Luca Gianazza

Independent researcher, professional engineer

APPLYING STATISTICS AND COMPUTER SCIENCE TO THE STUDY OF BIG COIN FINDS: AN ENGINEERING APPROACH

Abstract

Any large amount of data raises processing and interpretation issues. Coin finds, particularly hoards made of several thousand pieces, are no exception. In front of a great number of specimens, a comprehensive study, conducted with methods usually applied to small finds, becomes a difficult target to achieve. Statistics, as well as Computer Science, can provide important analysis tools and solutions allowing the researchers to extract relevant information from finds data. This contribution will examine how Statistics and Computer Science can support the work of numismatists. It will present at an introductory level what is still available today and what could become affordable hopefully not too far in the future, going through the major pros and cons. It will be shown how large and articulated amounts of data – from denominations of coins to the mints of origin, from image descriptions to weights and diameters – can be managed and organized in a smart way along with coin images into a structured information system.

The analysis will be carried out under an engineering perspective, always focusing on aspects such as application limits, implementation costs and the effort required in terms of human resources.

Keywords

Coin hoards, Statistics, Computer Science, point estimators, interval estimators, linked data, open data, semantic web, speech recognition software

STUDYING THE BIG COIN FINDS: FIGHTING AGAINST THE WINDMILLS?

A simple calculation highlights the complexities inherent in the study of large monetary finds. Let's assume we want to study the well-known Reka Devnja hoard, and to publish it in full. To do this, we assume that we have all the coins under optimal conditions, considering an eight-hour working day and a 250-day working year.

A count of the 81,096 coins – the surviving part of the hoard[1] – performed at the speed of one coin per second requires no less than 2.82 man-days of work.[2] For the registration of every single coin in a database where the essential characteristics are to be reported (description, legends, weight, module, axis orientation...) together with a photographic reproduction of both sides requires that a great commitment must be taken into account. Assuming to limit the storage time to just 10 minutes per specimen, it follows that the complete acquisition of the hoard requires 1,689.5 man-days, or 6.76 man-years. Finally, the publication of these materials in a book or a series of books according to the publishing standards generally proposed by the major magazines of the sector, such as the *Numismatic Chronicle*, would require no less than 2,700 pages and 4,050 plates.

These numbers could be sufficient to provide concrete evidence of the huge effort required to study one of the largest hoards ever found. New perspectives can however be added by associating the costs that should be sustained for each single man-hour employed for the study operations, expressed now in more prosaically budgetary terms. Estimating a charge of 300 euro per man-day (a conservative estimate, if we consider the operating costs of a large public structure that could support such an initiative) we obtain that the only count of the coins would lead to a cost of 850 euros, whereas their registration in a database over half a million euros. Much, much more difficult is any prediction of the costs associated with any print publication of the hoard without a precise editorial plan.

These data might be questionable, but it is difficult to think that they overestimate the time required for the study the hoard and thus the associated costs. In presenting the classification project of the coins of the Kunsthistorisches Museum in Vienna, Klaus Vondrovec spoke of average processing times of a single coin in the order of one hour. Furthermore, the digitalization of the approximately 108,000 coins composing the Misurata hoard, although it has been proceeding intensively for several

[1] Moushmov 1930; Moushmov 1934; Metcalf 2002; Paunov, Prokopov 2002, 48-50 n. 75. For a critical summary of the contents, see also the corresponding record on the *Coin hoards of the Roman Empire* portal (http://chre.ashmus.ox.ac.uk/hoard/3406 – URL last visited on February 1st 2019).

[2] Under the assumption that we are not making any mistake which obliges us to repeat the operation more than once.

years, has hitherto covered just about 80% of the specimens, and only a small fraction of them is now available to the public.[3]

In our example we implicitly worked under the assumptions that all the coins of the hoard were available to the scholars for the classification. We have excluded any costs arising from the recovery of materials from the findspot, their cleaning and restoration. We have not considered the risk that the project may undergo changes for a variety of reasons, such as budget reduction or the decision of a scholar to terminate his or her collaboration. All these elements can lead to an increase in the time (and, consequently, an increase of the associated costs) needed to undertake a comprehensive study of the coins. Employees' turnover, particularly, is a major factor, as it deprives the project of experience and familiarity with the mechanisms associated with the management of materials in the daily operations. Provided that the conditions for substitution can still exist, and we should not proceed with one less resource in the work team, obviously, with a slowdown that at this point would become structural.

But above all, there was no assessment of the problems, and therefore once again of the timespan for their resolution, linked to the accuracy of the data. We cannot assume that reading a coin and entering its data into a database are completely error-free operations. We can introduce additional controls (at the price of an additional effort, and therefore of an additional slowdown), but the probability of error can only be reduced to a tolerable value, never to zero.

Besides, we must keep in mind that in several cases (a not-insignificant number) the study of a large monetary find cannot start from the direct observation of the coins, but is only based on previous studies, where similar errors may have profoundly affected the reliability of the data that we now want to re-examine. The case of Reka Devnja's hoard is once more exemplary. Marguerite Spoerri has pointed out how the texts that in the past have presented the coins of this hoard do not propose an exact correspondence between the description of the specimens and their reference to the volumes of the *Description historique des monnaies frappees sous l'Empire Romain* by Henry Cohen. There is therefore a great difficulty, if not a clear impossibility, in reconstructing the exact contents of the hoard, and consequently in using this data effectively to carry out a more in-depth study.

For the Medieval and Modern ages the situation is further complicated by a lower level of knowledge as compared to e.g. the Greek and Roman world. The classification of coins may be more complex due to the greater fragmentation of the monetary context, which translates into a wider heterogeneity of the coins usually present in the hoard. This may require the involvement of very specific skills, often difficult to find, especially concentrated in one single person. In several cases we are also forced to confront a very unsatisfactory bibliography, obsolete or of poor quality, where

[3] Garraffo, Mazza 2015; http://www.tesorodimisurata.it (URL last visited on February 1st 2019).

primary issues such as the determination of the mint of a coin, its dating, when not its very name, cannot be unequivocally clarified.

The problems highlighted here are not restricted to the case of a single monetary hoard, however large, but can also be extended to the study of several finds in which the number of coins appears more manageable. A project aiming only at the minimal inventory of all the finds, like the *Inventory of Greek Coin Hoards* or those volumes of the *Medieval European Coinage* that have chosen to dedicate a specific chapter in the appendices to the monetary finds, must deal with a lot of complexities related to the volumes of data to be processed, their lack of accuracy, their dispersion in publications hard to come by, the obsolescence of the classifications proposed, particularly when not accompanied by illustrations.

Such a scenario may seem bleak and lead us to the conclusion that a complete and scrupulous study of a big hoard, as well as a large set of finds, cannot be considered anything but a chimera: unworkable because of the effort and the prohibitive costs, when not for the difficulty in finding the right skills that such an operation could require. In a context where such a considerable expense would hardly give an economic return, as happens for example in the private industry, where an expense (more precisely, an investment) is made with the aim of making a profit. In the study of a monetary find the costs would remain merely costs, they cannot be regarded as investments.

If we do not want to give up our goal, it becomes necessary to increase operational efficiency. We therefore need to develop methodologies to reduce effort and costs without compromising either data accuracy, nor the validity of the information that can be obtained from them.

Each hoard is unique, but the coins of which is composed are a serial product. And this is reflected in the common elements (i.e., repeated occurrences) not only inside a given hoard, but in different other finds. Going back to the example of Reka Devnja once more, we notice right away how many of the surviving specimens are showing identical characteristics of others. Exploiting repetitiveness to reduce redundancy represents the simplest way to minimize the time needed for the study. Why, for example, in a database should we repeat the entry of inscriptions, descriptions or other data for a given coin, when the same operation has already been performed for another coin completely identical to it?

But we can go further, taking the concept to the extreme up to give rise to a question that in some ways may seem paradoxical: why should we study the hoard as a whole when, because of the repetitiveness of the characteristics of the specimens present in it, we could focus on a subset of his coins only? Of course, the subset must have precise requirements. First, it must "small", so that it can be studied with adequate meticulousness in a reasonable amount of time. It must also be sufficiently "informative" to allow the researcher to extract from it all the considerations that could be derived from the study of the find in its entirety.

Today there are informatic, mathematical and above all methodological solutions (we remind here the concepts of *lean thinking* and *lean production*) which find an increasing application in the most disparate contexts, but not enough in the field of the Numismatics and Human Sciences in general, where the precepts of Digital Humanities are still struggling to find a great diffusion.[4]

In this paper we will try to examine the ways in which Statistics and Computer Science can meet the needs of Numismatics to improve the efficiency of the study of large coin finds without jeopardizing data accuracy and information that can be derived from them. The discussion will be conducted under an engineering perspective, therefore mainly oriented to contextualize tools and solutions made available by Statistics and Computer Science to the area of interest, with the aim of highlighting their potential and limits.

What presented here does not claim to be exhaustive. The discourse is extremely articulated: both Statistics and Computer Science are very vast subjects, with many facets, and their discussion in numismatic terms cannot in any way be contained in the few pages of an essay.

In presenting some concepts, particularly related to Statistics, we will have to implement considerable simplifications, to stress the most important points and offer formulas that can be applied immediately. Anyone wishing to engage in an in-depth examination of problems and theorems has a wide variety of publications at his disposal, in all the languages of the world.

In the section dedicated to Information Technology there will be no explicit reference to specific software. This is a deliberate operation, dictated by the awareness that Computer Science is evolving so rapidly that any indication in this sense would risk becoming meaningless long before any need for a new congress updating the discussions, the results obtained and therefore also the software proposed today. A technology or software that *today* appear to be essential for the development of any application in a specific context could result obsolete *tomorrow*. We therefore prefer to provide general but clear indications on the problems to be addressed rather than on the specific solutions that can be adopted today for their resolution, also in consideration of the fact that (obviously) only a full awareness of the problem can lead to an optimal solution.

[4] An overview of the relationship between Numismatics and Computer Science is provided in WIGG-WOLF 2009 and PETT 2015. The papers are proposing two pictures considerably different. If in the 2009 text the projects discussed mainly concern the implementation of relational databases, in the 2015 edition we can observe a wider diversification of research directions and a wider attention to the sharing of resources through the World Wide Web.

SO, LET'S TALK ABOUT STATISTICS…

Statistics applied to Numismatics is by no means a novelty. For decades we have observed its application in the study of coins, although in most cases with little awareness of its methodologies and its potential. The use of Statistics has mainly occurred in areas such as the search for the number of dies associated with a given issue, or the estimate of production volumes of a single die. We developed models and formulas that today are the basis of all the debates on the quantitative aspects of the monetary production of a mint, but the authors who actively contributed to their realization or their dispute were a very small fraction of the scholar numismatists, while most of the authors who were inclined to their adoption have opted for an uncritical utilization.

But Statistics has found an application – mostly unaware – even more extensive in all those situations in which graphs were drawn, averages were calculated, assessments were made on percentages or numbers of occurrences. All this, in fact, falls mainly in one of the branches of Statistics that goes under the name of *Descriptive Statistics*. Within Descriptive Statistics we can think grouped all the tools and methodologies for the analysis of a set of data aimed at obtaining new quantities that summarize the characteristics of a sample. If I have N coins, each weighing $n_1, n_2, \dots\ n_N$ grams, when I calculate the arithmetic mean $(n_1 + n_2 + \ \dots\ + n_n)\ /\ N$ I'm summarizing the characteristics of a set of N data in a single quantity, obtaining additional information. Likewise, if I sum up the weights of these coins into tables, for example counting how many of them weigh less than I_1 grams, how many between I_1 and I_2, how many between I_2 and I_3, etc ... I'm processing my data to get indicators that can give me a new summary of the N coins from which I started. The same happens if I count how many of these coins come from the mint Z_1, how many from the mint Z_2, ... Finally, if I decide to plot a graph with the weight distribution that I have previously summarized in a table, or a map showing the mints of origin, I am performing once more operations that fall within the field of Descriptive Statistics.

In the search for effective ways to study "big" monetary finds, the Descriptive Statistics shows relevant limits of use. Processing the data related to N number of coins implies that all these N coins have already been counted, measured, classified… It means that for all of them I have an ideally complete and accurate set of data. But it also means that the hoard has already been examined in its entirety, and therefore a potentially remarkable effort has been already spent on it.

More help can come from a different branch of Statistics, which goes under the name of *Statistical Inference*. It includes the processes of using data analysis to deduce the properties of a population starting from a reduced set of data extracted from it. If I have a find consisting of N coins, where N is too high a figure to allow the study of the whole set of coins, I could think of extracting only $M \ll N$ coins and make my evaluations on this smaller set of specimens (much more manageable), and from it inferring the properties of the starting set of N coins using the methodologies made available by the Statistical Inference. The effort needed to classify and study M coins is obviously lower than that required to study N (much lower if, $M \ll N$). If from these coins we are able to understand properties that can be extended with a sufficient degree of reliability to the set of N coins from which we started, we get an undoubted advantage from the operation.[5]

To do this, we must ensure that precise conditions are met. Conditions that, however, hardly occur in the context of our interest.

The problem is very effectively summarized by this sentence by Warren W. Esty: *Unfortunately, hoard data are not the ideal "experimental" data treated in statistics texts. Numismatic analyses are often complicated by small sample sizes and non-randomness, which may invalidate statistical conclusions*.[6]

It is unlikely that a hoard can be considered as the result of random sampling of coins in circulation. The coins present in a hoard have not been chosen by lot as balls from an urn, but rather tend to be the result of a precise selection among those available, in turn a subset of those actually in circulation.[7] The fundamental criterion of randomness at the basis of the constitution of the restricted sample is therefore not satisfied, with the result that the considerations that can be inferred could be nothing more than misleading information.

The two limits presented by Esty to the use of hoards in the study of larger coin populations – samples too small, not randomly extracted – may be overcome if the hoard is not the sample taken from population, but rather in itself constitutes the population to be investigated. From this population-hoard of N coins it would be possible to extract a real random sample of M coins, and on this sample apply the methodologies of Statistical Inference to obtain the desired information on the characteristics of a starting data set.

Before proceeding any further with the operation that we have set out to accomplish, it is necessary to introduce a mathematical notation intended to facilitate the presentation of the concepts related to Statistics.

[5] From a certain point of view, in this approach we can recognize the same principle that leads the scholar numismatists to examine the finds – regardless of their numerical consistency – and based on them make considerations (sometimes risky) about the monetary circulation of a given area and/or in a given epoch, or the production volumes of a given mint.

[6] Esty 2005, 173.

[7] Subset, however, influenced by external factors of different nature – geographic, economic, social – not easily identifiable, nor quantifiable.

Our hoard-population of N coins can be modeled as an array of N independent random variables $\mathcal{H}_N = (X_1, X_2, \dots X_N)$. Each random variable X_i represents a single coin, or more generally an object belonging to the hoard, and is characterized by a series of specific *qualitative* (e.g., mint, depictions, inscriptions…) and *quantitative* (e.g., weight, axis orientation…) properties.

A sample of M coins randomly extracted from this population may in turn be modeled as an array of M independent random variables $\mathcal{H}'_M = (X'_1, X'_2, \dots X'_M)$. Also in this case each random variable X'_i represents a coin, characterized by specific qualitative and quantitative properties similar to those observed in the hoard-population. In fact, $\mathcal{H}'_M$ is a subset of $\mathcal{H}_N$ ($\mathcal{H}'_M \subseteq \mathcal{H}_N$), and therefore X'_1 belongs to both $\mathcal{H}'_M$ and $\mathcal{H}_N$.

Assuming that the properties expressed by the coins in the sample $\mathcal{H}'_M$ correspond perfectly to those of the coins belonging to the hoard-population $\mathcal{H}_N$, I would be in a position to study the properties of the set $\mathcal{H}'_M$ and extend them to the set $\mathcal{H}_N$, thus succeeding in understanding the characteristics of the set of N coins simply by analyzing a group of $M \ll N$ coins of identical nature. In mathematical terms, this would be possible if the *probability distribution function* of the M random variables were identical for all the coins in the set $\mathcal{H}_N$ and in the set $\mathcal{H}'_M$.

But, of course, this is an assumption that does not appear to be verified in any real situation. Not all the coins in a hoard, in fact, have identical characteristics. Let's use the example of Reka Devnja hoard once more (N = 81.096 pieces) to evaluate the most immediate consequences of such a situation. More specifically, let's focus on the "types" represented in it by subdividing the coins according to the *Roman Imperial Coinage* (*RIC*). Each "type" will be characterized by properties in whole or in part different from any other "type" (for example, they may or may not share the same metal, the same iconography… but there will be at least one discordant element, the one that precisely leads the *RIC* to introduce two different reference numbers), while all coins belonging to the same "type" will have identical properties (see tables 1.a-b).

The most common type is represented by a *denarius* in the name of Julia Maesa *Augusta* (*RIC* 268), with 547 pieces. This is a number that is anything but small in absolute terms, but which represents only the 0.67 % of the coins in the hoard.

In the Reka Devnja hoard we can recognize over 3,300 distinct types. It means that each type is represented on average with approximately twenty copies. The ten most represented types contribute with just 3,913 specimens, equal to a modest 4.83% of the total. The assumption that it is possible to carry out a random sampling of $M \ll N$ specimens when such a small fraction is representative – even with approximations – of the N coins of the whole hoard looks totally implausible, be-

Table 1 – Most represented "types" in the Reka Devnja hoard (a) and occurrences of the same "type" (b) (*source*: http://chre.ashmus.ox.ac.uk/hoard/3406)

Authority	*Coin*	*Mint and date*	*Reference*	*Number of specimens*
Julia Maesa (*Augusta*)	Denarius	Rome (218/22 CE)	*RIC* 268	**547**
Faustina I (*Diva*)	Denarius	Rome (141 CE)	*RIC* 351a	**498**
Julia Mamaea (*Augusta*)	Denarius	Rome (225/35 CE)	*RIC* 343	**467**
Maximinus I Thrax (*Augustus*)	Denarius	Rome (235/6 CE)	*RIC* 14	**418**
Faustina I (*Diva*)	Denarius	Rome (141 CE)	*RIC* 344a	**390**
Marcus Aurelius (*Caesar*)	Denarius	Rome (145/60 CE)	*RIC* 429a	**338**
Julia Mamaea (*Augusta*)	Denarius	Rome (225/35 CE)	*RIC* 360	**319**
Julia Maesa (*Augusta*)	Denarius	Rome (218/22 CE)	*RIC* 271 or 272	**316**
Faustina II (*Augusta*)	Denarius	Rome (161/75 CE)	*RIC* 677	**311**
Faustina I (*Diva*)	Denarius	Rome (141 CE)	*RIC* 362	**309**
Julia Domna (*Augusta*)	Denarius	Rome (196/211 CE)	*RIC* 574	**300**

types with 200 to 299 specimens each	**28**
types with 100 to 199 specimens each	**146**
types with 10 to 99 specimens each	**1,322**
types with 2 to 9 specimens each	**1,156**
types with 1 specimens each	**689**

cause we would hardly give evidence to all the types represented by very few specimens, which make up the majority of the hoard.

The considerations would have been different if my hoard had been constituted by a much smaller number of types, in the order of few units, each represented with a sufficiently high number of specimens to be adequately present in my sample of $M \ll N$ specimens. But we know how the hoards tend to have a somewhat heterogeneous nature in terms of content, and a multiplicity of types what we should expect to find. We also know how often it is a single specimen out of N to provide the most important contribution for the dating of the complex, or for its precise characterization.

A sample of M coins therefore is unlikely to reflect the starting population constituted by our hoard of N coins if we consider how different the coins composing it can be: different nominals, different weighting standards, different typologies … This would require a proper mathematical model, but such a model may have a huge complexity, and cannot find a practical application unless introducing several simplifications. Otherwise the sample $\mathcal{H}'_M$ should be "very big" so that it is adequately representative of the starting population $\mathcal{H}_N$.

The idea of being able to work on a "small" fraction of the "big" hoard clashes with what just discussed, leading in the first instance to the conclusion that the concepts and methods of Statistical Inference cannot be used for our purpose. Such a statement is not entirely correct.

Thanks to the example of Reka Devnja we are now aware of the limits of application of these methods. An *exhaustive study of all the properties* of a hoard $\mathcal{H}_N$ of N specimens through a subset $\mathcal{H}'_M$ of $M \ll N$ random samples is not feasible due to the heterogeneous nature of the hoards: the variability of the types usually present in the finds is too wide to suppose that $\mathcal{H}'_M$ is characterized by the same properties of $\mathcal{H}_N$, and the probability that a single specimen that can radically change the interpretation of the hoard is not present in $\mathcal{H}'_M$ is too high.

If, however, the variability of occurrences of a *single characteristic*, or at most of a *reduced number of characteristics*, appears to be much smaller than what we observed in the Reka Devnja hoard, it becomes possible to carry out a statistical analysis on a small subset of specimens. Let's consider, for example, the metal a coin is made of, rather than its origin or not from a given mint. With the metal the situation is clear: a quick analysis of the color of the coin easily allows to discriminate between few categories (gold, high-quality silver, low-quality silver, copper / bronze). With mints, the number of options can increase considerably, but if we limit ourselves to the study of the most represented ones we can provide a rough estimate of the distribution of the mints inside the entire hoard. An estimate that, of course, will always have margins of error, but that can still bring important information for example in case of a preliminary study of a hoard that cannot be managed in its entirety, to be carried out quickly and at reduced costs.

The application of the methods of Statistical Inference in the areas described above (although much reduced with respect to the initial goal) requires starting from a *model*. It is necessary to identify the correct probability distribution function, namely that mathematical function that best represents the probability with which the specific property we want to study shows up. The property modeling must precede any consideration that can generally be labeled as "statistics". It is an operation that has inevitable degrees of subjectivity, but that experience, common sense and precise evidence obtained from other areas can help to carry out successfully.

Qualitative properties (e.g., metal, mint of origin, denomination...), that is to say discrete quantities, can be efficiently modeled through a *multinomial distribution*:

$$f(x_1, \dots, x_m; n, p_1, \dots, p_m) = \mathrm{P}(X_1 = x_1, \ \dots, X_m = x_m) =$$
$$= \begin{cases} \dfrac{n!}{x_1! \dots x_m!} \cdot p_1^{x_1} \cdot \dots \cdot p_m^{x_m}, \text{when } \displaystyle\sum_{i=1}^{m} x_1 = n \\ \text{0, otherwise} \end{cases}$$

where p_i expresses the probability that the i-th event between the possible m occurs. In the case of the mints mentioned above, it can be interpreted as the "probability that the coin extracted from the reduced sample of M specimens were minted by the mint i", where i may be Milan, Paris, Lyon… or any other relevant mint for that specific hoard.

With M "big" enough, the factorial term introduces significant complexities. In fact, its calculation requires computational tools far more powerful than a domestic personal computer, or at least the use of a normal approximation.

If the possible options are reduced to only two (e.g., "the coin belongs to the mint Z" and "the coin does NOT belong to the mint Z", $m = 1$) the previous formula is simplified and becomes what we call *binomial distribution*:

$$Bin(n, p)\colon \ f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0,1,2, \dots, n$$

If it were also $n = 1$, i.e. one single occurrence examined at a time and no longer n at the same time, the binomial distribution is further simplified, and we have the so-called *Bernoulli distribution*:

$$Bernoulli(p)\colon \ f(k, p) = p^k (1-p)^{1-k} \text{ for } k \in \{0,1\}$$

Quantitative properties (e.g., weight, module), that is to say continuous quantities, can instead be modeled through a *normal distribution*:

$$N(\mu, \sigma^2): f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$$

under the assumption, however, that all the M coins of the sample have homogeneous characteristics (e.g., coins all based on the same weight standard).

Alternative models, i.e. probability functions different from those proposed here, can naturally find a valid application in the study of monetary finds.[8] The choice of the models we wish to adopt cannot in any way ignore the specific properties that will be analyzed, and above all the consistency of the set of samples/coins that we intend to study. For example, we cannot think of using a normal distribution to model the representation of mints in a sample, whereas a multinomial distribution could be used for the study of quantitative properties if we group continuous measurements – expressible as a group of values potentially infinite – in a finite number of intervals (e.g., expressing the weight of the i-th coin as x_i, we can discretize the weight data by simply counting the number of samples that fall in one of the $K + 2$ intervals $x_i < h_o$, $h_0 \leq x_i < h_1$, …, $h_{K-1} \leq x_i < h_K$, $x_i \geq h_K$ in which I have chosen to divide the set of measured weights).

The adoption of a specific distribution function is not in itself sufficient to consider our operation of modeling complete. We need the parameters that appear in it (e.g., p_i for the multinomial distribution, μ and σ for the normal distribution) to be properly evaluated. And this is precisely what constitutes the most critical part of our research, since these parameters are unknown. Knowing exactly the values assumed by p_i in the distribution of the mints modeled by a multinomial distribution, for example, would mean knowing exactly the percentage of the N coins in my hoard minted by the i-th mint: I would have already carried out the complete study of the N coins of my find, and any use of Statistical Inference for my considerations at this point would be useless.

It is however possible to *estimate* these parameters using *estimators*, that is to say functions of the M random variables $X_1, X_2, \dots X_M$ representing the M coins of the "small" sample that I want to study.

For our purposes we can simplify the discussion by subdividing the estimators into just two categories: *point estimators* and *interval estimators*. *Point estimators* allow to determine a single value that can be taken as "best estimate" of a parameter θ_i $(1 \leq i \leq k)$ associated with my probability distribution function. *Interval*

[8] Consider for example the uniform probability distribution for the axis orientation (see PARISOT-SILLON, SUSPÈNE, SARAH 2014 for an example of detailed statistical analysis on this aspect).

estimators lead to the determination of a range of “plausible” values for the parameter θ_i $(1 \leq i \leq k)$.

For each distribution function it is possible to define several different point and interval estimators, each with peculiar properties that can make it more indicated than others in some areas of application and not in others. There is no estimator capable of make a “perfect” estimate, i.e. an exact, error-free indication of the quantity that we aim to estimate. As functions of random variables, these estimators are themselves random variables, and therefore linked to a probability distribution function. This means that each estimate is in turn subject to a probability. When we obtain an estimate $\hat{\theta}_i$ of the parameter θ_i, this $\hat{\theta}_i$ does not necessarily represent the exact value of the parameter θ_i, but rather an evaluation that can be intended – depending on the different meanings and the nature of the estimator itself – as “the best possible” , “the most probable” or “sufficiently accurate”. Which is the level of uncertainty behind these words, and therefore how “valid” this estimate will be, will depend on the chosen estimator and indirectly on some characteristics of the sample examined, first of all its size M.

It is not the purpose of this paper to examine in detail the properties of the estimators, nor to discuss the different options available for estimating the same parameter. It will be here sufficient to present some examples of punctual and interval estimators for the bernoullian, binomial and normal distributions that can be easily used by the scholars wishing to carry out a statistical study of a hoard or in general of objects that can be modeled by one of these probability distribution functions.

For simplicity, the proposed notation will reflect the one most widely used in literature, where n constitutes the size of the sample under investigation. For consistency with what has been discussed up to now, we must put $n = M$.

POINT ESTIMATORS

- bernoullian and binomial distribution

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

- normal distribution

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

INTERVAL ESTIMATORS

To define such a type of estimators, it is always necessary to set a *confidence interval* r (e.g., $r = 0.05$), i.e. our interval of values such that the probability that the quantity we want to estimate falls within it is equal to $1 - r$ (with $r = 0.05$, this probability is equal to 95%).

- binomial distribution

$$p \sim \left[\bar{X}_n - z_{1-\frac{r}{2}} \cdot \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}; \bar{X}_n + z_{1-\frac{r}{2}} \cdot \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$$

under the assumptions that $\bar{X}_n$ and $1 - \bar{X}_n$ are not close either to 0 or to 1, $n(1 - \bar{X}_n) > 5$ and $n\bar{X}_n > 5$.

- normal distribution

$$\mu \sim \left[\bar{X}_n - t_{n-1,1-\frac{r}{2}} \cdot \frac{\boldsymbol{S}_n}{\sqrt{n}}; \bar{X}_n + t_{n-1,1-\frac{r}{2}} \cdot \frac{\boldsymbol{S}_n}{\sqrt{n}}\right]$$

$$\sigma^2 \sim \left[(n-1)\frac{\boldsymbol{S}_n^2}{v_{n-1,1-\frac{r}{2}}}; (n-1)\frac{\boldsymbol{S}_n^2}{v_{n-1,\frac{r}{2}}}\right]$$

where

$\bar{X}_n$ is the sample average over n samples

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \cdots + X_n)$$

$\boldsymbol{S}_n^2$ is the sample variance over n samples

$$\boldsymbol{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

z_b is the b-th quantile of the normal distribution $N(0,1)$
$t_{a,b}$ is the b-th quantile of the t (Student) distribution with a degrees of freedom
$v_{a,b}$ is the b-th quantile of the χ^2 distribution with a degrees of freedom

For z_b, $t_{a,b}$ and $v_{a,b}$ there are tables that report the values in function of a, b, but there are also calculation tools easily accessible on the Internet or already implemented in the most popular spreadsheets.[9]

Some of these estimators may appear familiar. It is common in numismatic literature to come across estimations of the standard weight of a given coin issue starting from the sample average of the specimens observed in a hoard. The operation is pretty simple, but the associated inference is not necessarily correct, because it may not take into account the weight reduction due to the circulation, the selection of the heaviest pieces among those circulating for hoarding, and all those elements of uncertainty related to the process of minting (let's think about the concept of *remedium in pondere* and the control procedures on the weight of coins made by a mint before issuing a piece). Above all, in many cases there is a tendency to confer undue value on such an operation. By extending the concept, it would be like asserting that "since the sample average of the weights of a set of M coins is equal to μ grams, that specific type has been minted to a standard of μ grams": a totally arbitrary statement for a series of mathematical and numismatic considerations, which does not differ so much from some assertions that sometimes we find proposed in academic journals.

Other estimators, specifically those proposed for an interval estimate, may be a novelty in Numismatics. Reviewing the most important numismatic journals in the last twenty years, for example, I could not find relevant examples of interval estimation of a statistical quantity as part of the study of monetary finds.

To give an idea of the risks associated with the use of estimators, we can perform two simple estimations on a sample extracted from a population whose properties are no longer uncertain but known, and evaluate the differences between the starting point and the results obtained.

A simulation of a hoard of $N = 100{,}000$ coins has been made through a spreadsheet, assigning to each of them a mint of origin among five possible, here indicated for simplicity with the letters A, B, C, D and E (where E can also be seen as "the set of all the *other* mints", for example because they are poorly represented in the hoard to be treated with sufficient accuracy). Specifically, the simulated hoard consisted of 45,051 specimens of the mint A, 24,978 specimens of the mint B, 10,029 specimens of the mint C, 9,979 specimens of the mint D and the remaining 9,963 specimens of the mint E.

The random extraction of a sample of $M = 100$ coins[10] gave the following results:

[9] See the paragraph … *and now about Computer Science* for further details.

[10] The value of M has been chosen deliberately very small, just to give greater evidence of the risks that can be encountered when an estimate is made starting from a sample "too small" compared to N.

- n_A = 41 coins of the mint A
- n_B = 23 coins of the mint B
- n_C = 14 coins of the mint C
- n_D = 12 coins of the mint D
- n_E = 10 coins of the mint E

We choose to proceed by simplifying the complexities related to the estimators of a multinomial distribution, that is, evaluating each of the mints independently of the others, i.e. taking a reference mint and examining whether a given coin belongs or not to the mint in question. In other words, the presence of coins of the mint A is modeled through a binomial probability distribution function. In the case of the mint A, for example, the sample of N = 100 coins is split into two distinct groups: 41 coins belonging to the mint A, 59 coins NOT belonging to the mint A.

Repeating the same procedure for the other mints, we obtain an estimate of the composition of the hoard rather simple and intuitive,[11] equal to:

- $\hat{p}_{A,100} = n_A \ / \ M = 0.41$
- $\hat{p}_{B,100} = n_B \ / \ M = 0.23$
- $\hat{p}_{C,100} = n_C \ / \ M = 0.14$
- $\hat{p}_{D,100} = n_D \ / \ M = 0.12$
- $\hat{p}_{E,100} = n_E \ / \ M = 0.10$

By estimating the composition of the hoard of N = 100,000 specimens with a confidence interval of 95% (i.e., interval estimation with r = 0.05) we obtain:

- 41,000 ± 9,640 coins of the mint A
- 23,000 ± 8,248 coins of the mint B
- 14,000 ± 6,810 coins of the mint C
- 12,000 ± 6,369 coins of the mint D
- 10,000 ± 5,880 coins of the mint E

If we compare these results with the actual composition (known) of the hoard, we would be led to conclude that our estimation, despite the approximations we chose to adopt, was effective. However, if we evaluate the width of the confidence interval in relative and not absolute terms, we can appreciate with more clarity how wide it is, especially for the mints C, D and E, i.e. those represented by a smaller number of specimens:

[11] In the procedure just described, there are evident inaccuracies related to the implicit assumption of independence of the five causal variables. But the choice to operate in this way is once more intended, just to highlight the limits related to similar operations of estimation.

- 41,000 ± 23.5% coins of the mint A
- 23,000 ± 35.9% coins of the mint B
- 14,000 ± 48.6% coins of the mint C
- 12,000 ± 53.1% coins of the mint D
- 10,000 ± 58.8% coins of the mint E

The success of the estimation has had a cost given by a width of the interval of confidence so extensive (the wider, the smaller the representativeness of a mint) to make complex any evaluation of the result from a purely numismatic point of view.

With the second example we move away for a moment from the study of the great monetary finds to examine the risks associated with the estimation of the weight standards at the root of the issue of a specific coin. This is an operation that we frequently find in literature, in many cases conducted from specimens coming from finds. Based on what previously discussed, it presents several critical issues, primarily since the coins in a hoard are not necessarily the result of a random extraction from the circulating coins, but rather a selection of the "best" pieces to be hoarded.

We will work with the *ducatone* in the name of Vincenzo I Gonzaga (1587-1612) issued by the mint of Casale Monferrato (Piedmont, Italy). We know exactly its weight standard: 31.94 grams of theoretical weight, 0.25 grams of *remedium in pondere*. The documents confirm that these characteristics were never modified throughout the years of Vincenzo's government. We also know that the check of the exact correspondence to the weight standard before putting a coin into circulation was carried out on every single specimen: if a given coin was lighter than 31.94 – 0.25 grams or heavier than 31.94 + 0.25 grams, that coin was re-melted.[12]

Unless the (unavoidable) measurement errors made by the mint officers, we can assume that this control mechanism was sufficiently accurate to ensure that all the *ducatoni* put into circulation were within the desired range, i.e. that an erroneous issue of specimens of non-standard weight occurred with probability very close to zero. By adopting a model based on the normal distribution to describe the weight of *ducatoni* leaving the mint, it is intuitive to put $\mu = 31.94$ grams. Since in a normal distribution 99.7% of the events are in the interval given by $\mu \pm 3\sigma$, we can assume $3\sigma = 0.25$ grams (or, equivalently, $\sigma = 0.083$ grams).

We now perform four different random extractions from a set consisting of all the M *ducatoni* today known for which reliable weight data is available. The first time we extract 10 samples, the second 20, the third 50 and the fourth 100. The

[12] The author of this paper is conducting a specific research on this mint and has a large archive of documents and data about over 2,500 coins, including 110 *ducatoni* in the name of Vincenzo I Gonzaga.

calculation of the punctual and interval estimators for the expected weight tolerance leads to results that are summarized in table 2.

Number of samples	Sample average (grams)	Sample variance (grams^2)	Estimated weight (95% confidence)	Estimated tolerance (95% confidence)
10	**31.32**	**0.192893**	**31.00 / 31.63**	**0.09 / 0.64**
20	**30.94**	**1.778489**	**30.32 / 31.56**	**1.03 / 3.79**
50	**30.86**	**2.389776**	**30.42 / 31.30**	**1.67 / 3.71**
100	**30.75**	**2.757318**	**30.42 / 31.08**	**2.13 / 3.72**

Table 2 – Estimated weight standards of the *ducatoni* in the name of Vincenzo I Gonzaga

The estimates are quite unsatisfactory even as the number of specimens in the extracted sample increases: the weight estimate never returns an interval in which the theoretical value falls, while for the tolerance we obtain values so high that they do not provide significant indications.

In the search for a justification for these results we might think of the effects of wear due to circulation: once the coins were out of the mint, they would have been subjected to a progressive decrease in weight due their use, which could also have been noticeably different from coin to coin. This could explain the reasons behind the low estimate of the theoretical weight (every coin suffers from the effects of circulation, therefore its average weight decreases) and an interval for tolerance much wider than what is established by the *remedium in pondere* (the wear differs from coin to coin, thus potentially increasing the variability). But the deviations from the theoretical values appear too large, especially if we consider that the effects of wear must be estimated in the order of few hundredths of a gram even in the worst cases. We must rather attribute this situation to the heavy clipping suffered by a not negligible percentage of the known *ducatoni*, with reductions with respect to the theoretical weight sometimes close to eight grams.[13]

[13] Coin clipping of the *ducatoni* and extent of the reduction in weight are discussed in GIANAZZA 2017.

The clipping has introduced a very significant alteration in the distribution of the weights of the *ducatoni*. It is therefore not possible to look at all the known specimens as a potential "random sample" extracted from the population made up of all the N *ducatoni* issued in the name of Vincenzo I Gonzaga. The sample of M *ducatoni* known today can still be modeled with a normal distribution, but now characterized by parameters μ' and σ' different from μ and σ that characterize the weight standards of the *ducatone*, with $\mu' < \mu$ (due to clipping) and $\sigma' > \sigma$ (due to the weight reduction applied – if applied – in many different ways). Thus, it can have great deal of validity to estimate μ' and σ', but not μ and σ.

The examples just proposed deliberately represent extreme cases and have been chosen precisely because the estimates obtained could lead the scholar numismatists without a robust mathematical-statistical background to improper considerations.

Reducing the study of a hoard of N coins to a sample of $M \ll N$, from a certain point of view, can lead to a considerable gain in terms of time and costs. But from another point of view it introduces additional "costs" due to the "margin of error" of any estimation. The researchers who will have to choose whether to examine the hoard of N coins integrally rather than proceeding through a more limited evaluation of a subset of M samples will always have to keep in mind all the "cost" items, making his decision with a full awareness of all the possible pros and cons.

… AND NOW ABOUT COMPUTER SCIENCE

Statistical Inference leads to partial results, but still allows a first rough evaluation of some of the characteristics of my "big" hoard of N coins. Under the conditions of tolerating a margin of error, it gives a preliminary overview of its content starting from a small sample and – not negligible benefit – a precise quantification of the uncertainty due to the simplification introduced.

By contrast, Descriptive Statistics can be very useful in providing a summary description of the whole find, regardless of its actual size. Speaking of hundreds, thousands or even millions of specimens does not imply any difference about the methodology of the analysis that can be borrowed from Descriptive Statistics. What is essential is to have a complete and accurate set of data, in a format that allows an effective processing with the tools made available by Descriptive Statistics, able to produce a summary of the characteristics of my group of coins. This is diametrically opposed to what has been discussed so far: no longer a random sample of M specimens extracted from the "big" hoard of N coins, but the hoard of N coins as a whole, appropriately described, measured, illustrated.

Storing the whole set of data (including high-resolution photographs) that can be taken from a “big” hoard of N specimens is not a problem. Even considering N in the order of one million pieces, we would remain in the order of Gigabytes, completely manageable with economic storage systems.

The processing of this set of data can be performed with a domestic personal computer, relational databases or spreadsheets can handle all our requests without any problem. Furthermore, these tools usually include advanced statistical functions, such as the most important formulas of both Descriptive and Statistical Inference (e.g., quantiles relative to the normal, Student and chi-squared distributions mentioned above), together with grouping solutions, pivot tables, queries … that can support the researcher in the extraction of desired information without the need to write specific code or to use programming environments certainly more peculiar to Statistics (e.g., R language) but that can require advanced knowledge of coding. Many useful tools have freeware versions and offer what is necessary to carry out in-depth analysis of the “big” hoards at minimal costs.

Having at our disposal a complete set of data, in digital form and organized in a relational database or in a spreadsheet, also allow the use of tools offering an advanced graphic visualization of the data, mostly based on a *drag-and-drop* approach, which make it possible the extraction of desired information and provide an immediate representation on the screen.

We are talking specifically about *data analytics platforms*, developed mainly in the context of Business Intelligence, but which can be applied to datasets of any nature. These tools are basically made up of dynamic dashboards within which it is possible to arrange graphs, maps and tables (usually starting from predefined but highly customizable templates) capable of providing a compact view of the data set that we have chosen to connect, made available in the form of ODBC databases, OLE DB databases, local folders with “open” format files (e.g., CSV) or even web URLs.

We can also take into consideration the dozens and dozens of APIs, plug-ins and widgets available free of charge on the Internet that allow, in a similar way to what done by the data analytic platforms mentioned above, spatial representations and a dynamic processing – sometimes combining these two aspects together – of any data set.

In the world of Numismatics, we can already find examples in this sense, with projects that provide a map view of a data set or of a subset of it extracted through dynamic queries.[14] In all these cases, however, we return to the criticality discussed

[14] We can mention here the project *Coin Hoards of the Roman Empire* (http://chre.ashmus.ox.ac.uk/ – URL last visited on February 1st 2019) presented during this congress, but the panorama is much wider. For example, there are also open source solutions that combine the representation of simple or aggregated data on a map with a web framework based on R language.

at the beginning of this paper: the effort required to catalogue the coins of the find. Working efficiently in the creation of a data set in the appropriate format thus represents the core challenge to be faced in the study of any "big" hoard.

In the example of the digitization of the Reka Devnja hoard proposed at the beginning we assumed a time of acquisition of a single specimen of the order of 10 minutes. With a simple multiplication of this value by the number *N* of coins constituting the hoard it was possible to evaluate the total effort required to be able to get the data set in the desired format.

An effort expressed in terms of man-hour implicitly indicates the two main directions that can be followed for its reduction: decrease of the time required for each single entry, decrease of the costs associated to each human resource.

A minimization of the time for data entry without compromising completeness and accuracy can only go through an efficiency improvement of the related procedures. How this can be achieved cannot ignore the basic elements of *lean production*, and must therefore go through a reduction of everything that can cause an increase of the processing time of a single sample, from photographic digitization to cataloguing, to the overall management of the coin.

An example of how we can achieve such a goal in a rather economic way comes from the experience made by the author of this paper. An Arduino Uno board, a digital camera with integrated *tethered shooting features*, a weight sensor with an accuracy of one hundredth of a gram and a personal computer on which a *speech recognition system* was installed, were connected to each other. The camera had been mounted on a stand, facing downwards the surface of the weight sensor where the coin was located, in turn placed on a small cube of plexiglass to get rid of shadows.

The speech recognition system allowed to fill in the cells of a spreadsheet translating into text the characteristics of the coin that were said directly in the microphone integrated into the personal computer. Once the desired sentence was completed, a simple keystroke on the personal computer activated a specific macro, which in turn performed a measure of the weight of the coin on the weight sensor, activated the camera and paused, allowing the operator to change the side of the coin. A second keystroke resumed the macro with a new acquisition of the weight and a new camera shot. The two measures of the weight were compared, harmonized appropriately if necessary, and written in the desired cell of the spreadsheet. The two files containing the coin pictures were then renamed based on a unique identifying code created automatically starting from a specific cell of the spreadsheet and stored in a desired folder of the personal computer.

A check of the quality and correctness of the text in the spreadsheet did not highlight significant transcription errors. The speech recognition software used in this test has proven to be very reliable. Situations in which some specific terms (e.g., the name of the king “Berengar”) were not initially recognized were resolved by adding the word to the vocabulary of the tool. Finally, to increase the accuracy of the weight measurement, a reset mechanism of the weigh sensor was introduced after every single photographic acquisition.

With this system it was possible to reduce the acquisition time of a coin in digital format in the order of two minutes: five times less than the estimate from which we started.

What proposed here is one possible scenario for optimizing the acquisition process. We can imagine several other situations that allow to achieve the same result. A collaborative approach to digitization can lead to an increase in the number of contributors without a corresponding increase in costs for human resources, especially if performed free of charge by volunteers. This is the case of the *Portable Antiquity Scheme* (*PAS*)[15] by the British Museum or of the *Coin Finds* portal developed by the author of this paper,[16] where it is possible to enter the data related to a find and modify or integrate the existing ones. A solution of this nature has clear limits as regards the reliability of the entered data and the intellectual honesty of the contributors, but at least in the case of the *PAS* all this is well compensated by the existence of a centralized structure that manages the contributions and which in turn actively contributes to the archive, albeit at the price of all the costs associated with the structure itself.

A further approach may consist in the reuse of data already available in a digital format, taking advantage of the fact that it is very common to find several specimens with the same characteristics inside a given hoard or already present in other finds. In such a situation, the corresponding records of the hypothetical digital archive of our hoard would contain a series of duplicated data (e.g., issuing authority, mint, inscriptions, descriptions, metal…). Therefore, the development of methodologies that allow to reuse these data, or more generally the data present in other digital archives, would go precisely in the desired direction of a reduction in the acquisition time.

We have seen during this conference an example of such an approach in the presentation of the *Coin Hoards of the Roman Empire* project by the University of Oxford, where the connection with the *Online Coins of the Roman Empire* (*OCRE*) portal[17] managed by the American Numismatic Society allows to fill the records related to the description of a specific coin of the hoard simply starting from the reference number to the *RIC*, thanks to specific APIs made available by *OCRE*

[15] https://finds.org.uk/ (URL last visited on February 1st 2019).

[16] https://www.sibrium.org/CoinFinds/ (URL last visited on February 1st 2019).

[17] http://numismatics.org/ocre/ (URL last visited on February 1st 2019).

itself.[18] Unfortunately, this valuable solution is limited to the field of Roman imperial coins only. For the Roman provincial coins, we can point out a project still managed by the University of Oxford[19] in which, nevertheless, only a part of the data is easily reusable (but not through APIs), while similar solutions regarding other types of coins are not available.

Digital archives like *OCRE*, useable in a simple way through APIs or that make their data available in an "open" format, are certainly very precious resources with a view to reducing the time required to build a complete data set related to a hoard. There are dozens of projects to digitize coins and other items of numismatic interest, but only in a minority of them the archives are accessible to the public. Even less numerous are those in which data are made available in a format that can be easily reused.

This means that, even with an open data portal, it may be necessary to convert them to the desired format. For the most part, we're talking of structured data, based on a *schema* in which each field has a precise meaning. This is the case, for example, of relational databases, where the meaning of the field (i.e., of the data contained in it) is defined by a schema and therefore by the position inside it. On a conceptual level, the migration of data from one database to another must go through an operation of re-mapping data from one schema to another considering the semantics associated with each of these schemas. On a practical level, such an operation may require the development of solutions to adapt the data to the new target schema, for example through specific decoders (parsers) for such models, but at the price of an additional effort.

If we work with non-relational databases, we would not have these needs. The concept of "schema" is replaced here by that of *key-value pairs* (*KVP*). We move from a solution where the meaning of a field is defined by a rigid schema (e.g., the text "Sciscia" will be identified as the "mint of origin of the coin" because in *that* database I decided *that* field, in *that* precise position, will have *that* meaning) to one where the semantic component is a part of the datum itself (e.g., mint:Sciscia).

Data available in an "open" format in the key-value pair format (implicitly present in JSON, RDF and XML formats) can be put into a non-relational, schemaless database without the need to develop a converter because the data will be stored simply as a collection of key-value pairs. This does not mean, however, a zeroing of the effort required for their use. The costs, now, would simply be moved further downstream, during the interrogation phase of the non-relational database (my concept of "mint of origin of the coin" may have been implemented with several keys – "mint", "atelier_monétaire", "zecca" – due to data coming from

[18] http://numismatics.org/ocre/apis (URL last visited on February 1st 2019).

[19] http://rpc.ashmus.ox.ac.uk/ (URL last visited on February 1st 2019).

heterogeneous sources), and may require to write complex queries, not based on SQL (e.g. query on JavaScript).

Any operation of re-use and integration of data should be as simple and quick as possible. In an ideal scenario we should be able to use the data contained in databases of different nature without any need for their pre-processing. This is what I would expect if I were to operate in a semantic web, where precisely the data are linked to one another and simply accessible through their Universal Resource Identifier (URI).[20] Portals such as *Nomisma*[21] by the American Numismatic Society and *Zenon*[22] by the Deutsches Archäologisches Institut are the best-known examples in the numismatic world.[23] These projects are implementing the concept of *ontology* created by sir Tim Berners-Lee[24] and later standardized by the World Wide Web Consortium (W3C).[25] The basic idea is to have a distributed environment in which published documents are associated with metadata that specify their semantic context in a format suitable not only for interrogation, but also for interpretation and – ideally – for their automatic processing.

A situation in which a question like "which are the hoards containing coins of Carausius?" can be put to a search engine in natural language is the best that could be expected from such a solution. But this is clearly possible just if the totality of data is available on the web in the desired format, on accessible and stable machines. A fascinating scenario, very seductive, but strongly influenced once more by the availability of these data (if they were available, we would not be here to question how to speed up the digitization…) and the instability of the World Wide Web in the medium and long term. We constantly must deal with the risks related to the fact that we do not have the full control of the servers storing the data of our interest. We cannot assume that they will be accessible in any moment, and that they will be forever: the possibility that a project ends and that the corresponding portal ceases to exist on the web is anything but negligible, especially if the maintenance of a given web space is linked to the availability of budget and/or of a specific person.

One of the prerequisites of the semantic web is the stability of the resources pointed by the URIs. A server migration from an http:// to an https:// protocol, for example, would require the adaptation of some URIs[26] and changes to the standard, raising at the same time the first serious doubts about the reliability of the semantic

[20] Gruber, Heath, Meadows, Pett, Tolle, Wigg-Wolf 2014.

[21] http://nomisma.org/ (URL last visited on February 1st 2019).

[22] https://zenon.dainst.org/ (URL last visited on February 1st 2019).

[23] For a more extensive review, see Pett 2015.

[24] Berners-Lee, Hendler, Lassila 2001.

[25] https://www.w3.org/standards/semanticweb/ (URL last visited on February 1st 2019).

[26] This is what happened with the linked data published by the British Museum on its web portal.

web. But much more unfavorable situations can occur if a specific object pointed to by a URI is no longer available due to the termination of a project.

To all this we must add potentially critical aspects regarding the quality of the available data and the obsolescence of information systems. Entrusting to automatic processing systems the interpretation of inaccurate data can lead to the extraction of information which in turn is incorrect, and of no practical use. What should I expect from my search for hoards containing coins of Carausius if there are inaccurate attributions of coins to this usurper on the World Wide Web?

Today we have a somewhat paradoxical situation: we can read inscriptions on fragile 4,500-year-old terracotta tablets, but not magnetic or optical disks produced just three or four decades ago. And even when we can access these physical media, we may come across file formats that are no longer usable for a variety of reasons.[27]

The idea of storage systems able to resist unscathed from one generation to the next, is a pure utopia. Any software is designed with a life cycle, with precise development and maintenance plans, and like living creatures they are bound to have an end sooner or later. The hardware or software solutions we can choose to adopt today will not necessarily be usable tomorrow. Projects may end due to lack of funding, or because the main contributors decide to stop their participation. The risk of being at a certain point in the face of the impossibility of reading hard-earned data is therefore very concrete.

To have an example of what we should expect, it is sufficient to look back on what happened in a relatively recent past, leafing through the proceedings of the congress *Monete in rete* held in Bologna in 2003.[28] We will find several projects that have not had the announced developments, or that today are even closed. Some of the proposed software programs are no longer able to run on modern computers, or are no longer of interest. Of course, we will also find several references to the XML format, still widely used, but no traces of technologies that appear to be more

[27] The most famous example in this regard is given by the *BBC Domesday project*, consisting in the digitization of the Domesday Book made on its 900th anniversary. The digitization was carried out between 1984 and 1986, and the data stored on the LV-ROM accessible only through an Acorn BBC Master expanded with a SCSI controller and an additional coprocessor-controlled Philips VP415 “Domesday Player”, a specially produced laserdisc player, which already in the early 90s of the 20th century was no longer in use. Consequently, it was not possible to access a set of data digitized data just a few years earlier, and it was necessary to spend a significant additional effort for data recovery and maintenance. After a few years the problem occurred again, as even the new support soon became obsolete. In the numismatic field we can also mention the case of the software *SAXA* by the Italian Ministry of Cultural Heritage, now no longer developed nor accessible by modern computers, with the result that the data of some public collections archived with this software in the 80s-90s of the last century are today unusable.

[28] GIOVETTI, LENZI 2004.

appealing today. Nothing different from what we should expect will happen in another fifteen years for some of the concepts expressed in this paper.

WRAP-UP

As we have just discussed, an exhaustive study of a "big" monetary hoard should preferably not be separated from the investigation of all its coins. Working on a sample of $M \ll N$ specimens with the approach suggested by Statistical Inference introduces relevant simplifications, but at the same time important margins of error that, due to the extremely heterogeneous nature of any monetary find, can deeply influence its interpretation. In situations where, usually, a single specimen changes the dating of the whole monetary complex, or where only a very small fraction of the coins has unedited characteristics, the methods offered by Statistical Inference are not sufficiently accurate to guarantee to be able to detect the cases of greatest interest. They can be used for a preliminary evaluation of the material, but only for macroscopic aspects, that is to say a high-level analysis where the inevitable uncertainties do not alter the interpretation of find.

Being able to deal in a reasonable time within a large amount of data, difficult to manage if we adopt the approach that is generally used for finds of more modest size, is the basis for any subsequent processing, even the most complex.

The "big" quantities involved must not be considered as the evidence that it is impossible to obtain the desired results. They simply remind us that the analysis of such large finds can no longer be conducted with traditional methods. Smart methodologies must be developed instead, to maximize the efficiency of material management, i.e. allowing to increase the amount of data made available in a digital format with the same amount of time spent.

Even if we operate in the most efficient way possible, we should always keep in mind that any study of a "big" hoard is a long-term project: with limited resources – both human and economic – a digitization project can take years, if not decades, to be completed. In doing this, we must always operate with the awareness that IT solutions and therefore archiving and digitalization evolve at the speed of light, towards paths that today may appear clearly set out, but that tomorrow are likely to result nothing more than dead ends. Speech recognition systems, solutions for the automatic weight acquisition, data retrieval APIs, linked databases … are just some of the examples of what can be used to improve productivity. Other solutions, or processes, can still be developed based on the specific needs of a specific research project. In such a scenario, we need to be far-sighted, always keeping the digitized data in a flexible, open and accessible format that can be reused in the future.

We are faced with a very complex situation, constantly changing towards very different and unpredictable directions. However, what is clear is the need for humanists to work in a different way, questioning their certainties and dealing with new technological skills that now can give them a fundamental support in the examination of hoards otherwise “too big to study”.

Bibliography and abbreviations

Berners-Lee T., Hendler J., Lassila O. 2001, "The Semantic Web", *Scientific American* 284, 5 (May): 33-43

Esty W.W. 1997, "Statistics in Numismatics", in *A Survey of Numismatic Research 1990-1995*, ed. by. C. Morrisson, B. Kluge, A. Burnett, L. Ilisch, W. Steguweit, Berlin: 817-823

Esty W.W. 2003, "Statistics in Numismatics", in *A Survey of Numismatic Research 1996-2001*, ed. by C. Alfaro, A. Burnett, Madrid: 921-927

Garraffo S., Mazza M. 2015 (a cura di), "Il tesoro di Misurata (Libia). Produzione e circolazione monetaria nell'età di Costantino il Grande. Convegno Internazionale di Studi. Roma, 19-20 aprile 2012" (Testi e Studi di Storia Antica 27), Catania-Roma 2015

Gianazza L. 2017, "Coin clipping and monetary crisis: the case of the Italian ducatone", in *XV International Numismatic Congress Taormina 2015 Proceedings*, II, ed. by M. Caccamo Caltabiano, Roma-Messina: 1257-1260

Giovetti P., Lenzi F. 2004 (a cura di), "Monete in rete. Banche dati, CD-ROM e Internet nella numismatica italiana" (*ER Musei e Territorio – Materiali e Ricerche 4*), Bologna

Gruber E., Heath S., Meadows A., Pett D., Tolle K., Wigg-Wolf D. 2014, "Semantic Web Technologies Applied to Numismatic Collections", in *Archaeology in the Digital Era: Papers from the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA), Southampton, 26-29 March 2012*, Amsterdam: 264-274

Metcalf W.E. 2002, "The Reka Devnia hoard re-examined", in *Ritrovamenti monetali nel mondo antico: problemi e metodi. Atti del congresso internazionale, Padova 31 marzo-2 aprile 2000*, a cura di G. Gorini, Padova: 145-150

Moushmov N.A. 1930, "Une trouvaille de monnaies antiques près du village de Reka-Devnia (Marcianopolis)", *Aréthuse* 27: 49-50

Moushmov N.A. 1934, "Le trésor numismatique de Réka Devnia (Marcianopolis)", Sofia

Parisot-Sillon Ch., Suspène A., Sarah G. 2014, "Patterns in die axes on Roman Republican silver coinage", *The Numismatic Chronicle* 174: 91-109

Paunov E., Prokopov I. 2002, "An Inventory of Roman Republican Coin Hoards and coins from Bulgaria" (Glaux 15), Milan

Pett D.E.J. 2015, "Numismatics, Computers and the Internet", in *A Survey of Numismatic Research 2008-2013*, ed. by C. Arnold-Biucchi, M. Caccamo Caltabiano, Taormina: 761-773

RIC = Roman Imperial Coinage

Wigg-Wolf D. 2009, "Numismatics, Computers and the Internet", in *A Survey of Numismatic Research 2002-2007*, ed. by M. Amandry, D. Bateson, Glasgow: 720-726