Quantitative Methods
Module 5: Sampling

5.01 Sampling: External validity threats

External validity, or generalizability, refers to whether the hypothesized relation holds for other persons, settings and times. Just like with internal validity, there are a number of threats to external validity.

A **history** threat means that the observed effect doesn't generalize to other time periods. Consider a compliance study performed in the nineteen fifties in the US. Results showed that participants were willing to comply with highly unethical directions provided by an authoritarian experimenter. These results would probably be less extreme if we repeated the study nowadays, for example because people are more highly educated and less sensitive to authority than in the nineteen fifties.

A **setting** threat to external validity means that the observed effect only holds in a specific setting. In other words, the findings do not generalize to other environments or situations. Suppose we investigate the relation between violent imagery and aggression and find that children who watch a violent video are more aggressive afterwards in the school playground. A setting threat occurs if this effect depends on the surroundings, for example if children are *not* more aggressive when they play at home under their caregiver's supervision.

There are two setting threats associated with the ***artificiality*** of the *research* setting specifically. These threats are **pretesting** and ***reactivity***. A ***pretesting*** threat means that the observed effect is found only when a pretest is performed. This threat is closely related to the ***internal validity threat*** of ***testing***.
Say we investigate a new therapy for treating depression and use a pretest. Suppose the depression pretest makes participants realize how serious their problems are, and thereby makes them more receptive to the treatment. The treatment is effective, but only if receptiveness is increased by the pretest first. In this case internal validity is threatened because 'receptiveness' is missing from our hypothesis. External validity is also threatened, because the hypothesis will only apply to situations where a pretest is part of the setting.

The second artificiality threat is **reactivity**. A reactivity threat occurs when the participants or experimenter react to the fact that they are participating in a research study. Reactivity includes participant and experimenter expectancy and altered participant behavior, for example due to nervousness. This can cause the hypothesized relation to occur only in a research setting and not in a in a natural setting. Say we investigate a new method for teaching high school math. The researcher is present during the lessons and measures math performance in class. What if students work harder because they know they are being studied and this makes

the new method more effective? In a natural setting, without the researcher present, students might put less effort in their schoolwork, reducing the effectiveness of the new method.

Selection is a final and very important threat to external validity. A selection threat occurs when the hypothesized relation only holds for a specific subset of people or if the results in our study are biased due to over- or underrepresentation of a certain subset.
Suppose that in our study on a new depression therapy, we recruited participants who actively volunteered. Say we find that the therapy method is effective. Of course our volunteers might be more proactive about solving their problems than the average person. It is entirely possible that the method is *in*effective for people who are less proactive. The overrepresentation of volunteers might lead to an overestimation of the therapy's effectiveness.
Another example: Suppose we want to know people's opinion on women's right to vote and we interview people on a university campus. The sample is now so selective that it is highly unlikely that results will generalize to the general public's opinion.

What can we do about these threats to external validity? Well history and setting threats to external validity can be reduced by replicating a study in a different time or by repeating a study in different settings. In the case of threats related to the artificiality of the *research* setting specifically, this means repeating a study in a more natural environment. Replication can also reduce the threat of *selection* to external validity, in this case by repeating a study with different groups of subjects. But there is another way to reduce the threat of selection. I'm referring to **random sampling** of the research sample, also referred to as **probability sampling**.

5.02 Sampling: Sampling concepts

Some sampling methods offer better protection against the selection threat to external validity than others do. To understand why, you first need to be familiar with some basic sampling concepts. The two most important concepts are ***population*** and ***sample***.

The selection threat concerns the generalization of findings to other persons. Exactly what other persons are we referring to? All people in the entire world, or just people in our country or culture? Researchers should anticipate this question by defining their target ***population*** explicitly. The term **population** refers to the entire collection of people or groups to whom the hypothesis is supposed to apply.

Let's look at two examples of populations. Consider the hypothesis "Loneliness causes an increase in depression". This is a typical universalistic hypothesis. If the population is not explicitly mentioned, we infer the relation is assumed to hold for all people, in all cultures, in the past, now and in the future.

Another example: "Patriotism is steadily declining in the Netherlands over the last five years". This is at typical particularistic hypothesis. It is clear that this hypothesis applies to a specific country and to a specific time.

Let's assume for a minute that the target population for a hypothesis is clearly defined. How can we determine if the results generalize to this entire population? Well if we measure the entire population, then we're automatically sure the results hold for the entire population, everybody was measured.
Of course for universal hypotheses it's simply impossible to measure the entire population, because the population consists of all people everywhere, including all people who are long dead and all people who have yet to be born.
Even if the target population is smaller and well defined, it is almost always too complicated and too expensive to include the entire population in a study. This is why we take a **sample**: a subset of the population. The sample is used to represent or estimate a property of the population.

Of course it's possible that this sample does not represent the population accurately. Suppose we sample mostly elderly people in our study on the effect of loneliness on depression and we find a strong effect. The overrepresentation of a specific part of the population can weaken the study's external validity. Perhaps the strong effect of loneliness on depression is less apparent for young people. If our sample had been more representative of the entire population we would have found a smaller effect.
The same goes for our study of decreased patriotism. Suppose our sample consisted mainly of highly educated people working at a university. This might lead us to underestimate patriotic attitudes in the Netherlands. Our results will be biased.

We will consider different sampling methods and see how they deal with the selection threat to external validity. But before we can do so, there are some terms you need to become familiar with.

An **element**, or unit, is a single entity in the population. Together all the elements form the population. An element most often consists of one person, but of course it depends on your hypothesis. An element can also be a group, a school, a city, a union, a country; you name it.

A **stratum** is a subset of elements from the population that share a characteristic. In the population of currently enrolled students from the University of Amsterdam we can distinguish a female and a male stratum, for example. Of course we can identify many different strata that may overlap, for example male and female undergraduate and graduate students.

The term **census** refers to an enumeration or *count* of all elements in the population. The term can also refer to a situation where all elements in the population are actually measured; in that case, the sample consists of the entire population. The term census can indicate a 'national census': A nation-wide survey where demographic information on each inhabitant is collected. Of course in many western countries this census is

conducted virtually, by collecting information from government databases.

A final term that you need to be familiar with is the term **sampling frame**. A sampling frame is essentially a list of all the elements in a population that can be individually identified. A sampling frame can overlap with a census defined as an enumeration of the population. A sampling frame is more than a simple list of elements however.

A sampling frame provides a way of actually contacting elements. It could be a phonebook or a list of email addresses for all students currently enrolled at the University of Amsterdam, for example. Also, a sampling frame doesn't always include all elements of a population. This could be due to clerical errors or an outdated list. Ok, you now know the basic concepts necessary to learn about different sampling methods.

5.03 Sampling: Probability sampling

Probability sampling minimizes the selection threat to external validity. Before I discuss different types of probability sampling, let's consider the essential feature of probability sampling and how this feature helps to minimize the risk of systematic bias in our selection of participants.

The essential feature of probability sampling is that for each element in the sampling frame, the probability of being included in the sample is known and non-zero. In other words, some form of random selection is required where any element could in principle end up in the sample. To use probability sampling, we need to have a sampling frame: a list of all elements in the population that can be accessed or contacted. A sampling frame is necessary to determine each element's probability of being selected.

Now let's see why probability sampling minimizes the threat of a systematic bias in our selection of participants. Reducing systematic bias means reducing the risk of over- or underrepresentation of any population subgroup with a systematically higher or lower value on the property. Otherwise, our sample value will be unlikely to represent the population value accurately.

We've already seen a method to eliminate systematic bias in participant characteristics. Remember how we eliminated the ***selection threat to internal validity***? We used ***random assignment*** to get rid of systematic differences between the experimental and control condition. In the long run any specific participant characteristic will be divided equally over the two groups. This means that any characteristic associated with a systematically higher or lower score on the dependent variable cannot bias the results in the long run.

The same principle can be applied, not in the *assignment*, but in the **selection** of participants. We avoid a systematic difference between the sample and the population, by randomly selecting elements from the population. In the long run any specific participant characteristics will be represented in the sample, proportionally to their presence in the population. We call this a **representative sample**.
Suppose a population consists of eighty percent women. With repeated random sampling, we can expect the sample to contain eighty percent women in the long run. Each individual element has the same probability to be selected, and since there are more women, female elements will be selected more often.

Besides resulting in a ***representative sample*** in the long run, probability sample has another advantage. Probability sampling allows us to assess the **accuracy of our sample estimate**. Probability sampling allows us to determine, that with repeated sampling, in a certain percent of the samples, the sample value will differ from the real, population value by no more than a certain **margin of error**.

This sounds - and is - complicated. But it basically means that we can judge how accurate our sample estimate is in the long run. Given a certain risk to get it wrong, we can assess what the margin of error is on average, meaning by how much the sample and population value will differ on average.

Consider an election between conservative candidate A and democratic candidate B. We want to estimate the proportion of people in the population that will vote for candidate A as accurately as possible. Random sampling allows us to make statement such as this: If we were to sample voters repeatedly, then in ninety percent of the samples, the true, population proportion of votes for A would lie within eight percentage points of our sample estimate.
So if we find that sixty percent of our sample indicates they will vote for A, then we can say we are fairly confident that the true proportion will lie somewhere between fifty-two and sixty-eight percent. This interval is called a **confidence interval**. Of course this particular interval could be wrong, because in ten percent of the samples the sample value will lie further that eight percentage points from the true value. This could be one of those samples, so we can never be certain.

5.04 Sampling: Probability sampling - simple

There are several types of probability sampling. In this video I'll discuss the two simplest types: ***simple random sampling*** and ***systematic sampling***.

The most basic form of probability sampling is **simple random sampling**. In simple random sampling each element in the sampling frame has an equal and independent probability of being included in the sample. Independent means the selection of any single element does not depend on another element being selected first. In other words,

every possible combination of elements is equally likely to be sampled.

To obtain a simple random sample, we could write every unique combination of sampled elements on a separate card, shuffling the cards and then blindly drawing one card. Of course, if the population is large, then writing out all possible combinations is just too much work.
Fortunately, an equivalent method is to randomly select individual elements. This can be done using random number tables, still found in the back of some statistics books. But these tables have become obsolete; we can now generate random number sequences with a computer. For example, if our population consists of twelve million registered taxpayers, then we can generate a sequence of two hundred unique random numbers between one and twelve million.

Systematic sampling is a related method, aimed to obtain a random sample. In systematic sampling only the first element is selected using a random number, the other elements are selected by systematically skipping a certain number of elements.
Suppose we want to sample the quality of cat food on an assembly line. A random number gives us a starting point: say the seventh bag. We then sample each tenth bag, so we select bag number seven, seventeen, twenty-seven, etcetera. It would be much harder to select elements according to random numbers, say bag number seven, thirty, thirty-six, forty-one etcetera., especially if the assembly line moves very fast.
With this approach, each element has an equal probability of being selected, but the probabilities are not independent. Elements seventeen, twenty-seven, thirty-seven etcetera, are only chosen if seven is chosen as a starting point. This is not a real problem; it just requires a little more statistical work to determine things like the margin of error.

The real problem with systematic sampling is that it only results in truly random sample if there is absolutely no pattern in the list of elements. What if the assembly line alternately produces cat food made with fish and cat food made with beef? Let's say all odd-numbered elements are made with fish. In our example we would never sample the quality of cat food made with beef!
Of course this is an exaggerated example, but it illustrates that systematic sampling can be dangerous. A pre-existing list or ordering of elements can always contain a pattern that we are unaware of, resulting in a biased sample.

So systematic sampling only results in a truly random sample if it is absolutely certain that the list of elements is ordered randomly. We can make sure of this by randomly reordering the entire list. We could generate a sequence of random numbers of the same size as the list and then select elements from this list using systematic sampling.
Of course this is equivalent to random selection directly from the original list using random numbers. Unless we can be sure that the list is truly random, systematic sampling should *not* be considered a form of probability sampling; instead it should be considered a form of non-probability sampling.

5.05 Sampling: Probability sampling - complex

There are many sophisticated probability-sampling methods. I'll discuss two methods that go beyond the basic idea of random sampling, but are still relatively simple. These are **stratified random sampling** and **multi-stage cluster sampling**.

In **stratified random sampling** we divide the population into mutually exclusive strata. We sample from each stratum separately using ***simple random sampling***. The separately sampled elements are added together to form the final sample. Stratified random sampling is useful for two reasons.

First, it allows us to ensure that at least in terms of the sampled strata, our sample is ***representative***. This means subpopulations are represented in the sample in exactly the same proportion as in the population. With simple random sampling we can expect the sample to be representative in the long run, but due to chance, in any particular sample, strata might be over- or underrepresented.
Second, stratification is useful because it can make sampling more efficient. This means, all other things being equal, that we achieve a smaller margin of error with the same sample size. Stratifying only increases efficiency if the strata differ strongly from each other, relative to the differences within each stratum.

Imagine we want to sample the quality of cat food produced on an assembly line. The line produces cat food made with fish and cat food made with beef. Suppose the average quality of beef cat food is higher than that of fish cat food. Also, the quality varies relatively little when we consider each type of food separately. Under these circumstances we will obtain a more accurate estimate of the population's average food quality if we stratify on food type.
This is because quality is related to food type; even a small overrepresentation of one food type can distort our overall estimate of food quality. Stratifying prevents this distortion. If the quality does not differ between food types, then overrepresentation of one food type will not distort the overall estimate and stratification will not improve efficiency.

It is important to realize that stratified sampling requires that we know which stratum each element belongs to. If we can *identify* strata, then we also know their size. As a consequence, the size of our subsamples does not have to correspond to the size of the strata. We can calculate a representative estimate by weighing the subsamples according to stratum size.
Why would we do this? Well, suppose our stratum of fish cat food is relatively small, or is known to strongly vary in quality. In both cases our estimate of the quality of fish cat food might be much less likely to be accurate than that of beef cat food. It might be worth it to take a bigger sample of fish cat food, so we have a better chance of getting an accurate estimate. Of course this means over-representing fish cat food.

We can correct for this overrepresentation by weighing the separate estimates of fish and beef cat food according to their stratum sizes before

averaging them into an overall estimate of food quality. This way the sample value is representative, efficient and more likely to be accurate.

Let's turn to multi-stage cluster sampling, the final type of random sampling I want to discuss. **Multi-stage cluster sampling** allows us to use random sampling without going bankrupt. Consider sampling frames that consist of all inhabitants, students, or eligible voters in a certain country. If we were to randomly select elements from these frames we would have to travel all over the country. In most cases this is just too expensive.

A solution is to randomly sample in stages, by first selecting clusters of elements. Say we want to sample math performance in the population of all Dutch students currently in their third year of secondary education. We start by forming a sampling frame of all school districts; this is the first stage, where students are clustered in districts. We randomly select a very small sample of school districts. We can use stratification to make sure we include districts in urban and rural areas.
In the second stage we randomly select schools from the previously selected districts. Students are now clustered in schools. In the third stage third year math classes are randomly sampled from the previously selected schools. We could even include a fourth stage where students are randomly sampled from the previously selected classes. Stratification can be used in all of these stages.

Multi-stage cluster sampling makes random sampling feasible. But the margin of error is harder to determine, because the probability to be included in the sample is no longer the same for all elements, like it was with simple random sampling. Also, cluster sampling is usually associated with a larger margin of error, even if stratified sampling is used to increase efficiency. However, these disadvantages are generally more than outweighed by the reduction in cost and effort.

5.06 Sampling: Non-probability sampling

Probability sampling can be contrasted with **non-probability sampling**. In non-probability sampling some elements in the sampling frame either have zero probability to be selected or their probability is unknown. As a consequence, we cannot accurately determine the margin of error. It's also impossible to determine the likelihood that a sample is representative of the population.
There are several types of non-probability sampling. I'll discuss the four most common types: **convenience sampling**, **snowball sampling**, **purposive sampling** and **quota sampling**.

Convenience sampling, or ***accidental sampling***, is the simplest form of non-probability sampling. In convenience sampling, elements are selected that are the most convenient, the most easily accessible. For example, if I'm interested in investigating the effectiveness of online lectures on study performance, I could recruit students in courses that I teach myself. Of course this a highly selective

sample of students from a particular university in a particular bachelor program. Results will almost certainly be influenced by specific characteristics of this group and might very well fail to generalize to all university students in my country, let alone students in other countries.

So the risk of bias is high and we have no way to determine how closely the sample value is likely to approach the population value. Even so, convenience samples are used very often. Because sometimes, it's simply impossible to obtain a sampling frame. In other cases, the effort and expense necessary to obtain a sampling frame are just not worth it; for example when a universalistic, causal hypothesis is investigated.

Snowball sampling is a specific type of convenience sampling. In snowball sampling, initially, a small group of participants is recruited. The sample is extended by asking the initial participants to provide contact information for possible new participants. These new participants are also asked to supply contacts. If all participants refer new ones, the initially small sample can grow large very quickly.

Suppose we want to sample patients who suffer from a rare type of cancer. We could approach a patient interest group, for example, and ask the initial participants if they can put us in contact with other patients that they know through other interest groups or through their hospital visits. We continue to ask new participants to refer others to us, until the required sample size is reached.

Snowball sampling is very useful for hard-to-reach, closed-community populations. Of course all disadvantages of convenience sampling also apply to snowball sampling, maybe even more so, because there is the added risk that we are selecting a clique of friends, colleagues or acquaintances. These people could share characteristics that differ systematically from others in the population.

In **purposive sampling,** elements are specifically chosen based on the judgment of the researcher. A purposive sample can consist of elements that are judged to be typical for the population, so that only a few elements are needed to estimate the population value. A purposive sample can consist of only extreme elements, for example, to get an idea of the effectiveness of social workers working with extremely uncooperative problem families.

Elements can also be purposively chosen because they are very much alike, or reversely, very different, for example, to get an idea of the range of values in the population. Or, elements can consist of people who are judged to be experts, for example when research concerns opinions on matters that require special knowledge.
Purposive sampling is used mostly in qualitative research, so I won't go into further details here. Suffice it to say that purposive sampling suffers all the same disadvantages that convenience sampling does. The researcher's judgments can even form an additional source of bias.

Quota sampling is superficially similar to stratified random sampling. Participants in the sample are distinguished according to characteristics, such as

gender, age, ethnicity or educational level. The relative size of each category in the population is obtained from a national statistics institute, for example.
This information is used to calculate how many participants are needed in each category. So that the relative category size in the sample corresponds to the category size in the population. But instead of randomly selecting elements from each stratum, participants for each category are selected using convenience sampling. Elements are sampled until the quotas in all categories are met.

Although this approach might seem to result in a representative sample, all kinds of biases could be present. Suppose the choice of participants is left to an interviewer. Then it's possible that only people who seem friendly and cooperative are selected. If a study uses non-probability sampling, the results should always be interpreted with great caution and generalized only with very great reservation.

5.07 Sampling: Sampling error

The goal of sampling is to estimate a value in the population as accurately as possible. But even if we use the most advanced sampling methods, there will always be some discrepancy between our sample value - the estimate - and the true value in the population. The difference between sample and population value is generally referred to as error. This error can be categorized into two general types ***sampling error*** and ***non-sampling error***. In this video I'll only discuss the first type: sampling error.

It's important to keep in mind that the true value in the population is almost always unknown. If we knew the population value then we wouldn't need a sample. This also means that for any particular sample we cannot assess how large the error is exactly.

However, for ***sampling error*** it is relatively easy to estimate how large the error is. Let's look at sampling error in more detail and see how it works. If we would take an infinite number of samples from a population, then under certain conditions, the average sample value of all these samples will correspond to the population value.
But of course individual samples will result in sample values that are different from the population value. **Sampling error** is the difference between sample and population value that we would expect due to chance. We can estimate how large the sampling error is on average, if we were to repeatedly draw new samples from the same population. Note that this only works for randomly selected samples!

The average error, called the ***standard error***, can be estimated based on the values obtained in a single sample. We can then use the standard error to calculate a **margin of error.** You might think the margin of error tells us by how much our sample differs from the population at most. But we can't calculate between what boundaries the true population value lies exactly, because we are estimating the sampling error in the long run, over repeated samples. In the long run a ridiculously small or large value is always

possible. What we *can* say is that the population value will lie between certain boundaries *most* of the time.

This information is captured in a **confidence interval**. A confidence interval allows us to say that with repeated sampling, in a certain percentage of these samples, the true population value will differ from the sample value by no more than the **margin of error**.

Suppose we want to estimate the proportion of people that will vote for candidate A in an election. We sample one hundred eligible voters and find that sixty percent of the sample says they'll vote for A. We have to decide how confident we want to be. Let's say that with repeated sampling, we want the population value to fall within the margin of error at least ninety percent of the time. With this decision, we can now calculate the margin of error. Let's say that the margin of error is eight percent. This means we can say that with repeated sampling, the population value will differ from the sample value by no more than eight percent, in ninety percent of the samples.

Sampling error is related to the sample size. As sample size increases, sampling error will become smaller. Sampling error is also influenced by the amount of variation in the population. If a population varies widely on the property of interest, then the sample value can also assume very different values. For a given sample size, sampling error will be larger in a population that shows more variation.

Ok, so to summarize: sampling error is the difference between population and sample value due to chance, due to the fact that our sample is a limited, incomplete subset of the population. Sampling error is unsystematic, random error. It is comparable to the random error that makes a measurement instrument less reliable.
We can estimate how large the sampling error will be in the long run, which allows us to conclude how accurate our sample value is likely to be. This only works under certain conditions. One of these conditions is that the sample is a random sample from the population.

5.08 Sampling: Non-sampling error

Sampling error can be contrasted with non-sampling error. ***Sampling error*** is the difference between population and sample value due to the fact that our sample is a limited, incomplete subset of the population. ***Non-sampling error*** is the difference between population and sample value due to sources *other* that sampling error. Two major sources of non-sampling error are ***sampling bias*** and error due to ***non-response***. They are both related to the **sampling procedure**.

Sampling bias is a systematic form of error. Sampling bias is the difference between sample and population value due to a systematic under- or overrepresentation of certain elements in the population. Sampling bias occurs when

some elements have a much smaller or larger chance to be selected than was intended. Sampling bias can also occur when certain elements have no chance to be selected at all.
Suppose we want to estimate the proportion of people that will vote for candidate A in an election. Sampling bias could occur if participants were recruited on the street by an interviewer during working hours. This could lead to an underrepresentation of people who are employed full-time. If these people would vote for candidate A more often, then we would systematically underestimate the percentage of votes for candidate A.

The risk of sampling bias is eliminated, at least in the long run by using a probability sampling method. With non-probability sampling, the risk of sampling bias is strong. Sampling bias is comparable to the systematic error that makes a measurement instrument less valid, or less accurate.

Non-response is another source of error. Non-response refers to a lack of response to invitations or the explicit refusal to participate in a study. Non-response also includes participants who drop out during the study or participants whose data are invalid because they did not participate seriously, because something went wrong or they did not understand or failed to comply with some aspect of the procedure.
If non-response is random, then you could say that non-response results in a smaller sample and will thereby slightly increase the margin of error. But sometimes non-response is not random. Sometimes specific subgroups in the population are less likely to participate. If this subgroup has systematically different values on the property of interest, then non-response is a source of systematic error.
Suppose people with a lower social economic status are less likely to participate in polls and also prefer other candidates to candidate A. In that case we are missing responses of people that would not vote for A, which could lead to a systematic overestimation of the percentage of people that will vote for A.

Besides sampling bias and non-response, there are other sources of non-sampling error related to the sampling procedure. One example is an incomplete or inaccurate sampling frame, for example because the frame is out of date.

Apart from non-sampling error related to the sampling procedure, there are two other general types of non-sampling error.

The first type is error related to the **collection of data**. This type of error could be caused by errors in the instrument, such as poorly worded questions or untrained observers. Data collection errors can also be due to the errors in the procedure, such as giving inaccurate instructions, a failure of equipment or distraction by fellow participants during data collection.

A final source of non-sampling error lies in the **processing of data** after they have been collected. Data entry errors can be made for example when data is entered into a data file manually, or when responses need to be recoded or aggregated in the data file.

As you can see, non-sampling error includes systematic error such as sampling bias, systematic non-response error and systematic collection error due to faulty instruments or procedures. However, non-sampling error also includes random error, such as random non-response error, random data collection and random data processing errors.

One final remark: For random samples the sampling error can be estimated. The size of non-sampling error is much harder to assess, even in random samples. There are all kinds of sophisticated techniques available to assess sample bias and systematic and random non-response errors. Unfortunately, these techniques usually require rigorous sampling methods, large sample sizes and all kinds of additional assumptions.

5.09 Sampling: Sample size

The goal of sampling is to obtain the best possible estimate of a population value, within the limits of our budget and our time. Suppose we've decided on a sampling method for our study - preferably a probability sampling method if this is at all possible. The question now remains how many elements we need to sample in order to get an accurate estimate of the population value.

An easy answer would be "as large a sample as we can afford". Because as sample size increases, the margin of error will decrease. Accidental over- or underrepresentation of certain elements will be less extreme and will become less likely. In other words, a bigger sample is always better in terms of accuracy.

But this doesn't mean we should all collect samples consisting of tens of thousands of elements. This is because as the sample size grows, the decrease in the margin of error becomes smaller and smaller. At a certain point the cost of collecting more elements outweighs the decrease in the margin of error.
Say we want to estimate the proportion of votes for candidate A in upcoming elections. Suppose we have a sample of five hundred eligible voters. Then the error won't be cut in half if we double the sample to a thousand elements, the decrease in error will be much, much smaller.

Note that it is the absolute size of the sample that matters, not the relative size. It doesn't matter if we are estimating election results in Amsterdam, with slightly more than half a million eligible voters, or national elections with more than 13 million voters. As long as the samples are both randomly selected, the margin of error will be the same, all other things being equal. This seems very counter-intuitive, but it's true nonetheless.

Of course there are other factors to consider when deciding on sample size. The variability of the population is an important factor. Heterogeneity, or strong variation in the population on the property of interest, results in a larger margin of error, all other things being equal. If values in the population vary widely, then

a sample is more likely to accidentally over- or underestimate the true population value.

If the population is more homogeneous or similar, meaning it takes on a narrow, limited set of values, well then the sample value will automatically lie close to the population value. If a population is more homogeneous, we can sample more efficiently. This means, all other things being equal, that we can achieve a smaller margin of error with the same sample size. Or, conversely, we can obtain the same margin of error with a smaller sample.

If a probability sampling method is used we can determine what margin of error we are willing to accept, given a certain confidence level. We can say that we want our sample estimate of election results to differ by no more than five percent from the final results in ninety five percent of the samples, if we were to sample repeatedly.

We, or rather a computer, can now calculate exactly what sample size we need, to obtain this margin of error at this confidence level. This does require that we use random sampling and that we can estimate the variability in the population, for example based on previous studies, old census data or just a best guess if necessary.

I'll just mention one other important factor to consider when determining the sample size. It's a good idea to plan ahead and compensate for **non-response**. Non-response refers to elements in the sample that cannot be contacted, that refuse to participate, fail to complete the study or provide invalid responses.
If the response-rate can be estimated based on previous or comparable research, then we can take non-response into account and sample extra elements that will compensate for the expected loss of elements due to non-response.