Can we rely on non-probaility sampling?

Pascal Ardilly[*]

Sample surveys are based on either a probability sample or a non-probability sample. In the non-probability approach, the probability of a given individual being included in the sample generally depends on the value of the variable collected from that individual. This produces a particular error known as 'selection bias'. In the non-probability 'quotas' method, this bias is limited by structuring the sample according to certain variables that explain the measured phenomenon. However, a bias remains if those variables fail to account its whole variability. In order to fully justify the method, one appeals to an assumed behaviour of individuals, known as modelling. Other non-probability selection methods exist, such as the purposive selection method – reflecting the common perception of 'representativeness' - or volunteer sampling, particularly developed in recent years through 'Access panels'. In this last case, the selection bias can be significant, even considerable. Unfortunately, the bias cannot be reduced by increasing the size of the sample. Two striking examples – one relating to the vaccine uptake rate against the coronavirus, the other to the 1936 presidential elections in the United States – illustrate this phenomenon, known as the 'big data paradox'.

[*] Expert, Department of Statistical Methods, *Direction de la méthodologie et de la coordination statistique et internationale* (Methodology, Statistical Coordination and International Relations Directorate – DMCSI), INSEE.
pascal.ardilly@insee.fr

Statisticians inherently face problems in relation to making estimates. They seek to get as close as possible to various "quantities", for which the exact value is, in principle, not known. These quantities, referred to as "parameters of interest", are defined within a given population (individuals, companies, sales items etc.), which is generally very large, on the basis of individual quantitative or qualitative variables called "variables of interest". Most of the parameters are means or are constructed based on means (totals, proportions or variances). For example, a statistician is interested in the average income of people living in Brittany on a given date, the total annual turnover of Parisian bakeries or the change in average food prices between two consecutive months. In this context, the most accurate statistics are obtained through comprehensive data collection, therefore through a census. Since the cost of a census is generally a deterrent, in practice, sampling techniques are used that restrict the data collection to a sub-population, the sample. A distinction can be made between two main types of sampling techniques: probability sampling and non-probability sampling (Ardilly and Lavallée, 2017). The latter sometimes relies on samples of volunteers – the notorious "access panels"[1] – and very often uses the famous "quota" sampling method. Non-probability sampling is appealing because of the speed with which it can be implemented and the savings on resources, while adhering to quotas has a reassuring side. Admittedly, using sampling always entails risk, but with this method, a special degree of caution is required. Why, and on what grounds can it be criticised?

This article seeks to answer this question by highlighting the errors that non-probability sampling most often produces. In particular, the errors cannot be reduced by simply increasing the sample size. However, they may be eliminated if certain assumptions are accepted regarding the variables of interest under consideration. A paradigm shift means that it is possible to construct a theoretical framework that can generally be used to explain this.

▶ Probability and non-probability sampling: a methodological schism

In conducting a sample survey, statisticians distinguish between four distinct stages: sampling, collection, estimation and accuracy estimation. Sampling – except in the very special case of purposive units (see below) – is a major source of uncertainty: it is a case of designating the units from which information will be collected. The next stage is collection, which must be carried out in accordance with a number of instructions and almost always produces non-responses. Non-responses introduce a second source of uncertainty, which statisticians should seek to minimise. The next stage is estimation, a computational stage that uses an appropriate technique to aggregate the individual data collected in order to estimate the parameter of interest. The operation ends with the sampling and non-response error measurement, commonly known as the "accuracy estimation".

In a given population, the selection of any sample may or may not be random, and if it is random, it may or may not be possible to calculate the probability of obtaining the sample in question. The context in which sampling allows for control of the selection probabilities is when probability sampling is the method used. "Control" should be understood to mean

1 These are databases that contain a large number of people who volunteer to participate, under certain conditions, in surveys on a variety of topics.

that the sampling method used allows for a theoretical calculation of these probabilities. Otherwise, non-probability sampling is used.

The basic principles of probability sampling assign a central role to the sampling frame and the selection algorithm. The sampling frame is the exhaustive list of individuals, without duplicates, who form the population of interest. An algorithm, which is an objective rule (without human intervention) that is fully coded, is applied to this sampling frame to randomly select the individuals who will make up the sample. Under these conditions, the probability of each individual in the sampling frame belonging to the sample can be ascertained. This can then be used to determine a sampling weight, which is a key factor that reflects the number of units in the population that the sampled individual represents. Each individual's sampling weight is multiplied by their questionnaire responses and the result is added together across the sample to produce the expected estimates. In practice, there are cases of non-response, which are usually handled by correcting the weights. The numerical scale of this correction is important: it consists, in the most basic approach, in multiplying the weights by the inverse of the proportion of respondents in the sample drawn. A final step, called calibration, is almost always added, which consists in modifying the weights again – slightly this time – to improve the quality of the estimate (Ardilly, 2006; Lohr, 2021).

> ***Non-probability sampling, when random, is a selection method that does not allow the calculation of the selection probability of the samples or of the individuals in the population.***

Conversely, non-probability sampling, when random, is a selection method that does not allow the calculation of the selection probability of the samples or of the individuals in the population. This is not due to mathematical impotence on the part of statisticians but because, by nature, this type of selection is the result of a partly subjective process. In practice, this role is entrusted to interviewers or is performed on a voluntary basis by the participants, resulting in a loss of control over the sampling probabilities: it is perfectly possible to impose and supervise the way in which a computer selection program works, but this control is no longer possible when the selection is partly the result of human behaviour!

▶ Quota surveys, the standard methodology of non-probability sampling

The most common form of non-probability selection is performed "on the ground" by interviewers, face-to-face or by telephone, based on a set of instructions that attempt to reproduce, to the extent possible, a uniform probability mechanism in which all individuals have exactly the same chance of being drawn. The goal is to make this selection as random as possible, while avoiding giving preference to certain population categories. A natural way to reduce this risk is through adherence to quotas – hence the name "quota methods". The aim is to define sub-populations based on the modalities of a set of qualitative or quantitative variables (the "quota variables") divided into tranches, and to ask each interviewer to compile a sample in which the numbers belonging to these different sub-populations – quotas – are equal with what would be produced "on average" by equal probabilities (so-called "equiprobable") probability sampling. For example, the sample produced through

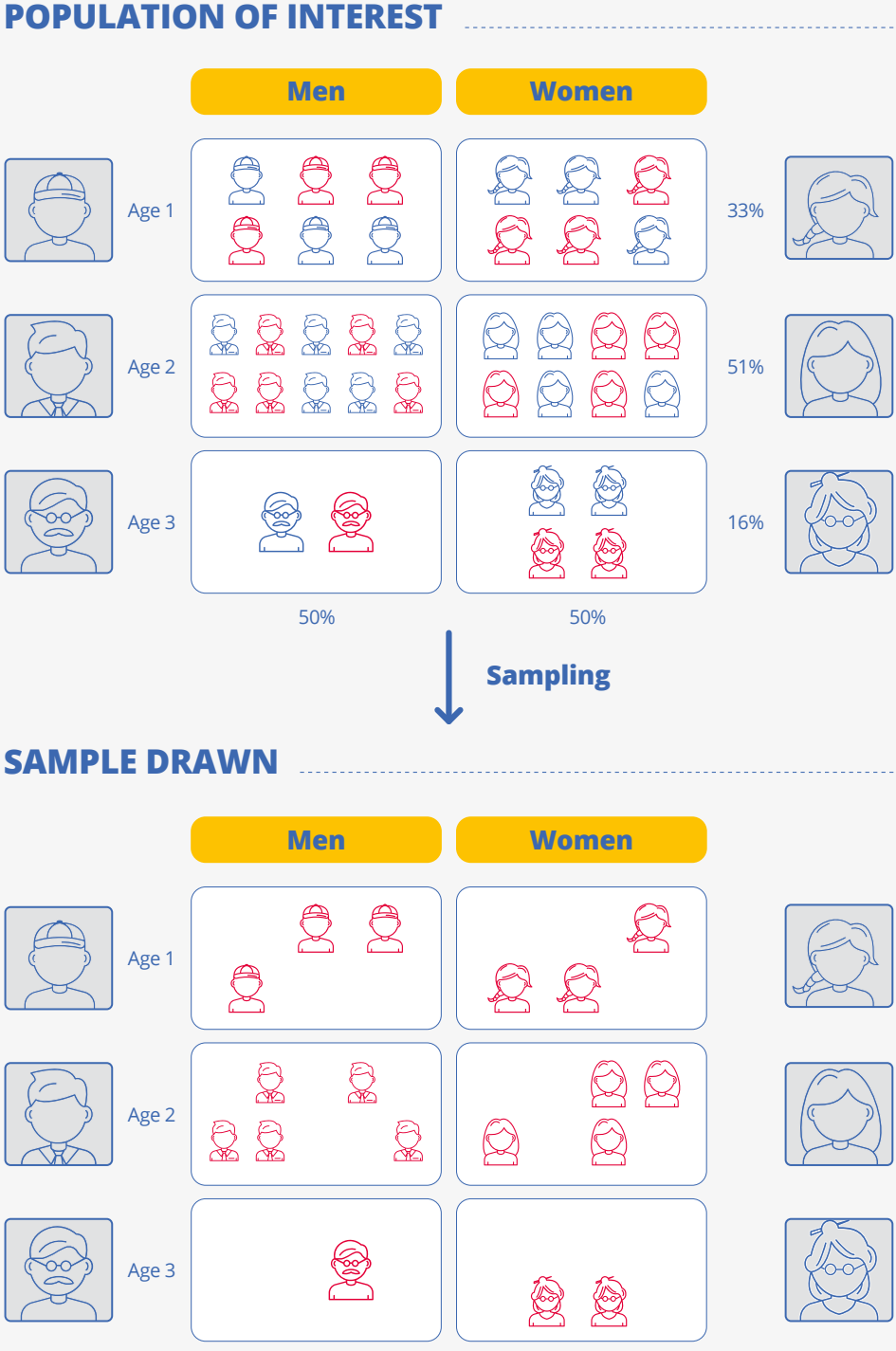
non-probability sampling is required to be composed of equal numbers of men and women, because this is the "average" sex structure that results from equiprobable random sampling. This is also the true structure of the French population based on this sex variable. This will prevent an interviewer from producing a sample that is too unbalanced in respect of the sex variable, which would move the selection process away from an equiprobable process. Most of the time, a set of quotas constructed through the simultaneous combination of multiple variables, such as sex, age and level of educational qualifications, is imposed ***(figure 1)***. This ultimately results in a sample that has the characteristics of a smaller-scale representation of the population of interest with regard to the quota variables. This method does not require a sampling frame, which is a considerable advantage, because the frames are often expensive to acquire and they can be covered upstream by the confidentiality of personal data.

▶ Compared to probability sampling, there is a dual handicap in terms of quality

Adherence to quotas is a necessary condition but one that, alone, is not sufficient for non-probability sampling to be comparable with equiprobable random sampling. To assess the nature of the risk, let us imagine a time use survey and impose quotas based on sex and age. The final sample will therefore respect the structure of the population of interest with regard to these two criteria. Interviewers, in the field, working face-to-face or by telephone, will select consenting individuals, and in principle they will do so during the day, at the times when the majority of working people practice their profession. Thus, it is highly likely that the sample will have a shortage of certain categories of workers, those who can be contacted early in the morning, late in the evening or sometimes only at night. Conversely, the sample will be "overloaded" with unemployed people, who are easier to contact during the day. Evidently, in this case, the issue is quite obvious, and this risk will be limited by taking action in at least three ways: the quotas will be enriched by adding at least one activity-related variable, interviewers will be asked to expand their collection schedules on weekdays and to work on weekends, and the collection period will be extended. The quotas will be enriched… provided that it is possible to determine variables that are sufficiently closely related to time use, provided that the structure of the population in relation to the modalities of these variables is known and provided that the constraints generated by the aggregation of quotas do not make collection unbearable for interviewers. Moreover, it will likely only be possible to extend collection schedules by a certain degree. Thus, even if the risks are significantly reduced, there will be no guarantee that there are not one or more hidden variables that explain time use but which are (unwittingly) managed in an unbalanced way by the network of interviewers. In contrast, probability sampling with equal probabilities has a significant advantage in eliminating this type of risk, as it produces a balanced sample "on average" regardless of the variable used.

> ***"There will be no guarantee that there are not one or more hidden variables that explain time use but which are (unwittingly) managed in an unbalanced way by the network of interviewers."***

▶ **Figure 1 - Quota Sampling**



In practice, the phenomenon of non-responses acts as an additional sampling stage, one that is not controlled by statisticians and which reduces the quality of the estimates. Both forms of sampling are affected by non-responses, but the collection that follows probability sampling requires multiple attempts to contact each sampled individual who cannot be contacted, until a drop-out threshold is reached. In contrast, with non-probability sampling, a sampled but non-responding individual is permanently ignored if their variable of interest is not immediately collected. The non-probability approach has a highly significant advantage in terms of cost but, once collection is completed, non-responses will have caused a significantly smaller imbalance in a sample obtained through probability sampling than in one obtained through non-probability sampling.

This insidious phenomenon is often ignored because non-responses are concealed in non-probability approaches: they are not quantified, they are almost never reported and they even seem non-existent for data users since the final sample is always the size initially required due to the way in which it is constructed. On this point, probability sampling offers a comparative advantage because it is possible to use a model to estimate the probabilities of response and to make corrections that limit the negative effect of non-responses; nevertheless, the (inevitable) imperfections of this corrective stage ultimately produce an estimation bias.

▶ Errors in sample surveys

Various types of errors affect sample surveys (INSEE Blog, 2022). We can identify four such error types.

The first error is *non-coverage*, which occurs when certain individuals in the population of interest cannot be sampled. In probability surveys, this is due to a possible lack of exhaustiveness of the sampling frame. In non-probability surveys, in the absence of a sampling frame, this type of error is more difficult to identify, but it is easy to imagine its effects. In particular, for face-to-face collection from natural persons, it is highly likely that some individuals will be consciously rejected by the interviewer – for example, because they are difficult to access or simply put the interviewer off immediately due to their appearance or unengaging behaviour. Indeed, when one has the choice of whom to interview, one naturally tends to approach individuals who seem "easy" to approach.

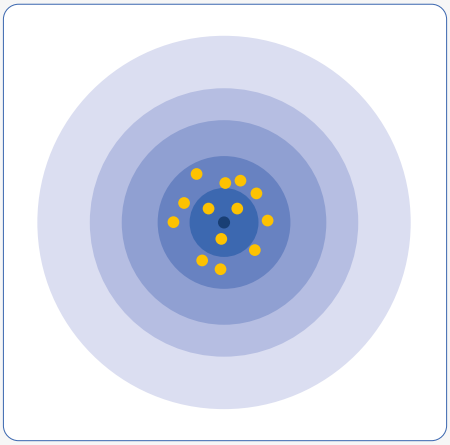
> ***If the mean of all estimates obtained from all these samples differs from the true value, it is said that there is estimation bias.***

The second error is *sampling error*. This error reflects the fact that the estimates produced are sensitive to the makeup of the sample and therefore cannot match with the "exact" value of the parameter of interest. Two components can be identified: bias and variance ***(figure 2)***.

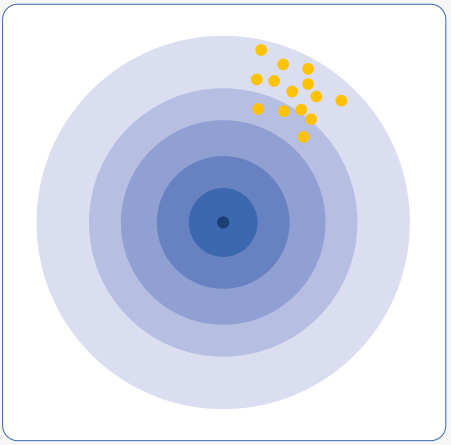
Let us assume that a large number of samples are drawn, using a given method, and that each sample produces its estimate. If the mean of all estimates obtained from all these samples differs from the true value, it is said that there is estimation bias. This may be due, for example, to systematic imbalances in the composition of the sample. Furthermore, the heterogeneity of the different estimates can be formalised through an indicator called sampling variance:

▶ Figure 2 - Bias and Variance

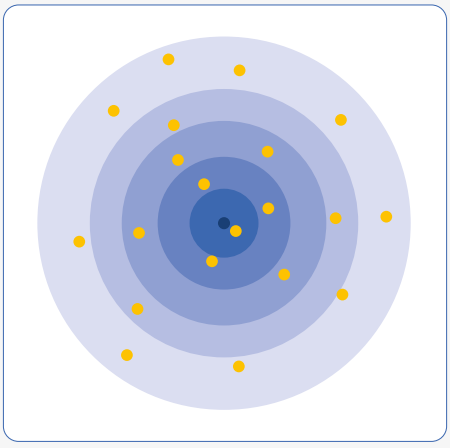
SCENARIO 1
No Bias and Low Variance



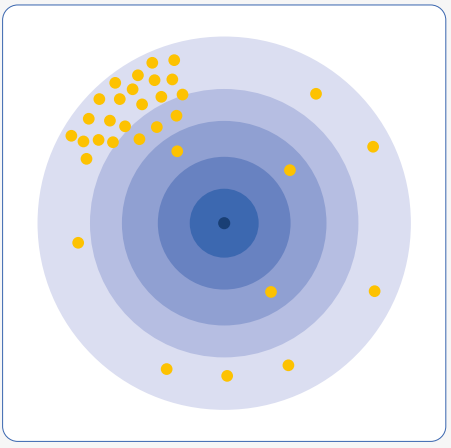
SCENARIO 2
Bias and Low Variance



SCENARIO 3
No Bias and High Variance



SCENARIO 4
Bias and High Variance



Key
The target covers all possible estimations.
The centre of the target represents the true value.
Each yellow dot represents an estimation and corresponds to a particular sample.

the greater the dispersion of the estimates, the greater the sampling variance and the lower the quality of the estimate will be. A high degree of variance therefore means that the estimate is highly dependent on the sample, which is obviously not desirable. In general, the variance of an estimate depends on three factors: the weighting system used – which essentially reflects the sampling method used – the size of the sample drawn and the level of non-responses. Regardless of the sampling method, variance decreases as the respondent sample size increases.

The third error, which has already been discussed, is due to *non-responses*, which come mainly from problems in locating the respondent (in probability face-to-face surveys), refusal by the respondent and inability to contact the respondent (in telephone surveys, for example).

Finally, there is *measurement error*, which occurs whenever the information collected does not conform to reality (for example, following a misstatement by the respondent – consciously or unconsciously – or an input error by the interviewer or even poor phrasing of the questions). For this error type, there is no reason to believe that the nature of the sampling has any particular effect. Measurement error is distinguishable from the previous error types in that it reflects a real human "failure": the other errors are more akin to imperfections for which responsibility lies with the context or which are simply random.

▶ The specific problem of bias in non-probability surveys

The quota method is sometimes criticised because it produces biased estimates. The fundamental origin of this bias is the fact that it is impossible to construct a weighting system that corresponds to both the sampling method used and the non-responses (as a reminder, the determination of the weight theoretically depends on the probability of being selected). In fact, since the probabilities of being selected are not controlled and nothing is known about non-responses, estimates from non-probability sampling are always constructed based on constant weights. Thus, to estimate means in the population of interest, simple means are calculated in the sample: this is because it is not possible to do anything else and this is the major weakness of quota surveys! Strictly speaking, it is possible to demonstrate ***(appendix)*** that the extent of the bias is determined by the correlation between the variable of interest and the probability of being selected of individuals in the population. It is fairly intuitive: using the example of the time use survey, one may fear that the probability of interviewing an individual will be higher the less that individual works. Probability sampling will not have this flaw, because the selection process will not be influenced by the nature of the respondent's activity – or if such is the case, it will be influenced in a perfectly controlled way. However, in a probability survey, non-responses generate a bias inherently; it is therefore important to seek to have the highest possible response rate.

> ***The extent of the bias is determined by the correlation between the variable of interest and the probability of being selected of individuals in the population.***

Removing the correlation between two variables that have no reason not to be correlated in a "natural" manner means creating the conditions to ensure that one of the two variables is constant. The first way to achieve this is to take action to ensure that the probability of being selected is constant.

This is precisely what non-probability sampling fails to do rigorously in practice, but statisticians naturally seek to get close to this ideal situation – which is obviously the situation with equiprobable random sampling. This is why it is absolutely essential to instruct interviewers to make this process as random as possible, so that the collection is as unselective as possible. In practice, these are common sense instructions, consisting in visiting various areas, not exclusively interviewing people in your neighbourhood, varying contact times and collection days etc. The second way of removing the correlation is to ensure that the variable of interest is constant. This method seems absurd at first glance, but it is nevertheless the one that produces the best justification for the quota method: its philosophy is supported by the model-based approach, set out below.

▶ The best way to statistically justify quota surveys

The specificity of sample surveys is related to the nature of the random sampling. In its traditional approach, which is also the default approach used today, it is actually the composition of the sample that is random, not the variables of interest. The data collected are thus considered deterministic, in other words fixed, known and provided by the respondent (except in cases of measurement error). And in this case, the estimate is tainted by bias and variance. In parallel, statistical theory has developed another approach, the so-called stochastic approach, which treats the data collected from a given individual as the result of a random phenomenon, exactly as if a lottery had decided on their values. This is another way of addressing the issue of parameter estimation, which provides a comfortable theoretical framework to justify the quota-based approach. The underlying idea is to link the value of the variable of interest to the modalities of the quota variables, with the former being a simple function of the latter, in this case a sum of values describing each modality. For example, it will be assumed that time spent on domestic chores is a function of sex (male/female), age (child/working age/elderly person) and activity status (employed/other), using these three variables and their modalities as quota variables to create the sample. Thus, knowing sex, age group and activity status, means that it is "almost" possible to determine the time spent on domestic chores. Under these conditions, it is fairly intuitive for only these variables to be important for determining the composition of the sample: since the other criteria do not count, or count very little, a possible imbalance in respect of them will not have an impact on the estimation. In this case, it is indeed necessary for the proportions of women, children, elderly people and employed people in the sample to reflect those in the population, but for the rest it does not matter: if the sample is also composed mainly of single rural people with little education, even in a grossly excessive way, this is nothing to be concerned about since the type of municipality, level of educational qualifications and marital status are not criteria that influence the time spent on domestic chores.

Such a state of mind therefore places complete trust in a relationship between variables, which constitutes an assumption that simplifies reality, which is precisely what we call a "model" in statistics.

Stochastic models also provide a very practical framework for calculating errors (Deville, 1991). These are no longer sampling errors but errors of a different nature since the randomness is what affects the values of the variables of interest. The starting point is always to assume that the model is correct, in the sense that the value of the variable of interest is on average equal to the sum of the values describing the modalities of the quota variables ***(Box 1)***. A fundamental principle follows: the method used to select the sample –

> ***The method used to select the sample – provided that it adheres to the quota constraints – is irrelevant.***

provided that it adheres to the quota constraints – is irrelevant and, as a direct corollary, it would appear that, when using the stochastic approach, there is no need to weight the sampled individuals (Smith, 1983). The use of a simple mean to estimate a true unknown mean is therefore fully justified.

It has been noted that the standard quota model consists in viewing the variable of interest as constant – with small random deviations – within each sub-population defined by combining the modalities of the quota variables. This is a highly restrictive assumption, especially since the quota variables are generally limited in number and the nature of their relationship with the variable of interest must be specific (in this case, additive).

▶ Box 1. The use of a model to justify the quota method

The use of a model - meaning a behavioural assumption - makes it possible to get around the composition of the sample. A particularly practical new paradigm emerges.

For the sake of simplicity, the context here involves two quota variables: sex and activity (whether the respondent is active or not). Modality i for the sex variable contributes to the formation of the quantity Y_k – for example, weekly time spent on household chores - *on average* at level a_i and modality j for the activity variable contributes a value of b_j to it *on average*. Which gives the following for any individual k in the cell (i,j) :

$$Y_k = a_i + b_j + \epsilon_k$$

where ϵ_k reflects the fact that knowledge of the cell (i,j) is not enough to numerically determine Y_k, or at least not precisely, because if the quota variables are correctly chosen, then ϵ_k will naturally be small (this is referred to as the "residual"). On average, the variable ϵ_k zero, which is the fundamental assumption made here, justifying the term "model". In fact, the ideal situation (with a small ϵ_k) is one where, when sex and activity status are known, it is possible to "almost perfectly" predict the time spent on household chores by any individual.

In this so-called "simple additive" model, Y_k is a random variable, just like ϵ_k, but the terms a_i and b_j are not random. The average defined in relation to the randomness of the model is called the "expected value".

The size of the sample in the cell (i,j) is $n_{i,j}$. The marginal sample sizes (respectively population sizes), which are also "quotas", are written as $n_{i,.}$ and $n_{.,j}$ (respectively $N_{i,.}$ and $N_{.,j}$). Compliance with quotas is essential, as it means imposing

$$\frac{n_{i,.}}{n} = \frac{N_{i,.}}{N} \text{ and } \frac{n_{.,j}}{n} = \frac{N_{.,j}}{N}$$

for every (i,j). In the case at hand, this requires the respective proportions of men and women in the population and in the sample to be equal. The same is also true for the proportions associated with the two specified activity/non-activity modalities. It is demonstrated that, in these conditions, and regardless of the sample drawn (which is essential!), the expected difference between the simple average $\bar{y}$ in the sample and the true average in the complete population $\bar{Y}$ is zero, reflecting the absence of $\bar{y}$ bias in this specific modelling context.

But this framework creates, *de facto*, (almost) zero correlations between the probability of being selected and the variable of interest, or, as previously stated, the conditions for a (very) low level of sampling bias[2] – so we have come full circle!

Since a model is the formalisation of an assumption, the risk is obviously found in having a false assumption, which would immediately generate an estimation bias in respect of the randomness of the model.

2 There is a similar principle for the correction of non-responses in surveys: bias occurs when there is a correlation between the variable of interest and participation in the survey, once certain explanatory variables (which play an equivalent role to quota variables) have been applied.

▶ Probability sampling or quota sampling?

This is clearly a key operational issue. Where a sampling frame is not available, needs must, as probability sampling is not possible. This is a fairly common situation, because sampling frames are very often confidential files created and held by public bodies, which cannot be disseminated. For natural persons, this is the case for the population census or tax files, for example. For companies, however, the Sirene business register is accessible to any user. Survey budgets must then be taken into account: probability surveys are significantly more expensive because they require sampled units. This requires more contact attempts and higher travel costs, if the collection method is face-to-face.

> ***Non-probability bias does not decrease as the sample size increases.***

Beyond these logistical and budgetary aspects, statistical quality considerations play a role in decision-making (Mac Innis, 2018; Brüggen, 2016; Shirani-Mehr, 2018; Forster, 2001). By construction, and this is an advantage of quota methods, adherence to quotas restricts the diversity of samples and this results in a reduction in sampling variance. However, as regards a disadvantage of quota methods, bias is a detrimental factor that does not affect probability sampling if non-responses (and non-coverage) are ignored, and it can then be verified, unfortunately, that non-probability bias does not decrease as the sample size increases. This leads to a trade-off between bias and variance. Looking at total sampling error, taking into account both bias and variance, it would seem that non-probability sampling is not suited to large samples. However, small non-probability samples may be preferable to an equiprobable random sample because the advantage in terms of variance exceeds the handicap of bias ***(Box 2)***. This principle is consistent with what is seen in practice: non-probability samples rarely exceed 2,000 units and their sizes are quite often around 1,000 units or fewer.

▶ Box 2. Comparison of probability and non-probability surveys in terms of sampling error

With regard to the sole criterion of statistical accuracy, the two types of survey examined in this article have different behaviours, particularly with regard to the effect of sample size. The main characteristics are presented below.

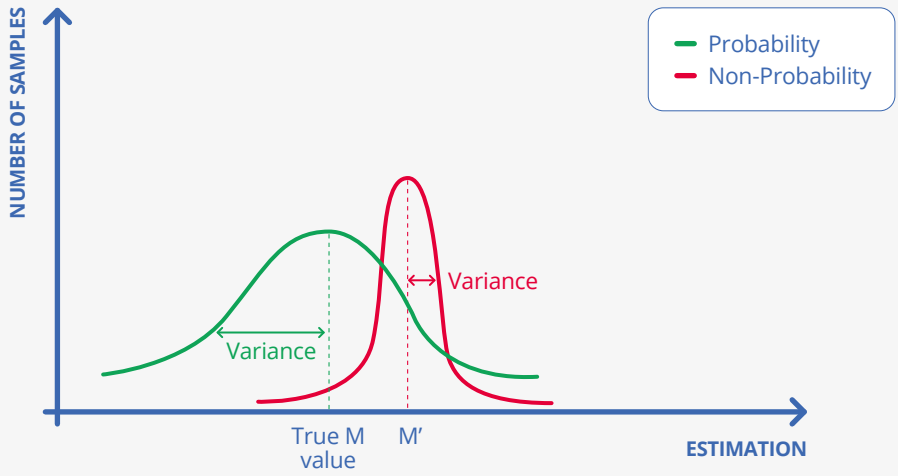


Figure 1 compares the distribution of estimations from non-probability sampling (red curve) and simple and equal probability sampling (green curve) respectively for a given and rather small sample size, such as a few hundred units, for example. These curves have Gaussian curve shapes ("bell curve"). The probability curve is centred on the true M value (no bias) and is more spread out than the red curve, which reflects a larger sampling variance. The red curve is centred around an M' value that differs from the true M value, meaning that the bias is $M'–M$.

▶ Other non-probability sampling techniques: purposive units, volunteer samples, and access panels

Non-probability sampling techniques go much further than quota surveys, which are only one modality. Let us now present three competing practices, the first of which (purposive units) is used little, while the third (access panels) is very popular.

Sampling does not necessarily mean random selection. Historically, it was not random selection that was used in the early stages of sample surveys, but rather techniques in which the individuals surveyed were conscientiously chosen – and ideally they were chosen wisely. The highly suggestive term "purposive choice" can be used to refer to samples defined without any involvement of randomness. This approach is entirely dependent on a model and is therefore scientifically deviant for statisticians who insist on avoiding models, because there is always sampling bias and the sampling variance loses its significance. The individuals surveyed are believed to be those who best represent the population as a whole, meaning that statisticians can speak of purposive units or even individuals who are

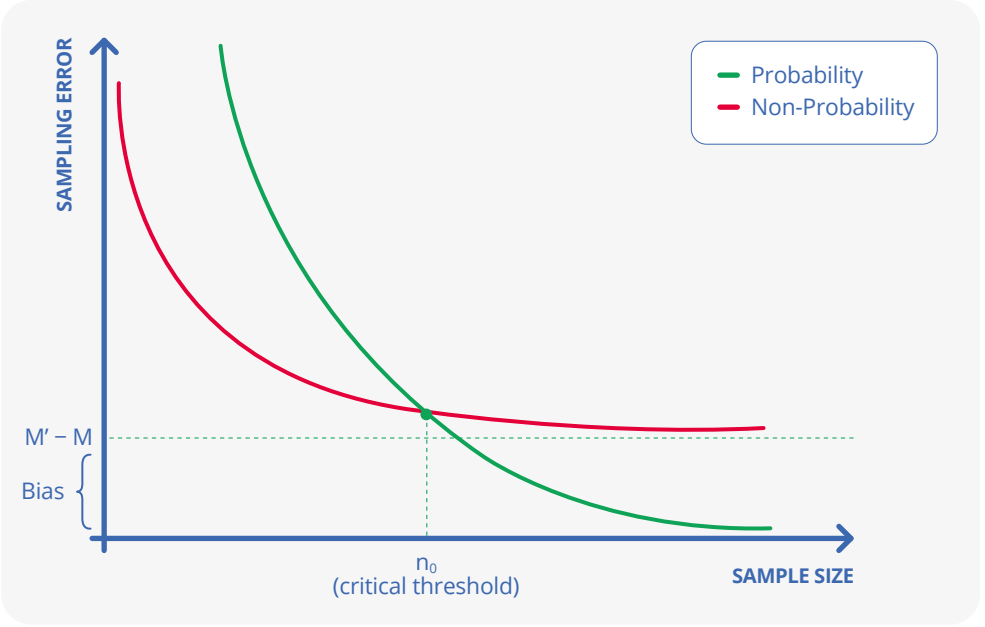


Figure 2 explains how sampling error, combining bias and variance, evolves in accordance with sample size. In the case of probability sampling (green curve), the error is significant in the zone for (very) small sample sizes and variance contributes greatly. Since bias is always zero and variance approaches zero as the sample size increases, the green curve falls to (almost) merge with the x-axis. The red curve is below the green curve in the zone for (very) small sample sizes, because the non-probability sampling has an advantage over simple and an advantage over simple random equal probability sampling (the one that assigns the same probability of being drawn to all samples of a given size). As bias is not zero in the case of non-probability sampling and is not (or not very) sensitive to sample size, the red curve is descending but becomes an almost straight line above the x-axis – positioned on the y-axis at the level $M'–M$, which is the value for sampling bias. The two curves must cross "somewhere" and the point at which they cross defines a critical sample size above which non-probability sampling is less effective than even the simplest form of probability sampling.

"representative" of the population. In the example of the household chores survey, a few cooperative individuals could be chosen for each combination of the modalities of the three explanatory variables selected. Again, since the selection method does not matter when using the model, the involvement of randomness contributes nothing. This method, which is used only in rare well-suited circumstances, retains its full value for very small samples, for which the variance would be enormous if their composition were left to chance. Typically, this approach can be used to define a sample of a few departments from which individuals are then drawn. This approach is also applied by INSEE in the selection of certain products and outlets monitored in the Consumer Price Index.

The sampling of volunteers refers to all situations in which part of the population is left to participate in a survey at their own initiative. Of course, surveys in France are first and foremost voluntary – even though most official statistics surveys are legally mandatory – in the sense that a refusal hardly ever has a significant penalty for individuals or companies. However, probability sampling, and non-probability quota sampling to a lesser extent, are performed in accordance with certain rules that seek to structure the sample in an

> ***Free of any framing, the sampling of volunteers has none of these virtues and has proven to be open to scientific criticism.***

efficient way upstream and the collection is based on principles that aim to preserve the composition of the sample drawn as much as possible. Free of any framing, the sampling of volunteers has none of these virtues and has proven to be open to scientific criticism. This category mainly includes online opinion surveys that call on consumers to give their opinions on a product or service. The resulting satisfaction indices are subject to considerable risks of bias since, in principle, this is a situation in which there is very strong correlation between the probability of participation and the variable of interest: for example, it is quite natural for a consumer who is dissatisfied with a meal in a restaurant to post a bad review online and, conversely, to post a very good one if he is highly satisfied. However, when faced with more standard service, will he make that effort?

The samples drawn from access panels are used in intermediate cases that combine volunteer sampling and quota sampling. This term refers to a set of practices that are probably quite diverse, but in many cases it is a question of creating a very large sample of volunteers upstream, which is managed on an ongoing basis and serves as a sampling frame to then, as and when needed, produce much smaller samples that respect the appropriate quotas – with the risks that we have just discussed in terms of bias. The fact that volunteers are compensated, in various ways, in return for their participation should not be overlooked and this is most likely not without consequences on the composition of the samples, whatever one may say. Access panels have the advantage of producing samples of individuals with a targeted profile at a lower cost, even allowing rare populations to be surveyed, but they are sources of data of variable geometry that can be highly opaque for users. Such situations should generate distrust: the lack of information regarding the methods is generally problematic and, in this specific case, increases the risk if the process of creating and managing the access panel, as well as its structure, are not explicit.

▶ The Big Data paradox: a recent catastrophic example...

We are accustomed to thinking that the more data there are, the better the statistics will be. This is false, even grossly false: the big data paradox is that it appears that quantity is no guarantee of quality (Meng, 2018), a theory that is demonstrated by the following two examples.

The first example concerns the very recent health crisis. In the United States, in 2021, three mechanisms (among many others) were designed to measure the vaccine coverage against coronavirus among Americans (Bradley, 2021). Two samples inspired by non-probability methods – the *Delphi-Facebook* (DF) and the *Census Household Pulse* (CHP) – were selected, while a third, designed by Axios-Ipsos (AI), had the characteristics of a probability sample. The DF sample, organised in weekly waves of 250,000 individuals, accumulated 4.5 million respondents between January and May 2021, taken among active Facebook users. The method of collection was (obviously) online collection. The CHP sample, drawn from a file of Internet addresses and telephone numbers, accrued 600,000 responses, also obtained online, over the same period. These two samples are drawn randomly from incomplete (and even largely incomplete for the DF) sampling frames and, above all, they are fully comparable to volunteer samples given their excessively low response rates (1% for the

DF and 6% to 8% for the CHP). For the AI sample, 10,000 people were interviewed over the period. It was drawn using probability sampling from a large pool, which was itself formed using probability sampling from an almost exhaustive sampling frame of postal address. This pool, which evolves over time, certainly contains people who volunteer to participate in various surveys and is therefore akin to an access panel (*Ipsos Knowledge Panel*), but in this case, it is a mechanism that maximises the probability components and is managed and controlled as a probability sample, while respecting good practices. The final AI response rate is 50%. The survey was conducted online, but Ipsos lent a tablet to all people who did not have access to the Internet. The situation was very favourable for assessing the performance of each mechanism because the actual vaccination coverage is available: indeed, the US Center for Disease Control and Prevention is a State agency that compiles vaccination statistics that reflect the reality on the ground. This is done with a time lapse, but the "true values" are nevertheless obtained. All samples are re-weighted – these are calibrations – so that certain socio-demographic structures are estimated perfectly. The results are disappointing: in May 2021, the DF process overestimated the true vaccination rate (which was 60%) by 17 percentage points, the CHP process overestimated it by 14 percentage points… and the AI sample proposed an estimate that proved correct! Using the collected data, it was verified that the weekly non-probability DF sample (250,000 individuals) produces estimates of statistical quality equivalent to that of a probability sample of… 10 respondents! A disaster that can be largely explained by a considerable imbalance affecting both major samples based on different criteria, in particular the level of education and ethnic origin. By comparing them with census data, it became clear that the DF and the CHP massively over-represent people with a high level of education (weighting of those with four years of higher education or more: 30% in the population, 36% in the AI, 45% in the DF and 55% in the CHP) while under-representing, although to a lesser extent, African-American people and, for the DF, Asian people. All this is due to the nature of the sampling frames, the method of collection and a very different non-response management strategy for the different surveys. However, it turns out that highly educated white people get vaccinated more than other categories of the American population. Sensing that pitfall, the CHP mechanism carried out calibrations with regard to ethnicity and level of education, thereby limiting the effects of the imbalance in the sample surveyed, but the DF did not. Although this has not been proven, there are also suspicions of harmful imbalances in relation to political opinion and the split between those living in urban and rural settings. This suspicion is well founded, because the AI sample has been calibrated in accordance with political opinion (partisanship) and in accordance with the category of municipality (metropolitan status), and this is not the case for the other two mechanisms, even though it is well known that these two variables have a significant effect on the propensity to get vaccinated (for example, people get vaccinated less in rural areas).

> ***We are accustomed to thinking that the more data there are, the better the statistics will be. This is false, even grossly false.***

...and an equally revealing precedent

The second example is taken from history (Antoine, 2005; Lusinchi, 2012). In the 1930s, in the United States, the media used to conduct survey operations known as "straw polls", which consisted of asking questions by post to people included in accessible files of various kinds containing personal data, such as lists of magazine subscribers, telephone service subscribers, vehicle owners or voters. The famous Literary Digest used this technique in 1936 to predict the winner of the presidential election, which saw Democrat Franklin Roosevelt face off against Republican Alfred Landon. At the same time, three forerunners – George Gallup, Elmo Roper and Archibald Crossley – used something that was quite new and bold, sampling that adhered to certain quotas: although their polling did not have the rigour of probability surveys, they nevertheless tried to diversify the respondent profiles as much as possible, and their structures based on several variables were "controlled". Each pollster had designed their survey and the three samples each included a few thousand or tens of thousands of individuals. However, the Literary Digest was proud to have collected two million responses from volunteers (the size of the sample used in reality is unknown – but it was very large), which allowed it to predict a very clear victory for A. Landon, with 57.4% of the vote. In contrast, the three pollsters announced a victory for F. Roosevelt. The result was unquestionable: Roosevelt won hands down with 61% of the votes. What happened? The individuals who received the letters from the Literary Digest were more educated and wealthier than the "average" American: you had to at least know how to read and write in order to be able to respond and subscribing to newspapers or possessing certain goods – telephone, vehicle etc. – showed a certain level of education, financial well-being etc. Those people were mostly in favour of the Republican Party.

These two examples illustrate the pernicious effects of inadequately controlled and inadequately corrected sampling. Thus, the big data obtained by the Literary Digest, the DF and the CHP paradoxically did not carry much weight in comparison to the much smaller but much better thought-out mechanisms used by Gallup and Axios-Ipsos. On the face of it, this is concerning because there can always be fears that an explanatory variable for the (often complex and multifaceted) phenomenon that the statistician wants to measure will not be taken into account either in the sampling process or in the estimation performed via calibrations. This may be due to ignorance, lack of knowledge of the actual structures or cultural or legal reasons. For example, in surveys conducted in France, ethnicity and political sensitivity – which can be thought to be correlated with a number of behaviours – are, in principle, rarely taken into account in the establishment of quotas.

It showed how reassuring it was to produce estimates resulting from sampling performed in a controlled mathematical framework, offering a minimum level of guarantees as well as quality measures for the estimates produced.

The Literary Digest affair certainly played a catalytic role in the development of the formalised theory of probability sampling, which dates back to that era. It showed how reassuring it was to produce estimates resulting from sampling performed in a controlled mathematical framework, offering a minimum level of guarantees as well as quality measures for the estimates produced.

By way of a conclusion

A sample is a complex multidimensional object with many facets. It can be very harmonious from some angles and very unsightly from others, such that in order to fully assess it, it must be possible to examine the sample from all sides. While the size of the sample of respondents is essential to the quality of a survey, another key to the problem lies in the role that is assigned to randomness. When the sample size is sufficient, the randomness generated by an algorithm has the advantage of considerably reducing the risk of distorting the sample in all respects, while the randomness attributed to humans throughout the selection process can be devastating. Entrusting non-probability sampling to experienced professional structures makes it possible to reduce the risk of bias. However, to remove controversy concerning the sample, statisticians may also be tempted to change the nature of the randomness: the randomness of the behavioural models, which reflects an assumption that simplifies reality, enables a paradigm shift, offering an attractive framework albeit one that is difficult to understand and that ultimately transfers the risks to errors in the specification of the model in question. Avoiding the need to subject the disseminated estimates to this leap of faith is a strong argument put forth by the Official Statistical Service to limit the use of non-probability methods to the extent possible.

Furthermore, experience with non-probability samples shows that it is important to avoid relying on the amount of information, which provides no protection against the risk of statistical disaster: this is the big data paradox... of which the journalists of the Literary Digest were among the first victims in history!

Finally, leaving aside the specific case of very small samples, for which the purposive unit technique is best suited, in terms of statistical efficiency and for a given sample size, the probability method is always preferable to the non-probability method. This is due to the fact that people know how to draw probability samples that adhere to quotas - before they are distorted by non-responses: this magical method is called balanced sampling (Deville, 2004). However, non-probability sampling does retain the undeniable advantages of speed of implementation and savings on resources.

▶ Appendix. The origin of bias in non-probability sampling

The selection bias resulting from sampling depends on the relationship between the variable of interest and the probability of being selected. It is structured as follows.

Using the example of a quota survey in a population of size N involving two quota variables, the modalities of which are identified by the indicators i and j respectively. Where Y_k is the value of the variable of interest for individual k and $\bar{Y}_{i,j}$ is the true means of this variable in cell (i,j). $\in_k$ can always be defined as meaning that: $Y_k = \bar{Y}_{i,j} + \in_k$ for any $k \in (i,j)$. The value of the probability of an individual k being selected is written as P_k. From this structure, it is possible to show that the sampling bias of the simple means calculated in the sample is:

$$\frac{1}{n} \times \sum_{i,j} N_{i,j}.\ Cov_{i,j}(P,Y)$$

where n is the sample size, $N_{i,j}$ is the size of the population in the cell (i,j) and $Cov_{i,j}(P,Y)$ is the covariance between variable P and variable Y in the population forming the cell (i,j), which gives:

$$Cov_{i,j}(P,Y) = \frac{1}{N_{i,j}} \sum_{i,j\ k \in (i,j)} (Y_k - \bar{Y}_{i,j}).\ (P_k - \bar{P}_{i,j})$$

$\bar{P}_{i,j}$ is the true means of P_k in the cell (i,j).

The covariance is positive when the variables Y_k and P_k vary in the same direction. It is negative if these variables vary in opposite directions. The covariance is 0 if the two variables are independent.

The latter case is the only one that removes the bias.

Contrary to what the appearance of the bias formula might suggest, bias is not sensitive to sample size: this is due to the fact that the probability of being selected P_k is always of the order of magnitude as the sampling rate n/N.

▶ Bibliography

- ANTOINE, Jacques, 2004. *Histoire des sondages*. Éditions Odile Jacob, 20 February 2004. EAN13 : 9782738115874.

- ARDILLY, Pascal and LAVALLÉE Pierre, 2017. *Les sondages pas à pas*. Éditions TECHNIP. ISBN 9782710811794.

- ARDILLY, Pascal, 2006. *Les techniques de sondage*. Éditions TECHNIP. ISBN 978-2-7108-0847-3

- ARDILLY, Pascal, CASTELL, Laura and SILLARD Patrick, 2022. Il y a sondage et sondage. In: *Blog Insee*. [online]. 25 July 2022. [Accessed 10 September 2023]. Available at: *https://blog.insee.fr/il-y-a-sondage-et-sondage/*.

- BRADLEY, Valerie C., KURIWAKI, Shiro, ISAKOV, Michael, SEJDINOVIC, Dino, MENG, Xiao-Li and FLAXMAN, Seth, 2021. Unrepresentative big surveys significantly overestimated US vaccine uptake. December 2021. In: *Nature.* Volume 600. [online]. [Accessed 10 September 2023]. Available at: *https://www.nature.com/articles/s41586-021-04198-4*.

- BRÜGGEN, Elisabeth, VAN DEN BRAKEL, Jan A. and KROSNICK, Jon, 2016. Establishing the accuracy of online panels for survey research. In: *site de CBS Statistics Netherlands*. Discussion Paper. [online]. 11 April 2016. [Accessed 10 September 2023]. Available at: *https://www.cbs.nl/en-gb/background/2016/15/establishing-the-accuracy-of-online-panels-for-survey-research*.

- DEVILLE, Jean-Claude and TILLÉ, Yves, 2004. Efficient balanced sampling: The Cube method. In: *Biometrika*. December 2004. Volume 91, No 4, pp. 893-912.

- DEVILLE, Jean-Claude, 1991. Une théorie des enquêtes par quotas. In: *Techniques d'enquête*. [online]. December 1991. Statistiques Canada. Volume 17, No N2, pp. 177-195. [Accessed 10 September 2023]. Available at: *https://www150.statcan.gc.ca/n1/pub/12-001-x/1991002/article/14504-fra.pdf*.

- FORSTER, Jonathan, 2001. Sample Surveys: Nonprobability Sampling. In: *International Encyclopedia of the Social & Behavioral Sciences*. Oxford, UK. Elsevier Ltd. Pp. 13467-13470.

- LOHR, Sharon L., 2021. Sampling: Designs and Analysis. In: *Texts in Statistical Science*. 30 November 2021. Éditions Chapman & Hall, vol 3. ISBN 978-0367279509.

- LUSINCHI, Dominic, 2012. "President" Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame? In: *Social Science History*. Volume 36, No 1, pp. 23-54.

- MacINNIS, Bo, KROSNICK, Jon A., HO, Annabell S. and CHO, Mu-Jung, 2018. The accuracy of measurements with probability and nonprobability survey samples. In: *Public Opinion Quarterly*. [online]. 31 October 2018. Volume 82, No N4, pp. 707-744. [Accessed 10 September 2023]. Available at: *https://academic.oup.com/poq/article/82/4/707/5151369?login=true*.

- MENG, Xiao-Li, 2018. Statistical paradises and paradoxes in big data: law of large populations, big data paradox, and the 2016 US presidential election. In: *The Annals of Applied Statistics*. [online]. June 2018. Volume 12, No 2, pp. 685-726. [Accessed 10 September 2023]. Available at: *https://statistics.fas.harvard.edu/files/statistics-2/files/statistical_paradises_and_paradoxes.pdf*.

- SHIRANI-MEHR, Houshmand, ROTHSCHILD, David, GOEL, Sharad and GELMAN, Andrew, 2018. Disentangling Bias and Variance in Election Polls. In: *Journal of American Statistical Association*. [online]. 25 July 2018. Volume 13, No 522, pp. 685-726. [Accessed 10 September 2023]. Available at: *https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1448823*.

- SMITH, Terence Michael Frederick, 1983. On the validity of inferences from non-random samples. In: *Journal of the Royal Statistical Society*. [online]. July 1983. Volume 146, No 4, pp. 394-403. [Accessed 10 September 2023]. Available at: *https://www.jstor.org/stable/2981454*.