph window is open.

The editor is very powerful and very intuitive and it allows you, for example, to easily add text and arrows to more clearly indicate what the graph is showing, as well as format the graph, axes, and titles. The Stata website says that you should get the hang of it in a few minutes as you just click on what you want to change; they are right. It's easy to get the hang of and makes many of the options available in the tabs unnecessary to start with as you can change them in the Graph Editor. Try creating basic graphs and then editing them. For a detailed text on Stata graphics, see Mitchell (2008).

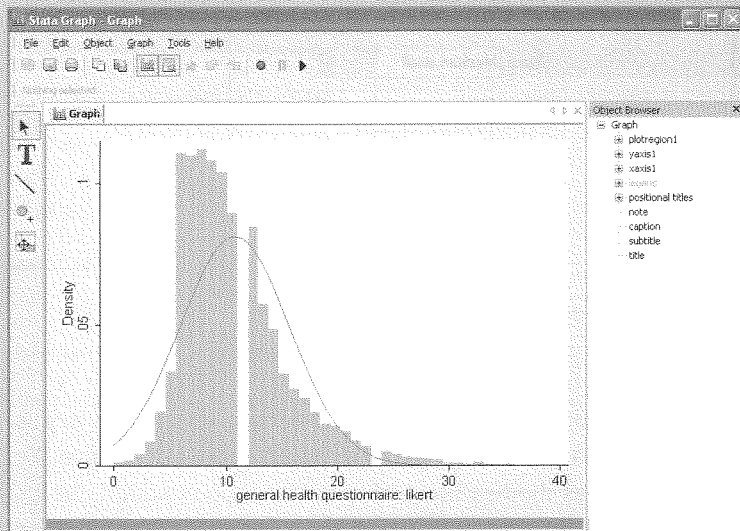
Box 5.6: Graphs and the Graph Editor

When you first create a graph it opens in a window called **Stata Graph - Graph** as shown below for a histogram of the GHQ. If this graph is what you want to save or copy then you can do this from the pull-down menus or using the icons on the toolbar. If you wish to use the Graph Editor then click on the  icon on the toolbar.



This opens the Graph Editor which looks like this. Note here that you cannot do any data analysis in Stata while the Graph Editor is

open. If you need to do this then save the graph (see Box 5.7) and return to editing the graph after your analysis. To edit an area of the graph either click on it with the cursor or click on the part in the list on the right-hand side.

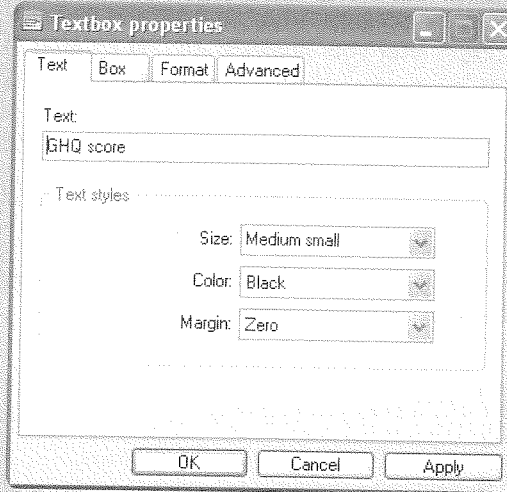


For example, if we wanted to change the X axis title we could (a) double-click on the title **general health questionnaire: likert** or (b) click on the + symbol next to **xaxis1** in the **Object Browser** on the right of the Editor. This expands to show the title in the tree, then double-click on **title**.

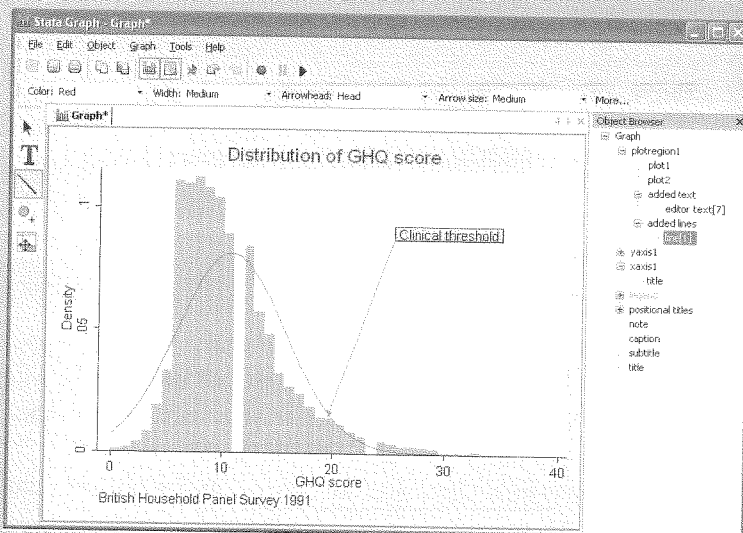


Both of these will bring up a new window where you can change the text, font, colour and position. If you are not exactly sure what the changes will look like we suggest you click on the **Apply** button which will keep this window open while you can see the changes in the graph. Once you are happy with the changes then click on the **OK** button to close this window.

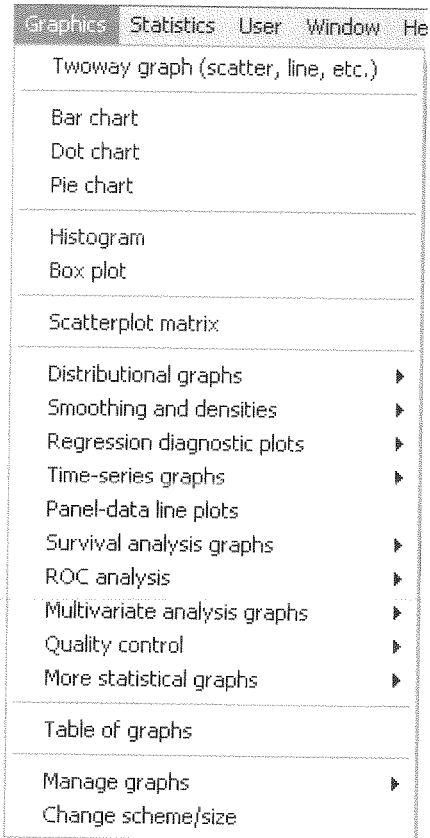




After changing the X axis title and adding a main title, caption and some text the final graph looks much more informative. You can see that as the text box and arrow have been added, the tree in the **Object Browser** has expanded to include **added text** and **added lines** which you can double-click on to edit. Above the graph on the left you can see a tab with **Graph*** on it. This is the name of the graph – at this stage we haven't saved it – and the * tells you that it has been changed since it was last saved.



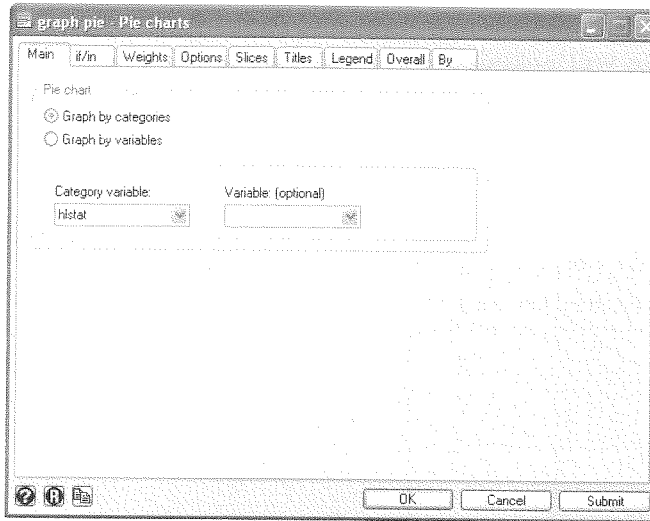
The **Graphics** pull-down menu has numerous choices, but in this chapter we will introduce some of the basic single-variable graphs: pie charts, box plots and histograms. The histograms in Stata cover both continuous and discrete variables. We cover two-way, or two-variable, graphs in Chapter 6.



PIE CHARTS

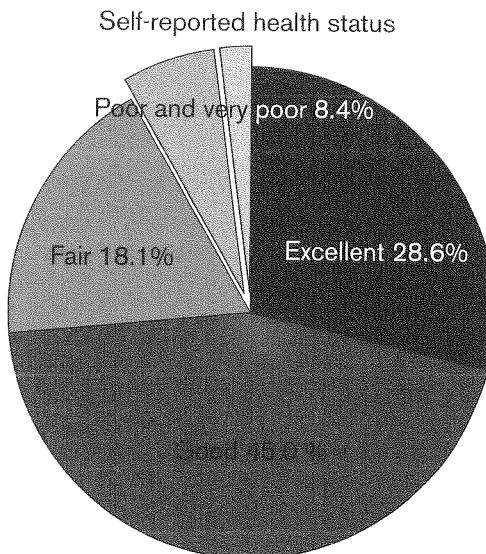
Graphics → Pie chart

This path will take you to a dialogue box called **graph pie - Pie charts** where, in the **Main** tab, you enter the variable name for which you want to graph the categories. Pie charts are most appropriate for variables that have relatively few categories as they are better displayed and interpreted. In this example we use the *hlstat* variable. We click on **Graph by categories** as we want to see the categories of the variable.



You can either type in the variable name or use the down arrow to scroll through the list of variables and select one. The **if/in** tab allows you to specify that the graph is to be based on a subset of cases. The **Titles** tab gives a number of options for titles and labels on the graph. The **Options** tab is where you can change the colour scheme.

The pie chart produced below is from these settings and some minor editing in the **Graph Editor**; it is still a little messy and would need tidying up before it is ready for a report.

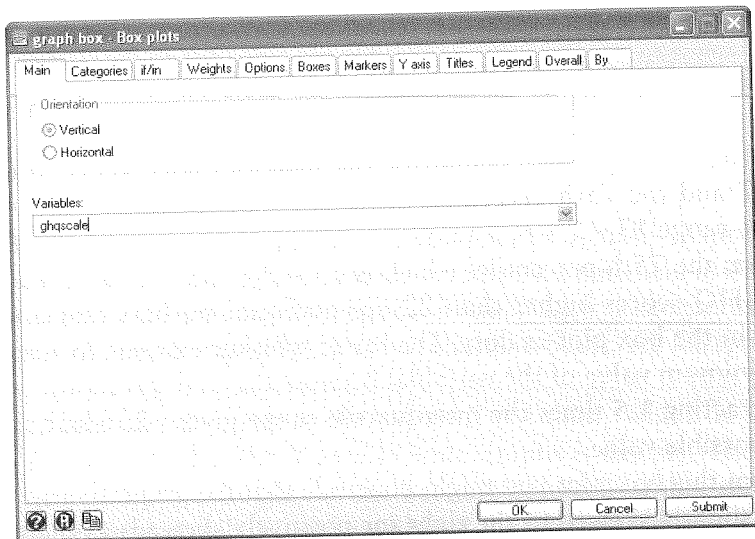


BOX PLOTS

Graphics → Box plot

Box plots (or box-and-whisker plots) are a good way to examine the distribution of a variable(s). The distribution is shown against an axis of the values of the variable. The box plot can either be vertical or horizontal depending on its use and your preferences.

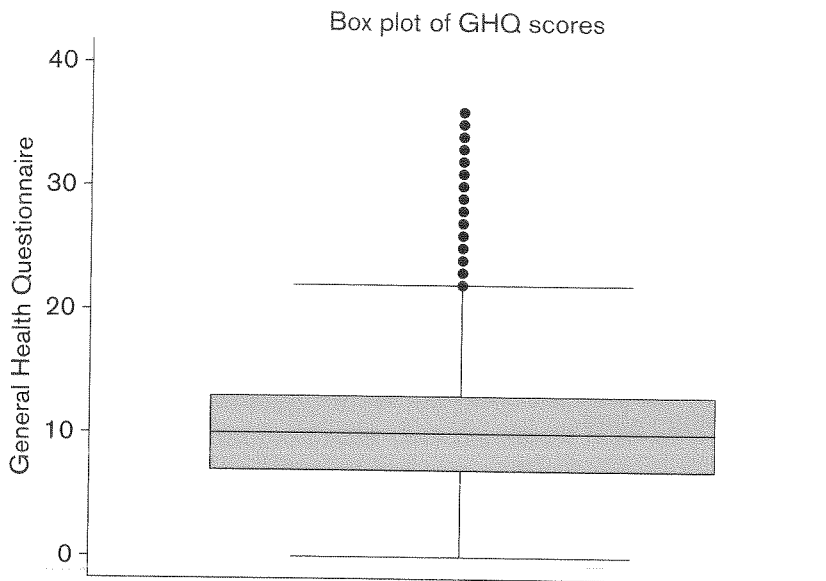
The pull-down menu path takes you to a dialogue box titled **graph box - Box plots** which has a **Variables** box where you can either type or scroll through and select variable(s). For now we concentrate on single-variable plots. There are other tabs in this window. The **Categories** tab is where you specify categorical variables to produce box plots to compare, and this is covered in more detail in the next chapter. The **if/in** tab allows you to specify that the graph is to be based on a subset of cases. The **Titles** tab gives a number of options for titles and labels on the graph. The **Options** tab is where you can change the colour scheme and choose how to treat missing values.



Box plots are normally used for interval level (or continuous) variables and are not appropriate for ordinal or nominal level variables. In this example we want to graph the variable *ghqscale* so we enter this in the **Variables** box and select **OK**.

The box in the box plot shows the inter-quartile range; that is, the values from the 25th to the 75th percentile. The line in the

middle of the box is the median or 50th percentile. The whiskers show the lower and upper adjacent values. These are the furthest observations which are within 1.5 times the inter-quartile range of the lower and upper ends of the box. If you wish to check these figures you can see the earlier results from the `su ghqscale, detail` command.

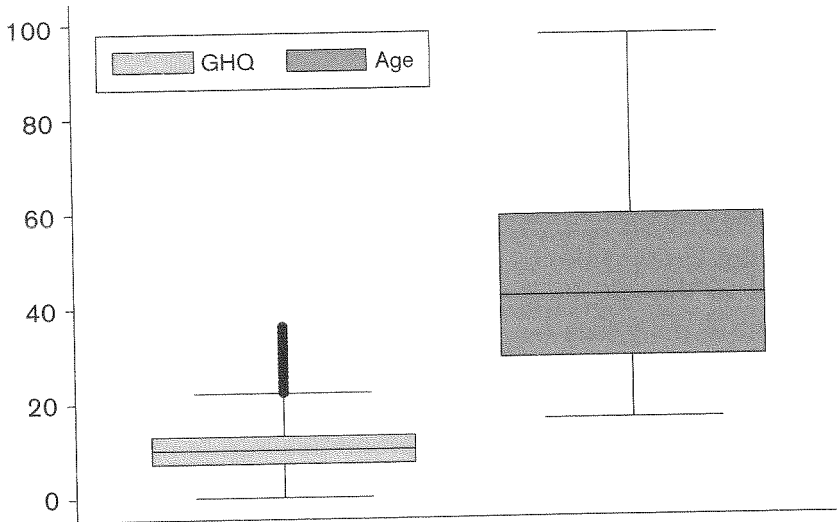


In this example the median is 10, with the 25th percentile being 7 and the 75th percentile being 13. This makes the inter-quartile range $13 - 7 = 6$. Therefore, the upper whisker is $1.5 \times 6 = 9$ from the 75th percentile, which is $13 + 9 = 22$. All cases that have GHQ scores higher than 22 are potential outliers and are shown in the box plot as dots. The lower whisker extends to zero (the minimum value of the variable) because the 25th percentile is 7; subtracting 1.5 times the interquartile range gives -2 , which is not a possible value.

From this box plot you might conclude that the *ghqscale* variable is slightly skewed to the right (or positively skewed) as there are a number of outliers above the upper whisker.

If you wish to produce a series of box plots for two or more variables then you can enter two or more variables in the **Variables** box on the **Main** tab. In the example below we have added the *age* variable to the *ghqscale* variable. Stata produces the two plots side by side. Note that Stata uses the variable labels as they are in the data set so good labels are useful, but with a quick

use of the Graph Editor you can easily change them into more meaningful labels if necessary. You need to be careful when choosing variables to graph together because if one has a very small range compared to the other then the plot will be so compressed that judging its distribution will be difficult.



HISTOGRAMS

Graphics → Histogram

Histograms are the most common way of visually inspecting the distribution of a variable. However, there is a minefield of discipline-specific terminology to negotiate when using histograms and/or bar charts (bar graphs). For bar charts in Stata, see Chapter 6.

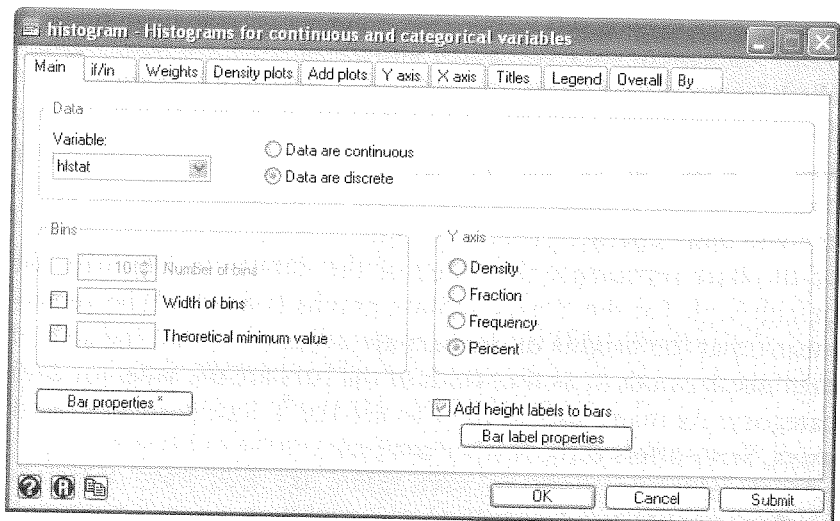
The main characteristic of a 'pure' histogram is that the area of the bars represents the value of the data, which is why the default scale for the Y axis in Stata graphs is density. The density means that the heights of the bars are adjusted so that the sum of their areas equals 1, as the width of the bars are the same for each category. As many users prefer to see the Y axis scaled in other ways, Stata offers three further options:

- Fraction: the height of all the bars equals 1.
- Frequency: the height of the bars is equal to the number of observations in the category.
- Percent: the height of all the bars equals 100.

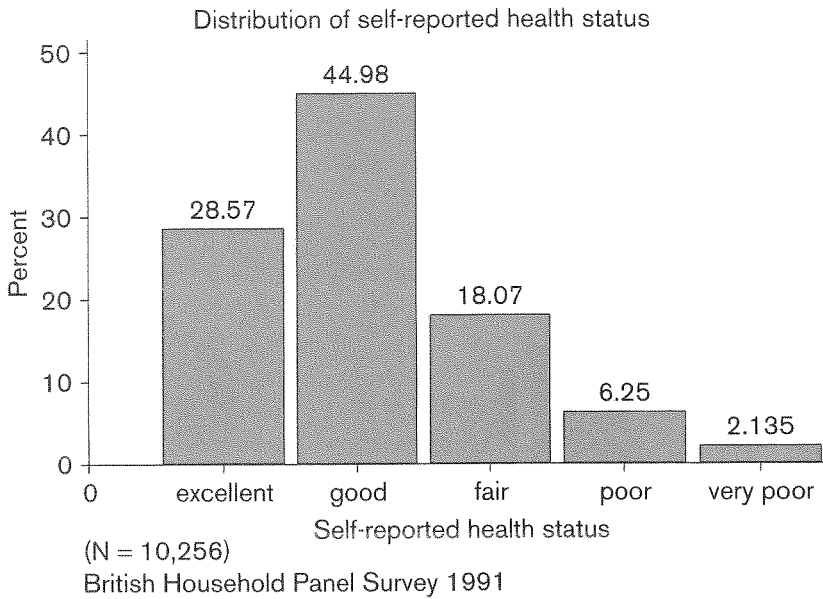
One of the main distinctions Stata uses in constructing histograms is whether the variable is continuous (interval) or discrete (ordinal or nominal). For continuous variables Stata can automatically determine the grouping of the values to display in the bars of the chart, but for discrete variables one bar for every category is shown. The number of bars used for continuous variables is an option that we come to later.

In the first example, we use a discrete variable *hlstat*. The pull-down menu takes you to a dialogue box called **histogram - Histograms for continuous and categorical variables**. In the **Main** tab you can either type in or scroll down the list of variables in the **Variable** box. To the right of this you select whether the data chosen is continuous or discrete.

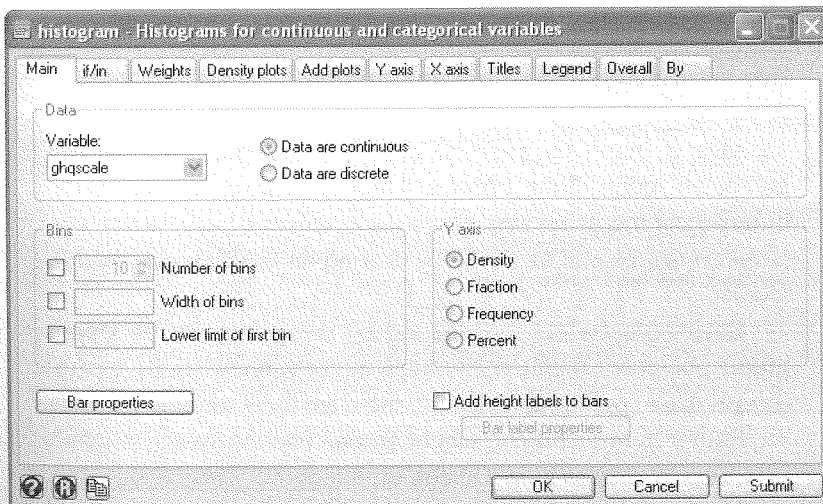
In this case we also want the histogram to display percentages instead of the default density scale on the Y axis so we select **Percent** in the **Y axis** box on the lower left-hand side. We have specified that 'height labels' are added to the bars. As we have asked for the bars to represent percentages, percentages will be shown. You can see that the **Bar properties** button on the lower left has an * on it, which shows that we have asked for some of the bar characteristics to be changed from the default settings. These are to do with presentation as we want the bars to be separated rather than touching, the latter being the default setting.



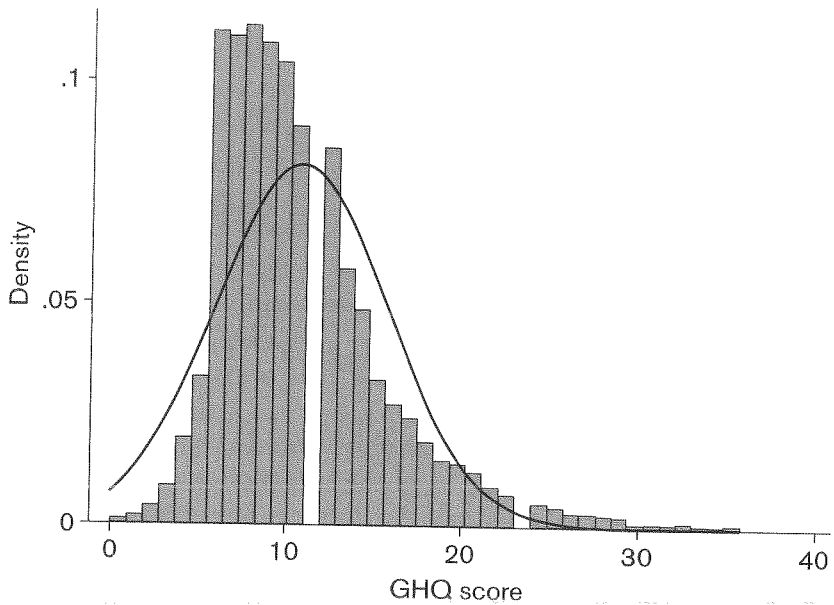
These settings followed by some editing in the Graph Editor produces the histogram shown below.



We follow a similar process in this second example but use a continuous variable *ghqscale* as in the box plot examples above. This time, after typing in or selecting the variable, select the **Continuous data** option in the **Main** tab. We leave the Y axis as the default density scale but we tick the **Add normal density plot** option in the **Density plots** tab. This will show a normal distribution for the same mean and standard deviation as the variable so we can compare the actual distribution with an expected normal distribution.



These options will produce the histogram shown below, which will need further editing in order to be of report quality. As you can see from the actual distribution against the expected normal distribution the variable *ghqscale* is slightly skewed to the right, confirming what we saw in the box plots.



Multiple graphs can also be combined into a single display (see Box 5.7).

Box 5.7: Saving and combining graphs

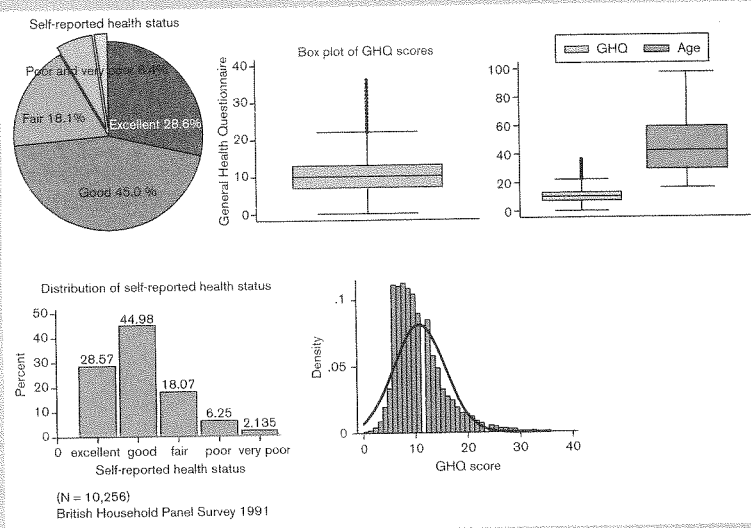
To save a graph for the first time use the **File** → **Save As** pull-down menu. Then choose where you want to save it by browsing in the usual way. Once the graph has been saved any further changes can be saved by clicking on the disk icon on the toolbar.

Saved graphs can be opened by using the **File** → **Open graph** pull-down menu in the main Stata window or by typing **graph use** "path/graphname" in the Command window.

Rather than displaying single graphs you may want to combine a set of graphs in one image for a report or publication. This

is easily done in Stata by using the **graph combine** command. Below is an example of some of the graphs we have created in this chapter combined into one image using the command:

```
graph combine pie.gph box.gph box2.gph ///
    histogram1.gph histogram2.gph
```



There are options to the **graph combine** command to choose columns or rows and positions. The really nice thing is that the new combined image opens in the Chart Editor so you can edit the image even more to ensure it's exactly what you want.

DEMONSTRATION EXERCISE

In Chapter 3 we manipulated the individual level variables and saved a new data set called demodata1.dta. In Chapter 4 we merged a household level variable indicating the region of the country onto the individual level data and saved the data with a new name demodata2.dta. In this chapter we analyse the variables we are going to use for their distribution, measures of central tendency and, for continuous variables, their normality.

First, we determine the level of measurement of all the variables:

<i>Variable</i>	<i>Level of measurement</i>
<i>female</i>	nominal
<i>age</i>	interval
<i>agecat</i>	ordinal
<i>marst2</i>	nominal
<i>empstat</i>	nominal
<i>numchd</i>	ordinal
<i>region2</i>	nominal
<i>d_ghq</i>	nominal
<i>ghqscale</i>	interval

Starting with the nominal and ordinal level variables, we simply tabulate these to examine the number of cases in each of the categories.

```
tab1 female agecat marst2 empstat numchd ///
      region2 d_ghq
```

```
. tab1 female agecat marst2 empstat numchd ///
      region2 d_ghq
```

-> tabulation of female

female indicator	Freq.	Percent	Cum.
male	3,914	47.95	47.95
female	4,249	52.05	100.00
Total	8,163	100.00	

-> tabulation of agecat

age categories	Freq.	Percent	Cum.
18-32 years	2,956	36.21	36.21
33-50 years	3,336	40.87	77.08
51-65 years	1,871	22.92	100.00
Total	8,163	100.00	

-> tabulation of marst2

marital status 4 categories	Freq.	Percent	Cum.
single	1,619	19.83	19.83
married	5,786	70.88	90.71
sep/div	569	6.97	97.68
widowed	189	2.32	100.00
Total	8,163	100.00	

-> tabulation of empstat

employment status	Freq.	Percent	Cum.
employed	5,575	70.89	70.89
unemployed	505	6.42	77.31
longterm sick	244	3.10	80.42
studying	224	2.85	83.27
family care	913	11.61	94.88
retired	403	5.12	100.00
Total	7,864	100.00	

-> tabulation of numchd

children 3 categories	Freq.	Percent	Cum.
none	5,182	63.48	63.48
one or two	2,443	29.93	93.41
three or more	538	6.59	100.00
Total	8,163	100.00	

-> tabulation of region2

regions 7 categories	Freq.	Percent	Cum.
London	898	11.00	11.00
South	2,504	30.67	41.68
Midlands	1,399	17.14	58.81
Northwest	849	10.40	69.21
North and Northeast	1,310	16.05	85.26
Wales	423	5.18	90.44
Scotland	780	9.56	100.00
Total	8,163	100.00	

-> tabulation of d_ghq

d_ghq	Freq.	Percent	Cum.
0	6,271	81.29	81.29
1	1,443	18.71	100.00
Total	7,714	100.00	

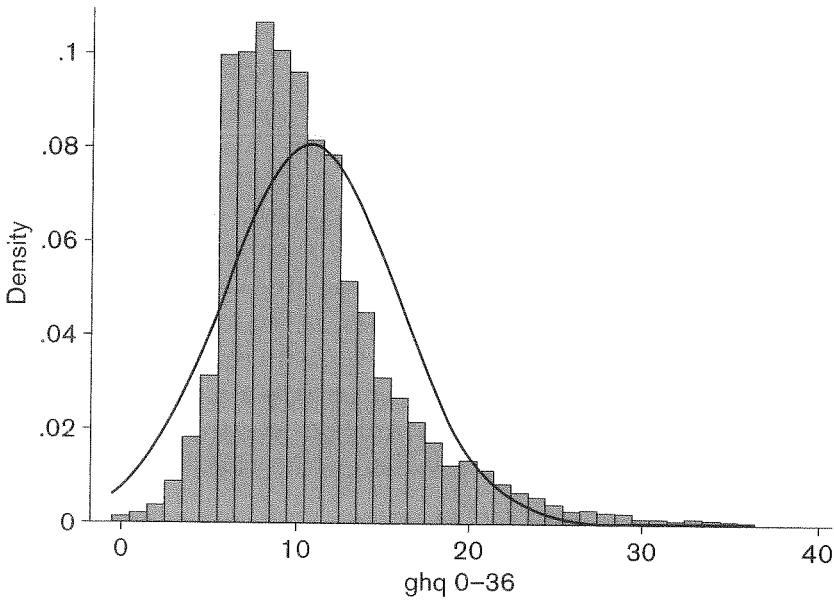
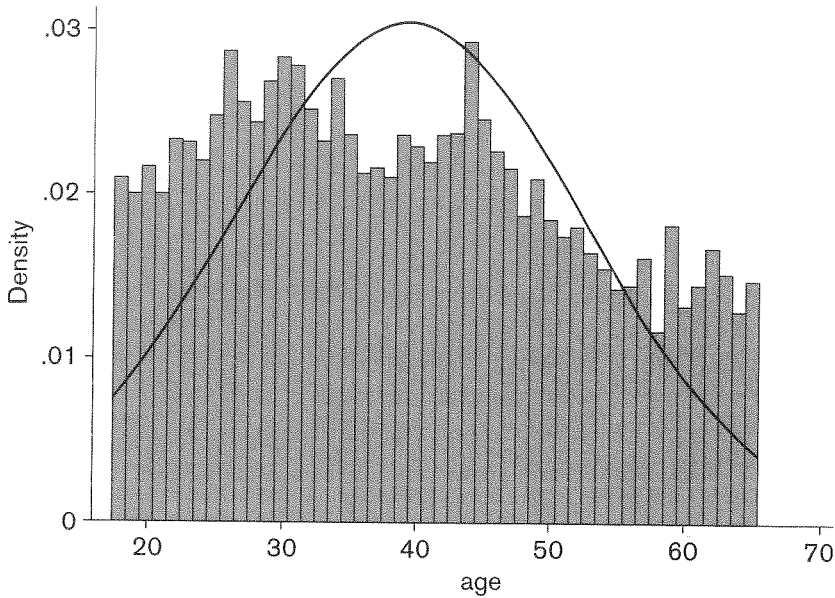
For the interval level variables we have a number of options when inspecting them. We start with simple descriptive statistics.

su age ghqscale

. su age ghqscale

Variable	Obs	Mean	Std. Dev.	Min	Max
age	8163	39.32733	13.08993	18	65
ghqscale	7714	10.78727	4.945154	0	36

Next, we use the pull-down menus to produce histograms of the two variables with expected normal distributions in order to compare them and judge any skewness. The histograms below show that the distribution of the *age* variable is rather flat compared to the normal distribution, with more cases in the younger ages than the older ones. The distribution of the GHQ variable (*ghqscale*) is more peaked than the expected normal distribution and slightly skewed to the right.



We now use the **tabstat** command to produce summary statistics for these two variables and then test their skewness and kurtosis using the **sktest** command (see Box 5.3).

```
tabstat age ghqscale, s(sk kur)
sktest age ghqscale
```

```
. tabstat age ghqscale, s(sk kur)

      stats |          age    ghqscale
-----+-----
skewness |    .2223762    1.364399
kurtosis |    1.981199    5.661401
-----+-----

. sktest age ghqscale
```

```
Skewness/Kurtosis tests for Normality
----- joint -----
Variable | Pr(Skewness) Pr(Kurtosis) adj chi2(2) Prob>chi2
-----+-----
      age |          0.000          0.000          .          .
ghqscale |          0.000          0.000          .          .
```

From the first part of the output you can see that the *age* variable has a skewness statistic of 0.22 which, being positive, indicates it is slightly skewed to the right, and a kurtosis statistic of 1.98 which, being less than 3, indicates that the distribution is flatter than normal. For the *ghqscale* variable the skewness value of 1.36 indicates that the distribution is skewed to the right and the kurtosis value of 5.66 indicates a distribution more peaked than normal.

Most of the techniques that we will use in this demonstration are robust to reasonable departures from normality such as these, but for the sake of this exercise we will consider what we might do if we wanted to transform the *ghqscale* variable to a distribution closer to normal. For distributions that are skewed to the right, one of the usual transformations used is the logarithmic (either natural logarithm or base 10). However, with this variable there is a problem in that it contains cases that have a score of zero. The logarithmic transformation of zero is not possible as it is minus infinity. You can see this if we transform the variable as it stands.

```
gen ln_ghq=ln(ghqscale)
su ghqscale ln_ghq
```

```
. gen ln_ghq=ln(ghqscale)
(459 missing values generated)

. su ghqscale ln_ghq
```

```
Variable |      Obs      Mean   Std. Dev.   Min      Max
-----+-----
ghqscale |    7714   10.78727   4.945154     0      36
ln_ghq   |    7704   2.283492   .4439032     0   3.583519
```

The new variable (*ln_ghq*) has 10 fewer cases than the original variable (*ghqscale*). This is because there were ten cases with a value of zero in the *ghqscale* variable. This can be checked by either:

```
tab ghqscale if ghqscale<2
```

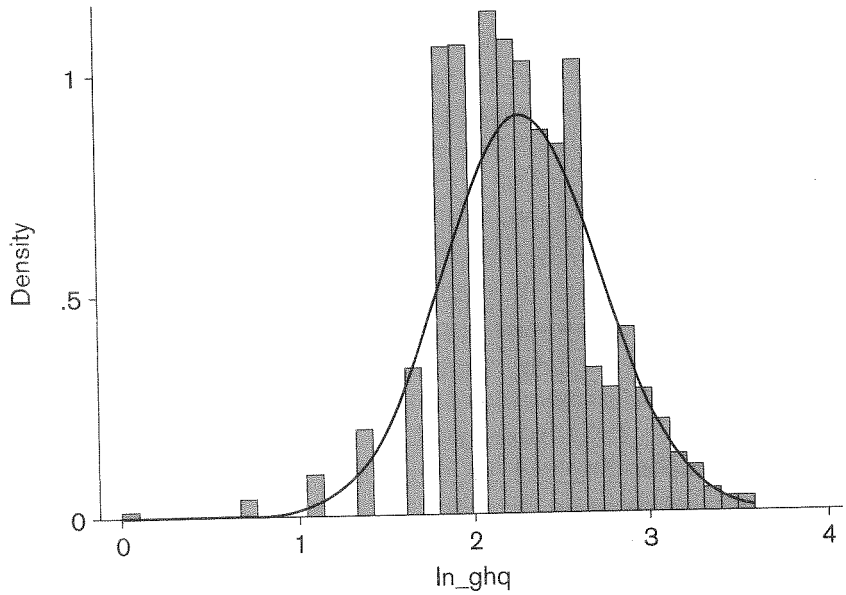
```
or
count if ghqscale==0
```

```
. ta ghqscale if ghqscale<2
```

ghq 0-36	Freq.	Percent	Cum.
0	10	40.00	40.00
1	15	60.00	100.00
Total	25	100.00	

```
. count if ghqscale==0
10
```

The new variable (*ln_ghq*) has a minimum value of zero because there are 15 cases in the original variable that have a value of 1 which has a logarithmic transformation value of zero. If we now graph the new variable we can see that it is closer to normal than the original variable.



The summary statistics and test for normality produce the following output.

```
tabstat ghqscale ln_ghq, s(sk kur)
sktest ghqscale ln_ghq
```

```
. tabstat ghqscale ln_ghq, s(sk kur)
```

stats	ghqscale	ln_ghq
skewness	1.364399	-.2796359
kurtosis	5.661401	4.491428

```
. sktest ghqscale ln_ghq
```

Skewness/Kurtosis tests for Normality

Variable	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
ghqscale	0.000	0.000	.	.
ln_ghq	0.000	0.000	.	0.0000

The variable has changed from being skewed to the right to being slightly skewed to the left (from 1.36 to -0.28) and the kurtosis has moved closer to 3 so it is not so peaked. However, we have 10 extra cases without a value in the new variable, and these cases will be lost in any analysis. So, what to do about it? It could be argued that the minimum value of zero in the original variable (*ghqscale*) was done for convenience and that a score of zero does not indicate an absence of poor mental well-being. Also remember that we recoded the original 12 items to 0–3 from their survey responses that were coded 1–4 and would have made a scale with values from 12 to 48. So, one possible solution would be to shift the distribution to the right by adding 1 to everyone's score so that the variable now has values from 1 to 37. Then we use the logarithmic transformation on the new scale.

```
gen new_ghq=ghqscale+1
gen ln_ghq_2=ln(new_ghq)
su ghqscale ln_ghq ln_ghq_2
tabstat ghqscale ln_ghq ln_ghq_2, s(sk kur)
sktest ghqscale ln_ghq ln_ghq_2
```

```
. gen new_ghq=ghqscale+1
(449 missing values generated)

. gen ln_ghq_2=ln(new_ghq)
(449 missing values generated)

. su ghqscale ln_ghq ln_ghq_2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ghqscale	7714	10.78727	4.945154	0	36
ln_ghq	7704	2.283492	.4439032	0	3.583519
ln_ghq_2	7714	2.38626	.4051969	0	3.610918

```
. tabstat ghqscale ln_ghq ln_ghq_2, s(sk kur)
```

stats	ghqscale	ln_ghq	ln_ghq_2
skewness	1.364399	-.2796359	-.2596882
kurtosis	5.661401	4.491428	4.946262

```
. sktest ghqscale ln_ghq ln_ghq_2
```

Skewness/Kurtosis tests for Normality

----- joint -----

Variable	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
ghqscale	0.000	0.000	.	.
ln_ghq	0.000	0.000	.	0.0000
ln_ghq_2	0.000	0.000	.	0.0000

At the start of this output you can see that Stata tells you that 449 missing values are generated in the new variable, which matches the number of cases that don't have a GHQ score in the original variable (*ghqscale*). This is checked by producing summary statistics where you can see that both the original variable (*ghqscale*) and the newly created variable (*ln_ghq_2*) have the same number of cases. The skewness and kurtosis statistics produced by the **tabstat** command indicate that the *ln_ghq_2* variable is slightly less skewed than the *ln_ghq* variable but more peaked as the kurtosis value is higher.

As with all transformations, there is a trade-off to be made between skewness and ease of interpretation. In this example we would probably use the *ghqscale* in its original form as the

techniques are robust to reasonable departures from normality. The transformed GHQ score variables (either \ln_ghq or \ln_ghq_2) are closer to normal but the interpretation of statistics further on in this example is less intuitive. For example, a difference of means between men and women would be a difference in the logarithmic GHQ scores rather than the raw scores constructed from the items. For the rest of this demonstration we will use the untransformed *ghqscale* variable.