# Chapter 3

# Measurement units

This chapter deals with important methodological decisions to be taken when conducting a study: the choice of units that will be observed or measured. This chapter is firstly a building block for subsequent chapters that will deal with descriptive and inferential statistics. Descriptive statistics (see chapter 6) are used to describe the general properties of our measurement units, such as their mean, or variability. Inferential statistics (see chapter 8) are used whenever we want to *infer* something from a sample, and declare the findings from that sample applicable to a larger set of research or measurement units, a so-called *population*. This chapter is also a building block for the analysis of data gathered through qualitative methods (discussed in chapter 7).

As in social science research, so also in empirical legal research we very often work with samples, as we generally do not have the time or resources to study an entire population. It is then important to ensure, as far as possible, that a sample resembles the population we are interested in.

In the following, we will first discuss the concepts of sampling and populations. We will discuss what constitutes a useful sample and what constitutes a less useful sample, in terms of representativeness. We list probability and non-probability samples, and discuss their pros and cons. We will also discuss the practical constraints in drawing samples. Next, we detail concepts such as representativeness, bias and noise, sample size and issues of nonresponse.

## 3.1 Populations and samples

The Netherlands has almost 18 million inhabitants. If I wanted to know how tall Dutch people are, I could set out with a measuring tape to measure all Dutch people, each and every one. This would take a very long time. By the time I am finished, quite a few will have died, new Dutch citizens will have been born, and perhaps I will be ready for my pension by that time too.

The entire population of the Netherlands is what we also call in a statistical sense a *population*: a universe of units that we are interested to know something about. We could also be interested to know the height of all female Dutch citizens, in which case

the population would consist of all female Dutch citizens. Or we might want to know the height of all adult Dutch citizens, in which case the population of interest would be the Dutch population aged 18 and over.

In this example, 'population' is taken very literally, in the sense that the statistical population is also literally a population: all inhabitants or citizens within a certain geographically defined unit or with a certain legally defined status. A population could however also be all defendants, or all tort claims in a defined jurisdiction, or all rulings. A population can even be tissue samples, or movies or burglaries or desserts. What matters is that there is a certain total reservoir of units of interest that we want to know something about. The people who together constitute the population, or the females or adults who do, or units, whether they be tort cases or light bulbs or desserts, are the *population members*.

Now in the example that we started with, it is clear that it would be an impossible task to measure all population members. It is clearly undoable. But even if a population were to have 'only' 1,000 members, assessing each and every member of that population would be a hefty task. There is more.

Because what statistics teaches us is that scrutinizing each and every population member is actually not necessary. If we operate wisely and according to certain rules, we can with reasonable precision say something about the entire population even if we investigate only a part of that population, a *sample*. Often, already a small part will do, like a 1% sample, or even less, depending on various issues that we will talk about later. So, basically, what statistical science teaches is that it is not necessary to scrutinize each and every population member to infer something about that population. In fact, within reasonable limits, we can get to know the population reasonably well through the study of only a selection of its members.

Sampling only a part of the population saves a lot of expenses. The other side of the coin is that – obviously – we are not absolutely certain any more of what we say. We have with our savings introduced *uncertainty*.

Statistics is the science that teaches us how to deal with that uncertainty. It gives us the rules and procedures, and in a sense the language in which we communicate how we managed the uncertainty in a sensible way, how uncertain we exactly are – and thereby how confident we are about the conclusions we draw from the sample about the population. Much more about this in chapter 8.

## 3.2 Representativeness and generalizability

Let us start again with an example. We want to know how Dutch citizens of voting age will vote in the upcoming general election in the Netherlands. In numerous countries, such questions are asked in polls right before elections, when politicians are bashed in TV shows and wise men and women air their views on why such and such party is losing and some other party has gained ground suddenly.

Now if we wanted to conduct such a poll, we could not possibly interview all Dutch citizens of voting age, there simply is no time. We therefore draw a sample. But how do we draw a sample? If I were to do door-to-door interviews in my neighbourhood I would definitely not get a good prediction of the average Dutch person's voting be-

haviour. My neighbourhood contains quite a few nice houses, no flats: it is a pretty, green, slightly upper-class area, with lots of academicians, some medical specialists and artists. I would predict perhaps, sampling only my fellow neighbourhood inhabitants, that the Dutch liberal party D66 would win, or the Green Party, or the Party for Animals. That would not reflect the voting behaviour of the average Dutch citizen. My sample does not reflect the Dutch population, but a specific segment – and while my conclusions about the types of citizens inhabiting my neighbourhood might be right, I would draw the wrong conclusion about the entire population. Similarly, if I were to interview only inhabitants of the Schilderswijk in The Hague, a neighbourhood where many citizens with a migrant background reside, or of the Bijlmer in Amsterdam, I would probably also not get an accurate picture of Dutch voting behaviour, nor if I were to carry out my investigations in Wassenaar, the Beverly Hills of the Netherlands.

Clearly, what we would want, and what all the options above are lacking, is that my sample would resemble the population, would reflect the population. We would want the sample's properties to be like the properties of the population. In statistics-speak we want a *representative sample*.

So, in this case we want the sample members to exhibit about the same voting behaviour as the population. We cannot check that, as we do not know the population's voting behaviour (that's namely what we're after!). Instead, we could check whether we have sample members with approximately the same age as the population, the same sex ratio, urban/rural composition, the same geographical spread, income, etc. as the Dutch population. See Figure 3.1. The idea behind this is that, if sample members resemble population members in various pertinent characteristics, we expect the sample's voting behaviour to then also resemble the voting behaviour in the larger population.

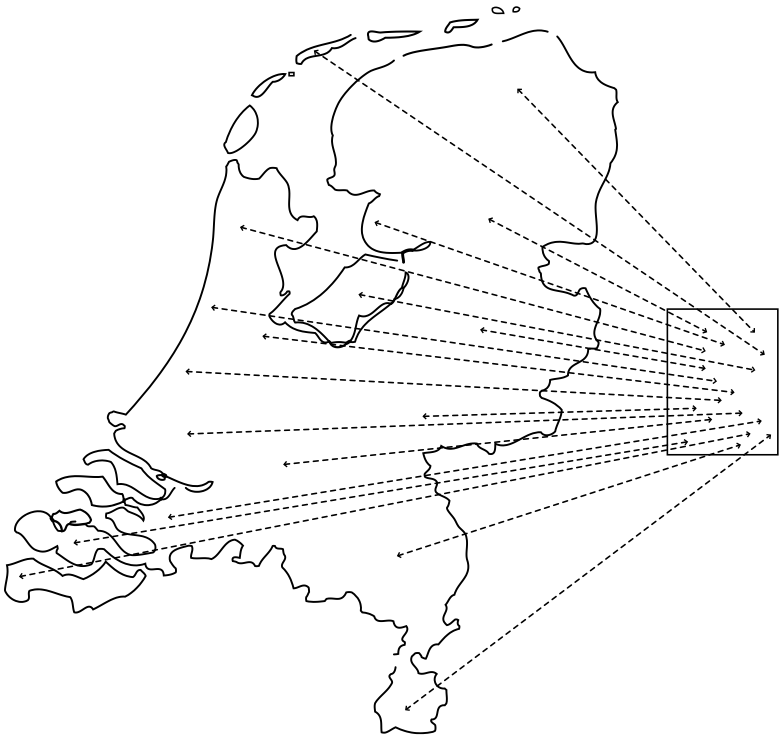
But how to achieve this? What are the pertinent characteristics for which we must ensure that the sample is similar to the population? How many such characteristics do we need? And how in practice could we achieve that? Luckily, these questions do not need to be answered nor do difficult procedures need to be followed. The best and easiest way namely to ensure that a sample's properties reflect those of the population is to draw that sample by chance, or: *at random*.

Suppose that we want to draw a random sample from the Dutch population. All Dutch inhabitants are registered at their municipality, and all these registrations form a list of the population members. This list is what we call the *sampling frame*. We can imagine random sampling now as follows: we could give every Dutch citizen of voting age a slip of paper with his or her name on it, we'd crumple up all those slips, throw them in a big pan, stir them really well with a very big wooden spoon, and then we'd take out – blindly – as many as we need for our sample. Such a sample is called a *random sample*: it is formed only by chance. Only chance determines whether someone ends up in the sample or not. For a random sample we may always assume that it is representative.

Once a sample is representative, it follows that we can *generalize* the sample properties to the population. So if we find that 36% of our sample members vote Labour, 14% Green, 35% Liberal, we conclude that 36% of Dutch voters would vote Labour, 14% Green and 35% Liberal.

Note that a sample is not just representative for any population, but is representative *for a certain population*. So a random sample of females of voting age is representa-

**Figure 3.1:** Example of spread of sample members in the Netherlands



tive for female Dutch voters, but not representative for all Dutch voters. And a random sample of all victims of sexual abuse in a compensation scheme is representative for all victims in the compensation scheme, but not representative for all victims of sexual abuse. A random sample of tort cases brought before a judge is not necessarily representative for all tort claims.

Dijksterhuis & Vels (2011) wanted to assess the opinion of Dutch citizens on the methods by which child maintenance is calculated according to Dutch law. They surveyed a random sample of clients of the National Bureau for Maintenance Contributions (LBOI); these are citizens who have to pay child or partner maintenance but who defaulted at some point. Of these the authors only sampled from those who had defaulted on child maintenance obligations. Their response rate was 22%. The authors found that 72% of respondents disagreed with the prevailing methods of calculation of maintenance. The report led to some political commotion and even to an initiative for a bill in which a new method of calculating child maintenance was proposed. However, the sampling frame was not made up of Dutch citizens, nor was it composed of citizens who had to pay some kind of maintenance, nor was it composed of all defaulters. It is therefore unclear how Dutch citizens regard maintenance calculation methods, as well as how the more narrowly chosen but perhaps more immediately relevant group of

those who pay (or receive) such maintenance would regard these methods. Given that the sample was composed only of citizens who had defaulted on child maintenance, it is very much doubtful whether the 72% that the study found is generalizable to all citizens who pay or receive maintenance, let alone the entire Dutch population.

Note that if we want to obtain generalizable statements from our research, it follows that the sample is by itself wholly uninteresting. No one is interested in some particular sample. The sample only starts to be of interest once it is representative for a larger entity, once it tells us something about the population we are interested in. If I have interviewed some sample of victims of sexual abuse by the clergy, then the properties of that sample are essentially unnoteworthy. If however that sample was drawn such that it is representative for a larger population, then suddenly my sample findings become relevant. So, if my sample of victims of sexual abuse by the clergy is representative for all victims of such abuse in compensation schemes, then it is interesting to note (and declare applicable to all such victims) that well over one in three had suffered repeated invasive sexual abuse or that only 40% report that the compensation scheme is adequate.

This applies most forcefully in quantitative research. In quantitative research we are interested in making generalizable statements. For that reason, sampling – and ensuring that one has obtained a good sample – is more important in quantitative than in qualitative research. As we said, in section 1.4.1 and later in chapter 2, a qualitative study generally focuses on a particular context, and generalization to other contexts is less aimed for and may in fact be regarded as unattainable given that contexts can be so different and context is inherently so important.

Note, secondly, that within the sample we have no uncertainty. I can with certainty measure (and no one can refute my finding) that the average age of victims of sexual abuse by the clergy whom I was able to interview because they had registered for a compensation scheme, is 53.5. This is a fact. If I attempt to generalize that finding to the population, I would state that the average age of all victims of sexual abuse by the clergy is also 53.5. However, now I am not so dead-sure. I have drawn a sample, so it might be that by chance I have a few too many younger victims because these are more likely to submit claims, or by chance a few too many older victims because these would be more willing to be interviewed. This means that while the sample mean age of 53.5 is a measurement, an irrefutable finding, the population mean of 53.5 is an *estimate*.[1] We will return later to the issue of whether and why the sample mean is a good estimator for the population mean.

## 3.3 Types of samples

So what kinds of samples are representative for the population, and what kinds of samples are not? In the literature two kinds of samples are distinguished: *probability samples* and *non-probability samples*. Drawing a probability sample, representativeness is

[1] We denote this by writing $M_X$ for the sample mean, and $\hat{\mu}_X$ for the estimate (indicated by the 'hat') of the population mean. By convention we write sample characteristics such as the mean with Latin literals (M, s, $r_{XY}$); whenever we refer to properties of the population it is the convention to use Greek literals (such as $\mu$ for the mean, $\sigma$ for the standard deviation, $\rho$ for the population correlation coefficient, etc.).

guaranteed; drawing a non-probability sample, representativeness is not ensured. This does not mean that a non-probability sample is always a useless sample, or by definition non-representative. Non-probability samples may very well be representative, but there is no guarantee.

### 3.3.1 Probability samples

As we argued above, samples are useful only insofar as they tell us something about the population we are interested in. In technical terms: a sample is useful only if it is representative for the population of interest. We already said that random samples may always be considered representative. But how do we assess whether a sample is representative?

The answer to this is fairly straightforward. A sample is representative if every population member had an equal chance of ending up in the sample. This is the litmus test of representativeness.

So, if I draw a sample from the province of Zuid-Holland in the Netherlands to investigate voting behaviour, the sample cannot be taken as representative for the Netherlands – those living outside of Zuid-Holland had a zero chance of ending up in the sample. All law students present at the ELS methods class of Monday morning are not a random sample of all law students at the university: those who have already passed will not sit in, nor will those who on a rainy Monday morning preferred their warm beds over a trip by bike through the rain to the university campus. That the latter is self-chosen and these students could have ended up being sampled does not matter. What matters is that the students who dragged themselves from their warm beds likely differ systematically from those who napped on: the ones in class are probably more motivated, or more in need of the ELS methods class than those who did not show up. The unmotivated or those unfearful of methodology therefore had a smaller chance of ending up in the sample. The sample members in that sense differ in a structural, systematic way from those who did not become sample members.

To know whether each population member really has an equal chance of ending up in the sample sometimes requires a bit of brain cell jogging. As an example, imagine that we carry out a survey in a certain country: we sample household addresses randomly, and at each address we interview that person older than 15 years of age who is the first to celebrate his or her birthday. Is this a random sample of all inhabitants? Obviously not, as those under 15 years of age cannot participate. Is it then a random sample of all inhabitants over 15 years of age? One might be tempted to say yes, as households have been sampled randomly, and as within each household again in a sense randomly a household member is chosen. The sample is not random however, as members of larger households have a smaller chance of ending up in the sample: someone living in a 6-person family has an a priori chance of $\frac{1}{6}$ to be selected, whereas someone in a 2-person household has a chance of $\frac{1}{2}$ to end up in the sample. By constructing the sample this way, we will end up with too many singles, and too few people from large families: persons living in large families are *underrepresented* in the sample, and persons living in smaller families are *overrepresented*.

Four kinds of probability samples are distinguished, and four kinds of non-probability samples. We briefly discuss them, starting with probabilistic samples and ending

with the non-probabilistic ones. It should be said ahead that in many studies, mixtures of probability samples are employed.

(1) *Random sample*

The first kind of probability sample is the random sample, also known as the *simple random sample*. A random sample is a sample where sample members are drawn as if we had a huge cooking pot, into which we toss a lot for each population member, stir very well, and draw lots blindly, irrespective of any properties of the sample member. Such a sample one also encounters as a 'simple random sample', as an 'a-select sample' or as a 'flat' random sample.

The simple random sample is not used very often though. The reason for this is firstly that drawing such samples requires quite a bit of administrative infrastructure: if there is no such thing as a pre-existing sampling frame, it is an immense task to draw up a list of population members and then select a few of them randomly. Secondly, if we draw a random sample, we are left in fate's hands as regards the representation of certain segments of the population.

We will clarify this with two examples. Suppose that I draw a sample of 100 registered cases from the police databases; police databases in most European countries contain thousands of cases so it would be logical to draw a sample. I would expect to end up in my sample with 85 male suspects and 15 female suspects as that is the distribution of male and female offenders in the population of suspects. However, it might very well be (as chance decides who ends up in the sample) that I end up with 25 females and 75 males. This is what we call *sampling error*. And even though these sampling errors are due to chance, these deviations can be problematic: suppose for instance that I end up in the above example with 5 females and 95 males. In that case, I have really too little to go on if I wanted to say something about female offenders – 5 is a very small number to draw any conclusions from. So, in this case it might be wise not to let chance have its way entirely: it might be better to try to exert some control over the representation of the group of female offenders in the sample.

Similarly, suppose that I want to assess fear of burglary in a neighbourhood. If I were by chance to select 25 houses in that neighbourhood, I could end up with all houses close to the shopping centre, or relatively many on the edge of the neighbourhood (areas with typically a relatively high risk of burglaries and therefore also a higher fear of being burglarized). Or I might end up with disproportionately many apartments. Also here, it might be desirable to enforce a certain spread over the various areas and types of houses in the neighbourhood.

Lastly, and most importantly, strictly random samples may be expensive. If I were to draw a simple random sample to interview 170 Dutch citizens out of all 17 million inhabitants, it would probably mean that I'd do seven interviews in Rotterdam, ten in Amsterdam, four in Utrecht or The Hague, and that I'd need to travel to another 100 or so municipalities to interview the remaining approximately 135 respondents. That is very time-consuming and costly. It is also unnecessary, as we will see below. For all these reasons, researchers often choose to use more efficient random samples than the simple random sample.

(2) *Systematic sample*

The first alternative to the simple random sample is the systematic sample. Suppose that we are interested in fear of being burglarized and we are investigating a newish neighbourhood in a big city. The neighbourhood has 250 standard family houses (in which we assume that each time 2 adults live). If we want to draw a 10% random sample of these adults, this means that we could visit 50 houses and in each house interview one adult. This would give us a sample of 50, which is 10% of all 500 adults living in this neighbourhood.

The systematic sample gives an efficient framework for achieving this, and it works as follows. All houses are first numbered, from 1 to 250. We next draw at random a starting number, say 137, meaning we start from house 137 on our map. Starting from house 137, we draw a track through the neighbourhood, following the layout of houses, and we take every 5th house (as we want to survey 50 out of 250 houses), ring the doorbell and ask to interview someone, for instance the first adult to celebrate his or her birthday. We move up, stepwise at regular intervals, to house 250 and next finish sampling from house 1 up to house 137. As the first number has been drawn at random, and as after this number all other selected houses depend on this starting number, each house has the same chance of ending up in the sample and so does (given that we assume two adults live in each house) every adult. This is a systematic sample, and it is a random – as well as an efficient – sample.

Two conditions must be met for a systematic sample to be random and representative. First, the starting number must be chosen randomly. Second, the physical layout of the neighbourhood, or in more general terms, the ordering of sampling units on the list, must not have the same *periodicity* as the sampling interval. To understand this, inspect Figure 3.2, which gives a hypothetical example of a neighbourhood plan. In this neighbourhood, houses have been built in blocks of 5. Suppose that our random starting number is 15. There are 30 houses in this example, and suppose we want a sample of 6 houses, this means that we would have to sample each 5th house. However, if we start at number 15, and if we interview someone in every 5th house, this would mean that in our sample we would end up with houses at the corner end of a block only. And, obviously, the risk of a burglary is larger in a corner house than in a house in the middle of a block. This means that we end up with a sample in which people who live in corner houses (which may be assumed to be special) are overrepresented and people who live in ordinary in-between houses are underrepresented. Thus, the sample is not representative for the neighbourhood.

So, in this case it might be better to change the periodicity of the sample and for instance sample every 4th house (in which case you end up with a larger sample), or every 6th house (in which case you end up with a smaller sample), or three times sample every 4th house and the next three times sample every 6th house (in which case you end up with a sample of 6 as planned). See Figure 3.3 which gives the last solution.

**Figure 3.2:** Example neighbourhood layout for systematic sample with inadequate periodicity

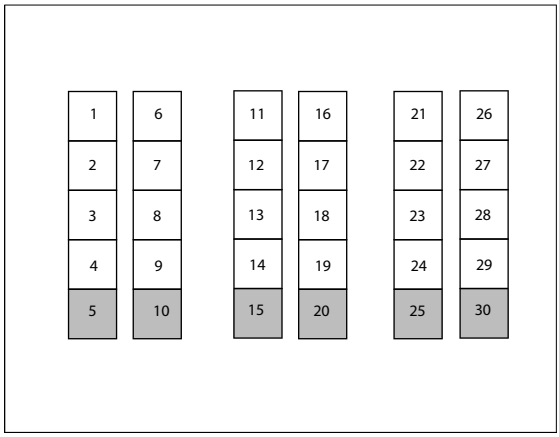
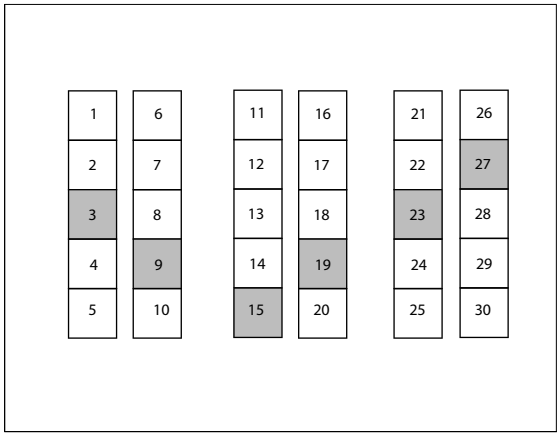



**Figure 3.3:** Example neighbourhood layout for systematic sample




(3) *Stratified sample*

Let us return to the example that we gave above, of male and female police suspects. We said that if we drew a flat random sample of 100 respondents it might occur by chance that we ended up with only 5 female suspects. Five respondents is too small a number to draw meaningful inferences from: statements such as 40% of female suspects had committed a violent offence are based on just two respondents – had it been not 2 but 3 females who had committed a violent of-

fence the percentage would have been 60%! So, if we want to make meaningful inferences about this smaller segment in the population, we have to ensure that we have a sufficient number of females.

There are two ways to go about this. Sticking with the example, we could randomly draw 85 male respondents from the population of male suspects and similarly draw – at random – 15 female suspects from the population of female suspects. In this way, we have forced the sample to have the same proportional representation of males and females as we know the population has. The males and females are considered a 'layer' or *stratum*, and from each layer a sample is then drawn of the requested size. This kind of sampling is called stratified sampling.

Obviously, 15 female respondents is not a large number to make stable inferences about. It might actually be best if we had 50 males and 50 females, so that we could investigate both groups of suspects with the same level of detail and precision. So, what we could also do is to randomly draw 50 female suspects from the population of female suspects, and another 50 males from the population of male suspects. This kind of stratified sampling is called *disproportionate stratified sampling* – as opposed to the previous kind which constituted proportionate stratified sampling. Obviously, now the ratio of males to females has been changed, which means that – while we can carry out computations within each layer – we cannot compute averages over the strata just like that. If we wanted for instance to know how many offences suspects have committed on average, we cannot use the sample mean for that. There are too many female suspects in the sample, and the sample is not representative for the population of suspects anymore. This can however be solved by *re-weighting* the data. Suppose for instance that men have committed on average 2 previous offences, and women on average 1. We can then compute the mean as $2 \times 85/100 + 1 \times 15/100 = 1.85$.

Stratified sampling is often used when the population under investigation is heterogeneous in the sense that it consists of several groups of unequal sizes. If we want to study each group with similar precision, the smallest groups are oversampled, and for presenting sample findings the results 'weighted back' afterwards.

(4) *Cluster sample*

Cluster samples (also referred to as two-stage sampling) are used very often. The reason for that is that they save a lot in terms of time and money, and they can also be carried out very well when precise information about the population distribution is lacking, when we have no sampling frame.

To illustrate how a cluster sample works, suppose that we want to interview prisoners – we are for instance interested to know how they judge the prison climate. In the Netherlands, on any given day, about 10,000 persons are detained. There are 24 penitentiary institutions. Some are larger, some are smaller. There is a centralized database of detainees that keeps track of admissions, transferrals to other institutions, and discharge, so there is a sampling frame. This means that

in principle it would be very easy to draw a simple random sample. If we wanted a sample of 200 detainees, we could simply number the detainees in the database and draw them using random numbers. However, if we were to do that, it would mean that we probably would have to visit almost each and every penitentiary institution in the Netherlands.

A much more efficient way is to first randomly sample penitentiary institutions, and next draw random samples of detainees from those selected institutions. Supposing that each institution is of about equal size, this procedure would result in a random sample: each institution has an equal chance of being selected for the sample, and within each institution again each detainee has an equal chance. Obviously, this is very efficient: if we want to interview 200 detainees, we could simply randomly select 5 detention facilities, and from each interview 40 detainees. That would save a lot in travelling cost and time. One could of course argue that in reality detention facilities are not likely to be of exactly equal size (so that detainees in smaller facilities would have a bigger chance to end up in the sample), but even that can be accommodated: we can simply give each detention facility a sampling chance proportional to its size.

Cluster samples have huge advantages, in terms of feasibility and cost. But they come at a price too: in research methodology – as in many areas of life – for every gain usually a price is paid. This has to do with the fact that every time we sample, we introduce sampling error: we might in a flat random sample – by chance – find relatively tall Dutch people, or tort cases with relatively high claim amounts. Such sampling error is all in the game: we know that the price we pay for not investigating each and every population member is uncertainty. We know that when we generalize the findings in the sample to the population, there is always a margin of uncertainty: sampling error.

However, if we now sample twice (as we do in our example cluster sample: once over penitentiary institutions and again over detainees within those centres), we essentially accumulate sampling error. It might for instance be that by chance we have selected those institutions in which detainees have very few distractions and there is a relatively harsh climate, and in second instance that we have selected – by chance again – within those institutions those individuals who have a bleak outlook on life by nature. The fact that we sample in two steps exposes us twice to sampling error. And sampling error in the first instance is in a sense 'multiplied' by sampling error in the second instance. For that reason, results from cluster samples are more susceptible to sampling error, and one could say more 'volatile'.

Such added volatility, or as it is called *noise*, is especially likely to occur if members of clusters resemble each other. To understand what happens, imagine that we have a cluster sample of m clusters, with b respondents per cluster (so that we have b $\times$ m sample members, generally written as N). Now, if we add one extra sample member, from one of our clusters, this gives us in principle more information on the population (we now observe more population members). However, suppose now that cluster members are very much alike. Then, adding a sample member from a cluster that is already present in the sample doesn't give that

much extra information. Loosely speaking, adding one extra person from the same cluster gives us a bit more of what we already knew.

Practice has revealed that cluster members indeed often share characteristics. Population members from the same cluster are more alike than population members from different clusters. This implies that if we draw a cluster sample of, say, 100 respondents from 5 clusters (so 20 from each), we will have less information about the population than if we had drawn 100 respondents in a flat random sample from the population.

This unfortunate consequence of cluster sampling is referred to as the *design effect*. This design effect, usually abbreviated as DEFF, is expressed in a number that gives the increase in variability of our estimates (and therefore our imprecision) due to the fact that a cluster sample has been chosen – relative to the variability that we would have had, had we chosen a simple random sample. So, if DEFF equals 3, this means that the variability is thrice that of a simple random sample of the same size. If DEFF is 1.40, this means that the variance is increased by 40% – as DEFF is a ratio, a DEFF of 1 would indicate no difference.[2] As a rule of thumb, we find DEFF values higher than 3 problematic.

Now why exactly does this design effect lead to imprecision of our estimates as we said, or differently said, to variability of our estimates? Let us assume that we are interested to measure prisoners' moods. We want to know whether they are depressed or not. If we were to draw a simple random sample of, say, 200 prisoners, we might find that their average depression score is 5. If we were to draw a new random sample of 200 prisoners, that outcome might differ a bit: it might be 4.9, or 5.1, or even 5.3 or 4.7. Chance will give us a different sample each time, with slightly different results. That is the uncertainty we have to accept.

We now draw instead of a flat random sample (where we would need to travel to many different penitentiary institutions, all over the country), a cluster sample. We draw by chance (proportional to their size) two penitentiary institutions, and we interview prisoners from those two institutions. It might be (chance is at work) that we drew two institutions with harsh regimes, bad food, located far away from public transport. This affects prisoners' moods. So we find a mean of 8.3: the prisoners are on average quite depressed. If we were to re-do the cluster sampling, and we again selected just two institutions from which we then interview prisoners, we might, however, by chance end up with two glitzy new facilities: good food, closer to a metro station (so more family visits), open regimes. Prisoners feel much better now, and we end up with a sample mean of 2.2. Depending on what clusters we accidentally sample, we may get widely different sample averages.

[2] DEFF is computed as DEFF $=1 + \rho \times$ (b-1) with $\rho$ the so-called intra-class correlation, a measure that reflects the extent to which members from the same cluster resemble each other, and b the cluster size. Note that if $\rho$ is 0, in which case cluster membership does not mean sample members have similar scores, then – as it should – DEFF becomes 1, i.e. there is no difference in variability between the cluster sample and an ordinary random sample. Note also that if b = 1, in which case the clusters have size 1 so that effectively we have a random sample, DEFF also – as should be the case – becomes 1.

What these examples show is that in a cluster sample, the selection of just a few clusters is in fact a tricky business: we are much more vulnerable to ending up with one or two particularly nice or particularly nasty facilities. As a flat random sample generally makes us travel to many facilities, that risk is very much mitigated then.

Hypothetically, if we were to re-do our study 100 times, and draw a flat random sample of prisoners 100 times, the mean depression scores from those samples would not vary much. If however, we were to re-do our study 100 times with cluster sampling, the means of these 100 samples would likely vary – considerably – more. This is expressed by saying that cluster sampling gives more 'noisy' results. This noisiness of cluster samples relative to regular simple random samples is expressed using DEFF.[3]

Many surveys use cluster samples. This is mainly for pragmatic reasons. Unfortunately, key variables of interest are often clustered. Prisons differ in their regime, law firms differ in corporate culture, courts differ in the likelihood to award damages, municipalities differ in the strictness with which they check on recipients of benefits.

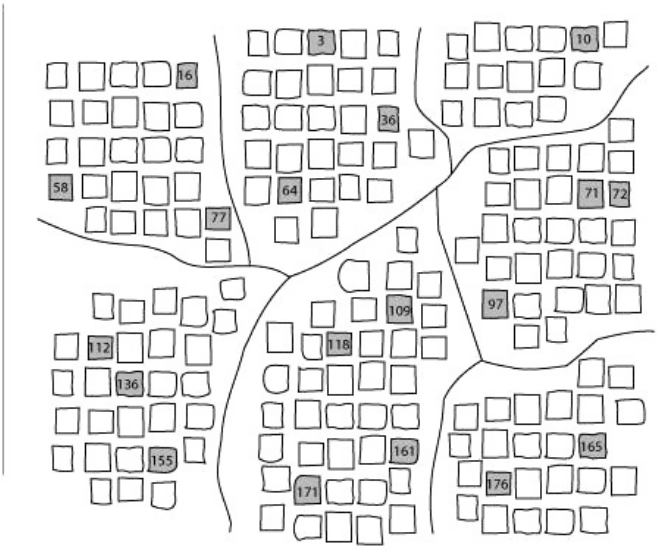
DEFF measures are computed *per variable*. So, separate DEFFs can be computed for claims, for sentence length, for socio-economic status. This implies that the DEFF may be high for one variable but need not be so for another.

It is generally advised to aim to employ at least 30 clusters, to keep the design effect within acceptable limits. If then 7 units per cluster are assessed, this results in a sample size of a little over 200. In empirical legal studies this is often not possible: we may for instance have only 20 prisons or only 17 courts from which to sample. Researchers then often impose some kind of stratification over the clusters to ensure spread over pertinent properties, such as by choosing courts from rural as well as urban areas, from the north and south of the county, etc.

The figures below illustrate simple random sampling, systematic sampling and cluster sampling for a hypothetical study into trust in the judiciary, within a village. As can be seen, both the random sample (Figure 3.4) and the systematic sample (Figure 3.5) have an $N$ of 19; the cluster sample (Figure 3.6) in this example simply selects two clusters and thus leads in this example to a much larger sample size (obviously we could have chosen to select 10 sample members from each). Note that for both the simple and the systematic sample the spread over the village is pretty nice: for the systematic sample the area in the bottom left is less well represented but such things are bound to happen by chance. The cluster sample has a nice geographical spread as well. Of course what could easily have happened is that we could have ended up with two adjacent clusters, which may be less desirable as they may be alike in terms

[3] Some researchers use instead of DEFF a measure called DEFT, which is nothing other than the square root of DEFF: DEFT $= \sqrt{\text{DEFF}}$. Given that DEFF relates to the variance in the sample, DEFT relates to the standard deviation (viz. section 6.3.2). DEFT can be interpreted as the increase in the standard deviation relative to the standard deviation had a simple random sample been drawn. All in all, whether DEFF or DEFT is used, both are indicative of the precision-price you pay for the money you save in drawing a cluster sample.

**Figure 3.4:** Fictional example of a random sample in a village




of being less or more affluent neighbourhoods (a property associated with trust in the judiciary). However, as we need to sample randomly to ensure that our sample passes the litmus test, we cannot force this to be so.

### 3.3.2 Non-probability samples

If we cannot draw a representative sample, we can always draw a non-representative one. Reasons to draw a non-representative sample may be that there is no sampling frame. For instance, we may want to interview illegal residents, a category of persons that typically is not registered. A second reason is that we are not interested in generalization per se: we may be aiming to understand the rationales that judges employ when they have to take decisions in divorce procedures where one parent has accused the other parent of sexual abuse of a child. If we aim to answer such a research question, generalization is a less prominent goal: rather, we would want to have in-depth discussions with judges and lawyers, possibly witness a number of court hearings, interview psychologists and child protection services staff. Our aim is rather to find a reasonable representation of types of cases and be able to find sufficient well-informed actors in these cases so that we can uncover the manner in which judges process information, reason and subsequently take decisions. Other reasons for not drawing a representative sample may be that we do not have the time or resources to do so. Or the study may be just a first pilot study, or researchers may simply be lazy, sometimes combined with there being no need to generalize to a population: we may for instance be studying phenomena deemed universal in people, such as physiological reactions.

**Figure 3.5:** Fictional example of a systematic sample in a village

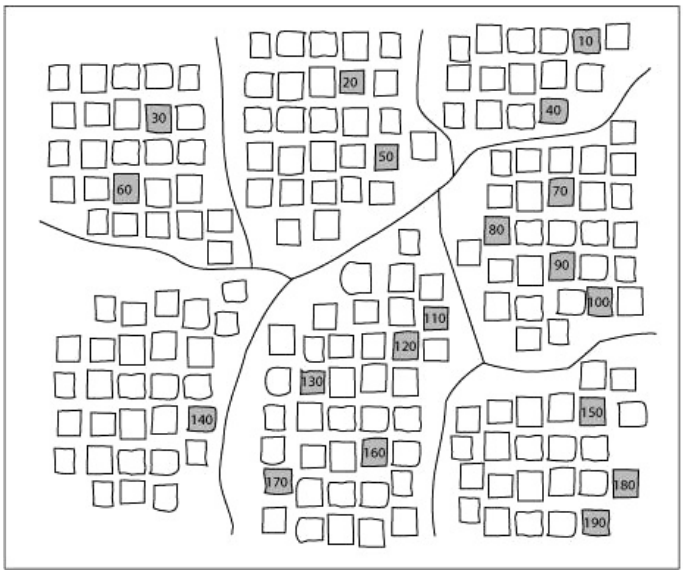
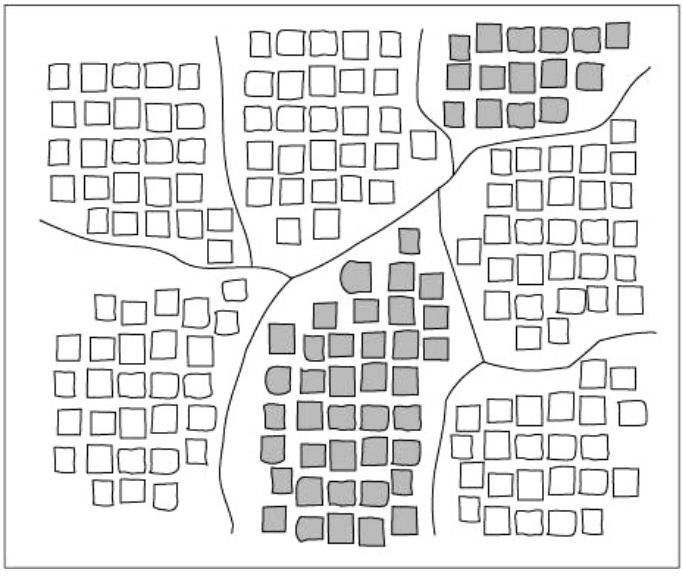



**Figure 3.6:** Fictional example of a cluster sample in a village



It is not the case that non-probability samples are never representative. Some non-probability samples may in fact very well be adequate or even representative. We however do not have the *guarantee* of representativeness that we have when we draw a probability sample, so we cannot rely on statistical sampling theory (see chapter 8).

We will now discuss four generally distinguished types of non-probability samples. There are clear 'quality' differences in terms of how likely they are to be representative. The least sophisticated is clearly the convenience sample. For some situations, the snowball sample approaches a probability sample and has good generalizability properties. Again, if one's aim is not generalization, many of these samples may be perfectly adequate for the research purposes.

(5) *Convenience sample*

The first type of non-probability sample that we discuss is the *convenience sample*. It is also encountered under different names, namely *accidental sample*, *availability sample* or *haphazard sample*. These terms make clear that the convenience sample is not a very sophisticated type of sample. When one draws a convenience sample, one simply selects for the sample whomever is available, whomever is 'at hand'. The best example of a convenience sample is the samples that consist of university students: university researchers have such sample members in arms' reach, in large numbers and on the cheap. In that sense, one could also say that the convenience sample is a lazy sample. For that matter, the term 'accidental sample' is perhaps less apt, as it seems to imply that some chance or random phenomenon is in operation to form the sample.

Convenience samples can be drawn in many contexts. We already mentioned students, who are very often used in smaller-scale experiments. Other examples are interviews with visitors to a shopping mall, or sending questionnaires to the readership of a journal.

Clearly, given the manner in which it is drawn, the convenience sample offers no perspective whatsoever of representativeness. If representativeness is important and a convenience sample has been drawn, it is very important to describe the manner in which the sample was drawn and detail in what way the sample might be representative for (a part of) the population, and in what way biases might have occurred.

This does not mean that convenience samples are always a bad choice. If a questionnaire needs to be tested, or a pilot study carried out, or an exploratory study, a convenience sample may be a perfectly efficient choice. In such cases there is no need for representativeness – so why go to all the trouble? Also, it may be that we are testing universal human properties that are not affected by social class or intelligence, such as visual memory. Or we may be investigating the association between a certain instruction A and B, and behaviour during an experiment, in which case we do not want to generalize to a population but are interested in comparing the group that had instruction A with the group that had instruction B, within the sample. Our first aim is then not so much external, but internal validity (see section 2.6). Also, it may be the case that the easily available respondents, such as students or shoppers at a supermarket, are precisely the

ones a researcher in interested in, for instance for studying student jobs or sexual harassment in fraternities and sororities, or for studying shoppers' awareness of consumer information on products.

(6) *Quota sample*

The second type of non-probability sample we discuss is the quota sample. A quota sample is a non-probability sample that attempts to mitigate some of the drawbacks of the convenience sample by assuring that the distribution of relevant properties in the sample is similar to their distribution in the population. So, as an example, assume again that we want to study 100 police suspects. We do not have a lot of time and/or money so we interview all suspects who have been held by the police and then released: when they leave the police station we ask them whether we may interview them. Obviously, this is not a representative sample: some suspects are transferred to jail after arrest, or to psychiatric wards and we miss out on those. However, given that we know that 15% of police suspects are female and 85% male, we can ensure that our 100 interviewed suspects have the same gender distribution, by interviewing 15 females and 85 males. Once we have interviewed 85 males, we then interview only females to arrive at a quota sample that has the same gender distribution as the population.

The quota sample is like a stratified sample in the sense that per stratum a desired number of sample members is drawn: the researcher attempts to guarantee that the relative representation of certain properties of sample members is as in the population. The manner in which that is achieved is however much less rigorous than in a regular probabilistic sampling procedure: there is no random sampling within strata. For that matter, the quota sample is also often referred to as the non-probabilistic penchant of the stratified sample.

(7) *Purposive sample*

In a purposive sample, sample members are selected because they have certain desirable properties. For instance, they may be mediators employing a specific negotiating strategy in divorce conflicts, or members of a particular international corporate law firm. These categories of people have especially relevant, valid, in-depth, unique information that is of use to the researcher interested in the novel strategies employed by these mediators or in corporate culture in big international business law practices. In a sense, they are selected as 'unrandomly' as possible – they are targeted to provide certain information. Such respondents are also sometimes referred to as *key informants*, as they literally hold the key to the inaccessible topic or sphere. In that sense, this sample is also sometimes referred to as a *targeted* sample.

Purposive samples are used very often in qualitative research. Flood used ethnographic methods to study the workings at a large Chicago law firm. In order to overcome issues with client–attorney privileges, the firm hired him on a temporary basis. Flood thus sat in on meetings, had access to files. Earlier, he had similarly done ethnographic fieldwork with clerks, working with them and participating, and observing their work. Unaccustomed to heavy drinking, he also

felt he had to join in as the clerks went to drink (heavily) in pubs, as he felt he had to participate in order to be one of the bunch, to hear gossip and stories "despite the consequences of my own physical collapse. (...) I had to learn to drink and behave in ways that were unfamiliar to me" (2005, p. 44).

At times, the purposive sample is one (N = 1), such as one trade union, or one law firm. In other instances, qualitative researchers speak repeatedly with several persons. Who should be selected for inclusion in the purposive sample then, and when is a purposive sample 'sufficient' or of sufficient size? In general one seeks for a purposive sample persons who are willing to talk, accessible, who have a good overview of what one wants to study either themselves or who give that overview in complementarity with the other sample members. That raises the issue of how one can be sure that the persons interviewed provide that overview. To decide on that, only qualitative criteria are available.

Rubin and Rubin (1995) state that generally in qualitative research two criteria should be used. The first criterion is whether the set of interviews or observations give a complete overview of the phenomenon under study. As a researcher, obviously one cannot be sure about that, so one has to use common sense, reasoning, and the face validity of the data to decide on this. The second criterion is 'saturation', or 'theoretical saturation'. What is meant by that is probably best illustrated with the following. When conducting qualitative research using a purposive (or any) sample, there will usually come a point where new respondents do not add new information. Each new informant tells you as researcher things you already learned from others, reiterates what earlier interviewees said, and as such at most confirms what you already knew. This is what is meant by saturation: the picture becomes clear at a certain point, and adding new respondents does not add to your knowledge. This is also sometimes referred to as *theoretical validity*. This type of sampling up to saturation is also referred to as *theoretical sampling*.

To enhance the significance of the findings, and to prevent 'cherry-picking', some researchers advise even purposely selecting cases that may deviate from the preliminary conclusions of the research, a practice that is called 'negative case analysis' or 'deviant case analysis'. This shows that sampling in qualitative designs is of a different nature than in quantitative designs: adding such a 'cherry-picked' negative case to the sample might even be considered as violating all rules of randomness! We return to this topic in chapter 7.

This shows that while in probabilistic samples a researcher can from the outset in a sense 'compute' the required sample size (for instance 30 clusters $\times$ 7 respondents, totalling 210 respondents), such advance planning is less feasible or perhaps even simply impossible when carrying out qualitative research with a purposive sample. It is harder to assess ahead of time how many interviews one needs: one simply does not know beforehand at what point respondents will not give new information. As such, planning is also harder in qualitative research. This shows also how qualitative research is steered by the empirical reality one encounters: it may be that the picture is clear after 20 interviews, but it may equally take 30 or 40 interviews. It should be noted that sampling in qualitative

studies is often constrained by practical matters, such as the amount of time that is available for the study or the amount of money. Given that qualitative data are generally quite an investment in terms of ordering the data for analysis, large samples may also become too big for a researcher to have oversight.

A purposive sample is in some cases the only solution for some types of research questions. If respondents are suspicious or anxious, or difficult to access, a solution may be to interview knowledgeable informants: *key-informant sampling*. Key informant sampling can be very efficient: interviewing just a few gives an overview over many. However, the view of these few key informants may be biased, and second, one cannot be sure that the key informants actually do know a lot about the phenomenon or about the inaccessible others – the researcher can never check anyway...

Augusteijn, Bijleveld, & Pemberton (2022) interviewed professionals to investigate whether and if so in what situations compliance with victims' rights is lacking in Dutch criminal proceedings. The authors analysed interviews conducted with a total of 26 professionals. All were either employed in the criminal justice chain, with the police, the prosecution office, at the courts, with Victim Support Netherlands, the Violent Offences Compensation Fund, or Restorative Justice Netherlands, or outside of the criminal justice chain working for either the Victim Support Fund, as a victim support lawyer or personal injury lawyer, or as a scientific researcher. The interviewed professionals were selected in a stratified manner, so that for each relevant part of the justice chain, one or more professionals could be consulted, as well as for each relevant and knowledgeable party outside of that chain. The interviewed professionals were also sampled through 'snowballing' (see next section), that is, through referral by earlier interviewees.

The authors state that their findings – even though they attempted to arrive at a representative sample – should be interpreted with caution. Although they tried to paint as balanced a picture as possible through stratified sampling, it cannot be ruled out that some selectivity occurred in the selection of respondents, and that, for example, more 'critical' persons responded to their request for an interview, or especially more critical persons were referred to them. As a result, it may be that their results are biased towards revealing situations where compliance is lacking.

Findings from a purposive sample are – by definition – not generalizable. In many cases of qualitative research, that does not constitute a limitation. Firstly, one may be interested in one particular group of persons. Another way of formulating this is by saying that if one uses a purposive sample, one has actually studied the entire population of interest, i.e. the one particular youth group, the one particular set of professionals. Secondly, in qualitative research the aim is mostly not to generalize simple facts, such as prevalences or correlations, to additional population members, but to interpret the phenomena that have been studied. That interpretation gives the framework to understand the world around us.

(8) *Snowball sample*

The snowball sample is an important type of non-probabilistic sample. It is often used for investigating inaccessible populations, or groups who are extremely shy or mistrusting. Examples are drug addicts, people who work in the illegal labour market, prostitutes or – indeed – criminals. It is also sometimes referred to as *respondent-driven sampling* (RDS), although respondent-driven sampling is actually a special kind of snowball sampling.

When drawing a snowball sample, one starts with a certain number of accessible respondents (also sometimes called 'starting-respondents' or 'zero-stage respondents'). After interviewing each respondent, one asks the respondent whether s/he knows one or more other respondents who meet eligibility criteria and who might also be interested in being interviewed. The starting respondents are then asked to encourage these respondents to get into contact with the interviewer; sometimes a small fee is offered. When these 'stage-two' respondents have been interviewed the same procedure is repeated, and so forth. When the respondents are extremely shy, the zero-stage respondents themselves may also be asked to conduct the interviews with the stage-two respondents. The snowball sample indeed functions as a snowball, onto which more and more clings with each turn.

Snowball sampling is a workable and relatively cheap way to study groups of persons who would be unsurveyable otherwise, that is, members of a so-called *hidden population*. It is a powerful method in the sense that with a small number of steps – in general – good coverage over the entire population is reached. Heckatorn (1997) has shown that in the USA – under certain conditions – six steps or waves are sufficient.

A disadvantage of snowball sampling is that it may be lengthy: it usually takes a while before respondents have traced new respondents, after which it takes a while before these contact the researcher, etc. In addition, members of hidden populations may not lead the most mundane organized lives and for that reason not be the most punctual.

A second disadvantage of snowball sampling is that its success is dependent on the choice of 'zero-stage respondent'. If this person has only a small circle of acquaintances, or does not instil a lot of trust, the trail will quickly run cold. For that reason Wright et al. (2001) chose for their study of burglars as a zero-stage respondent someone who was a burglar himself, that is, someone absolutely unassociated with the criminal justice system and who was most likely to generate trust in the next wave of respondents.

In snowball sampling it is assumed that all members of the population are sufficiently connected to be able to reach each population member eventually. If that is not case, for instance if there are isolated subgroups (in a sense on social 'islands'), then those on other 'islands' will not be reached by snowball sampling. The method does not succeed in getting the zero- and subsequent stage respondents to leave their islands. To minimize that risk, it is generally advised to use various zero-stage respondents, preferably from different segments of the population. As said, a precondition for any kind of snowball sampling is that

respondents know each other. For that matter, drug addicts are a good candidate for snowball sampling as they are likely to know each other from drug dealers, methadone buses, etc. Persons committing social security fraud on the other hand are less or not likely to know each other.

A variation of snowball sampling is so-called *respondent-driven sampling*, in which a respondent receives coupons that s/he can distribute among prospective new respondents. As soon as a successful interview with a new respondent has been completed, both the old and the new respondent receive some kind of bonus. The advantage is that both the old and the new respondent have an incentive for interviews actually to be completed: only then is the remuneration awarded. The second advantage is that the identity of the new respondent remains hidden until the moment the actual interviewing takes place.

Snowball sampling is an accepted method of sampling. It produces – with sufficient numbers of waves – good coverage of the population. However, persons with a large social network have a larger chance of ending up in the sample, and lonely or isolated persons a smaller chance. This means that snowball samples have a certain bias towards 'more sociable' respondents. Secondly, respondents after the zero-stage are likely to resemble those who entered them into the sample, as friends and acquaintances tend to be like each other too. This means that we will also have cluster effects in a snowball sample. In a statistical sense this implies that the estimates of the pertinent properties we are interested in will also be in a sense noisy.

## 3.4 Bias and noise

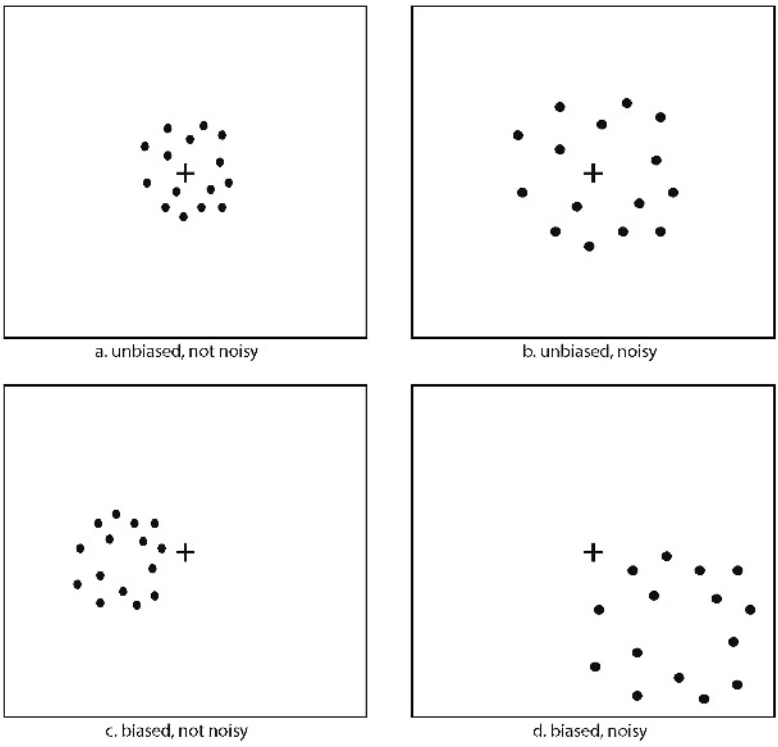
In the previous paragraphs we have at several places encountered the terms 'bias' and 'noise'. What is meant by these terms?

Roughly speaking, noise reflects variability. If we are trying to assess citizens' trust in the judiciary, and if there is a lot of variation around the population mean in the samples that we can draw (for instance because we worked with cluster samples), then we have a *noisy* sample. The sample members' averages vary a lot around the true population score that we are interested in. On average, though, they may provide a good estimate of the population mean.

Suppose however that sample means do not reflect the population mean: sample means are consistently lower than the population means, or consistently higher. Then our sample means are *biased*. The sample averages then differ systematically from the population mean: they are most of the time lower, or most of the time higher. Even if we are not bothered by noise, the sample mean is not very useful: it does not give us a valid estimate of the population mean. This means that a biased sample is not a representative sample: it does not reflect the properties of the population.

Does that mean that a noisy sample is therefore always better than a biased sample? Yes and no. A biased sample is not very useful as it does not tell us the properties of the population. A sample that is only noisy gives at least (if we were to draw a large number of samples) on average an unbiased estimate of the population mean. A noisy

**Figure 3.7:** Visual representations of bias and noise




sample however generates problems of its own: in practice we do not draw a large number of samples and repeat our analysis many times. We draw one sample and report our findings. So if we have a noisy sample, our estimate may still be very far 'off'. We will discuss these and related issues much more formally in chapter 8 below.

For now, it is important to remember that both noise and bias are a nuisance. Samples can be noisy, samples can be biased and obviously samples can be noisy as well as biased. Ideally we would want samples to be both unbiased and not noisy. See Figure 3.7.

Figure 3.7 is a visual aid to understand the concepts of bias and noise; each picture should be viewed as a bull's eye that we try to hit with a pistol or with darts. The bull's eye represents the true population mean that we want to estimate, and the holes made by the bullets or darts represent the sample measurements that should aid in arriving at an estimate of that population mean. In this figure, the upper left situation (a) is one where there is little bias and little noise: the cross '+' indicates where the bull's eye/population mean is, and we can see that the hits/sample measurements, indicated as dots, cluster closely around the population mean and would on average equal the

population mean. In the upper right situation (b) there is more noise: the measurements spread much more around the population mean but, on average, the measurements tell us what the population mean is. In the lower left situation (c), there is bias: the average of the sample measurements is clearly 'off'. The measurements deviate systematically from the population value. In the lower right situation (d), finally, there is bias as well as noise: the measurements are systematically to the right of where they should be, and they also vary a lot.

## 3.5 Sample size

So far, we have talked of samples and populations. As we said, in many cases the population is simply too large to investigate member by member, and we therefore have to make do with samples. We therefore 'draw' a sample, a group of members of the population, ideally at random, and what we find in the sample we declare to be generalizable to the population. Obviously, samples are smaller than the population. But how small may a sample be, or how large exactly should a sample be?

Having said that, it is not the case that sample size is always determined by strict iron-fist rules. In practice, a lot also depends on the amount of money and time available, and on the type of question asked. If we conduct qualitative research (see chapter 2, earlier in this chapter and chapter 7), sample sizes depend more on 'saturation' and 'coverage'. Qualitative researchers can decide while their research is ongoing whether they do or do not need additional sample members to study. Qualitative samples vary in size between 1 and a few hundred (and seldom more). Also because qualitative research is generally much more time-intensive, large sample sizes are extremely rare. Sample sizes of 20 to 30 are quite common. Quantitative researchers often work with samples of between 100 to 300, to in some cases 500, or even more respondents. There are various statistical issues that may aid in determining the desired sample size for investigating a certain research question with a certain precision (we will return to such issues in chapter 8). A few basic guidelines may help in understanding issues of sample size.

Firstly, very generally speaking, the larger the sample – assuming it was drawn randomly – the more the sample can tell us about the population. This is logical. Suppose that a sample constitutes about 1% of the population. Anything we measure about that sample, say its mean, may fluctuate quite a bit depending on the sample we accidentally happen to draw, and – although this is not very likely – any statistic, such as the mean, can theoretically even be quite far off the mark. Suppose now however that the sample size increases, to for instance 10% of the population. Again, theoretically speaking, it is still possible that the mean we find for the sample is quite different from the mean in the population, but as we are measuring so many more people, it can only be closer to the population mean. Next, suppose that our sample contains 50% (or even 60%, 70% – or even 90% of the population): the sample mean now approaches the population mean more and more as we are examining larger and larger chunks of the population! It is at 50% highly improbable that the sample mean would be very far off the real population mean, and it becomes almost impossible when we have observed 80% or 90% of respondents. Thus: the larger the sample, the better that sample helps us

to conclude something about the population. The larger the sample, the more precisely will our inferences about the population be. In terms of noise: larger samples are in principle less noisy.

This is of course intuitively true: if we investigate more members of the population, we know more about the population, and then we are able to make more precise statements about that population. Adding more sample members adds more information and can thus only improve upon what we do. This however does not help us much in practical situations. Because – while larger is better – exactly how large is then large enough? As mentioned above, time and money often dictate sample size. This is so for qualitative and for quantitative research.

Even though we thus cannot be very precise as to sample size yet, a second general rule is that the number of so-called 'disaggregations' one would like to make also determines the required size. What is meant by that?

Suppose that one is interested to know the contacts that people have had with the courts, as a litigant, plaintiff, defendant, accused or victim or any other mode. One could survey 1,000 persons, and find that 40 persons have had at least one contact the past year. But now suppose that one would want to disaggregate this by men and women, and not only by men and women, but also by age – investigating young men and young women and older men and older women. Now, assuming age and gender are distributed evenly, this would mean that there would be 10 young men, 10 older men, 10 young women, etc. Suppose now that I would also be interested to know the differences between persons with low and high levels of education. Then, the groups would be divided up once more, into groups of 5 persons – generally referred to as 'cells'. And suppose now that I also have reasons to believe that urbanity plays a role: then my groups would dwindle even more. Clearly any differences between the groups now become shaky: what if I find that 2 out of 5 young highly educated women have had a court contact the past year, against 4 out of 5 young highly educated men? How confidently can I report that this chance is twice as high from one group to the other? Suppose I have encountered only 1 young highly educated woman with a court contact the past year: the chance would then have increased fourfold. Within such small groups, one man or one woman extra or less can greatly affect the percentages, so that we may arrive at sweepingly different conclusions, depending on just one or two sample members. The findings are *unstable*.

Clearly, here the fact that one wants to compare so many groups makes it necessary to start out with a larger sample: one would not want to make too strong statements based on just one or two sample members having had a contact with the courts or not. In general, as a rough rule of thumb it is assumed that the number of respondents per cell, i.e. per group one wishes to compare, should not be less than 30. This means that if one wants to investigate just two cells, 60 respondents would be an absolute minimum; for comparing four groups, 120 respondents would be needed, and if one wants to make as many subdivisions that one would be comparing 16 different cells, 480 sample members would be needed.

Thus, summarizing this section, we note that for qualitative studies, sample size is more often guided by 'information gain' or saturation, meaning that sample size will be determined as the research proceeds. Also for practical reasons, sample sizes tend to be smaller in qualitative than in quantitative research. For various viewpoints

on desirable sample sizes in qualitative research, see Baker & Edwards (2012). For quantitative studies, bigger samples give a more precise estimate of the properties of the population. For statistical reasons we prefer to work with sample sizes from 100 and up, but this is a rule of thumb. We will return to this issue in chapter 8. Last, we note that if we want to investigate in more depth and want to compare several subdivisions of the sample, for reasons of stability we need a larger sample. This applies to both qualitative and quantitative studies.

## 3.6 Sample nonresponse and representativeness

Above we said that samples are a part of a population. Instead of investigating all population members, we investigate only part of the population. In choosing that part, our sample, wisely, we hope to have the sample resemble the population. Drawing random samples guarantees representativeness. Obviously, there is uncertainty as we have not investigated each and every sample member, but as we stated, statistics will help us deal with that uncertainty in a sensible and accepted way. Thus, the path to quantitative research seems not too hard: we simply draw a random sample from a population, interview or otherwise assess the sample members, do some statistical wizardry, and we can tell our captivated readers what's likely going on in the population!

Life is unfortunately not that uncomplicated. Suppose that in some quantitative study, we have been able to compile a list of sample members, drawn at random from the population. Therefore we have a random sample and we assume we can generalize the sample results to the population. We start approaching sample members, ringing their doorbells or telephoning them, asking them to participate. Two complications will quickly arise. First, we will be unable to contact all respondents: some will simply never be at home or never answer their phone. Secondly, not all those we establish contact with will consent to participate. Some will say that they are too busy, are not interested, do not like the funding agency for the research. Some may simply be too ill to collaborate. These two mechanisms generate what is called sample *nonresponse* or sample *attrition*.

In most practical situations, sample nonresponse occurs. Depending on the topic of the study, the infrastructural possibilities, the persuasive skills of interviewers and the like, response rates in surveys these days generally hover between 20% and 40%; higher response rates are rare. This means that if we start out aiming for a sample of 100 respondents, we may end up with just 40 completed interviews: a so-called *retention rate* of 40%, and a so-called *attrition rate* of 60%. One might be tempted to think that this is not a real problem, as we could simply draw a larger initial sample of say 250, and then end up with 100 interviewed respondents, reaching our target.

Unfortunately, this does not solve the problem that nonresponse generates. The problem is namely not simply that we are left with fewer respondents, the problem is that nonresponse is generally not accidental. It is not a coincidence that certain respondents do not end up in our realized, interviewed sample. It is the vulnerable and the elderly who are too ill to be interviewed, it is the mistrustful, those who are afraid to talk to strangers or the busy bees with 80-hour work weeks who refuse to talk to us. The final realized sample thus generally contains an overrepresentation of younger and

elderly persons, of persons from less urbanized areas, and an underrepresentation of the elderly, of townspeople, of people employed in very busy management positions. In addition, those who refuse to be interviewed generally have more extreme perceptions of the topic under study. Thus it is likely the fervent religious opponents of abortion as well as those who believe that abortion is simply anyone's right and nobody else's business, who will be underrepresented in our survey of citizens' views on abortion.

All in all, even if we start out with a randomly drawn list of sample members, the non-random attrition process will make us end up with a non-random selection of the original random sample. Formulated more loosely: nonresponse messes up the representativeness of a sample. One might be tempted to think that this is a particularly problematic phenomenon when doing surveys with 'live' people to be interviewed, who can be ill and who may say 'no'. Attrition however also plays a role when studying for instance court files or treatment dossiers. Court files of defendants who have their case up for review are typically 'travelling' and not to be found in the archive, the treatment files of recidivists may have been requested for inspection by the investigating psychiatrist or psychologist. Dossiers of withdrawn claims are cleaned earlier than those of cases that were taken to court. Thus, also here, it is the particular, atypical files that will be missed and a non-representative part of the original sample that will be left for inspection by us.

Nonresponse is essentially irreparable. One can inspect the resulting sample, in a so-called nonresponse analysis, and hope that it resembles the population on background characteristics (if one knows them) such as age, gender, type of claim, geographical origin and the like. If there are no serious differences, that is, if the realized sample resembles the population on such background characteristics, then that is more comforting than if we found there were differences. This 'background variables check' however does not tell you whether the nonresponders differ from the responders on the key variables of interest, central to the main research question.

If there are differences between the realized sample and the population, an option would be to 're-weight' the data. What is meant by that is best illustrated by an example. Suppose that it turns out that our realized sample has fewer females than there are in the population. We could then (assuming that the females who did respond are representative for all females – something which is unlikely) give the females' answers more weight than the males' answers, by counting their answers double or something else proportional to their underrepresentation. This solves the problem only if we have too few respondents of one kind. If we miss one kind of respondent altogether (for instance, we were unable to reach any pensioners), there is no way we could correct for this by re-weighting, as we do not have pensioners at all... There is then no way we can know what kind of answers to multiply and add to the existing ones.

Re-weighting has disadvantages as well, however. The easiest one to imagine is that findings become unstable, as they become heavily dependent on the few respondents of the underrepresented category who have been assessed. Suppose that females are underrepresented and that we have only one female respondent, and 50 males. In that case, we would in a sense have to 'multiply' this lone female's answers by 50 and add them to the dataset. The single female who is in our dataset now determines what all cloned females in our dataset report. Had we by chance interviewed another female, the 'female response' could have been wildly different.

Nonresponse rates vary per topic that is studied and per type of study objects (paper or electronic sample members such as case files generally do not generate high nonresponse rates). Nonresponse rates can be so high that generalization to the population becomes increasingly unrealistic. We already mentioned the study by Dijksterhuis & Vels (2011) where the response rate among clients who had defaulted on child maintenance payments was 22%, and the nonresponse rate therefore 78%. Especially when the topic is sensitive, response rates as low as 2% have been encountered. Response rates of 40% to 50% are generally perceived as acceptable, even though then one should always check to what extent the nonresponders differ from the responders.

It should be noted that in so-called longitudinal studies where respondents are followed over time, any so-called *drop-out* that occurs accumulates over waves. So, if we miss, say, 20% of our intended sample at the first measurement wave, meaning we retain a really good 80%, experiencing the same drop-out at wave 2 implies that we will then be left with 64% of our original respondents. Should such *attrition* occur similarly at wave 3, we would then be left with just under 50% of our original sample, and so forth.

By far the best is to prevent nonresponse. This is achieved by careful introduction of the study, good training of interviewers, revisiting of respondents, offering respondents a monetary incentive, and more, which we will detail in chapter 4 below.

# Chapter questions

1. Argue whether all cases registered with the prosecution constitute a population, whether all witnesses testifying in international criminal proceedings constitute a population, and whether all medical malpractice cases brought before a disciplinary council do so (section 3.1)

2. What is a sampling frame? Argue whether all tort cases decided upon by the supreme court in a certain country are a representative sample for all tort cases in that country (section 3.2)

3. Order the four types of non-probability samples according to the extent to which you might regard them as being representative for the population from which they were sampled (section 3.3)

4. What is meant by saturation in the context of qualitative sampling? (section 3.3.2)

5. What is negative or deviant case analysis? (section 3.3.2)

6. Order the four types of probability samples according to the extent to which they would generate biased or noisy estimates of properties of the population (section 3.3 and section 3.4)

7. Give two reasons why nonresponse constitutes a problem in empirical research (section 3.6)