

Chapter 5

Types of Studies - I.

Descriptive Studies: Case Reports, Case Series, Cross-sectional Studies

Descriptive studies address the task of providing explanations about phenomena by assembling groups of subjects that represent either a non-diseased general population or those who share features which might predispose them to particular outcomes. Descriptive studies are useful for studying patterns of diseases and the characteristics of subjects with defined outcomes. They are also used for studying health care utilization patterns and for deciding about resource allocation.

There are several design options, each with its own advantages and shortcomings. The commonest designs are case reports, case series and cross-sectional studies. Each type provides information about persons (*WHO* has the disease?), place (*WHERE* is the disease more or less common?), and time (*WHEN* is it occurring?). Descriptive studies suffer from not having an adequate comparison group as is the case with case reports and case series, or by being unable to define the temporal relationship between an exposure and disease as is the case with cross-sectional studies. Hence they are not able to test hypotheses. For that analytic studies are needed. On the other hand the identification of descriptive characteristics of subjects is the first step in the search for determinants of a disease, or of risk factors. Descriptive studies are usually the first step in formulating hypotheses and deciding about designs of future studies.

Case Reports

These are detailed presentations of a single case. They are a useful device by which unusual diseases or unusual presentations of disease are brought to the attention of the scientific community.

Case reports serve the following functions:

1. Since they report rare events, they serve as a rich source of ideas for further research about disease frequency, risk, prognosis and treatment. They often trigger more detailed studies. The first report of vertical transmission through breast milk of HIV infection from a recently infected mother to her baby appeared as a case report. Other studies followed and eventually the risk has come to be quantified as 14% over and above that of transmission in utero or during labour. The effects of thalidomide on the foetus, foetal alcohol syndrome, and halothane toxicity were studied in this manner.
2. Case reports also serve to describe the mechanisms of disease or of treatment by providing detailed accounts of clinical findings, laboratory studies, or treatment. In this way, case reports have

contributed a great deal to the understanding of genetic, metabolic and physiologic bases of a number of diseases.

3. Case reports often describe the joint occurrence of unusual events. These serve as hypotheses to be tested by means of other studies.
4. The commonest use of case reports is for describing the side effects and complications of drugs. Since each individual is unique, several thousand people must be treated before the rarer type of side effects is noticed.

Case Series.

Case series comprise a description of a group of individuals with a particular disease. It is a common way of describing the clinical picture in an uncommon presentation of an illness. For example, measles occurring in children who are suffering from a malignancy. (*BMJ 1987;295:15-18*). Another example is that of clinical manifestations of tuberculosis in infants. (*Arch.dis.chidh. 1993;69:371-374*).

Case series suffer from a number of drawbacks and caution must be exercised in planning. The following are the main features to watch:

1. They contain, of necessity, information acquired over time. Health workers change, nursing patterns differ, and treatment regimens may also change. The details with which history has been recorded can vary over time. The same also applies to details of laboratory and other investigations.
2. The patients in a given series ought to be similar in one way or another in order to arrive at meaningful conclusions.
3. All the patients in a defined period must be included in order to avoid unintended bias.
4. A comparison may or may not be involved. If comparison is being made then fairness of comparison becomes an important issue.

Case series describe the clinical manifestation of a disease at one point in time. Because a time dimension is absent their value as a means of studying the cause-effect relationship is limited. They also suffer from the absence of a comparison group.

Case series can advance scientific understanding, but they can also mislead. In order to enhance the scientific value of case series the following precautions need to be taken:

- All cases seen in the given time period must be included. Assumptions based on selected cases can be treacherous. Precise inclusion and exclusion criteria must be defined at the outset and scrupulously followed. In addition, what was done to each patient, their progress, and the outcome in each case must be reported.
- In reporting results the influence of all variables should be carefully accounted for.

- A protocol which has been developed at the outset and tested, should be employed to record all information systematically. Thought should be given beforehand to ways of handling missing values, drop-outs, out migrations, and other contingencies. The end point of the study should be determined in advance.
- When series have been accumulated over a long time, as is likely to happen with rare conditions, temporal drift is a major pitfall. Referral patterns change over time, diagnostic criteria may become revised, more modern techniques may come into use, and treatment regimens may change.

Goals in Case Reports and Case Series

In Chapter 4 the three main goals of descriptive studies were listed as describing:

- i). the characteristics of the members of the group being described.
- ii). the distribution of the characteristics among the members of the group.
- iii). the expression of the characteristics in a summarized form to aid memory.

These three goals are served by keeping the following in mind:

1. The criteria for inclusion of cases in the study must be carefully developed.
2. A systematic search for the available universe of subjects who have the condition of interest must be made to find all those who are eligible. A diagnosis is after all a label. Some people are likely to be wrongly labelled, and others with the disease may not yet have the label put on them, because the presentation is atypical. Clinical and other criteria of case definition should be carefully considered.
3. Elements of bias in the available universe of study subjects must be assessed.
4. Rules for classifying observations must be developed and consistently applied.

These requirements are examined further below:

Criteria for Inclusion

Knowing how the population was defined is important because the purpose of descriptive studies is to communicate observations. For example, the subjects may be a group of patients in a given institution who happened to receive a particular disease label. As stated earlier a difficulty arises when potentially eligible patients in whom the labeling has not been done get excluded, or those wrongly diagnosed (labeled) included. Hence precise and replicable case definitions need to be established before the subjects are assembled.

Systematic search of the available universe for eligible subjects.

An ideal descriptive study is population based and searches every known individual in the general population for characteristics that would make him or her eligible for inclusion in the study group e.g. pooling of subjects with certain diseases from a number of institutions. In small settings the

advice is to first establish case criteria and then search the universe of patients for those who fit the criteria.

Assessment of bias in the available universe of patients.

Groups assembled according to defined criteria may still differ from one study to another, e.g. by age, sex or degree of advancement of the disease. Systematic errors that lead to selected groups being different are biases. An important cause in hospital based studies is the referral pattern. Cases referred to hospital tend to be the more complicated or advanced forms of a condition out of a large universe of patients in the general population. For example, among children with febrile convulsions attending general practitioner clinics a much smaller number go on to develop recurrent non-febrile seizures later on as compared to those in hospital series. This is because only those with more severe or prolonged or unusual type of febrile convulsions get referred to hospital.

Consistent application of rules for classifying observations.

Once the subjects have been identified the methods of observation need to be rigorous. For interviewing and for making extracts from records there should be independent collaborators different from the main investigator. Biopsies, slides, x-rays and results of laboratory tests should be reassessed by impartial observers instead of relying on filed reports. This helps to minimize bias because subjective judgment is being made. Here there are three basic requirements viz.

- i). A good record system in order to find patients.
- ii). Patients classified in ways relevant to the question being addressed.
- iii). Collaborators are needed who are capable of making independent assessment of laboratory reports and other results.

Cross sectional studies

As a group cross sectional studies fall between purely descriptive studies and those that can be used for testing hypotheses. They are largely used for descriptive purposes, for example in surveys of various sorts, and for measuring prevalence. Cross sectional studies are generally carried out on the general population, and so disease can be classified by personal characteristics like age, sex, race, education, socio-economic status and place, or time. Cross-sectional studies are also useful for describing the clinical spectrum of a disease. For example, a cross sectional study of diabetes may be used to study the proportion who have retinal, renal or cardiovascular complications. In modified forms cross sectional studies are sometimes employed to do some analytical work. If one is looking for etiology or a cause-effect relationship then cross sectional studies are more suited for diseases of slow onset and long duration e.g. asthma, mental disorders, osteoarthritis, and so on.

In cross-sectional studies all observations are made on a single occasion. The distribution of variables in the sampled population is analyzed, and inferences with regard to cause and effect are made from association between variables designated as predictor and outcome variables by the investigator. However, there is one important difference. In cross-sectional designs the choice of

'predictor' and 'outcome' variables is made arbitrarily by the investigator rather than the study design. That is the reason why a cause / effect relationship can only be speculative. Cross-sectional studies are appropriate for simply describing attributes (variables) and their distribution pattern in the sampled population.

In analytical work cross sectional studies are useful for studying the effects of variables like gender, socio-economic class, and ethnic background on health. With these variables the difficulty about deciding temporal relationship does not arise. People are born with these attributes, and so they can safely be considered as predictor variables.

In studies of etiology researcher like to begin with cross sectional studies, because disease is often due to a number of causes, and to identify all possible causal factors we need to study systems as they function in nature. Cross-sectional designs are convenient for examining a network of causal links. Promising clues can then be followed up by other types of study designs. Thus, cross-sectional designs can be included as the first step in a cohort or intervention study at little additional cost for defining the demographic and clinical characteristics of the subjects at baseline. At the same time associations of interest between variables can be examined.

Cross-sectional designs do provide an important descriptive statistic, namely the prevalence rate. Another useful analytic statistic viz. Relative Prevalence can also be obtained in cross-sectional studies. Relative prevalence is the ratio of the prevalence of outcome in subjects classified by their level of predictor variable. It serves as a good approximation of relative risk provided the risk factor is not influencing the duration of the outcome. We have already seen that in cross-sectional studies exposure and disease get measured together, and so it is not possible to determine whether exposure preceded the outcome or resulted from it. Moreover, the design limitation is such that the investigator must study the existing (prevalent) rather than new (incident) cases. Those who died or got cured do not get included. This results in the data reflecting the determinants of survival as well as of etiology. The effect of a risk factor on disease duration can be easily mistaken for effects on disease occurrence.

It is often possible in cross sectional studies to compare two or more naturally occurring groups with regard to one or more variables. Such studies are sometimes described as Correlational studies. For example, in the study of first time mothers in a squatter area of Brazilian city (*J.Trop.Ped.1990;36:14-19*) it was possible to compare teenage mothers and older mothers with regard to a number of variables which could provide clues of antecedents of teenage pregnancy. In such cases where cross sectional studies are used to explore a particular problem by means of comparing naturally occurring groups it is often necessary to control for extraneous variables like age, income, social class or education. It may be found that groups vary on several variables other than the ones chosen by the researcher. If a significant difference is found the researcher may offer a number of alternative explanations, but these are all conjectural hypotheses, and do not unequivocally explain causation. The study does in a way provide an early clue to exposure/outcome association. From this can arise important pointers to the possible causes and further studies.

By comparing naturally occurring groups the researcher is able to identify interrelationship among clinically important variables. The strength of the association can be expressed numerically. Even when strong relationship is found there is no evidence that one variable is causally related to

another. It must also be pointed out that one explanatory variable does not necessarily account for all the variation seen in the outcome variable. For example, in a study of birth weights the characteristics of mothers of babies weighing less than 2500g may be compared with those of mothers of heavier infants. A correlation between birth order and weight at birth may be demonstrable. But birth order may not be the only factor of importance. Maternal nutrition, illness during pregnancy, and several other factors may also be operating. Another example is that of the association between prevalence rates of oesophageal cancer in different populations and the habit of chewing tobacco. A strong correlation may be found between the two. But there may also be other factors associated with the occurrence of oesophageal cancer. The strength of correlational studies lies in the fact that they can be done quickly and cheaply often using information already available from data which are collected routinely, like for example the Household Survey. But another word of caution may be sounded! The limitation is that the data is about populations rather than individuals, and one cannot link exposure with disease in individuals. The data represent average exposure levels in the population rather than actual individual values. Secondly, it is not possible to control for possible confounding factors. It may appear from correlational studies that a possible association exists but a more complex relationship between exposure and disease may get masked.

Surveys are a form of cross sectional study aimed at describing accurately the characteristics of a given population. They are often used for assessing attitudes, opinions or beliefs of persons concerning health related issues. The information is obtained by means of questionnaires and interviews. Surveys are also useful for obtaining information about demographic characteristics, and for obtaining data about populations with regard to diet and consumption patterns of different groups, utilization of health care, prevalence of health problems like hypertension, anaemia, protein-energy malnutrition, emotional problems, drug use patterns, and so on. Surveys are discussed in detail in Chapter 9.

Statistics obtained from surveys provide an overview of the state of health, illness and treatment patterns (e.g. *Lancet* 1994;344:1675-8). They give an insight into prevalent causes of death or of health needs. Patterns observed in the data like significant differences between groups or interrelationships between variables can become the basis for hypotheses or theories concerning causes of illness in a community. Clinical decisions may then be based on information regarding symptoms, signs, laboratory abnormalities or the frequency of outcome like death, disability, or improvement. Similarly, health service policy decisions may be made in a more informed manner being based on the frequency of various ailments rather than empirically.

Hypotheses formulation from descriptive studies

As stated earlier, descriptive studies are often used as a first probe into studying a problem. Hypotheses are then formulated to serve as bases for future studies. The logical approaches to such hypothesis formulation are:

1. Method of difference. If the frequency of a condition is markedly different in two groups, or under different sets of circumstances then the disease is likely to be caused by a factor that differs between them.

2. Method of agreement. If a single factor turns out to be common to a number of circumstances in which a given disease occurs with a high frequency then the factor is most likely associated with the disease.
3. Concomitant variation. If the frequency of a factor varies in proportion to the frequency of a given disease then association between the two is very likely.

The hypotheses generated are usually about "Person" (e.g. age, sex, socio-economic factors, religion, occupation, and so on.), about "Place" (e.g. urban/rural; geographical comparisons between countries and within countries etc.), and "Time". (e.g. seasonal variations or outbreaks). With regard to "Time", secular trends over years are sometimes helpful for studying the epidemiology of chronic diseases. But then true variations must be carefully differentiated from changes occurring because of improved diagnoses, accuracy in enumerating the population at risk, changes in survival because of improved treatment (e.g. Acute lymphoblastic leukaemia), and changes in the age distribution of the population.

Sample selection in cross sectional studies

Prevalence studies and surveys are all about proportions. Since a proportion has a numerator and a denominator it is necessary to be clear about "cases" and "population". At the outset the question as to what is a case? should be addressed, and a proper case definition determined. With regard to population, one generally means population at risk. In field studies the term population means the people resident in a defined geographical area. In clinical studies it is the population of a ward or a clinic. Here the term "population" relates to clinical settings where the information was collected.

As we saw in Chapter 4 sample selection is usually done by exposure, outcome, or other criteria.

The selection of subjects can have major influence on conclusions. We have already considered the difference between population based and hospital based studies with regard to febrile convulsions in children. Hence it is important to find out who in the population becomes a study subject. Studies can also suffer if the subjects do not co-operate. As a rule of thumb drop out rates of less than 20 per cent are acceptable. Anything more than that should arouse suspicion, and one should be asking why those who declined participation or dropped out of the study did so. Moreover, in the case of some topics people who agree to participate may differ from those who do not. They may be better educated, from a particular social class or age group, with more advanced disease, or other characteristic because of which they are happy to continue or participate in the study.

When sample selection is by exposure the level of exposure is measured by questionnaires (relying on the memory of the subject), clinic records (relying on the accuracy of history taking), laboratory tests (relying on the laboratory procedures and the sensitivity/specificity of the tests), and by physical measurement (relying on the accuracy of the instrument used as well as the observer).

The main weakness in selecting the sample by history of exposure lies in ascertaining the duration of the exposure, or its intensity. The former relies on the memory of the respondent, and the latter can change over time.

When sample selection is by outcome cross sectional studies measure prevalence outcome. Therefore, they are best suited for chronic, non fatal conditions because subjects get excluded in the event of death, drop-out, out migration or cure.

When cross-sectional studies are carried out sequentially on the same population information about trends is built up on the assumption that representative sampling of the same parent population takes place. If the composition of the population changes, or sampling frame or method are changed, or if the response pattern changes then the results of the study are influenced by these changes rather than representing the true health behavior.

Analysis of results

Analysis of data depends on the research question being asked, the method for sample selection, and the measurement scales used for the different variables.

In prevalence studies when continuous variables are used the results are expressed in terms of mean (or mode or median) and the standard deviation. This is a useful way of summarizing the data since mean \pm 1 SD indicates that 65% of the values lie within the range. Mean \pm 2SD expresses 95% and Mean \pm 3SD describes 99% of the values obtained in the survey. When nominal and ordinal measures have been used proportion is the measure used for describing the results.

If continuous variables have been used then mean outcomes are compared between the groups as defined by their exposure status. If categorical variables have been used then rates are compared between the groups and a measure of risk may be calculated.

If evidence of association is being sought between variables it can be measured by means of appropriate statistical tests. But association does not necessarily mean cause and effect. Evidence that the dependent variable varied after the independent variable is more problematic in cross sectional studies. Time sequence may be assessed by asking the subjects about it, but this relies on memory. However, if causal relationships are hypothesized using race, age, sex, and so on as independent variables it is reasonable to assume that they preceded the outcome.

The next step is to search for evidence that the association is not spurious. One strategy is to consider similar evidence for variables which might produce spurious association, and subject it to the same analysis. Even then there may be some unknown variable which could be the true cause, and which has not been included in the research design.

If sample selection has been by exposure the analysis is similar to cohort study, and when selection has been by outcome the analysis is similar to case-control study.

Bias in cross sectional studies

There can occur sampling bias, information bias, confounding bias, matching bias, and stratification bias. An important form of bias to which cross-sectional studies are prone is the reverse causality bias. Let us take for example a study about obesity and osteoarthritis. If the prevalence of obesity is found to be more frequent in people with osteoarthritis one may be tempted to say that because arthritic people are less mobile they are more likely to put on weight. But the truth may be that obesity led to osteoarthritis by putting greater strain on the joints.

All these forms of bias should be carefully considered in designing studies, in conducting them, and in the interpretation of the results. The basic steps for avoiding bias are to have well defined criteria for being a case and for the population. Both cases and noncases should be from an unbiased sample.

Prevalence surveys by their very nature include only those available at the time the study is conducted ; in other words those who are alive and diagnosable. Bearing in mind that diagnoses are nothing but labels, some cases may not yet have been labelled, and some wrongly so. Prevalence is dominated by those who are able to survive their disease without losing its manifestations. For example, neurological deficit after an acute catastrophic illness like encephalomyelitis. Many are likely to die during the acute episode, and so one does not get a true picture. Prevalence is also affected by the average duration of the disease. Rapidly fatal diseases will be missed e.g. kwashiorkor.

Modifications of cross-sectional designs.

Trend Design. Two or more cross sectional studies are implemented at different times in the same population. Here the assumption is that representative sampling has occurred on each occasion. For causal search the trend in one explanatory variable is compared with trends in one or more outcome variables e.g. trend in smoking and cancer of the lung, or prohibition of promotion of bottle feeding and the trend in prevalence of malnutrition. If one of the variables changed after the other it provides evidence regarding time sequence.

The trend design offers an advantage over simple cross sectional design by providing evidence of temporal sequence as well as association. But from the practical angle it may not be feasible to collect data at short intervals to pin point changes in trends. Moreover, changes are often gradual and occur over several years during which other contaminating influences are possible. Trend designs have been applied to census data and other health data where information is gathered routinely. Sometimes time sequence can be identified.

Spurious association is a problem. The strategy for dealing with it is to collect information on other possible explanatory variables likely to cause spurious association, and account for them in the analysis.

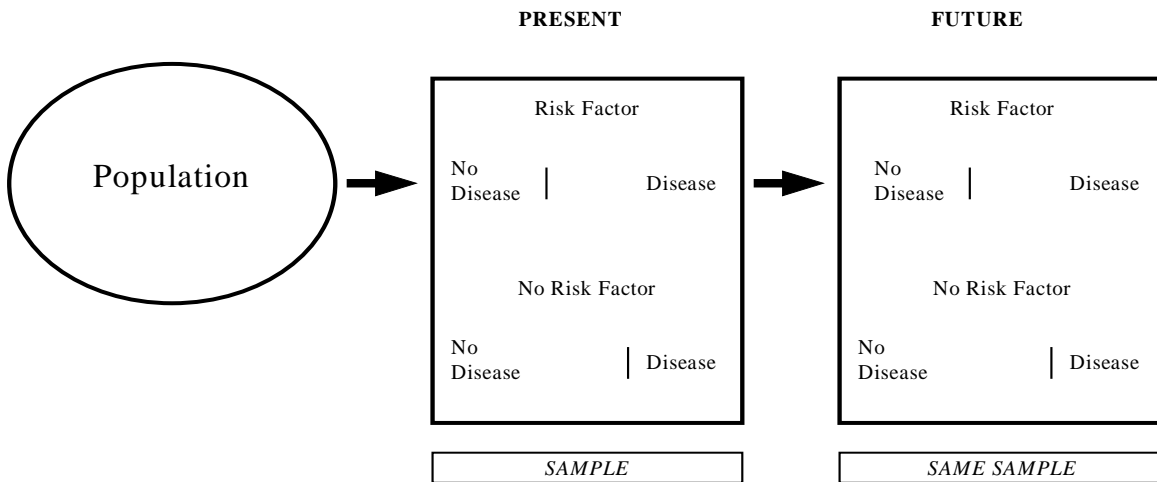


Figure 5.1: Panel Design.

Panel Design. This is also called Repeat Measure Design. In this kind of study information is gathered from the same subjects in a population at different times. Sets of data are linked by cases. Changes in one variable helps to identify changes in other variables. A good example is measuring growth in height of children and recording an increase in scoliosis.

Trend studies only show that over time two variables vary with one another. In contrast the panel (or repeat measure) design provides more direct evidence that the outcome variable varied after the explanatory variable. The panel design is not entirely immune to spurious association, because there may be confounders. The main problem with panel designs is attrition. It is very difficult to ensure that all subjects will continue to participate in the study with full enthusiasm.

Another form of repeat measure study is often called the Pseudo-cohort study. This is often used in preparing growth standards in children. Many growth charts have been prepared by measuring subjects and then 'smoothing out' the mean weights and heights to present as one growth curve. The results can be made more accurate if each subject is measured twice at an interval of say six months or one year. This provides additional data on growth velocity and the "smoothing out" of the curve is that much more efficient.

Appendix 5.1

1). Calculation of sample Size

Formula for calculation of sample size for studies involving comparison of two means

$\bar{X}_1 - \bar{X}_2$ = Difference between the means

Sd_1, Sd_2 = Standard deviations of the two means

u = Power of the study

v = Significance level

$$\text{Minimum sample size} = (u + v)^2 (Sd1^2 + Sd2^2) / (\bar{X}_1 - \bar{X}_2)^2$$

Values of u	Power of the study
0.84	80%
1.28	90%
1.64	95%
1.96	97.5%
2.33	99%

Values of v	Significance level
1.64	10%
1.96	5%
2.23	2.5%
2.58	1%

2).

In studies where the expected effect size (E) and the standard deviation (S) of the outcome variable are known, and the groups are unequal

The number of subjects required can be calculated from

$$\left[\frac{1}{q_1} + \frac{1}{q_2} \right] S^2 (z_\alpha + z_\beta)^2 + E^2$$

where

q_1 = proportion of subjects in group1

q_2 = proportion of subjects in group2

For a two-tailed test

$z_\alpha = 1.96$ for $\alpha = 0.05$

$$z_{\alpha} = 1.645 \text{ for } \alpha = 0.10$$

For a one-tailed test

$$z_{\alpha} = 1.645 \text{ for } \alpha = 0.05$$

$$z_{\beta} = 0.84 \text{ when } \beta = 0.20 \text{ (Power of 80\%)} \text{ and } z_{\beta} = 1.282 \text{ for } \beta = 0.10 \text{ (Power of 90\%)}$$

For two equal sized groups the table below provides a ready guide for sample **size per group**

E/S	One-tail	α	0.00			0.02			0.05		
	=	=	5			5			5		
	Two-tail	α	0.01			0.05			0.10		
	=	=	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
0.10			3563	2977	2337	2599	2102	1570	2165	1713	1237
0.15			1584	1323	1038	1155	934	698	962	762	550
0.20			891	744	584	650	526	393	541	428	309
0.25			570	476	374	416	336	251	346	274	198
0.30			396	331	260	289	234	174	241	190	137
0.40			223	186	146	162	131	98	135	107	77
0.50			143	119	93	104	84	63	87	69	49
0.60			99	83	65	72	58	44	60	48	34
0.70			73	61	48	53	43	32	44	35	25
0.80			56	47	36	41	33	25	34	27	19
0.90			44	37	29	32	26	19	27	21	15
1.00			36	30	23	26	21	16	22	17	12

Table for sample size required when using the correlation coefficient (r)

r	One-tailed Two-tailed β	$\alpha =$	0.00			0.02			0.05		
			5	5	5	5	5	5	5	5	5
		$\alpha =$	0.01			0.05			0.01		
		$\beta =$	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
0.05			7118	5947	4663	5193	4200	3134	4325	3424	2469
0.10			1773	1481	1162	1294	1047	782	1078	854	616
0.15			783	655	514	572	463	346	477	378	273
0.20			436	365	287	319	259	194	266	211	153
0.25			276	231	182	202	164	123	169	134	98
0.30			189	158	125	139	113	85	116	92	67
0.35			136	114	90	100	82	62	84	67	49
0.40			102	86	68	75	62	47	63	51	37
0.45			79	66	53	58	48	36	49	39	29
0.50			62	52	42	46	38	29	39	31	23
0.60			40	34	27	30	25	19	26	21	16
0.70			27	23	19	20	17	13	17	14	11
0.80			18	15	13	14	12	9	12	10	8

Sample size for a descriptive cross-sectional study of a continuous variable

W = desired width of confidence interval and S = standard deviation

	Level		
	Confidence		
	90%	95%	99%
W/S			
0.10	1083	1537	2665
0.15	482	683	1180
0.20	271	385	664
0.25	174	246	425
0.30	121	171	295
0.35	89	126	217
0.40	68	97	166
0.50	44	62	107
0.60	31	43	74
0.70	23	32	55
0.80	17	25	42
0.90	14	19	33
1.0	11	16	27