ISR Italian Sociological Review
University of Verona, Italy
ISSN 2239-8589 | Editor in Chief Paola Di Nicola
Managing Editor Debora Viviani

How Does the Error from Sampling to Big Data Change?

Cleto Corposanto, Molinari Beba

How to cite
Corposanto, C., Molinari, B. (2022). How Does the Error from Sampling to Big Data Change?. [Italian Sociological Review, 12 (7S), 665-684]
Retrieved from [http://dx.doi.org/10.13136/isr.v12i7s.576]

[DOI: 10.13136/isr.v12i7S.576]

1. Author information
Cleto Corposanto
Department of Law, Economics and Sociology, Magna Graecia
University of Catanzaro, Italy

Beba Molinari
Department of Law, Economics and Sociology, Magna Graecia
University of Catanzaro, Italy

2. Author e-mail address
Cleto Corposanto
E-mail: cleto.corposanto@unicz.it

Beba Molinari
E-mail: beba.molinari@unicz.it

3. Article accepted for publication
Date: May 2022

How Does the Error from Sampling to Big Data Change?

Cleto Corposanto*, Beba Molinari**

Corresponding author:
Cleto Corposanto
E-mail: cleto.corposanto@unicz.it

Corresponding author:
Beba Molinari
E-mail: beba.molinari@unicz.it

Abstract[1]

In this article the authors aim to present a series of considerations, regarding the research carried out in the last 8 years, which starting from Big Data have posed different methodological problems related on the one hand to sampling and on the other to the conception of error in the scientific field. More precisely, the contribution will be divided into two macro areas of discussion.

In the first part we will discuss sampling, with particular attention to break-offs and drop-outs and the relative response and cooperation rates, in order to understand how much these rates can still be valid in web 2.0 contexts. But at the same time we should ask whether it still makes sense to speak of probability sampling when in the hard sciences only a few cases are used in experiments, often less than a hundred.

Further reflections concern the determination of a statistical representativeness which, especially online, can sometimes be overcome by an effective sociological representativeness.

The second part of the contribution will be devoted to the discussion regarding biases and how the error can bring a series of further complexities in a pandemic reality.

---

* Department of Law, Economics and Sociology, Magna Graecia University of Catanzaro, Italy.

** Department of Law, Economics and Sociology, Magna Graecia University of Catanzaro, Italy

[1] This paper is a joint effort. Cleto Corposanto wrote the introduction, the first, the second and the third paragraph, while Beba Molinari wrote the fourth, the fifth paragraph and the Conclusions.

In this regard, the authors are convinced that an interpretative turning point must be made in the discussion that takes place around the error considered in the "science of discovery".

Keywords: Big Data, sampling, error.

1. Introduction

More than a year after the institutionalization of COVID-19 disease, we can pause to reflect on the information and the related sources of communication that characterize our time during the pandemic, with a particular focus on the methodological aspects.

The information channels have remained substantially the same as in the pre-pandemic period, but the interests of Italians have changed quite clearly in the lockdown period and in those to follow.

Istat (Italian National Statistical Institute), in the report to the Parliament for the Department for Family Policies (Istat, 2020, 2021), underlined the role of television and social relations, defining them as the two great pillars of Italian families in the lockdown period and later.

The need to inquire about any new treatments, vaccines, the trend of the curve and the related predictions, means first of all dealing with science.

It is appropriate to wonder, today more than ever, what science is. We have been accustomed to live with this lemma since the beginning of our school education.

According to the Treccani Encyclopedia, science is:

> The set of disciplines based essentially on observation, experience, calculation, or which have as their object nature and living beings, and which make use of formalized languages.
>
> In particular, modern science represents the set of knowledge as it took shape in its hierarchical structure, in its institutional and organizational aspects, since the scientific revolution of the 17th century. [...] Subsequently, the role of science has been gradually strengthened from both the social and institutional point of view and the methodological and cultural one, as well as science has become one of the aspects that best characterize, also due to its many technical applications, the contemporary world and the cultural values which it expresses.

The relevance of social aspects in science proper had evidently already emerged in the early 80s in the studies carried out by Garfinkel and his collaborators (Garfinkel, Lynch, Livingston, 1981) within which ethnomethodology was used in order to find out, through the researchers'

recordings, how much the relational aspects could affect the choices made during the performance of the activities in the laboratory.

The search for information that distinguishes this historical period is characterized by the unconscious constant search for technical-scientific information aimed at combating COVID-19.

It is, indeed, the subject of common discussion among friends, relatives, work colleagues, the debate, in a more or less competent way, regarding vaccines, healthcare and medical protocols, but also lung ventilators, the vaccination plan, masks, etc.

Let us therefore try to understand how we can, as social actors, comprehend the multitude of information of medical-scientific relevance that is provided to us by the media and, at the same time, extricate ourselves from the many discordant information coming from the world of science itself.

In this context, we aim to deepen some aspects that are worthy of interest related to Data Quality, with a focus on how web 2.0 can make changes compared to traditional sampling procedures, and in particular to a new form of interest that has not only emerged in the scientific-academic world, but has also spread among a large portion of the population, regarding the error in the scientific field about the predictions on COVID-19 cases, we all had to deal with (Corposanto, Molinari, 2020; 2021).

We are accustomed to consider the so-called hard sciences as incontrovertible, but it's not. The new media allow us to redesign the traditional canons thanks to which we approach the field of scientific research, acquiring new aspects with which to relate, not only as a new object of study, but also as an integral part of the researcher's toolbox.

2. From Data Quality to error detection in web surveys

Discussing Data Quality means, first and foremost, concentrating part of the speech on the construction of the data itself, therefore on the plausibility of the adopted detection tool designed in relation to the object of study (Marradi, 1989).

It is precisely for this reason that, in the last 8 years, the authors have paused to reflect on the plausibility of new online search tools which allow us to analyze large amounts of information collected within databases and gathered by data mining (Corposanto, Molinari, 2021).

In these contexts, data reliability becomes a cross-cutting aspect not only with respect to sampling, but also and above all to the possibility of running into some errors.

We should, therefore, ask ourselves to what extent online techniques can be subject to bias, reminding us that most of the traditional tools in the researcher's toolbox are not exempt from such risks.

When we talk about bias, there are many aspects to consider and they vary depending on the tool which has been taken into account; so we should dwell on the reliability of the fidelity of data in its dimensions related to the the sincerity of the answer, to the congruence of meaning and, in the case of techniques requiring an a priori ranking of the answers, also to their exhaustiveness (Corposanto 2000, Corposanto 2004). It is not necessary to say that these are all aspects that are linked to the construction of the data of the so-called traditional techniques, which are scientifically accepted, so it is good, in this context, to start from these assumptions.

But let's make a small clarification: we often discuss Data Quality but, in web 2.0 contexts, does the term "Data" still make sense? Or does the connotation change depending on the context, the type of information and the source, i.e. the data warehouse[2] from which they have been extrapolated?

In this case, in our opinion, considering that the level of information changes depending on the type of analysis to be carried out and even more depending on the application programming interfaces (API) used to carry out the analysis, Data Quality is constantly changing. This change happens because the nature of the web itself changes very quickly: even the software we are accustomed to using, tomorrow may already be obsolete and allow a lower level of information than new applications. In light of these considerations, it would be more appropriate to pause to reflect on the potential of data warehouses and the possible Mash-ups of data, i.e. the multiple combinations that can be made with data of different nature and source.

Those considerations done, when we talk about bias we use to attribute a purely negative connotation to it, but this is not always true; in fact, unexpected influences should not always be perceived as "mistakes", i.e. "errors" that contaminate the quality of the recorded information and thus make the information itself to be considered as unusable. We could, instead, think of biases as a space-time continuum, in which possible distortions can actually modify the Data Quality to varying degrees and, in some cases, allow us to discover new aspects to which not much weight had been given, becoming non-negligible areas of research for the study of the phenomenon.

[2]The data warehouse is a sort of second level of a "database" within which there are a series of information oriented to individual subjects, whether they are users, consumers and / or patients, integrated with multiple databases from which it "imports" specific information previously identified, and therefore of interest to the programmer / researcher.

In this regard, it is easy for data from the web to be affected to varying degrees by distortions that we could distinguish into two macro categories: on the one hand, we have the possibility of incurring errors due to the tool preparation, as it happens in traditional research contexts, in other cases, instead, the biases derive from big data and from the data warehouses themselves, which can alter any information through different digital formats; therefore, it will be up to the researcher to understand these differences and "transform" the format of the data in its useful form, aimed at the analysis that is intended to be carried out.

When you think of a questionnaire you mainly use the word "design" because creating a questionnaire is not a simple thing: in addition to the "rules" that we all know, which are illustrated within the many manuals of methodology of social research, online platforms for web surveys make this phase even more complicated[3] with the risk that, in the wrong hands, the tool may be used the wrong way and, even worse, that the results of the study may be misunderstood.

Given these aspects, we could consider web surveys as an evolved form of paper questionnaires, because the forms of analysis do not change, as well as the "basic rules" with which the questionnaire itself is designed; however, it is significantly different in the type of sampling, in its numerosity and also in the data analysis phase.

An even more innovative way is to use as a means of administration the APPlications (APPs) designed for smartphones and tablets, so that users must first answer a short questionnaire and then access the application itself. Also of this second opportunity we have already discussed on several occasions (Corposanto and Molinari, 2015; 2016). It is an alternative that, even today, is not used as much as it could, but it is part of the idea that the web allows us to make different use of platforms designed for other purposes; it will be up to the researcher to get involved and understand how to take advantage of it.[4]

But let's go with order and let's tackle in the next paragraph the aspects related to sampling, with particular attention to break-offs and drop-outs and the relative response and cooperation rates, in order to understand how much these rates can still be valid in web 2.0 contexts.

---

[3] For a detailed definition of the strengths and weaknesses of web surveys, see the following article: Corposanto C., Molinari B. (2014), Survey e questionari online?, in C. Corposanto., A. Valastro (a cura di), *Blog, fb e twitter*, Giuffrè, Milano.

[4] As for the Applications, we have presented the methods of use and the search results in two essays being published, entitled: *Web research e salute: quando le App le usano per mangiare* and *Ditelo con un'App. Comunicare oltre.*

3. Probabilistic or non-probabilistic sampling, that is the problem

First of all it is necessary to make some clarifications with respect to the concept of "probability" in relation to the sampling process.

When we talk about sample representativeness, for the OECD (2022) there is a confusion of the term intense in the lexical sense. In the widest sense, a sample which is representative of a population. Some confusion arises according to whether "representative" is regarded as meaning "selected by some process which gives all samples an equal chance of appearing to represent the population"; or, alternatively, whether it means "typical in respect of certain characteristics, however chosen[5].

Let's try to make order and illustrate the definition we have adopted. In particular we have preferred a definition that is commonly used in the branches of science by all those disciplines that consider 'sampling'[6] as an important step for their studies, is the following: "a sample is said to be probabilistic if, and only if, all the units considered are chosen by sampling, frame at random". One of the objectives of this contribution is to investigate whether or not it is possible to adopt this definition also for studies conducted through web-surveys.

However, it is not sufficient to examine the concept of probability without taking into account the notion of 'random sampling' (Cochran, 1953); a clear and typical example is undoubtedly the lottery, where, by definition, each unit of the population has an equal chance of being selected. A debate on representativeness, intended as considering a part to represent the whole, has come up around this definition: a synecdoche that has interested, for years, those studies conducted with statistical inference.

It is one thing if we consider sampling as a simple extraction from an urn but it is a different matter if we follow the same procedure by interviewing people that, unlike the dice boxes of the lottery, may refuse to contribute to their task. Besides, it should be pointed out that a 'random sampling' with individuals is statistically representative only if the population is well known in its entirety and a list has been provided.

Under these circumstances, it is quite evident that carrying out the inference procedure on the outcomes obtained may result rather difficult. As a matter of fact there are some objective issues to take into account from the very

[5] A Dictionary of Statistical Terms, 5th edition, prepared for the International Statistical Institute by F.H.C. Marriott. Published for the International Statistical Institute by Longman Scientific and Technical.

[6] Fabbris (1989), De Carlo e Robusto (1999), Wonnacott (1969), Henry (1990), to name a few, without claiming to be in any way exhaustive.

first stages of the research, when methods and techniques aimed at scrutinizing the considered case of study are still under definition. In that respect it would be advantageous to broaden the approach adopted and try to understand what limits but also opportunities the network may bring.

On account of this, our objective is to investigate the *break off* phenomenon and draw attention to the response and cooperation rates emerged from our survey.

Sampling in the social sciences has been, and still today, the subject of much debate. The discussion revolves around the representativeness of the sample, from the statistical to the social sciences the leap is wide and often this aspect is forgotten in the methodology of social research. It is therefore important to understand whether sampling in research carried out through the aid of the web has more strengths than sampling conducted in so-called classical research contexts (Corposanto, Valastro, 2014; Molinari, 2014). In this broad debate, the main discussion revolves around the web-surveys carried out through the aid of dedicated online platforms. In this regard, the main risk is to confuse the tools made available by social media, such as Facebook and Twitter, which allow online surveys to be carried out, with the latest generation platforms specially designed to carry out social research activities. There are in fact some of these platforms set up by research bodies and universities. We therefore compare these platforms with the older tools with which to administer an online questionnaire, the most commonly known as CAWI[7].

It is often easy to confuse their function by thinking of the only diffusion through the social-media; indeed, Facebook, Twitter and blogs are not the only way to administer an online survey[8]: the modality adopted to extend a survey online will impede at first glance a series of possibilities which will enable us to draw the sample of those willing to contribute to the study. Naturally, the process by which it is common to make a web-survey in context 2.0. is quite different from CAWI, from those questionnaires sent by e-mail in different electronic formats (pdf, word, Excel). What is more, the email is just one among many instruments used to contact and inform people of the link referring to the website, wherein it is possible to participate in the survey.

Through the latest platforms, access to the link may be restricted to a certain number of people through the use of a password communicated to all who will take part of the sample. Accesses through the link can be also monitored in real time. Besides, there would be the possibility to get to know

[7] The web interviews (CAWI - computer-assisted web interviewing) have distant roots and were born towards the end of the seventies (Hayslett, Wildemuth, 2004).
[8] Web surveys (CAWI - Computer Assisted Web Interviewed) date back to the late seventies (Hayslett, Wildemuth, 2004).

additional information through the tracking number of the user's computer (IP): this kind of information will appear directly to the researcher among the recorded data and will further limit access from the same computer to a second survey[9].

The first step is supposed to include a list of names and e-mail addresses to let the researcher carry out his sample according to the object of study. It is important to state that surveys can be administered through several other tools. With the introduction of social networks, not only is it possible to quickly disseminate a survey using a simple copy and paste function but there is now the opportunity to set up surveys using online platforms. It follows that it will not be possible to know the sampling frame a priori just as it is not possible to know in advance "the people who are going to see a museum". Besides, it is essential to note that a web survey is not necessarily linked with non probability sampling procedures: the statistical inference changes depending on the mode of operation carried out by the researcher. Yet, the sample representativeness index tends to decrease in all those cases when the researcher interferes and forces somehow the sampling procedures agreed in advance, passing from probability sampling to non probability sampling.

In this regard, it is worth reminding that the inference concept is based on two basic requirements: Representativeness and Randomness, both called into question by Marradi (1997) who stated that randomness does not contribute to drawing the representativeness of a sample and vice versa. Accordingly, the concept of inference, as we are used to studying, will automatically lapse. In this respect, the central limit theorem[10] is undoubtedly of considerable importance: Although such considerations are the result of concepts derived from classical research, we assume they might be favourable to the web 2.0 concept, where the number of compilations is substantial, to the extent that over 1,000 statistical questionnaires may be compiled in a few days thanks to effective

[9] Many of these decisions are related to the internal configuration of the platform, which varies according to the objectives which led the creator of the site to enter the market. In this regard it is remarkable that the increasing use of such platforms occurs in the relevant market research. However, according to the European Society for Opinion and Marketing Research (ESOMAR), which is one of the most renowned companies conducting social and market research at European level, there might be the risk of misusing such tools. As a consequence, if each passage of the survey is not well carried out, the privacy of individuals will be infringed. This was debated within the code of ethics drawn up by the International Chamber of Commerce.

[10] The central limit theorem states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population (with mean μ and variance $\sigma/\sqrt{n}$), where *n* is the number of samples.

communication channels. Among the surveys under consideration, for instance, the sample size lies roughly between 1,000 and 3,000 cases for a population of 62,000.

4. Break-off and drop out, passing through the sampling error

We could therefore say that there are no web surveys without dedicated online platforms, because the tools made available by social networks to carry out online surveys do not meet the basic requirements of traditional questionnaires. In addition to these requirements, it is possible to calculate some trends that we often forget to investigate: break-off and drop out.

Here for the sake of brevity we propose a practical example of these trends, what is of extreme interest is the fact that in eight years of research the percentages calculated in different studies are always similar and differ by no more than ten percentage points. Particular attention was given to a study on food transgressions of people suffering from eating disorders associated with chronic diseases that occurred in the last two years: both studies provide a sample which estimates about 3,000 individuals for a total population of 62,000 people with diagnosed diseases.

To calculate the trends of the answers that emerged from the online compilation, we follow what is indicated in the good practices defined in the document presented by the America Association for Public Opinion Research (AAPOR), which not only states the methods by which partial and complete interviews are used to be distinguished, but it also provides five calculation methods aiming at the analysis of contact, response and cooperation rates as well as the break-off phenomenon. Such *standard definitions* are particularly suited for *person to person* interviews, postal surveys (named person), and for all those surveys which can provide the names of their own respondents. Along with this statement, we decided to consider the above mentioned rates because, as it was previously mentioned, we had the great opportunity to obtain a list of names from our survey. Based on this methodology, the aim of our analysis is to investigate how such measures can work effectively in the era of Web 2.0.

The first trend concerns the level of completion of our survey, for which we adopted the standards proposed by Frankel (1983) and shared by AAPOR. In this respect, they consider "break-off" those interviews with less than 50% of responses. Interviews providing between 50% and 80% of answers are said partial whereas those whose response rate stabilizes over 80% are said complete. The variables which are considered crucial to the objective of the survey are 22 out of 50.

The following table highlights the specificities occurred between the two surveys; the first three entries report the classification by AAPOR, adoptable without limitations for web-surveys, whilst the "dropout" and "other" entries need some clarifications since they are directly linked to the web.

The "dropouts" entry has been regarded as the equivalent of NR, meaning "Non Respondents", since for a web survey there is no way of knowing, except in an approximate way, the precise number of people who have not been contacted. However, thanks to the 'big data" provided by the platform, after having traced all IP addresses, it was possible to identify those who opened the web page to complete the survey. Precisely, after having carefully read the presentation page of the study for at least one minute, these people finally decided not to proceed with the completion of the online questionnaire, thus closing the page. As a consequence, they have been considered the equivalent of "non respondents" with reference to AAPOR criteria. Finally, the "other" entry includes those who accessed the main page, where all the information about the aim of the survey is provided. Subsequently, they opened the page to complete the survey but for reasons related to restrictions on their own computer or server, they were not able to fill out any item.

TABLE 1. Interviews level of completion – I and II survey.

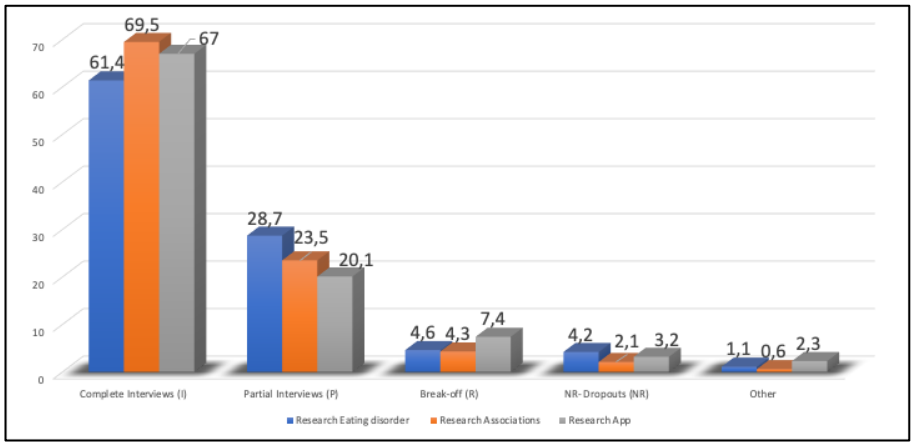
	FIRST SURVEY	**SECOND SURVEY**	**FIRST SURVEY**	**SECOND SURVEY**
	V.A.	%	V.A.	%
Complete Interviews (I)	2034	67.8	1753	61.4
Partial Interviews (P)	715	23.4	819	28.7
Break-off (R)	145	4.8	130	4.6
NR- Dropouts (NR)	88	2.9	119	4.2
Other (O)	19	0.6	32	1.1
TOTAL	3001	100.0	2853	100.0

Besides the variances that occurred between the first and the second survey, it is exceptionally interesting to compare these percentages by taking into account what has been discussed by Bichi (2007: 128). This text enabled us to obtain a first approach to the document proposed by AAPOR, with regard to the interviews carried out through classical methods, hence without any intervention from the web, where it must be acknowledged that in May 2006, as published on the website of Agcom, on a total of 49 surveys conducted, 15,62% of interviews were successfully concluded, whereas the average of non responses was 73,3%. These data seem to be very different from what comes to light from our studies. We might dare to say reversed, if compared to the trends of 'Paper- and-Pencil' (P & P) surveys, even though among the 49

surveys, 3 of them nearly achieved the percentage resulting from our study. However, it can be assumed that, as in our case, the people interviewed were particularly motivated to complete the questionnaire. In this regard, we take a further step and try to verify if the involvement of such a large number of respondents is directly connected to this specific survey or if, in terms of compilation, there might be the possibility to obtain further comparable trends in other different cases.

The following analysis shows the results of two studies. The first one is related to community life, the second one to the use of applications aiming at a better knowledge of food. As a result, the percentages of complete and partial interviews, as well as break-off were found to be similar[11].

CHART 1. trend in the level of completeness of the interviews per study – percentage values.



From the analysis of the graphics there is some evidence to demonstrate that the percentages are not significantly different. The fluctuations vary by a maximum of 8.1% between the complete interviews conducted on the study concerning eating disorders and those carried out

In fact, the analysis shows the outcomes of a single list, where it is easy to suppose that participants have not only a keen interest in the survey about eating disorders but also in similar themes affecting their well-being besides their health.

[11] 1,432 people participated in the survey concerning community life, while 1,850 expressed their level of satisfaction as to the applications used by diagnosed people. Both surveys did not exceed 20 questions.

In a study conducted 2 years after the one on eating disorders concerning the consent to school life of parents and upper secondary school students[12], it emerges that, even in this research, the percentages are very close to each other. For this second study, born with other purposes and without any claim to statistical representativeness, it would not make sense to calculate the response and cooperation rates, while it is possible to calculate them for the survey conducted on eating disorders. In addition, it can be assumed that we are facing a case of non-probability sampling by virtue of the procedure by which the sample was set and its representativeness 'ex-ante' (Marradi, 1997).

From the analysis of the graphics there is some evidence to demonstrate that the percentages are not significantly different. The fluctuations vary by a maximum of 8.1% between the complete interviews conducted on the study concerning eating disorders and those carried out to investigate the most commonly used Applications by diagnosed people. To demonstrate the various rates adopted, we recommend reading the article *Chasing a Dragonfly on the Lawn* paper written by the authors (2015).

The fourth report of AAPOR demonstrates four different types of rates.

For each one different calculation formulas have been provided. Naturally, it is desirable to use the most appropriate one along with the available data of the considered survey. Among the diverse possibilities we found a suitable formula for each rate, which met our specific demands[13].

As a *Minimum Response Rate* we adopted RR5 calculated with the following formula.

$$RR5 = \frac{I}{(I+P) + (R+NC+O)}$$

The choice of the previous formula helps to determine the choice of the subsequent rates. Therefore, for the cooperation rate we decided to adopt

$$COOP\ 4 = \frac{(I+P)}{(I+P) + R}$$

As to the dropout rates we used

$$REF3 = \frac{R}{(I+P) + (R+NC+O)}$$

[12] 1,005 students enrolled in the first, second and third year replied to the questionnaire.
[13] Abbreviations correspond to those listed in table 1 and have been directly gathered from the fourth AAPOR report.

Last but not least, the contact rate

$$CON3 = \frac{(I+P) + R + O}{(I+P) + R + O + NC}$$

The resulting data shown in the table below are extremely important: the higher the level of involvement in the subject of the study, the greater the propensity to reply to the questionnaire, especially when it comes to the health of a person or that of their family.

Let us return to the example of eating disorders where we highlight that the various rates are equivalent to each other, there are no major differences between the first and the second survey.

TABLE 2. Interview percentage rates – I and II survey.

	FIRST SURVEY	**SECOND SURVEY**
RR5	67.7%	61.4%
COOP4	95.0%	95.2%
REF3	3.8%	4.5%
CON3	97.1%	95.8%

Finally we propose a further step forward, we can analyse the errors within the sample, in order to better understand how much deviation occurs between real and theoretical sampling, given that most of the variables which make up our survey are not metric. In this respect, we used the formula with regard to Bernouilli's theory of finite population sampling, and we selected a level C of the confidence interval equal to 1,96, that is 95% of probability that the value does not fall outside the interval.

$$e = k \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}}$$

Letters can be replaced by the following values:

K is the level of confidence for the confidence interval, which is 1,96;

N represents the number of observations in the population, which numbered 62,000 people;

n refers to the number of observations in the sample, that in the second survey was estimated at 2,853 people

Pq represents the percentage by which a parameter may occur as well as its opposite (or not present)

The sampling error calculated is lower than we expected. It is equal to 1,78% for all those parameters attributed to 50% of the sampled subjects.

By following the same formula of the sampling error, we try to verify if the percentage remains constant even as regards the previous study, where the only difference that was found was in the size of the sample, since the selected units numbered 3,001 instead of 2,853. The sampling error, in this case, is equivalent to 1,74% with only 0,04 deviation points between the second and the first survey.

In the light of what emerged, the authors believe that the researcher is based on the researcher's ability to understand that randomness and representativeness are not synonymous, the first relates to the procedure for extracting the cases, while the second to the outcome of the procedure (Kruskal, Mosteller, 1980).

The two concepts can be further distinguished by shifting the attention to the concept of representativeness ex ante and ex post (Marradi, 1997). In this case it is much easier to recognize how difficult it is to think of a representativeness ex ante for web-surveys. Although the methodological precautions adopted have been presented in the third paragraph, it seems much more arguable to consider the concept of representativeness as a continuum rather than a dichotomy. Given these considerations, we can assume that the obtained samples from both surveys about eating disorders due to chronic diseases reflect the so-called representativeness ex-post. In fact, the access to the frequency distributions provided by some institutional sources gave us a remarkable opportunity to compare the population with the sample[14]. Indeed, the findings reflect the demographic as well as the territorial variables published by the Ministry of Public Health.

Moreover we decided to investigate the break-off phenomenon, as well as the response, cooperation and contact rates besides the so-called dropouts. In this respect, it can be concluded that, due to the high percentage rates obtained, it may still be possible to consider the same calculation procedure but it would be better to avoid a comparison between online and P&P surveys.

5. Between science and pandemic: the relevance of the error

Having discussed the various aspects of web surveys in web 2.0 contexts, let's enter into the merits of the relevance that the concept of science has

[14] The analysis is limited to the frequency distributions and does not take into account the association among variables, which, for this specific case, are relatively unknown (Marradi, 1997).

assumed with the advent of pandemic and, even more, how much the possibility of error has entered the common debate.

We can therefore borrow the vision of science provided by Kuhn in 1962, particularly appropriate, even today, to define the current historical period.

In *The structures of scientific revolutions* (1962) Kuhn distinguishes between periods of normal science and periods of scientific revolution, with respect to which there would be no linearity between one and the other, but real leaps forward, determined precisely by some discoveries that trigger real scientific-academic revolutions.

The authors therefore believe that if we adopt a conception of science understood in a very broad sense, we can consider that the pre-pandemic period was substantially characterized by the so-called normal science where, apart from some specific cases, there was the consolidation of reference paradigms accepted by the scientific community to which they belong, which have been largely questioned because of the pandemic, with more or less bright and harsh tones. This debate has entered our homes through the great media, while once scientific diatribes were the prerogative of "insiders" through conferences, lectio magistralis and publications.

Such a heated debate is in stark contrast to a "scientist" vision of intellectual knowledge with which the general public is accustomed to being confronted in the media, in which knowledge is usually presented as indisputable and as the foundation of other areas of knowledge, including political, ethical and economic knowledge.

It is quite obvious from the ongoing scientific debates and, even more, from the multitude of discordant information, even a few hours away one from the other, that we are not in a period of normal science. COVID-19 has forced us to increase research in different fields of study and citizens have rediscovered the importance of research carried out in the laboratory.

It was a succession of more or less understandable explanations of two experimental macro areas: on the one hand we have listened to discussions regarding the possible forms of treatment for COVID-19 among which we remember, from the early stages, the media debate on hyperimmune blood, which today has been replaced by a discussion oriented to the curative power of monoclonal cells; the second major area of experimentation concerns the fight against the pandemic through vaccines. The latter are divided into two macro categories, as we have learned to know from the information circulating in the media: viral vector vaccines (Vaxzevria by AstraZeneca and Janssen by Johnson & Johnson) and mRNA vaccines (Pfizer and Moderna).

What does not emerge from traditional scientific dissemination (papers, articles, books, etc.), nor from the media, and therefore does not reach the general public, is the negotiating side of the research that takes place within the

laboratory (Knorr Cetina, 1981) and that, apart from specific studies, remains anchored to a process buried by the homologation of scientific conventions concerning the drafting of papers.

How many times, in the past year, have we heard news based on scientific research published in well-known trade journals, denied a few hours later by the publication of a further study that refuted previous research.

Those considerations done, it is evident that we are in a period of scientific revolution in many different fields and areas of study. It is indisputable that the first innovation in the scientific field is, from a certain point of view, the acceleration of the timing related to research processes.

On the other hand, the media have responded to the problematic nature of the moment through the reduction of pain mediatization in favour of greater research and consequent dissemination of scientific information.

Information or, rather, the ability to transmit knowledge, therefore requires a fundamental clarification: the way of conceiving error in the scientific field.

We are accustomed to consider the so-called hard sciences as incontrovertible, but it's not.

We have already discussed the fact that the negotiating power of those who make up the research team plays a considerable role not only in the choices made, but also in the understanding of the results, an aspect that, moreover, as already mentioned, is the subject of Garfinkel's study, but not only (Latour, 1983; Knorr Cetina, 1981).

Let's consider another point of view, the one belonging to the researcher who intends to study his own theory.

We are therefore in the field of discovery: the researcher could ask himself a substantial question, that is, if the detection tool he intends to use, which is already supplied in the scientific field, can really measure his object of study, or if it does not measure the object of study along with some other variable, or even if it measures something completely different that has nothing to do with the object of research.

These considerations were made by the physicist Joseph Weber in 1969 during his study concerning gravitational waves. Such questions are very reminiscent of the medical-scientific discussions, also spread by the media, about the reliability of COVID-19 antigen tests, from oro-pharyngeal to serological ones.

Joseph Weber is not the only "hard science" scientist who asked himself questions of this kind; albeit in a different way, even Feynman questioned the reliability of the research, as it was presented in a recent article published by Corposanto (2021) on *Mimì*, the cultural insert of *Il Quotidiano del Sud*.

We must, therefore, ask ourselves if the so-called disturbing elements and / or errors are really recognizable, but even more if researchers are really ready

to question results that contrast with the reference paradigms consolidated by solid schools of thought.

6. Conclusive considerations

With the arrival of the pandemic we found ourselves constantly looking for scientific information (Istat 2020, 2021), more than we would have ever thought before. On the other hand, the media have practically always proposed the same in-depth topics: COVID-19 has monopolized the information aimed at the general public, both in the press and on television (Mazzoli, Menduni, 2020). In this context, communicating science means distorting the traditional conception of scientific communication in which at the extremes we find science on the one hand and the public on the other, while the media become the means of dissemination in charge (Bucchi, 2019: 132). This model, extremely simplistic, does not take into due consideration what has been discussed so far, i.e. the relevance of the error in the scientific field and the erroneous belief of the reference paradigms' indisputability.

It is good therefore to distinguish between expert knowledge, which in times of pandemic is attributed to epidemiologists, virologists, infectiologists, immunologists, biologists, microbiologists, up to general practitioners, and non-expert knowledge, represented by the general public, i.e. social actors.

It is necessary to immediately point out that, according to the authors, there is no prevarication of one knowledge over the other.

Expert knowledge is not always scientifically accurate: we have already discussed the possibilities of error and the very ability of the individual researcher and / or the research team to know how to recognize and consider it as such. Let's think, for example, about the inability of the scientific world to predict the trend of the contagion curve, with respect to which some political decisions have been taken regarding the opening / closing of commercial activities with considerable economic implications.

By its very nature, non-expert knowledge is not equally accurate; however, the common feeling sometimes has intuitions that anticipate scientific research, as happened for the degree of contagiousness outside the Coronavirus. From the first months after the denomination of the disease, it seemed evident that the contagiousness was much greater indoors and residual outdoors. A year later, in fact, even science is wedded to this intuition: according to what has been published by the Health Protection Surveillance Centre, an Irish research institute, only one person in a thousand is infected outdoors.

The public receives continuous discordant information from the media, at the expense of a continuous search for balance between expert and non-expert knowledge.

Instead of considering knowledge as a dichotomy, let's try to shift the focus to another aspect and think of knowledge as a continuum (Cloitre, Shinn, 1985) in which information should pass through several levels, listed below:

- intra-specialist knowledge in which the language used is particularly complex and set up for professionals only;
- inter-specialist knowledge in which an intermediate language is used, which is oriented to scientific aggregators such as journals like *Nature* and *Science*;
- pedagogical knowledge disseminated through a typically textbook language that provides basic skills;
- popular knowledge in which a very simple language is used, which is aimed at the general public.

The information provided to us by the media ranges from an intra-specialist level used by epidemiologists, by doctors of different types, i.e. expert knowledge, to a more popular level and vice versa, depending on the reference media, television broadcast, documentary, etc.

We even find ourselves listening within a single television and / or radio broadcast, different levels, at a distance of a few minutes from each other, depending on the expert who takes the floor, without forgetting that non-expert knowledge is not only represented by the general public, but many times it is an expression of the world of policy makers as politicians, trade associations' spokesmen, etc.

This situation generates social complexity, a constant information confusion linked to an arena of debate that is too wide, and that disagrees in language, interests, and purposes.

We should therefore ask ourselves whether such a situation is at the expense of the ability to convey knowledge. This is an aspect that, moreover, has been underestimated and taken for granted for too long, considered since the fifties as a simple "transfer" of knowledge, forgetting skills, the exclusive prerogative of the media (Lewenstein, 1995).

It is necessary to point out that knowledge, information and science during discovery periods are constantly renegotiated and shall not be regarded as paradigms.

That is why recognizing error not only becomes an integral part of our lives, but it turns into a sort of "antibody" that supports us in understanding what is really happening as a result of COVID-19.

References

American Association for Public Opinion Research (2006), *Standard Definitions: Final Dispositions of Case Codes and Outcomes Rates for Survey*, 4th edition, AAPOR, Lenexa (KS).

American Association for public opinion research (2014), *Social Media in Public Opinion Research: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research*, Lenexa (KS).

Bichi, R. (2007), *La conduzione delle interviste nella ricerca sociale*, Roma, Carocci editore.

Bucchi, M. (2019), *Scienza e società. Introduzione alla sociologia della scienza*, Milano, Raffaello Cortina Editore.

Cloitre, M., Shinn, T. (1985), Expository practice: Social, cognitive and epistemological linkages, In T. Shinn, R. Whitley (eds.), *Expository Science. Forms and Functions of Popularization*, pp. 31-60, Dordrecht-Boston, Reidel Publishing Company.

Cochran, W. (1953), *Sampling Techniques*, New York, Wiley.

Corposanto, C. (2000), *Tecniche del sondaggio d'opinione*, Trieste, Lint Editoriale.

Corposanto, C. (2004), *Metodologia e tecniche non intrusive nella ricerca sociale*, Milano, FrancoAngeli.

Corposanto, C. (2021), Scienza, il piacere della scoperta si prova anche nell'errore da svelare, in *Mimì, inserto culturale del Quotidiano del Sud.*

Corposanto, C., Molinari, B. (2021), Il virus e la società dell'accelerazione, in R. Grimaldi., A. Gallina, (eds.), Milano, FrancoAngeli (Forthcoming).

Corposanto, C., Molinari, B. (2020), Dai Big Data alla valutazione passando per la metodologia della ricerca sociale, In S. Gozzo, C. Pennisi, V. Asero, R. Sampugnaro (eds.), *Big Data e processi decisionali. Strumenti per l'analisi delle decisioni giuridiche, politiche, economiche e sociali*, Milano, Egea.

Corposanto, C., Molinari, B. (2016), Say it with un App, *Journal of Advanced Statistics*, 1(2), 52-6.

Corposanto, C., Molinari, B. (2015), Chasing a dragonfly on the lawn, *Science Innovation*, 3(4), 39-45.

Corposanto, C., Molinari, B. (2014), Survey e questionari online?, in C. Corposanto, A. Valastro (a cura di), *Blog, fb e twitter*, Milano, Giuffrè.

Corposanto, C., Valastro, A. (2014), (eds), *Blog, fb e twitter*, Milano, Giuffrè.

De Carlo, A. N., Robusto, E. (1999), *Teoria e tecniche di campionamento nelle scienze sociali*, Milano, Led Edizioni.

Fabbris, L. (1989), *L'indagine campionaria. Metodi, disegni e tecniche di campionamento*, Roma, Carocci.

Garfinkel, G., Lynch, M, Livingston, E. (1981), The work of a discovering science constructed with materials from the optically discovered pulsar, in *Philosophy of the Social Sciences*, 11, 131-158.

Hayslett, M.M., Wildemuth, B.M. (2004), Pixels or pencils? The relative effectiveness of Web-based versus paper surveys, *Library & Information Science Research*, 26, 73-93.

Henry, G.T. (1990), *Practical Sampling, vol 21*, Newbury Park, Sage publications Ltd.

Institute for Social and Economic Research (2020), *Understanding Society: Waves 1-10, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009, User Guide,* 29 October 2020, Colchester, University of Essex.

ISTAT (2020), Reazione dei cittadini al lockdown, in *Statistiche Report*, Roma.

ISTAT (2021), Comportamenti e opinioni dei cittadini durante la seconda ondata pandemica, in *Statistiche Report*, Roma.

Knorr Cetina, K. (1981), *The manufacture of Knowledge: An Essay on the constructivist and Contextual Nature of Science*, Oxford, Pergamon.

Kruskal, W., Mosteller, F., (1980), Representative Sampling, IV: The History of the Concept in Statistics 1895-1939, *International Statistical review*, XLVIII, 169-195.

Kuhn, T.S. (1962), *La struttura delle rivoluzioni scientifiche*, Tr. it. Einaudi, Torino, 1978.

Latour, B. (1983), Give me a laboratory and I will raise the world, In K. Knorr Cetina, M. Mulkay, (eds.), *Science Observed*, pp. 141-170, London, Sage.

Lewenstein, B. (1995), Science and the media, In S. Jasanoff, G. Markle, J.C. Peterson, T.J. Pinch (eds.), *Science Technology and Society Handbook*, pp. 343-359, Thousand Oaks, Sage.

Marradi, A. (1989), Casualità e rappresentatività di un campione nelle scienze sociali: contributo a una sociologia del linguaggio scientifico, in A. Mannheimer (ed.), *I sondaggi elettorali e le scienze politiche. Problemi Metodologici*, pp. 51-13, Milano, FrancoAngeli.

Marradi, A. (1997), Casuale e rappresentativo: ma cosa vuol dire?, In P. Ceri (ed.), *Politica e sondaggi*, pp. 23-87, Torino, Rosenberg & Sellier.

Mazzoli, G., Menduni, E. (2020), *Sembrava solo un'influenza. Scenari e conseguenze di un disastro annunciato*, Milano, FrancoAngeli.

Molinari, B. (2014), La piattaforma online come strumento di rilevazione e fonte di possibili scenari interpretativi, in *Salute e Società*, XIII, n.3., 103-117.

Wonnacott, T.H. (1969), *Introductory Statistics*, Hoboken, John Wiley and Sons.