Statistical sampling for record and archive groups

A practical guide

Ine Coessens · Wim Heirman

July 17, 2020

Abstract We introduce statistical sampling for the record manager/archivist, its various options and how to correctly apply these techniques with modern aids. Using the mathematical concept of confidence intervals, we propose a statistically sound method for determining sample size and quality. Illustrated by means of examples and graphs, this guide aims to show statistical sampling as a valuable option when defining a successful record and archive management strategy.

Keywords Statistics · Sampling · Selection · Records · Archives

1 Introduction

This guide is an instrument for the record manager/archivist (hereinafter the 'professional') willing to perform sampling according to statistically sound principles. It consists of an introduction into what statistical sampling is, the options that are available and how they should be implemented. For each option, the pros, cons and potential risks will be discussed and illustrated with examples and graphs. Based on this information, the professional should be able to take a well-considered statistical sample of the record or archive group[1] (hereinafter the 'archive').

Other sampling and selection techniques should also be considered and may be executed on the basis of an in-depth study of the archive's history and context. Choosing between one or several techniques will depend on a number of intellectual and practical considerations. When it comes to sampling, the expenses related to the preservation of large-scale archives as well

[1] A record/archive group is defined as "a body of organisationally and functionally related records established on the basis of provenance with particular regard to the administrative history, complexity and quantity of the records/archives of the agency, institution or organisation involved" [Evans et al(1984)].

as the inaccessibility of huge portions of records or files contained in these archives (hereinafter 'items') for future research will have a decisive impact on the decision-making process [Hull(1981)].

2 Definitions

In statistics, sampling is considered to be a process that provides quantitative and qualitative information about the whole by examining only part of it [Som(1973)]. The information obtained is considered representative for the whole [Evans et al(1984)].

In the record and archive management context, sampling provides quantitative and qualitative information on an archive by analysing only the part that is put forward for permanent preservation. This part, the *sample*, can be obtained by sampling based on the theory of probability (random sampling) or on subjective grounds (non-random sampling).

Random sampling: A sample selected on the basis of probability theory. This sampling method allows each item to have an equal chance of being part of the sample, resulting in a sample that reflects a diversity equal to the one of the entire archive.

The sample will allow for extrapolations and valid statements to be made about the archive to which the sample once belonged.[2] A random sample is also capable of providing a reliability estimate of the sample, as it indicates the confidence interval or the extent of error due to sampling. Lastly, this method allows to abstract a 'designed' sample which sets the sample's size or its reliability, depending on what is required [Som(1973)].

> *Example:* An archive containing 15,000 medical files is subjected to random sampling. Taking into account the preservation costs involved, only 1,000 files can be retained. Out of each 15 files, only one will be kept, chosen at random by the professional. The frequency of relative occurrence of the illnesses shown in the sample will be similar to the frequency initially present in the entire archive. If for instance 10% of the medical files in the sample relate to the flu, a similar degree of occurrence (10%) would have been the case for the entire archive.

Non-random sampling: A sample based on subjective grounds. Sampling choices are subject to personal decisions and result in a bias that will prevent any measure of the sample's reliability or its extrapolation to the initial archive [Som(1973)].

[2] An extrapolation allows statements to be made about the unknown, based on what is known. In this context, the items that are part of the sample will allow statements to be made on the entire archive (which – apart from the sample – has been eliminated and for which information is no longer available).

Example: In the medical files archive, 1,000 files are selected on the basis of two criteria: (1) files relating to famous patients and (2) files containing a unique medical case. It is clear from the outset that files answering one or both criteria will dominate the sample. A historical study of the most prevailing diseases in the sample and in the archive (by means of extrapolation) is impossible.

Random and non-random sampling methods can be used and combined by the professional, provided the resulting samples are taken, documented and kept intellectually separate for each method to ensure the statistical validity of the random sample [Cook(1991b)].

3 Case study

We will illustrate the potential and limits of each statistical sampling method with a fictitious case study. The case study covers 15,143 invoices for linen fabrics produced by the textile merchants Bethune & Fils (Courtrai, West-Flanders, Belgium) during 1737–1799 [Adriaenssens(2016)]. The invoices' analysis by PhD student Annik Adriaenssens gave us a good understanding of the archive and its content in preparation of the hypothetical sampling exercise.

In particular, the analysis allowed us to identify the following types of metadata as important for the exercise:

1. Linen length: By comparing the lengths of linen pieces mentioned in the invoices, a distinction could be made between the different kinds of linen fabric.
2. Linen colour: In line with the customer's demands, the linen kept its original colour ('écru'), was dyed or bleached. The cheapest linen was dyed, the more expensive linen bleached. Two bleaching methods were frequently used by Bethune & Fils: bleaching with ashes, named 'menagebleek', 'halve bleek' or 'waterbleek' and a more intense additional bleaching with buttermilk, named 'melkbleek'.

4 Sampling criteria

Before taking a statistical sample, the professional should first verify whether the criteria for taking such a sample are met. In our case: can a statistical sample be taken from the linen invoices produced by Bethune & Fils?

To answer this question, two additional questions should be raised and answered:

1. Are the items part of a uniform and closed archive?
2. Is at least one reliable random variable present in this archive?

1. For a sample to be representative for the entire archive, the items contained in this archive should be part of a closed unit and should contain similar information/metadata.

Case study: The archive is closed and contains the same type of documents with similar metadata – invoices produced by Bethune & Fils between 1737–1799.

2. A random variable is a type of metadata for which each item has a numerical value. By means of these values, in particular on the basis of their average and standard deviation, the internal diversity of this type of metadata can be calculated. The internal diversity obtained in turn determines the quality of the sample as a function of its size, provided a reliable random variable has been used. A reliable random variable is characterised by uniform and comparable values, fit for extrapolation to the rest of the archive. Using an unreliable variable will result in extra diversity on top of the internal diversity. This extra diversity will eventually prevent reliable statements to be made on the basis of the internal diversity.

   Case study: Within the archive, several random variables are present. One of them is the length of linen pieces mentioned in the invoices. For each invoice, at least one length value is available to calculate the internal diversity of the linen pieces' length. To meet the requirements of a reliable random sample, the length values have to be uniform and comparable to one another. As a mixture of French and Flemish ell was used to indicate the linen pieces' length, all lengths were first converted to the most frequently used unit of length (Flemish ell). If more than one piece of the same linen type was mentioned on an invoice (e.g. 11 pieces of 'melkbleek' bleached linen), only the minimum and maximum length were recorded in the data file created by Annik Adriaenssens (e.g. 59–63 Flemish ell). The average of both values was considered to be the length of all linen pieces mentioned in the invoice (here 61 Flemish ell), so as to obtain uniform and internally comparable values to calculate the internal diversity.
   Another, more obvious candidate random variable is the price of linen pieces mentioned in the invoices. Again, at least one price value was available per invoice. However, to enable a comparison of the price values, prices would have had to be converted into a common currency. In the invoices, two currencies were used: Flemish pounds and Brabant guilders. To correctly convert the price values, both the impact of the exchange rate and the inflation or deflation of the linen prices over time would have had to be included into the conversion process. Each of these factors is nevertheless difficult to quantify and adds a large portion of additional extra diversity. Any statement on the basis of this random variable would therefore have been extremely difficult, if not impossible.

The case study examples demonstrate that the existence of multiple candidate variables does not necessarily mean that all of them will end up being useful and/or reliable. The ideal random variable does not exist, but, based on an in-depth analysis of the types of metadata available in

the archive, the professional should be capable of rejecting the least reliable options. After having identified a reliable variable, the professional should also be capable of providing reliable statements on the quality and size of the sample.

5 Representative samples

Before sampling, two essential questions must be answered:

1. What sample size is required?
2. How large must an archive be to make sampling viable?

Ideally, the sample approximates the complete archive as closely as possible. Practical and financial constraints usually limit the amount of material that can be retained. It is up to the professional to find the right balance between the representativeness of the sample and its size.

Literature does not indicate a clear quantitative relation between a sample's size and its representativeness. Cook states that the quality of the sample depends on the absolute sample size (e.g. 1,000 items) [Cook(1991b)]. For large-scale archives, this is correct: once the sample size exceeds a certain threshold, it is likely to include a sample of all different types of items. At this point it no longer matters whether the initial archive contained one million or ten million items, and hence whether the sampling ratio was 0.1% or 0.01%. For small-scale archives, Cook recommends using a table from Bell Telephone, which suggests a sample size corresponding to the total archive size.[3] Again no quantitative indication of the sample's representativeness is provided for; the table only offers *low*, *average* and *high* sample sizes. Even more problematic is that these tables are not constructed for random sampling, but to solve a very different problem: to find defective or substandard parts in a large shipment of goods.[4] The table indicates how many parts must be tested to ensure that all parts of the shipment are of a given minimum quality. The mathematical background used to construct this table hence assumes a very small number of deviant items (the defective goods), hidden inside a very homogeneous group (those goods of acceptable quality). When applying these statistical formulas to a different use case, namely that of items with diverse properties from which we want to extract a representative sample, their validity is no longer guaranteed. It is understandable that Cook, in an age when accessible compute power was insufficient to run statistical analyses, had to rely on existing (albeit not applicable) tables. A modern professional with access to ample compute

[3] "The Bell Telephone MIL-STD 105D Sampling Plan" [Bell Telephone(1963)], reproduced in [Cook(1991b)], p 46.

[4] The MIL-STD 105 standard was initially defined for use in military acquisitions and is a type of acceptance sampling. Before accepting a shipment of goods, a random sample of products is tested comprehensively. When all goods in the sample pass the test, it is deemed most likely that no (or an acceptably low number of) items in the shipment are defective and that the shipment can be accepted. The size of the sample is defined by the tables of the MIL-STD 105 standard and depends on the size of the shipment and on the probability that a shipment with too many defective parts is still accepted.

power[5] and standard software packages such as Microsoft Excel no longer has this excuse.

A more recent example is the "Sampling Manual" provided by the Dutch National Archives [Nationaal Archief(2020)]. This website and accompanying Excel worksheet provide a formula to calculate the required sample size, as well as an automated way of generating the random numbers needed to make the actual selection. Their method is based on a commonly used formula to calculate the required sample size for doing surveys which typically have a yes/no answer. While the statistical model used on this website is much more applicable to archival sampling than that underlying the MIL-STD 105D Sampling Plan, archives can have a much more varied behavior than survey results. Moreover, the National Archives' method requires the archival professional to provide a value for the internal 'spread' of the archive but contains no guidelines on what this parameter exactly means or how its value can be obtained. When applying the proposed default value of 50%, a worst-case scenario is assumed which leads to a generic sample size rather than one tailored to the intrinsic variation of the specific archive sampled.

5.1 Computing confidence intervals

In this guide, we propose to determine a sample's representativeness by using the mathematical technique of *confidence intervals*. This technique is based on a piece of metadata (the random variable) and a derived metric (usually the average) which can be determined for both the sample and the initial archive. The mathematical background and formulas needed to compute the confidence interval are included in Appendix A. In this section we provide an intuitive description of the confidence interval and show how to use it. Appendix B includes instructions that allow the professional to compute confidence intervals for his/her own archive using standard computer software such as Microsoft Excel.

The confidence interval for a given random sample indicates the interval (lower and upper bound) in which the average of the chosen metadata can be found with a given probability. Hence, this interval expresses the accuracy of the extrapolation that a researcher will be able to make based on the sample.

> *Case study:* The average length of linen pieces was 74 Flemish ell. When taking a random sample of 1,000 invoices, the confidence interval – computed with the methodology outlined in Appendix B – states that "the average length, with a probability of 95%, is situated between 73 and 75 Flemish ell".
>
> In this statement, three elements are relevant:
>
> 1. The size of the sample, here 1,000 invoices.

[5] Today's average laptop computer is significantly more powerful than the supercomputer that was needed in 1963 to construct the tables of the MIL-STD 105D standard.

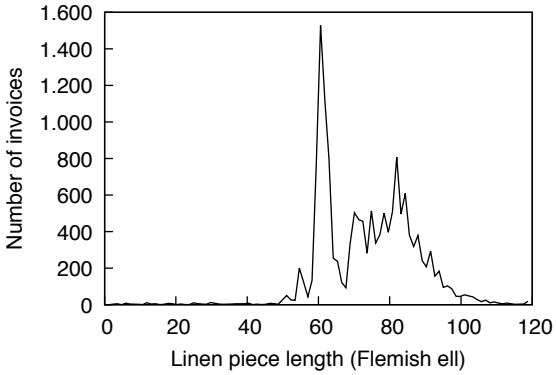

Fig. 1 Distribution of length for all linen pieces.

2. The interval itself, here 73–75 Flemish ell. Usually the size of the interval is expressed relative to its centre. In our case the interval is 1 Flemish ell above and below the average of 74 Flemish ell, which implies a relative error of 1.35%. This means that, based on the sample, we can estimate the average length of all linen pieces in the initial archive with an error of at most 1.35%.
3. The probability of the statement being correct, here 95%, indicates the probability that the average length over the initial population does indeed fall within the given interval. Statisticians see 95% as an acceptable probability. Higher values can be applied as well but these make the interval wider and reduce the practical relevance of the statement.[6]

The confidence interval allows the professional to make an informed decision about the sample's representativeness by means of objective criteria. It also allows for a more informed way of choosing the sample's size, thereby enlarging the sample until the confidence interval provides acceptably tight bounds.

When it comes to computing the confidence interval, it should be pointed out that the size of the interval does not solely depend on the size of the sample. It also depends on the internal diversity of the archive (the standard deviation σ in Appendix A): a large internal diversity yields a larger confidence interval and implies that the extrapolation that can be made from the sample will be less reliable. Given a fixed sample size, the quality of the sample will therefore be lower for those archives that show more internal diversity. To obtain the desired level of quality, a larger sample size will be needed.

[6] With 99% probability the confidence interval would become 72.7–75.3 Flemish ell while with 99.99% probability the margin lies between 69.6–78.4 Flemish ell. These examples illustrate that large errors are unlikely. They also show that if one wants to use a confidence interval with an extremely high probability, this interval becomes so wide that it no longer provides a useful error bound.

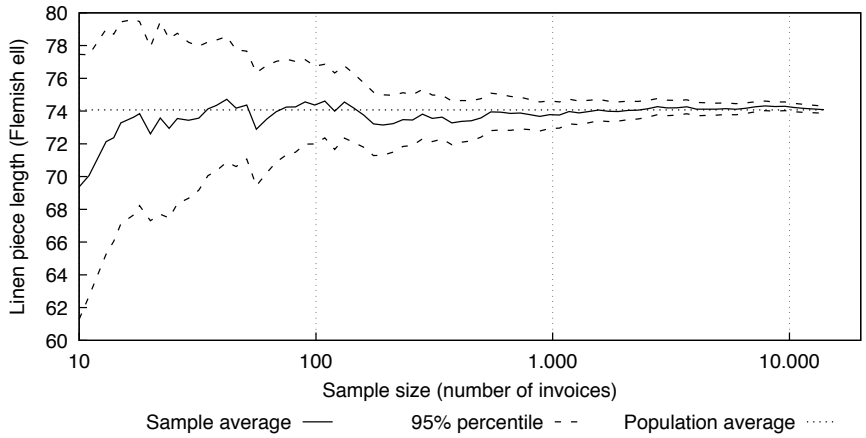

Fig. 2 Confidence interval at different sample sizes.

Case study: In total, 15,143 invoices were analysed. For each invoice we retained the type of fabric and the piece length. The distribution of lengths is shown in Figure 1.
Next, we simulated a number of potential sample sizes and plotted the confidence interval (95%) of each sample in Figure 2. The size of the sample starts at ten invoices and is gradually increased until the sample covers the entire population. The horizontal dashed line indicates the average length of the entire population, 74 Flemish ell. The solid line shows the sample average for each sample size. The dotted lines indicate the bounds of the confidence interval.
The graph shows that with a small sample of less than 100 invoices, the sample average fluctuates significantly between the different samples. The variation between invoices is such that, when adding a single invoice to a small sample, we most likely added an invoice that significantly differs from the already chosen pool of invoices. The confidence interval confirms this at the mathematical level. We can therefore state that the sample average is unlikely to be representative for the population. At a sample size of 500 invoices, the confidence interval shrinks significantly to about 1 Flemish ell. This means that a sample of 500 invoices will, with a probability of 95%, yield an average length that is at most 1 Flemish ell away from the average of all invoices.

Note that this graph was constructed with knowledge of the entire archive. We knew the population average to compare against and were able to try out many different (including very large) sample sizes – something that in reality will usually not be possible.

One should however keep in mind that the confidence interval is solely computed based on information that is part of the sample; it does not need data pertaining to non-selected invoices. Only those metadata that are part of

the (proposed) sample should be collected. Depending on the metadata chosen and the difficulty to collect these, using the confidence interval technique may result in a significant time saving. This technique also allows one to start off with a relatively small sample and to only collect metadata for the items part of the sample. Based on the confidence interval of that sample, one can evaluate the representativeness of the sample and extend it if needed. This incremental approach allows the sample to be progressively extended until the quality required is achieved or until it reaches its largest possible size.

6 Sampling methods

Apart from determining the sample size required to obtain a representative sample, one or several methods must be selected to choose those items to include into the sample. This guide only covers random sampling, i.e. those methods in which selection of a given item is purely based on chance.

Two types of sampling methods are discussed here: simple sampling and stratified sampling.

6.1 Simple sampling

Three different types of simple sampling will be discussed: simple random sampling, simple systematic sampling and simple semi-systematic sampling.

In the following sections we will use these symbolic notations:

n	Sample size
N	Population size
p	Sampling ratio n/N; the probability of a given item to become part of the sample
k	*Skip* $1/p$; to obtain a sampling ratio of $p = 1/k$, one item needs to be selected out of every k items

6.1.1 Simple random sampling

The most straightforward type of sampling is simple random sampling. This method treats all items equally (there is only one class of items, hence *simple*) and each item has the same probability of being chosen (*random*). When dealing with homogeneous archives, this method is the most appropriate one to use and provides the best chance to obtain a representative sample.

Once the decision has been taken to use simple random sampling with a given sampling ratio, the question arises how to organise the sampling process in practice. Let us assume a sampling ratio p, computed as the ratio of the desired sample size n to the size of the total population N. The ratio $p = n/N$ is a number between zero and one and indicates the probability of a given item

to be selected. For an archive with total size $N = 100$ and sample size $n = 5$, this probability will be $p = n/N = 5/100 = 0.05$ or 5%.

Random sampling solely relies on chance, so the decision to add a particular item must be made independent of any other decisions to avoid correlation between items. One way of achieving this is to assign a random number between zero and one to each item, and to include an item if its random number is below p. Another technique is to place all items in random order[7] and to construct the sample using the top n items. The latter technique supports incremental selection by first selecting the top n items, then computing the confidence interval of the sample, followed by adding more items from the existing list if needed.

Both techniques require the generation of a unique random number for each item. Today, this can easily be generated with computer software (see Appendix B). Historically, generating large amounts of good random values was more cumbersome and lead to the use of precomputed lists of random numbers.[8] These lists were often reused however and introduced certain systematic effects into the selection.

6.1.2 Simple systematic sampling

A popular sampling technique that is often found in literature is systematic sampling [Cook(1991b)]. In contrast to random sampling, systematic sampling uses a systematic mechanism to select those items to include into the sample. With a sampling ratio of $1/k$, every k^{th} item will be added from the list, starting at a random position. For a sampling ratio of one in ten and random starting point 3, this means that the items with rank 3, 13, 23, etc. will be selected.

The biggest issue with this technique arises when the items themselves exhibit a systematic effect, thereby making the sample non-representative for the whole. When an archive for example consists of daily accounting summaries with one per working day, the selection of items 3, 13, 23, etc. will only pertain to Wednesdays. This effect is called *aliasing*. When handling items pertaining to working days, aliasing can be avoided by using a sampling ratio that is not a multiple of five. The professional can however never rule out other systematic behaviour of which (s)he is not aware.

6.1.3 Simple semi-systematic sampling

The main disadvantage of pure random sampling according to Cook is that large gaps can occur in the sample (which he calls 'missing pockets') [Cook(1991a)]. Since simple random sampling is done exclusively at random, one can indeed not prevent that the majority of the sample contains items from the start or

[7] This sorting is only done at the intellectual level (for instance in an Excel worksheet that lists all items); a physical reordering of the archive is not needed.

[8] One such extensive list is for example included in the RAND Corporation's "A million random digits with 100,000 normal deviates" (1955) [McKay(1978)].

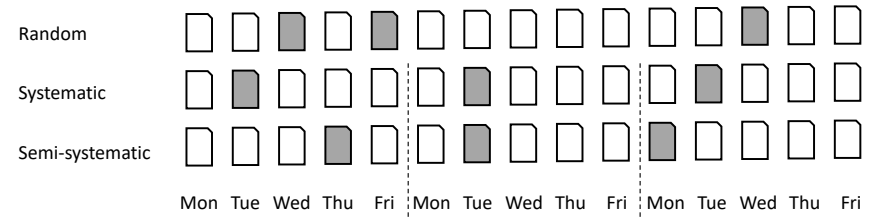

Fig. 3 Illustration of simple sampling methods.

end of the archive. With systematic sampling this problem is avoided, as the number of consecutive non-selected items is always equal to $k-1$.

Due to the potential problem of aliasing, naive systematic sampling is not a reliable solution either. The ease of implementing systematic sampling compared to the cumbersome random sampling with large lists of random numbers has been invalidated thanks to modern computing power. The missing pockets concern also does not seem to be a problem in practice. If the archive is homogeneous, it should not matter whether the sample consists of items taken from the start, middle or end. The occurrence of large gaps is also rather small: using a sampling ratio of one in ten, the probability of finding a gap of 100 or more consecutive unselected items is less than one in 35,000.[9]

In those cases where large gaps should be avoided, for instance for archives consisting of a time series where each period should be covered, a variant of systematic sampling can be proposed that does not cause the aliasing problem. Assuming a sampling ratio of one in k, the archive is subdivided into groups of k consecutive items. In each group, a random item is selected. This selection process requires a set of random numbers between 1 and k and one random number per group. Semi-systematic sampling has the advantage that there are at most $2\cdot(k-1)$ consecutive non-selected items, which places an upper bound on the size of the missing pockets. This happens when the first item of one group and the last item of the next group are chosen.

Unlike pure systematic sampling, this variant does not exhibit periodic behaviour that may concide with periodic effects in the archive itself. Compared to random sampling however, this technique is less straightforward to implement in computer software such as Microsoft Excel and does not easily lend itself to incremental selection.

6.1.4 Example

Figure 3 illustrates the simple sampling methods described earlier, with a sampling ratio of one in five. Random sampling allows for an equal probability of sampling each item and excludes any correlation. Both systematic and semi-systematic sampling split the archive into groups of five. Systematic sampling

[9] Random sampling with ratio p can be modelled as a Bernoulli process. The probability of n consecutive negative experiments is $(1-p)^n$ [Hogg and Craig(1978)]. With $p=0.1$ and $n=100$, this probability is $1/37,649$.

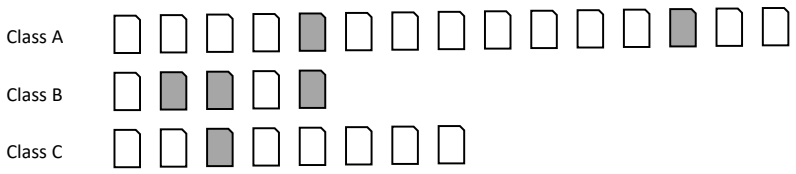

Fig. 4 Illustration of stratified sampling.

selects a random starting point, here the second item, so the sample consists of the second item from each group. All selected items unfortunately pertain to a Tuesday and illustrate the risk of aliasing. Semi-systematic sampling selects a different item in each group and results in a sample with a better distribution of the week days.

6.2 Stratified sampling

A homogeneous archive has relatively little internal variation, which means that the confidence interval will quickly reduce to acceptable levels when applying incremental selection. In practice, however, it is common to have a number of distinct classes that significantly differ from one another but that are internally more homogeneous. It is also possible to have a largely homogeneous population which contains a small number of highly deviant items (for example medical records pertaining to rare diseases).

These types of cases require the application of stratified sampling. With this technique, the population is first divided into a number of classes (groups or strata),[10] based on a parameter chosen by the professional (for instance, the type of disease). Next, simple random sampling is applied to each class. The sample size is determined for each class individually, by using incremental sampling. The classes do not necessarily have the same sampling ratio. Choosing a proper classification parameter will allow the classes to be more homogeneous and will also require a much smaller sampling ratio to obtain a representative sample. By strategically exploiting stratification it is possible to create multiple small samples that are more reliable than one large sample, even if the total amount of selected items is lower.

When applying stratified sampling, it is important to document the selection procedure and to indicate the initial size of each class in relation to the total population. Other elements worth documenting include the sampling ratio used for each class and the class each item in the sample belonged to. Only on the basis of this knowledge can the future historian reconstruct the distribution of the entire population.

[10] A stratum is a group that is part of the population and that is internally homogeneous. Different strata can be heterogeneous with respect to one another [Som(1973)].

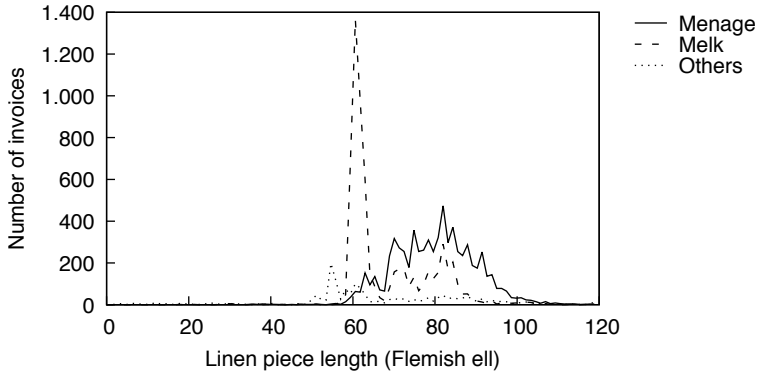

Fig. 5 Distribution of length for each group of linen pieces.

6.2.1 Example

Figure 4 illustrates stratified sampling. The archive is subdivided into classes A, B and C. Class A contains 15 items and is relatively homogeneous, so a sampling ratio of 2/15 is used. Class B contains only 5 items but is much more diverse so a ratio of 3/5 is used. Class C contains 8 items and is again very homogeneous, allowing for a ratio of just 1/8. The sample now contains 2 items from class A, 3 from class B and 1 from class C. Only when the initial size of each class is documented during the sampling procedure can the historian conclude that class A was the largest group in the initial archive, not class B, which has most items in the sample.

6.2.2 Case study

To illustrate stratified sampling, we again refer to the linen sales of the textile merchants Bethune & Fils between 1737 and 1799. Figure 1 displays a large peak in the distribution of sold linen pieces around 60 Flemish ell and an increase between 70 and 90 Flemish ell. The collection of pieces is not very homogeneous, as is confirmed by a relatively wide confidence interval when taking a sample of fewer than 1,000 invoices. We therefore opt to stratify the invoices and to group them into the classes 'menage' (46% of all invoices), 'melk' (43% of all invoices) and 'other' (remaining 11% of the invoices). This 'other' class consists of 'écru' and finished products such as 'nappe' and 'serviette'.

When we plot the distribution of piece lengths but now split them by the three classes, we get Figure 5. We see that the peak at 60 Flemish ell is caused almost exclusively by the 'melk' class, while the 'menage' and 'other' classes are responsible for the increase between 70 and 90 Flemish ell.

Next, we apply incremental selection to each class individually. We start with a sample size of ten invoices per class and increase the sample size whilst plotting the confidence interval for each sample. Using 1 Flemish ell as bound

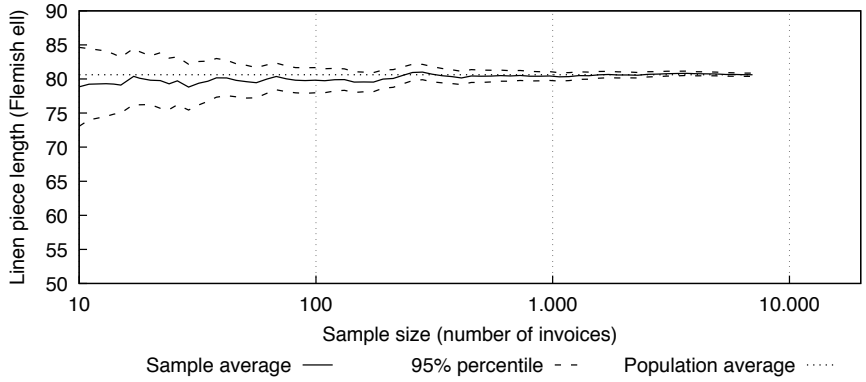

Fig. 6 Confidence intervals, 'menage' class.

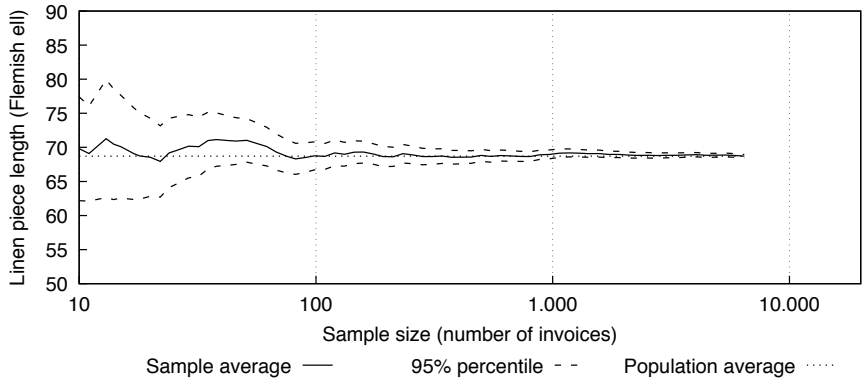

Fig. 7 Confidence intervals, 'melk' class.

to denote a reliable sample, simple random sampling requires an overall sample size of 500 invoices (see Figure 2). When applying stratified sampling, the 'menage' and 'melk' classes need a sample size of around 300 invoices each as they have a much smaller internal variation (Figures 6 and 7). The 'other' class is more diverse and has a confidence interval that is more than 1 Flemish ell wide, even at a sample size of 1,000 invoices (Figure 8). Stratified sampling in this case suggests a larger sample size of 1,600 invoices, but will also allow for a more accurate extrapolation for each type of fabric.

7 Practical considerations

Before concluding this guide, three particular situations should be highlighted that might occur when applying the sampling methods explained earlier to an archive. In particular, what should the professional do when several items in the archive relate to one another, when part of the archive is missing, or

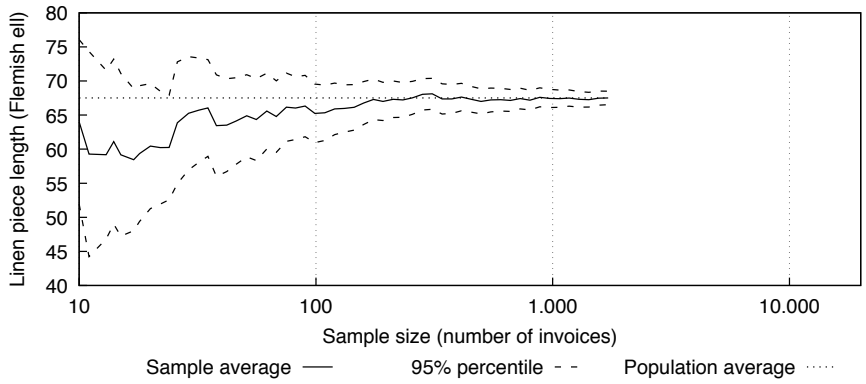

Fig. 8 Confidence intervals, 'other' class.

when a sampling method should be combined with other sampling or selection methods and decisions?

7.1 Related items

When a statistical sample is taken, each item of the archive has an equal chance of being chosen. When several items relate to one another however, one is often tempted to sample a certain number of items and to add the related items to the sample afterwards. At this stage items no longer have an equal chance of being chosen, which renders the sample invalid. Each item relating to other items has a higher chance to become part of the sample than other non-related items. If each item has a probability p to become part of the sample, those items relating to other items of the archive will see their own probability p be complemented with an additional probability p for each of the l items with which it is related. This results in a total probability of $p \cdot (l + 1)$.

To prevent this from happening, the archive should first be divided into clusters. Each cluster contains all items related to one another. As these clusters no longer relate to each another, the sample can be executed at cluster level. This time, each cluster has an equal probability of being chosen. A valid sample is taken and valid extrapolations can be made towards the initial archive.

> *Example:* In a hospital, each department has individual patient files. When a sample is taken, it is possible that for those patients having a file in multiple departments only part of the patients' files will be captured by the sample. Such situation would prevent future researchers from having a complete view on certain patients' physical and mental health conditions. If the additional, unselected files would be added to the sample afterwards, those files belonging to patients visiting multiple departments would be overrepresented which would invalidate the

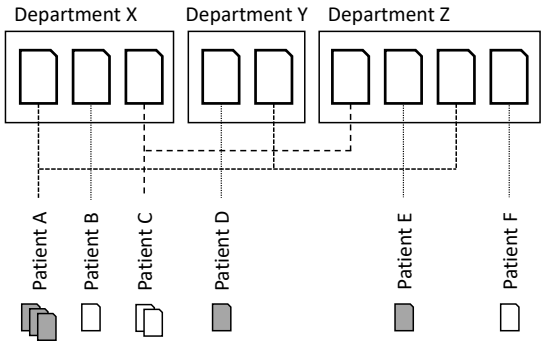

Fig. 9 Illustration of related items. The archive consists of a number of files produced by multiple departments. Each patient can have a file in multiple departments. The files are first clustered by patient, followed by sampling at patient level.

sample's representativeness. To prevent this, all files relating to a particular patient are first grouped together in an individual patient cluster, followed by sampling at patient (rather than file) level (see Figure 9). When following this procedure, each patient (and all files relating to this patient) will have an equal chance of being part of the sample.

7.2 Sampling an incomplete archive

When an archive is no longer complete, a sample can still be taken. The remaining part of the archive will then have to be considered the entity from which the sample will be taken. By means of sampling, the entity will be further reduced according to statistically valid principles that will also allow extrapolation to the initial entity, i.e. the remaining archive to which sampling was applied.

This being said, can we consider an incomplete archive as a sample in itself, provided we know its initial size? Can extrapolations be made that provide a reliable picture of the initial archive? In most cases, the answer to these questions will be negative. Obtaining a statistically valid sample requires the selection on the initial archive to have been random. In other words, each item should have had an equal probability of being lost.

When looking at the events responsible for the loss of part of the archive however, most of them do not meet the conditions required to obtain a random sample as there is a significant (topic or spatial) correlation between the items lost. Examples of such correlation include correspondence with a particular individual that was kept at a particular location and was never transferred to the main archive, water damage affecting patient files of patients with last names T–Z because they were kept on the lowest shelf of an underground archive or the intentional destruction of a select number of files pertaining to a specific event.

7.3 Combining statistical sampling with other sampling or selection methods

It is possible to combine sampling based on statistical principles with other (subjective) sampling and selection methods and decisions. One should however remain cautious. A well-executed random sample does not contain an overrepresentation of certain classes (or, when using stratification, the classes are at least documented). For those items added by means of non-statistical methods, the risk of overrepresentation cannot be eliminated. Selecting items based on medical peculiarity for example will even promote the selection of items non-representative for the whole and will prevent any extrapolation towards the initial archive.

It is therefore important to strictly separate the statistical sample from the subjectively added items and to document the methods applied and the decisions taken. An intellectual separation suffices.

8 Conclusion

In this guide, we outlined the pros, cons and potential risks of random sampling methods and demonstrated how to execute these in line with the underlying mathematical principles.

In addition, we have tried to show that statistical sampling can be combined with other (subjective) sampling and selection methods and decisions, provided each method and decision is well-documented.

References

[Adriaenssens(2016)] Adriaenssens A (2016) Van laken tot linnen: de textielhandel Bethune & Fils, tweede helft achttiende eeuw: een analyse op basis van het bedrijfsarchief. PhD thesis, Universiteit Gent

[Bell Telephone(1963)] Bell Telephone (1963) The Bell Telephone MIL-STD 105D sampling plan

[Cook(1991a)] Cook T (1991a) The archival appraisal of records containing personal information: a RAMP study with guidelines. Paris: United Nations Educational, Scientific and Cultural Organization

[Cook(1991b)] Cook T (1991b) "Many are called, but few are chosen": Appraisal guidelines for sampling and selecting case files. Archivaria 32

[Evans et al(1984)] Evans FB, Himly FJ, Walne P (1984) Dictionary of Archival Terminology. München: K.G. Saur Verlag KG

[Hogg and Craig(1978)] Hogg RV, Craig AT (1978) Introduction to mathematical statistics. New York: Macmillan

[Hull(1981)] Hull F (1981) The use of sampling techniques in the retention of records: a RAMP study with guidelines. Paris: United Nations Educational, Scientific and Cultural Organization

[McKay(1978)] McKay E (1978) Random sampling techniques: a method of reducing large, homogeneous series in Congressional papers. American Archivist 41

[Nationaal Archief(2020)] Nationaal Archief (2020) Handleiding steekproeven. https://www.nationaalarchief.nl/archiveren/waardering-en-selectie/handleiding-steekproeven, accessed: 2020-06-06

[Som(1973)] Som RK (1973) A manual of sampling techniques. London: Heinemann Educational Books Ltd

A Size of the sample

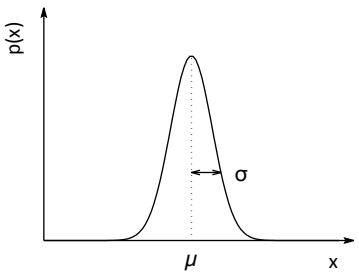

Fig. 10 The probability distribution of a normal distribution. For the entire archive, this distribution is centred around the average μ and has a width related to the standard deviation σ.

This appendix provides the mathematical foundation to calculate the quality of a given sample when using the selected random variable. We will use the following symbols:

x	Value of the random variable for a given item
x_i	Value of the random variable for the ith item in the sample
μ	Average value of x for the entire archive
σ	Standard deviation of x for the entire archive
n	Number of items in the sample
N	Size of the entire archive
$\overline{x}$	Average value of all x_i in the sample
s	Standard deviation of x_i in the sample

To start, we assume a normal distribution with the probability $p(x)$ of the random variable to have a value x that is shaped as a Bell curve (Figure 10).

After applying sampling to the population (i.e. the entire archive), we again determine the probability distribution of the sample and compute the average $\overline{x}$ and standard deviation s (given x_i as the values in the sample and n the sample size):

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum x_i \\ s &= \sqrt{\frac{1}{n}\sum (x_i - \overline{x})^2} \end{aligned} \tag{1}$$

To allow the sample to be an accurate representation of the population that can be used to make valid conclusions about the initial archive, we desire the sample average $\overline{x}$ to approach the population average μ as closely as possible. The central limit theorem states that the sample averages $\overline{x}$ taken from different samples will themselves be normally distributed, with an average μ

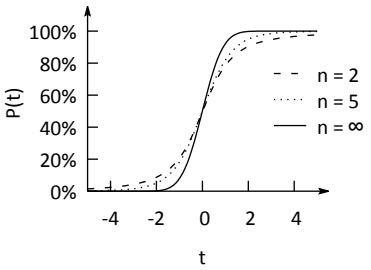

Fig. 11 Cumulative distribution of Student's t-distribution.

and a standard deviation $\sigma/\sqrt{n}$. This theorem remains valid (i.e. $\overline{x}$ remains normally distributed) when x itself is not normally distributed, provided that the values x_i are chosen independently [Hogg and Craig(1978)]. This means that when the sample is sufficiently large, the sample average will approach the population average.

It also allows us to construct a probability distribution that indicates how far the sample average is removed from that of the population. To this end the t-test can be used that is based on the variable t given by:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \tag{2}$$

The value of t hence indicates how far the sample average $\overline{x}$ is removed from an unknown, correct value μ (the population average), expressed as a factor of $s/\sqrt{n}$ which can be computed from the sample. The variable t itself is distributed according to the Student's t-distribution.

Figure 11 illustrates this distribution. When $n = 5$ for instance, the probability of $t \leq 2$ is approximately 95%. Using the original definition of t (Equation 2), this implies that the probability of $\overline{x} \leq \mu + 2 \cdot s/\sqrt{n}$ equals 95%, or that the probability that the sample average has a positive error is larger than $2 \cdot s/\sqrt{n}$ is 5%. Usually both positive and negative errors are considered and a preset total confidence level is used (for example 99%). For a sample size of $n = 5$ we can see that $t < -4.03$ with a probability of $p = 0.5\%$, or that $t > +4.03$ for $p = 99.5\%$. Taken together both statements say that, with a 99% confidence level, $|t| < 4.03$.

Lastly, we can use the values obtained for t to estimate the error on the sample average $\overline{x}$. To this end, we use the equation given by the t-test:

$$|\overline{x} - \mu| < t \cdot s/\sqrt{n} \tag{3}$$

We can also compute the confidence interval to calculate the bounds inside which the sample average $\overline{x}$ will lie given the chosen probability:

$$\mu - t \cdot s/\sqrt{n} < \overline{x} < \mu + t \cdot s/\sqrt{n} \tag{4}$$

The value of t depends on the chosen confidence level (usually 95%) and on the sample size n. Computing t is rather complex and was previously taken from

precomputed tables,[11] but today the value can easily be computed by using software such as Microsoft Excel for arbitrary values of p and n. To calculate these values, the built-in functions `T.DIST` (computes the probability for a given value of t) or `T.INV.2T` (when determining t, given a probability) are available.

In relation to the confidence interval, we would also like to know how we can make the sample as accurate as possible. Equation 4 shows that the quality of the sample (how close $\overline{x}$ approaches μ) depends on the absolute size of the sample n (through the factor $\sqrt{n}$ or the value of t) but also on the standard deviation of the sample s (the intrinsic variability of the population). The latter factor cannot be controlled by the professional but should nevertheless be taken into consideration. Archives with a large internal variation require a larger sample to ensure that, within reasonable bounds, the sample remains representative. The method outlined in this appendix can be used to decide whether the expected sample error is acceptable (given the practical bounds of the sample size), or whether a larger sample size should be considered.

Note that thus far we have only considered the absolute size of the sample n, but not the size of the total population or the sampling ratio. The statistical models used here assume an infinite population to simplify any mathematical calculations.

For very large populations the ratio between the sample size n and the population size N will be so large that the assumption of an infinite population is accurate, so the confidence interval computed earlier is a good approximation. Conversely, when the sample covers a very large fraction of the total population or the entire population, the sampling error will approach zero.

For moderate sampling ratios it is still possible to find aberrant values in the fraction of the population that did not make it into the sample, but as soon as we increase the sample size the probability of this happening decreases. Equation 4 offers a good approximation for small sampling ratios. For large sampling ratios, a correction factor can be applied to Equation 1 which reduces the size of the confidence interval:

$$s = \sqrt{\frac{1}{n}\sum (x_i - \overline{x})^2} \cdot \sqrt{\frac{N-n}{N-1}} \tag{5}$$

In practice, this correction is applied only when the sampling ratio is higher than 5%.

B Confidence intervals in Excel

Computing confidence intervals and applying sampling is relatively straightforward when using software such as Microsoft Excel:

1. The starting point is the creation of an overview of all items in the archive, in which each item is uniquely identified by means of a code. In our example this code is recorded in column `A`.

[11] See for instance [Som(1973)], p 339.

2. For each item, we also need the numerical value of the chosen random variable in column `B`.
3. In column `C`, the `=RAND()` formula is applied to assign a random value between 0 and 1 to each item.
4. To prevent Excel from regenerating the random numbers each time the sheet is recalculated, copy and 'paste as values' the column `C` values in the same column.
5. Sort the sheet in ascending order by the random value in column `C`.

Once sorted, the sample can be constructed by taking the topmost n items. The size of the sample n does not need to be fixed yet. Instead, a number of different sample sizes can be simulated together with the corresponding confidence interval by means of the following technique (see Table 1 for an example Excel worksheet):

1. Compute the average and standard deviation for the chosen sample size (i.e. the topmost n items). For $n = 100$ this yields `=AVG(B1:B100)` for the average and `=STDEV(B1:B100)` for the standard deviation in columns `E` and `F`.
2. The value of t can be computed by means of the t-distribution, the sample size n and the confidence level. For a confidence level of 95% the probability of t exceeding the threshold is 5%, or 0.05 with the formula `=T.INV.2T(0.05, 100-1)` (column `G`).
3. The error is calculated with $t{\cdot}s/\sqrt{n}$ or with the sample size, standard deviation and t-value in cells `D7`, `F7` and `G7` of Microsoft Excel: `=G7*F7/SQRT(D7)` (column `H`).
4. The confidence interval is the sample average plus or minus the error `=E7-H7` and `=E7+H7` (the lower and upper bound; columns `I` and `J`).
5. Lastly, the relative error can be computed as `=H7/E7` (column `K`).

Note that the random variable value is only needed for those items that are part of the (potential) sample. Where determining these values is not straightforward, it suffices to measure the value for the first n items. We can in other words assume an initial sample size of 100 items and measure their random variable, compute the resulting confidence interval and decide whether the resulting sample is sufficiently representative. If not, we can increase the sample size to 200, measure the random variable for items 101–200 and repeat the process. This way we do not only execute the selection incrementally, but also perform an incremental analysis of the archive.

(a) Example Excel worksheet

	A	B	C	D	E	F	G	H	I	J	K
1	Item#	Length	Random value	**Population**							
2	9881	84	0.0001495	*n*	Average	Std. deviation					
3	5982	83	0.0002044	15143	74.07	13.27					
4	10898	79	0.0002474								
5	6922	53.25	0.0003674	**Sample**					Confidence interval		
6	12498	82	0.0004078	*n*	Average	Std. deviation	*t*-value	Error	Lower	Upper	Rel. error
7	4953	64	0.0005148	5	76.25	12.99	2.78	16.13	60.12	92.38	21.16%
8	3014	62.5	0.0007487	10	69.38	11.31	2.26	8.09	61.28	77.47	11.66%
9	7769	61.5	0.0007685	15	73.28	11.13	2.14	6.16	67.12	79.45	8.41%
10	2118	62.5	0.0008038	20	72.61	11.34	2.09	5.31	67.30	77.92	7.31%
11	6126	62	0.0008259	50	74.32	11.88	2.01	3.38	70.94	77.69	4.54%
12	11963	77	0.0009720	100	74.35	11.86	1.98	2.35	72.00	76.71	3.16%
13	7686	79	0.0009779	1000	73.74	13.52	1.96	0.84	72.90	74.58	1.14%
14	13252	88	0.0009906								
15	7287	75.5	0.0010857								
16	9185	86	0.0011589								
17	998	62	0.0012145								
...											
15144	3604	62	0.9997647								

(b) Excel formulas used to compute (a)

	D	E	F	G	H	I	J	K
1	**Population**							
2	*n*	Average	Std. deviation					
3	15143	=AVERAGE(B2:B15144)	=STDEV(B2:B15144)					
4								
5	**Sample**					Confidence interval		
6	*n*	Average	Std. deviation	*t*-value	Error	Lower	Upper	Rel. error
7	5	=AVERAGE(B2:B6)	=STDEV(B2:B6)	=T.INV.2T(0.05,D7-1)	=G7*F7/SQRT(D7)	=E7-H7	=E7+H7	=H7/E7
8	10	=AVERAGE(B2:B11)	=STDEV(B2:B11)	=T.INV.2T(0.05,D8-1)	=G8*F8/SQRT(D8)	=E8-H8	=E8+H8	=H8/E8
9	15	=AVERAGE(B2:B16)	=STDEV(B2:B16)	=T.INV.2T(0.05,D9-1)	=G9*F9/SQRT(D9)	=E9-H9	=E9+H9	=H9/E9

Table 1 Example Excel worksheet and formulas used to compute confidence intervals with simple random sampling and different sample sizes.