2. Drop-out
More of one type of person may drop out of one of the groups. For example, those less committed, impatient For example in a study of malaria in pregnancy, 500 pregnant women were enrolled into the study design to follow them up for a period of 12 months. At the end of the study about 100 women have drop-out due to various reason as lack of interest, unwillingness to submit sample, miscarriage and even maternal death.

3. History
Events that happen to participants during the research which affect results but are not linked to the IV. For example in a study to compare cocaine drug usage and relaxation occurrence, Participants were put into two groups. One group was given cocaine, while the other group was not. Among the control group were persons who secretly took the drug unknown to the research to make them relax.

4.Reliability of measures and procedures
Unreliable variables or inconsistency in variables can invalidate the study.

5.Using a design of low power
In particular, a small sample size may have insufficient power to detect a real effect even if it is there. As a result, the researcher claims the manipulation had no effect when in fact it does (Type 1 error); he just couldn't pick it up. As well, different statistical tests have varying sensitivity to detect differences.

6.Order effects
If we measure something over a series of trials, we might find that a change occurs because we are becoming bored, tired, disinterested, and fatigued, than at the beginning of the series. "Counterbalancing" is a way of overcoming this problem in repeated measures designs.

7.Multiple tests of significance
The more significance tests you conduct on the one set of data, the more likely you are to claim that you made a significant finding when you should not have. You will be capitalising on chance fluctuations.

Threats to external validity

Two types of external validity follow.

1.Population validity
To what population can you generalise the results of the study? This will depend on the makeup of the people in your sample and how they were chosen. If you have used young, middle class, above average intelligence, different students as your subjects, can you really say your results apply to all people?

2.Ecological validity
Laboratory studies are necessarily artificial. Many variables are controlled and the situation is contrived. A study has higher ecological validity if it generalises beyond the laboratory to more realistic field settings.

Ethical issues in Research

Ethical issues are very important in research these days. Universities are required by law to have Human (and Animal) Ethics Committees which vet and oversee all research that is conducted under the University's name. Their main responsibility is to check for issues in the study that might interfere with a participant's right to not participate, or with possible harm, deception, or embarrassment to participants.

A major issue is that potential participants are clearly informed about what they are agreeing to participate in. Also, even if a participant agrees to participate they have the right to withdraw their participation at any time without prejudice. These two issues need to be spelt out on the Consent form.

Consider the earlier example in which we deliberately withhold treatment to participants so as to test a drug for success. This has **ethical implications** as to whether withhold treatment to participants may result in more injuries. For the study to be approved by a Research Ethics committee, you would have to brief the participants afterwards. You would also have to nominate a source of counselling or assistance if a participant was adversely affected. It might be unlikely, but consider what might happen if only one person is dies. Ethics committees need to see a very good reason for deception before they will allow it. They will on occasions but the benefit of the information obtained needs to clearly outweigh the risks associated with deception and considerable care has to be taken to prepare for problems that might occur.

Most often we conduct research without bothering about ethics. Ethic implies that you take considerable and deliberate steps to ensure that subjects are not exposed to undue harms. For instances if you are investigating the prevalence of infectious diseases, it is important that positive cases are treated or referred to hospital for treatment. It is unethical not to do this.

Collection of data

Sources of Data

Having settled on a design and an operationalisation of your construct or constructs, the next step is actual data collection. But where do you get the data? And from whom should the data be collected?

In experimental design, our data comes from measurements and observations that we will be taking. In non-experimental design, your data should come from the participants that are both available to you and relevant to the question you are studying.

Supposing we were interested in nematode parasites of plants, we would randomly pick plants from the population (garden) to be in the study. By randomly picking plants, you know that there is no way any of the results could be attributed to the biases in your selection. Of course, this is next to impossible, even if your population was quite small.

Having established sources of the data, the next question is how to get the data. You could have plant squashed and nematode extracted from it. It all depends on what you are studying. How you conduct your study is entirely up to you, so long as you do it convincingly, and within ethical considerations.

Samples and populations

The following is an adapted extract from Tabachnick and Fidell (1989):

Samples are measured in order to make generalisations about populations. Ideally, samples are selected, usually by some random process, so that they represent the population of interest. In real life, however, populations are frequently best defined in terms of samples, rather than vice versa; **the population is the group from which you were able to randomly sample**.

Sampling has somewhat different connotations in non-experimental and experimental research. In non-experimental research, you investigate relationships among variables in some predefined population. Typically you take elaborate precautions to ensure that you have achieved a representative sample of that population; you define your population, and then do your best to randomly sample from it.

In experimental research, you attempt to create different populations by treating subgroups from an originally homogeneous group differently. The sampling objective here is to assure that all subjects come from the same population before you treat them differently. Random sampling consists of randomly assigning subjects to treatment groups (levels of the IV) to ensure that, before differential treatment, all samples come from the same population. Statistical tests provide evidence as to whether, after treatment, all samples still come from the same population. Generalisations about treatment effectiveness are made to the type of subjects who participated in the experiment.

Sampling methods

In non-experimental research there several different types of sampling methods: random sampling, stratified sampling, quota, cluster, and systematic sampling.

Sample size

Finally you need to consider how many people you need for your survey. The number of people you target depends on the trade off between costs and benefits. You need enough people to answer the questions you are asking in a valid and reliable way, so, in general, the more the merrier. But the more you recruit the more it is going to cost you in terms of money and time. So you need to be able to determine some minimum number.

Sample size determination

General formula for calculating sample size

Z^2 (p)(1-p)DEFF)/d^2

Where Z = 95% confident interval usually 1.96

Where p = estimate prevalence from previous studies or 50% if there is no study

Where DEFF = Design effect, usually assumed to be between 1-5 depending on the type of sampling methods (1=random sample) (2-sysmatic sampling)

Where d = desired precision

Quantitative and qualitative approaches

Most of the research you will encounter is of the quantitative type and that is what we will be dealing with in this unit. In such research, we rely on measuring variables and comparing groups on those variables, or examining the strength of the relationship between two or more variables. The belief here is that objectivity in the data collection process is paramount. Whoever was repeating this study or using the same instruments and methods would get approximately the same numbers.

However, another branch of research uses more qualitative approaches. These approaches employ more subjective approaches and frequently use interviews, focus groups, or single case designs that lack objective measurement or have restricted generalisability. However, these methods are becoming more widely used these days as analysis methods improve and people search for better ways of gathering data about a problem. Focus groups recruit six to eight participants into a group in which the researcher has a structured set of questions which direct the discussion in the group to the research question. Usually, the whole discussion has to be tape-recorded or video-recorded and all interactions transcribed. The researchers then have to go back over the transcripts and extract the information they need. This can be a quite subjective, laborious, and costly process, but with a standardised set of guidelines, specific training, and greater familiarity with the technique, the considerable richness of these methods has been able to be tapped.

Questionnaire design

A widely used aspect of research in the social sciences is the **self-report questionnaire**. Personality traits, disorders, illnesses, abilities, recollections, and all sorts of other mental states are often assessed by self-report. As such these questionnaires constitute important research tools. They are in fact *measuring tools* for quantifying the amount of some psychological substance. We want to be able to say that someone has more of this substance (or construct) than someone else (a between comparison) or we want to be able to say that one particular person has more of this substance (construct) at one time than they do at some other time (a repeated measures comparison).

Measuring tools therefore need to be valid (i.e., measure what they are intended to measure) and reliable (i.e., get the same answer each time you are measuring the same thing). If our measuring tool is inaccurate than we cannot make precise claims about the psychological construct we are interested in because we cannot be confident that we know exactly how much of the construct is actually there. Once a measuring instrument has been tested and improved and tested again in terms of reliability and validity it is considered *standardised*. A standardised test can then be used in a research project confident about the reliability and validity of the test as a measuring tool.

The development and evaluation of a new survey instrument can be a long and involved process. But essentially, the construct to be measured needs to be clearly defined and the questions or items chosen to measure the construct need to be carefully chosen and tried on a few people for clarity and ambiguous wording. Next, the trial questionnaire needs to be given to a moderately sized sample and all the responses examined for how well they satisfy various criteria such as internal consistency called **pre-testing**.

Types of questions

The construction of a new survey needs to consider how certain information is to be gathered. **Attitudinal Scale**: Measure the degree of variable of interest (Strongly Agree, Agree, Disagree, Strongly Disagree). Scales are used in social studies to measure variable that are non-quantitative. Scales are capable of converting non-quantitative variable such as colour, taste to quantitative variables.

Yes/No: Allows for one option

You have to decide whether to use **closed** and **open-ended** questions. There are also other types such as the semantic differential.

Finally the survey needs to be set out clearly and unambiguously. Respondents need to know what is being asked of them. You need to take into account the amount of time you are asking people to spend on your survey.

Response rate

Another important consideration is response rate. If you give out 300 surveys and only get 50 back, you have a poor response rate (16.7%). You then have the major problem of how representative of the population of interest is your 50 returned questionnaires? Are the characteristics of the 50 who returned the survey somehow different to the 250 who did not return the survey?

You need to maximise response rate. There is a whole literature out there about enhancing response rate. But the most important principle is to maximise the rewards or benefits for the person responding while minimising the costs for the person. So, things like including stamped addressed envelopes for the return of the survey and making the survey short and interesting, with clear instructions need to be considered. Also, the survey should not look formidable, with small print, many arrows and boxes, and verbose instructions. Many researchers include small tokens or payments for completing the survey such as pencils or pens, tea bags, face washers, money (some of the ones I have received!).

Coding of data

The final phase in the data collection process is to convert the observations and measurements you have made about a variable (construct) suitable for computer software (eg SPSS) to understand. This is referred to as **coding**.

In SPSS (but not necessarily so in other analysis packages) each row in the database represents a different example or case of the thing being measured. In the Social Sciences these objects are usually people, but they might also be rats. The columns in the database represent the variables that have been measured about each object. For each person we might measure their gender, their birth order, their age. Each of these things is a variable and until some particular analysis is undertaken they are all dependent variables.

For some variables such as age, or attitudinal scale response, we can simply enter the number we measured. For gender or ethnicity or religious affiliation we have to invent a coding system. Such a system is therefore completely arbitrary. For gender we might use a 0 for female and a 1 for male, or we could use a 2 for female and a 1 for male. The main thing to remember is that we need to enter numbers. SPSS will allow you to enter "male" and "female" (by changing the Type of variable from Numerical to a String), but you are very limited in the sorts of analyses you can then conduct. For religious affiliation we might decide to use a 1 for Catholic, 2 for Anglican, 3 for United, 4 for other, and 5 for No religious affiliation.

Missing values

A common problem with research and particularly survey research is that some people do not answer every question. Some questionnaires have over 500 items and quite often somebody will miss answering a question. What do you do with such data? Do you throw it out or do you find a way to use what you have?

The general answer is to code the missing values in a particular way and then tell SPSS that any occurrences of that code are to be treated as missing values. I tend to use a "-1" because this usually stands out clearly from the other data and you can get a sense for how random or non-random the missing values are by glancing over a printout of the data base. Other authors will suggest a 9 or 99, as long as the code you use cannot be a real number. To code a particular number as missing you need to look under the Define Variables procedures and then under Type. The large data base for Assignment II has one missing value.

Open ended questions

Coding closed-ended questions is relatively easy. But a problem occurs when you try to code open-ended questions. An open-ended question might ask, "What is your occupation?" and you want to be able to use this information in your analyses. First you could code every different occupation with a different number. This would work but you would only be able to carry out very limited analyses. A better approach would be to group occupations according to some dimension. It might be socioeconomic status (SES) for example. So we might have three levels of SES: low, middle, high (and therefore use the numbers 1, 2, and 3 respectively). Then when you come across "truck driver" you decide to put this in "Low" and give it a 1. A response of "doctor" might be put into "High" and be given a 3. You can see that this is very arbitrary and open to quite different results depending upon who is doing the coding. So obviously you need to find standardised ways or unambiguous ways of doing the coding. A good option is to get two people to do the coding and see how well they agree (inter-rater or inter-coder reliability). If the coding system you have is clear and consistent you will get a high correlation and if it is poor you will get a low correlation.

Data Analysis

Descriptive statistics describe patterns and general trends in a data set. In most cases, descriptive statistics are used to examine or explore one variable at a time. However, the relationship between two variables can also be described as with correlation and regression**. Inferential statistics test hypotheses** about differences or relationships in populations on the basis of measurements made on samples. Inferential statistics can help us decide us if a difference or relationship can be considered **real** or just a chance fluctuation.

Statistics can be viewed as a means of finding order and meaning in apparent chaos. At the end of the data collection phase of a research project, really all we've got is a bunch of numbers with no apparent order or meaning. The first phase of data analysis involves the placing of some order on that chaos. Typically the data are **reduced** down to one or two descriptive summaries like the **mean** and **standard deviation** or **correlation**, or by **visualisatio**n of the data through various graphical procedures like histograms, frequency distributions, and scatterplots.

As an example, we will be using a hypothetical data set. The study was concerned with the phoning habits of university students. The data come from a questionnaire where respondents were asked how many calls they make on a particular day. The variable we will examine is the number of calls respondent reported making in a day. 177 people have provided data on this variable.

Question: "How many calls did you make today?"

Respondents: UNAAB students

Data: 1, 0, 2, 4, 5, 1, 1, 1, 2, 1, 1, 4, 1, 1, 10, 2, 6, 1, 1, 1, 1, 1, 1, 2, 5, 2, 1, 1, 6, 2, 4, 2, 1, 4, 1, 0, 1, 5, 0, 1, 0, 1, 1, 4, 2, 1, 1, 0, 1, 3, 1, 1, 3, 1, 4, 1, 0, 1, 1, 1, 0, 1, 8, 1, 15, 1, 1, 1, 2, 3, 1, 4, 3, 3, 1, 1, 2, 1, 1, 1, 1, 1, 4, 3, 2, 1, 0, 0, 1, 2, 1, 2, 3, 1, 7, 1, 2, 2, 1, 1, 1, 1, 0, 2, 0, 2, 1, 3, 0, 0, 2, 1, 1, 0, 1, 2, 1, 1, 2, 2, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 2, 0, 1, 7, 1, 0, 1, 2, 2, 7, 2, 8, 0, 1, 3, 14, 1, 1, 0, 1, 3, 3, 1, 1, 4, 1, 1, 2, 1, 1, 0, 1, 1, 3, 3, 1, 0, 0, 1, 2, 5, 0, 6, 2, 2

Frequency Distributions

From the data, a large number of questions can be answered. For example, how many students had only 1 call on that day? What was the maximum number of calls? Did more students abstain from making calls than who did not? What percent of students who responded had fewer than 5 calls that day?

Some of these questions can be answered easily and others with more difficulty. For example, to answer the first question, you'd have to count up all the ones. To answer the second question, you would search through the data to find the maximum. How easy this is depends on how distinctive the maximum is and how many numbers you have to search through. The answer to the third question would require first counting up how many students responded with zero and how many responded with other numbers. The last question would require knowing how many 0s, 1s, 2s, 3s, and 4s there are in the data compared to higher numbers.

Frequency distributions are a way of displaying this chaos of numbers in an organised manner so such questions can be answered easily. A frequency distribution is simply a table that, at minimum, displays

how many times in a data set each response or "score" occurs. A good frequency distribution will display more information than this although with just this minimum information, many other bits of information can be computed.

Frequency distributions usually display information from top to bottom; with the scores in either ascending or descending order (SPSS displays the data in ascending order unless you tell it otherwise). In Output 4.1, the variable has been named "phonecall" and the range of possible values for this variable is displayed in the left-hand column. The number of times that score occured in the data set is displayed under "Frequency." So 83 of the 177 respondents made only one phone call that day, which is the answer to the first question. This is derived from the "83" in the "1.00" row. You can see that the most frequent response was "1" with "0" and "2" occurring next most frequently and about equally often. Interestingly, 3 people had 10 or more calls that day. This is derived by noting that there is only 1 "10" response, 1 "14" response, and 1 "15" response, which sums to 3 responses greater than or equal to 10.

In general, it would more useful to answer these questions with proportions or percentages. It is quite easy to convert these absolute frequencies into proportions or percentages. A **proportion,** sometimes called **relative frequency**, is simply the number of times the score occurs in the data, divided by the total number of responses. So the relative frequency for the "3" response is 13/177 or about .073. Notice that the relative frequency, while not displayed in this frequency distribution, is simply the percent divided by 100. So the relative frequency for "0" is 0.158. Relative frequencies, like proportions, must be between 0 and 1 inclusive.

A more meaningful way of expressing a relative frequency is as a **percentage**. This is displayed in SPSS under the "Percent" column (Ignore the column labelled "Valid Percent"). As can be seen, 15.8 percent (from 100*28/177) of the students who responded didn't make a phone call that day. Because the percentages have to add up to 100, we know then that 100%-15.8% or 84.2% of students who responded reporting having made **at least one** phone call that day. Thus, the answer to the third question is "No." Based on the responses we received, more students made phone calls that day than not making call at all.

Note the difference between reporting absolute values and reporting percentages. If we simply report that "3 people had made more than 10 phone calls for that day" we are very limited in drawing **generalisations** from this. We don't know if "3" is a small number or a large number. We can't draw any inferences about the general population from this information. It all depends on the total number in our sample. If the total was 177 as here, then we can conclude that about 1.7% of the student population has made more than 10 phone calls that day. If the total was 30, then we would have a completely different story! Whenever we conduct research we always interested in **drawing inferences from our sample to the population** at large.

Output 4.1 Summarize Frequencies

phonecall

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	28	15.8	15.8	15.8
	1	83	46.9	46.9	62.7
	2	29	16.4	16.4	79.1
	3	13	7.3	7.3	86.4
	4	9	5.1	5.1	91.5
	5	4	2.3	2.3	93.8
	6	3	1.7	1.7	95.5
	7	3	1.7	1.7	97.2
	8	2	1.1	1.1	98.3
	10	1	.6	.6	98.9
	14	1	.6	.6	99.4
	15	1	.6	.6	100.0
	Total	177	100.0	100.0	

Output 4.1 The result of using SPSS Summarise Frequencies on the "number of calls" variable.

Another useful statistic, which can be derived from a frequency distribution, is the "**cumulative percent**". The cumulative percent for a given score or data value corresponds to the percent of people who responded with that score or less than that score. So 79.1 percent of the respondents had made **no more than** 2 phone calls. If you defined a "frequent caller" person as someone who had more than 5 calls per day, then you would claim that, from these data, 6.2 percent of UNAABITES students could be called frequent callers (notice the generalisation). This comes from the fact that the cumulative percent for 5 is 93.8%. That is, 93.8% of students had 5 or fewer sexual partners last year. So 100.0% - 93.8% or 6.2% of students had more than 5 sexual partners.

Histograms and bar charts

A **histogram** is a graphical way of presenting a frequency distribution. It is constructed by first selecting a number of "intervals" to be used. The choice is between reducing the information sufficiently while still providing enough variability to picture the shape of the distribution. Most computer programs that construct histograms will allow you to select the number of intervals, as well as their width. If you don't tell the computer how many intervals to use, it will make the decision based on the data it has. In Figure 4.2 you will find a histogram produced by SPSS of the phone calls data in Figure 4.1.

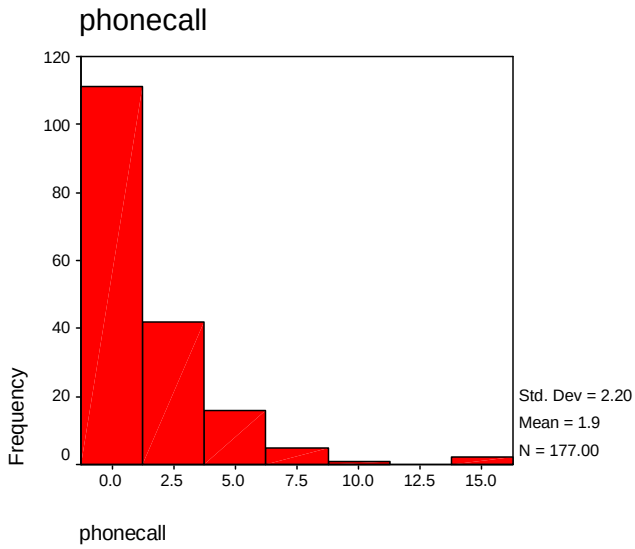

Figure 4.2 A histogram produced by SPSS for the "number of phone calls per day" variable.

First of all, notice that in this histogram, there are 7 intervals. The numbers on the X axis (also called the "abscissa") correspond to the midpoints of the interval. Halfway between adjacent intervals are the **real limits** of the interval, which determine where a particular data point gets "counted" in the histogram. For example, notice the third bar in this histogram. The midpoint is 5. The lower real limit is half way between 2.5 and 5, or 3.75. The upper real limit is between 5 and 7.5, or 6.25. So by convention, any score in the data set equal to or greater than 3.75 and LESS THAN 6.25 gets assigned to the "5" bar. The "Y" axis (also called the "ordinate") displays the **frequency** or number of times a particular piece of data in the data set falls into that interval. So, for example, you can see that 16 respondents in the data set reported having between 3.75 and 6.25 phone calls that day (i.e., 4, 5, or 6). Now, this might seem somewhat silly given that number of phone calls must be an integer (i.e., a discrete variable). For this reason, histograms are best used with data where non-integers are actually possible. Regardless, this histogram does summarise the information in Figure 4.1 quite well. From the frequency distribution (Output 4.1) we know that most people reported having either no phone call or one phone calls that day. The histogram does reflect this (the numbers 0 and 1 occurred 111 times). We also know from the frequency distribution that increasingly fewer people reported having many phone calls. The histogram also reflects this.

Sometimes histograms are constructed with relative frequencies or percentages or proportions on the Y-axis. Because of the close relationship between counts or raw frequencies and relative frequencies and percentages, the interpretation of the frequency distribution remains the same.

Bar Charts. A graph very similar to a histogram is the **bar chart**. Bar charts are often used for qualitative or categorical data, although they can be used quite effectively with quantitative data if the number of unique scores in the data set is not large. A bar chart plots the number of times a particular value or category occurs in a data set, with the height of the bar representing the number of observations with that score or in that category. The Y-axis could represent any measurement unit: relative frequency, raw count, percent, or whatever else is appropriate for the situation. For example, the bar chart in Figure 4.3 plots the number of people making calls.

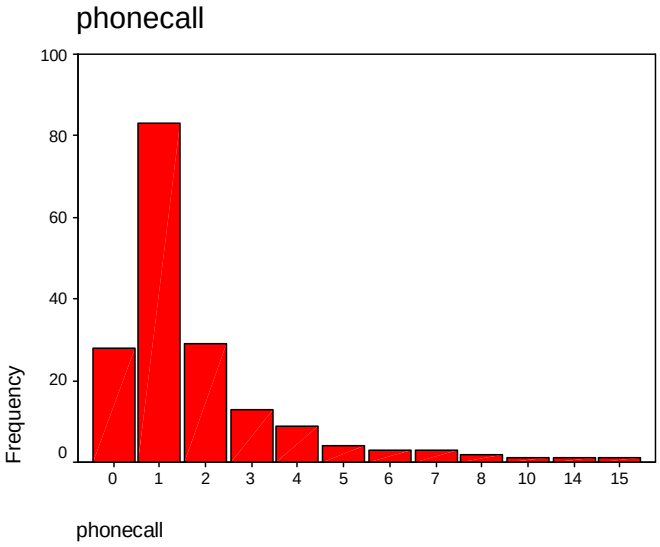

Figure 4.4 A bar chart of the "number of calls per day" variable.

Most computer programs that generate bar charts will treat each quantitative score as a category. What this means is that the bar chart may not space out the scores appropriately on the X axis of the chart. As you can see in Figure 4.4, SPSS ignores the fact that there are no 9s and no observations with values between 11 and 13. As a result, it places "8" and "10" right next to each other, and then places "14" next to "10." It simply treats these scores with no observations as impossible. As a result, looking at a bar chart can give a misrepresentation as to the **shape** of the distribution. Second, if there are many unique scores in the data set, each of which occurs infrequently, a bar chart may provide no additional information than could be obtained from just looking at the data set. For example, imagine a bar chart of the following data: 4.3, 6.5, 1.2, 6.9, 4.1, 0.4, 6.1, 3.6, 1.6, 2.3. There is only one of every score. So a bar chart would provide little information because it would just display 10 bars equal in height (i.e., with a height of 1).

The shape of a distribution

Symmetry. A distribution of scores may be symmetrical or asymmetrical. Imagine constructing a histogram centred on a piece of paper and folding the paper in half the long way. If the distribution is **symmetrical**, the part of the histogram on the left side of the fold would be the mirror image of the part on the right side of the fold. If the distribution is **asymmetrical**, the two sides will not be mirror images of each other. True symmetric distributions include what we will later call the **normal distribution**. Asymmetric distributions are more commonly found.

Skewness. If a distribution is asymmetric it is either **positively skewed** or **negatively skewed**. A distribution is said to be positively skewed if the scores tend to cluster toward the lower end of the scale (that is, the smaller numbers) with increasingly fewer scores at the upper end of the scale (that is, the

larger numbers). Figure 4.2 is an example of a positively skewed distribution, the majority of people report 0, 1, or 2 phone calls that day and increasingly few report more.

A **negatively skewed** distribution is exactly the opposite. With a negatively skewed distribution, most of the scores tend to occur toward the upper end of the scale while increasingly fewer scores occur toward the lower end. An example of a negatively skewed distribution would be age at retirement. Most people retire in their mid 60s or older, with increasingly fewer retiring at increasingly earlier ages. A graphic example of a negatively skewed distribution can be found in Figure 4.5.

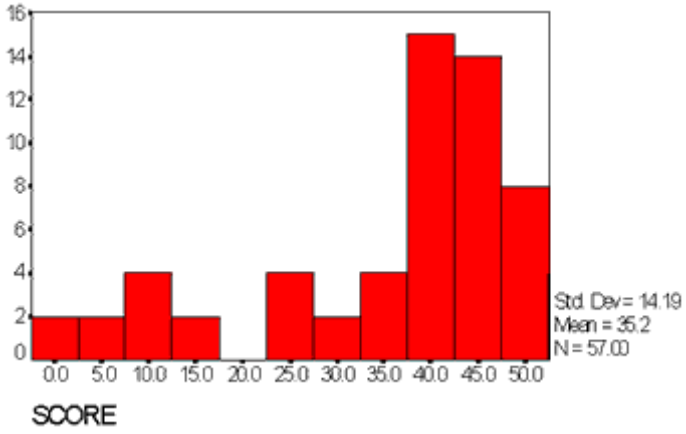

Figure 4.5 An example of a negatively skewed distribution

SPSS example

If we select Frequencies from SPSS on the "number of phone calls" variable and also select the appropriate statistics (as shown in the SPSS Screens and Output booklet), you will find the following output.

Output 4.2 Summarize Frequencies with statistics options.

Frequencies

Statistics

SEXPARTS

N	Valid	177
	Missing	0
Mean		1.8644
Std. Error of Mean		.1651
Median		1.0000
Mode		1.00
Std. Deviation		2.1960
Variance		4.8224
Skewness		3.076
Std. Error of Skewness		.183
Kurtosis		12.778
Std. Error of Kurtosis		.363
Range		15.00
Minimum		.00
Maximum		15.00

Output 4.2 The results of selecting Statistical options within the SPSS Frequencies procedure.

Output 4.2 shows many numerical descriptive measures for the "number of phone calls" variable. If a distribution is **not** skewed, the numerical value for "Skewness" is zero. The fact that it is positive (3.076) in the output above, shows that the variable is positively skewed.

Kurtosis. Another descriptive statistic that can be derived to describe a distribution is called kurtosis. It refers to the relative concentration of scores in the center, the upper and lower ends (tails), and the shoulders of a distribution. In general, kurtosis is not very important for an understanding of statistics, and we will not be using it again. However it is worth knowing the main terms here.

Note, that these numerical ways of determining if a distribution is significantly non-normal are very sensitive to the numbers of scores you have. With small sets of scores (say less than 50), measures of skewness and kurtosis can vary widely from negative to positive skews to perfectly normal and the parent population from which the scores have come from could still be quite normal. Numerical methods should be used as a general guide only.

Modality. A distribution is called **unimodal** if there is only one major "peak" in the distribution of scores when represented as a histogram. A distribution is "**bimodal**" if there are two major peaks. If there are more than two major peaks, we would call the distribution **multimodal**. An example of a bimodal distribution can be found in Figure 4.6.

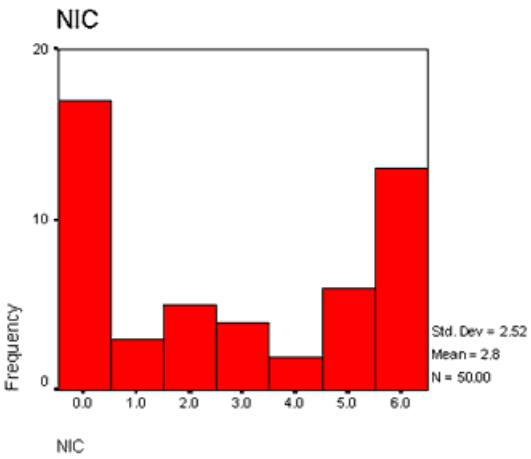

Figure 4.6 An example of a bimodal distribution. The figure shows the frequency of nicotine use in the data base. Nicotine use is characterised by a large number of people not smoking at all and another large number of people who smoke every day.

Central Tendency

In the last section we explored various ways of describing a data distribution. Our focus then was on determining the frequency of each score in the data set, deriving percentages, and visualising and describing the shape of the distribution. While these are all important in the description process, **numerical descriptions** of **central tendency** and **variability** are much more important. The former gives a single description of the average or "typical" score in the distribution and the latter quantifies how "spread out" the scores are in the distribution.

Measures of central tendency

Measures of central tendency, or "location", attempt to quantify what we mean when we think of as the "typical" or "average" score in a data set. The concept is extremely important and we encounter it frequently in daily life. For example, we often want to know before purchasing a car its average distance per litre of petrol. Or before accepting a job, you might want to know what a typical salary is for people in that position so you will know whether or not you are going to be paid what you are worth. Or, if you are a drinker, you might often think about how many cigarettes you smoke "on average" per day. Statistics geared toward measuring central tendency all focus on this concept of "typical" or "average." As we will see, we often ask questions in science revolving around how groups differ from each other "on average". Answers to such a question tell us a lot about the phenomenon or process we are studying.

Mode. By far the simplest, but also the least widely used, measure of central tendency is the **mode**. The mode in a distribution of data is simply the score that occurs most frequently. In the distribution of phone calls data, the mode is "1" because it is the most frequently occurring score in the data set. If you have had only one phone call that day, it would be reasonable therefore to say that you are fairly typical of UNAAB students (or at least of those students who responded to the question). Importantly, you can’t necessarily claim that "most" UNAAB students had only one phone call that day. From the frequency distribution, notice that actually fewer than half of the respondents reported having only one phone call. **So "most" students reported having something different to 1 phone call**. Still, "1" was the most

frequent single response to this question, and so it is the mode or **modal** response. In some cases, however, such a conclusion would be justified.

Recall that one way of describing a distribution is in terms of the number of modes in the data. A unimodal distribution has one mode. In contrast, a bimodal distribution has two. Now this might seem odd to you. How can there be more than one "most frequently occurring" score in a data set? I suppose statisticians are a bit bizarre in this way. We would accept that a distribution is bimodal if it seems that more than one score or value "stands out" as occurring especially frequently in comparison to other values. But when the data are quantitative in nature, we would also want to make sure that the two more frequently occurring scores are not too close to each other in value before we would accept the distribution as one that could be described as "bimodal." So there is some subjectivity in the decision as to whether or not a distribution is best characterised as unimodal, bimodal, or multimodal.

Median. Technically, the **median** of a distribution is the value that cuts the distribution exactly in half, such that an equal number of scores are larger than that value as there are smaller than that value. The median is by definition what we call the 50th percentile. This is an ideal definition, but often distributions can't be cut exactly in half in this way, but we still can define the median in the distribution. **Distributions of qualitative data do not have a median**.

The median is most easily computed by sorting the data in the data set from smallest to largest. The median is the "middle" score in the distribution. Suppose we have the following scores in a data set: 5, 7, 6, 1, 8. sorting the data, we have: 1, 5, 6, 7, 8. the "middle score" is 6, so the median is 6. Half of the (remaining) scores are larger than 6 and half of the (remaining) scores are smaller than 6.

To derive the median, using the following rule. First, compute (n+1)/2, where n is the number of data points. Here, there are 5, so n = 5. If (n+1)/2 is an integer, the median is the value that is in the (n+1)/2 location in the sorted distribution. Here, (n+1)/2 = 6/2 or 3, which is an integer. So the median is the 3rd score in the sorted distribution, which is 6. If (n+1)/2 is not an integer, then there is no "middle" score. In such a case, the median is defined as one half of the sum of the two data points that hold the two nearest locations to (n+1)/2. For example, suppose the data are 1, 4, 6, 5, 8, 0. The sorted distribution is 0, 1, 4, 5, 6, 8. n = 6, and (n+1)/2 = 7/2 = 3.5. This is not an integer. So the median is one half of the sum of the 3rd and 4th scores in the sorted distribution. The 3rd score is 4 and the firth score is 5. One half of 4 + 5 is 9/2 or 4.5. So the median is 4.5. Here, notice that half of the scores are above 4.5 and half are below. In this case, the ideal definition is satisfied. Also, notice that the median may not be an actual value in the data set. Indeed, the median may not even be a possible value.

The median number of phone calls per day is 1. Here, n = 177, and (n+1)/2 = 178/2 = 89, an integer. So in the sorted distribution, the 89th data point is the median. In this case, the 89th score is a 1. Notice that this doesn't meet the ideal definition, but we still call it the median. It certainly isn't true that half of the people reported making fewer than 1 phone call, and half reported making more than 1. Violations of the ideal definition will occur when the median value occurs more than once in the distribution, which is true here. There are many "1"s in the data.

Computing the median seems like a lot of work. But computers do it quite easily (see Output 4.2). In real life, you'd rarely have to compute the median by hand but there are some occasions where you might, so you should know how.

Mean. The **mean**, or "average", is the most widely used measure of central tendency. The mean is defined technically as the sum of all the data scores divided by n (the number of scores in the distribution). In a sample, we often symbolise the mean with a letter with a line over it. If the letter is "X",

then the mean is symbolised as $\bar{X}$, pronounced "X-bar." If we use the letter X to represent the variable being measured, then symbolically, the mean is defined as

$$\frac{\sum X}{n}$$

For example, using the data from above, where the n = 5 values of X were 5, 7, 6, 1, and 8, the mean is (5 + 7 + 6 + 1 + 8) / 5 = 5.4. The mean number of phone calls reported by UNAAB students who responded to the question is, from Figure 4.1, (1 + 0 + 2 + 4 + . . . + 0 + 6 + 2 + 2)/ 177 = 1.864. Note that this is higher than both the mode and the median. In a positively skewed distribution, the mean will be higher than the median because its value will be dragged in the direction of the tail. Similarly in a negatively skewed distribution, the mean will be dragged lower than the median because of the extra large values in the left-hand tail. **Distributions of qualitative data do not have a mean.**

While probably not intuitively obvious, the mean has a very desirable property: it is the "best guess" for a score in the distribution, when we measure "best" as LEAST IN ERROR. This might seem especially odd because, in this case, no one would report 1.864 phone calls, so if you guessed 1.864 for someone, you would always be wrong! But if you measure how far off your guess would tend to be from the actual score that you are trying to guess, 1.864 would produce the smallest error in your guess. It is worth elaborating on this point because it is important. Suppose I put the data into a hat, and pulled the scores out of the hat one by one, and each time I ask you to guess the score I pulled out of the hat. After each guess, I record how far off your guess was, using the formula: error = actual score - guess. Repeating this procedure for all 177 scores, we can compute your mean error. Now, if you always guessed 1.864, your mean error would be, guess what? ZERO! Any other guessing strategy you used would produce a mean error different from zero. Because of this, the mean is often used to characterise the "typical" value in a distribution. No other single number we could report would more accurately describe EVERY data point in the distribution.

Choosing a measure of central tendency

With three seemingly sensible measures of central tendency, how do you know which one to use? Not surprisingly, the answer depends a lot on the data you have and what you are trying to communicate.

While the mean is the most frequently used measure of central tendency, it does suffer from one major drawback. Unlike other measures of central tendency, the mean can be influenced profoundly by one extreme data point (referred to as an "outlier"). For example, suppose one additional respondent answered that he (or she?) made 200 phone calls that day (!) The inclusion of this person would increase the mean from 1.864 to 2.977. Using the mean as our definition of "average" or "typical," we would conclude that UNAAB students are frequent caller quite a bit more than we would conclude if this person was not included in the data.

The median and mode clearly don’t suffer from this problem. With the "200" person included in the data, the mode would still be "1", as would the median. So the mode and median tend not to be influenced much, if any, by one or two extreme scores in the data.

There are certainly occasions where the mode or median might be appropriate, but it depends on what you are trying to communicate.