A Short Research Guide for Designing Representative, Proportional, and Random Samples of Papers to Gather Scientometric Data of Research Fields

[1]Manuel Goyanes
[2]Olga Blasco-Blasco

[1]Carlos III University, Department of Communication, Spain. Email: mgoyanes@hum.uc3m.es (corresponding author)
[2]University of Valencia, Department of Economics, Spain

Abstract

Gathering representative scientometric data from papers (such as the gender of the first author, funding, number of co-authors per paper, authors' affiliation, etc.) is fundamental to examining the state and evolution of research fields. However, despite the need to design representative, proportional, and random samples of research publications to provide generalizable scientific findings of research fields, a simple step-by-step methodological protocol is surprisingly missing. This article addresses this gap and provides a guide for meta-research scholars on how to gather and process data from the Journal Citation Report (JCR) to design random samples proportional to the journals' number of publications and representative of the fields under examination. The study contributes to advancing the scientific study of research patterns in science by presenting a research guide that covers all the methodological steps to provide reliable scientific findings based on the design of representative, proportional and random sample sizes.

Keywords: representative sample, proportional sample, random sample, scientometric, sample size, research patterns, research fields.

A Short Research Guide for Designing Representative, Proportional, and Random Samples of Papers to Gather Scientometric Data of Research Fields

Despite the many values associated to “good” or “sound” science, many scientists would agree that providing reliable evidence that informs society to solve the most pressing problems is the intended meaning of most scientific endeavors (Chalmers, 2013; Goyanes, 2020). To do that, the sciences typically rely on robust methodologies and technical procedures aimed at reporting the most accurate findings to tackle empirical, social, and theoretical quandaries (Bourdieu, et al., 1991; Goyanes, 2017). In this regard, one of the most relevant technical procedures, among many others, is the design of representative samples to infer reliable conclusions of the population under examination (Grafström & Schelin, 2014; Omair, 2014).

It is widely known by laypeople and largely addressed by statisticians (Serdar et al., 2021) that representative samples enable the generalization of findings, warranting the inference of results to the target population, saving costs yet losing certain levels of accuracy due to measurement errors (Jenkins & Quintana-Ascencio, 2020). Despite the paramount importance of sample sizes in warranting reliable results, several challenges may impede their proper execution, such as unknown populations, budget allocations, or sampling errors (Lakens, 2022). In social sciences, the design of representative samples is significantly relevant to examining and understanding public attitudes, perceptions, and behaviors of a given target population, particularly in survey research (Taherdoost, 2017). In other fields of study, such as biology, economics, or genetics, the appropriate design of samples is also crucial to infer reliable causal findings (Hart et al., 2013; Gauderman, 2002). Likewise, scientometric research is also in need of sample designs that accurately represent their target population (Goyanes et al., 2023), for instance, when measuring different research patterns and evolution of scientometric variables of academic fields (), and therefore provide reliable results that can be generalizable.

Some typical examples may be easily presented. For instance, prior research on scientometrics has largely focused on the examination of gender representation of the first author of papers published in different research fields (Anderse et al., 2020; Pinho-Gomes et al., 2020; Denby et al., 2020). This analysis, however, can focus on particular cases (such as the five most important journals in a given field), or the field as a whole (all journals in a given field). Likewise, research has also examined the number of authors per paper in specific journals and fields (Sacco & Milana, 1984; Plummer et al., 2023). In the former case, a sampling strategy would focus on gathering first the population of papers of the journals under study to then compute the sample size, while in the last, the sampling strategy would consider gathering the population of the papers published in the field (i.e., all journals) to then compute the sample size. In this paper, we focus on the design of sample sizes that are representative of the field, although the same strategy may be followed to analyze other target populations. In what follows, a simple step-by-step guide is presented to manage, process, and compute a representative, proportional and random sample to infer reliable findings of scientometric data, such as the above-mentioned, coming from journals and papers of different research fields and ranked in the JCR.

Designing a Representative, Proportional and Random Sample: Example for the Field of Communication

This guide focuses on the design of representative, proportional and random samples based on data coming from the Journal Citation Report of the field of Communication in 2022, but a similar approach may be followed to examine other fields and years. As data for the population is coming directly from the JCR ranking, the sample is only representative of journals indexed by this ranking and included in this field ($N_{journals}$ = 96).

Step 1: Selecting the Field(s) of Research and Year(s)

The first step is selecting in the JCR ranking (or other) the field of study and year to proceed to compute the representative sample. Depending on the number of fields and the number of years the researcher is interested in examining, the sample size may accordingly vary. For practical reasons, the study will only focus on one field (Communication), and one year (the most recent available data, 2022), but the same procedure could be implemented to examine multiple fields and years.

Step 2: Generating the Population of Articles for Each Journal in the JCR

To design a sample size representative of the target population, the target population needs to be known first. In this case, the target population is the sum of all articles published in all journals indexed in JCR in 2022. Accordingly, N is the population size, in this case, the total number of articles. N_i is the size of each stratum, i.e., the number of articles for each journal:

$$, \qquad (1)$$

where K is the number of journals indexed in JCR in 2022. Although not perfectly accurate[1], the population could be computed by extracting this information directly from the JCR, by examining each journal in the field of communication and checking the “content metrics” section, the source data, and the “total citable items”. Specifically, this information is revealed as “number in JCR Year 2022 (A)”, by considering either articles, reviews, or both combined. In this guide, only articles will be computed. Alternatively, this information can be also gathered directly by signing into the JCR

[1] There may be a time gap or discrepancy between articles published each year and their inclusion in an issue, which may be later. The JCR lists the articles published in a giver year, no matter the issue of inclusion.

platform, and customizing the information displayed. Specifically, “total Articles” in the checklist of customized features needs to be checked. Since the sample needs to be representative of the field and proportional to the journals’ research output, the output for each journal needs to be computed. Accordingly, the population or articles is the sum of the research output (i.e., total publications) of all journals (see Table 1 as example, first column), Expression (1). In this example, the population of articles from all journals ranked in the JCR list in communication in 2022 is N_{articles} = 5,045.

<Insert Table 1>

Step 3: Computing the Representative Sample Size

Since the population in this example is finite (N_{journals} = 96, N_{articles} = 5,045), the sample size can be easily computed by many online sample size calculators, including the following parameters: population (N_{articles} = 5,045), confidence level (typically 95%), and margin of error (typically 5%). The mathematical expression for computing the sample size in finite and large populations, (n) that allows to warrant a 95% confidence (1-α) and a maximum margin of error (e) of 5%, computed with the population proportion in which p = q = 50, will be:

(2)

Where:

N is the population size.

z is the value of the normal distribution (0,1) or standardized normal distribution. As the confidence level is set at 95%, the value of $z = 1{,}96$.

e is the sampling error.

p is the percentage of success.

q is the complementary percentage.

In the example of the field of communication, the calculations yielded a sample size of 358 articles with the above-mentioned parameters. This sample size is the base for the scientometric analysis of the field. Consequently, any findings related to the research questions of the study (such as the gender representation of first authors in the field, the number of authors in the field, the funded/not funded articles in the field, etc.) to be considered representative should keep this sample size.

Step 4: Computing the Representative and Proportional Sample Size

If the analysis wants to adjust the sample size to the proportions of the research output of each journal (since journals, as represented in Table 1, have different levels of publications), a proportional sample size must be designed. If not, a random sample (explained in the next step) of 358 articles can be randomly selected directly from the population of articles from all journals. In this example, a representative and proportional random sample is computed. For the sake of simplicity, the guide addresses first the representative and proportional sample size (which is optional) and follows-up with the random selection of articles.

Again, for computing the representative and proportional random sample of the field of communication, some information needs to be known. These elements are the population of articles in the field ($N_{\text{articles}} = 5,045$), the representative sample size of articles in the field ($N_{\text{sample\_size}} = 358$), and the total number of papers published by a given journal (i.e., $N_{\text{Communication\_Methods\_and\_Measures}} = 11$). It is considered that the sample size of each proportional stratum is directly proportional to the size of each stratum, that is,

(3)

Considering, for example, the sample size for *Communication Methods and Measures,*

Knowing this information, computing the representative and proportional sample for *Communication Methods and Measures* will be (N = 1), resulted from multiplying the total number of papers published by the journal ($N_{\text{Communication\_Methods\_and\_Measures}} = 11$) by the representative sample ($N_{\text{sample\_size}} = 358$), and dividing it by the population ($N_{\text{articles}} = 5,045$). As the accurate result of this computation is 0.78, and the unit of analysis is the paper, no less than 1 paper can be examined. Accordingly, the results should be rounded up in cases when decimals are above 0.5 and should be rounded down when decimals are below 0.5. As seen in Table 1 the discrepancy between the representative sample and the adjusted representative and proportional sample as a result of rounding up or down according to the decimals, is only 3 papers in favor of the adjusted representative and proportional sample (361-358).

Step 5: Computing the Representative, Proportional, and Random Sample Size

Once the dataset is representative to the field of research and proportional to the journals’ research output, the next step is to randomly select the papers that will ultimately be the basis of the scientometric analysis. This random selection means that all articles from the journals under examination have the same chance of being selected for the study.

To satisfactorily compute a random sample of papers coming from the representative and proportional sample, two elements are needed: the list of articles for each journal, and a random number generator, which is easy to find online.

To have the list of articles for each journal in a given year, each journal should be examined extracting this information, once again, from the “content metrics” section, “source data”, and the “total citable items” of the JCR platform. Specifically, in the total citable items, the column “Combined (C)”, offers the possibility to directly export this data to a manageable software, such as .XLS or .CSV. In this selection, “articles” must be checked while “reviews” must be unchecked, if the analyses only focus on article types. Once the dataset is downloaded, every entry (i.e., paper of the journal), should be assigned with a numerical identification as reflected in Table 2 for *Communication Methods and Measures*.

<Insert Table 2>

The reason why every entry (i.e., paper) should have a numerical identification is because it will be the basis for the random selection of papers. For the case of *Communication Methods and Measures,* the minimum numeral identification is 1, while the maximum is 11, corresponding to the number of articles published by the journal. Accordingly, in the (online) random number generator, the minimum number will be assigned a value of 1 and the maximum will be 11. According to this calculation (reported in Table 1), *Communication Methods and Measures* only have one paper in the representative and proportional sample. Therefore, only one random number should be generated in the random number generator. In the case of this example, the random number generated was 6, meaning that entry number six in the dataset (i.e., “Inter-annotator Agreement Using the Conversation Analysis Modelling Schema, for Dialogue”) should be included in the final representative, proportional and random sample for the field of communication. The same procedure should be implemented for the rest of the journals to compute the final representative, proportional and random sample of 361 articles.

Conclusion

This research guide represents a step forward to design representative and proportional random samples of research fields to gather scientometric data. The step-by-step protocol is intended to guide scholars in the design of scientometric projects aiming at producing reliable results that can be generalizable to the target population (i.e., the research field). Tested in the field of communication and based on the JCR list and the journals’ research output of this ranking, the guide can also be implemented to examine multiple fields across several years. Likewise, other rankings or journal lists (such as Scopus) could be considered by applying a similar procedure.

References

Andersen, J. P., Nielsen, M. W., Simone, N. L., Lewiss, R. E., & Jagsi, R. (2020). COVID-19 medical papers have fewer women first authors than expected. *elife*, *9*, e58807

Bourdieu, P., Chamboredon, J. C., & Passeron, J. C. (1991). *The craft of sociology: Epistemological preliminaries*. Walter de Gruyter.

Chalmers, A. F. (2013). *What is this thing called science?* Hackett Publishing

Denby, K. J., Szpakowski, N., Silver, J., Walsh, M. N., Nissen, S., & Cho, L. (2020). Representation of women in cardiovascular clinical trial leadership. *JAMA Internal Medicine*, *180*(10), 1382-1383.

Gauderman, W. J. (2002). Sample size requirements for association studies of gene-gene interaction. *American journal of epidemiology*, *155*(5), 478-484.

Goyanes, M. (2017). Desafío a la investigación estándar en comunicación: Crítica y alternativas. Barcelona: Editorial UOC.

Goyanes, M. (2020). Against dullness: on what it means to be interesting in communication research. *Information, communication & society*, *23*(2), 198-215.

Goyanes, M., Demeter, M., Grané, A., Tóth, T., & de Zúñiga, H. G. (2023). Research patterns in communication (2009–2019): testing female representation and productivity differences, within the most cited authors and the field. *Scientometrics*, *128*(1), 137-156.

Grafström, A., & Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, *41*(2), 277-290.

Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A., & Kocher, J. P. (2013). Calculating sample size estimates for RNA sequencing data. *Journal of computational biology*, *20*(12), 970-978.

Jenkins, D. G., & Quintana-Ascencio, P. F. (2020). A solution to minimum sample size for regressions. *PloS one*, *15*(2), e0229345

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267.

Pinho-Gomes, A. C., Peters, S., Thompson, K., Hockham, C., Ripullone, K., Woodward, M., & Carcel, C. (2020). Where are the women? Gender inequalities in COVID-19 research authorship. *BMJ global health*, *5*(7), e002922

Plummer, S., Sparks, J., Broedel-Zaugg, K., Brazeau, D. A., Krebs, K., & Brazeau, G. A. (2023). Trends in the Number of Authors and Institutions in Papers Published in AJPE 2015-2019. *American Journal of Pharmaceutical Education*, *87*(2).

Omair, A. (2014). Sample size estimation and sampling techniques for selecting a representative sample. *Journal of Health specialties*, *2*(4), 142.

Sacco, W. P., & Milana, S. (1984). Increase in number of authors per article in ten APA journals: 1960–1980. *Cognitive Therapy and Research*, *8*, 77-83.

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia medica*, *31*(1), 27-53.

Taherdoost, H. (2017). Determining sample size; how to calculate survey sample size. *International Journal of Economics and Management Systems*, *2*.

Tables

Table 1. *Total publications by journal in the field of Communication in 2022 with the representative and proportional sample and the adjusted representative and proportional sample*

Journal name	Total Articles	Representative and proportional sample	(Adjusted) Representative and proportional sample
Communication Methods and Measures	11	0.78	1
Science Communicatin	22	1.56	2
Journal of Communication	41	2.90	3
Political Communication	31	2.19	2
Journal of Computer-Mediated Communication	31	2.19	2
International Journal of Advertising	65	4.61	5
Communication Research	33	2.34	2
Journal of Advertising	44	3.12	3
Comunicar	36	2.55	3
Telecommunications Policy	102	7.23	7
Digital Journalism	109	7.73	8
Social Media + Society	129	9.15	9
New Media & Society	252	17.88	18
Human Communication Research	32	2.27	2
Mobile Media & Communication	34	2.41	2
Policy and Internet	45	3.19	3
International Journal of Press-Politics	55	3.90	4
Journal of Health	78	5.53	6

Communication			
Media Psychology	37	2.62	3
Information Communication & Society	101	7.16	7
Public Relations Review	83	5.88	6
Profesional de la Información	114	8.08	8
Public Understanding of Science	71	5.03	5
Health Communication	188	13.34	13
Communication Theory	15	1.06	1
Journalism & Mass Communication Quarterly	51	3.61	4
Information Society	25	1.77	2
Public Opinion Quarterly	38	2.69	3
Journal of Public Relations Research	26	1.84	2
Media Culture & Society	112	7.94	8
International Journal of Conflict Management	41	2.90	3
Journal of Broadcasting & Electronic Media	43	3.05	3
Media and Communication	91	6.45	6
Journalism Studies	105	7.45	7
Mass Communication and Society	54	3.83	4
Psychology of Popular Media	66	4.68	5
Journal of Children and Media	25	1.77	2
Journalism	119	8.44	8
Cyberpsychology-Journal of Psychosocial Research on Cyberspace	57	4.04	4
Games and Culture	58	4.11	4
Convergence-The International Journal of Research into New Media Technologies	108	7.66	8
Journal of Social and Personal Relationships	195	13.83	14
International Journal of Business Communication	43	3.05	3
Research On Language and Social Interaction	18	1.27	1
Chinese Journal of Communication	27	1.91	2
Communication & Sport	42	2.98	3

Environmental Communication-A Journal of Nature and Culture	59	4.18	4
Journal of Information Technology & Politics	31	2.19	2
Asian Journal of Communication	23	1.63	2
Communication Monographs	12	0.85	1
Journal of Advertising Research	24	1.70	2
Management Communication Quarterly	40	2.83	3
European Journal of Communication	36	2.55	3
Media International Australia	58	4.11	4
Journal of Applied Communication Research	56	3.97	4
Discourse & Society	46	3.26	3
Journal of Business and Technical Communication	15	1.06	1
Journalism Practice	123	8.72	9
Journal of Language and Social Psychology	29	2.05	2
Television & New Media	51	3.61	4
Social Semiotics	48	3.40	3
Discourse Context & Media	40	2.83	3
Communication and Critical-Cultural Studies	30	2.12	2
Written Communication	26	1.84	2
International Communication Gazette	36	2.55	3
Discourse & Communication	37	2.69	3
Journal of Media Ethics	23	1.63	2
International Journal of Public Opinion Research	41	2.90	3
Discourse Studies	38	2.69	3
Feminist Media Studies	175	12.41	12
Journal of Media Psychology-Theories Methods and Applications	43	3.05	3
International Journal of Communication	124	8.79	9
Communication Culture & Critique	36	2.55	3

IEEE Transactions on Professional Communication	27	1.91	2
International Journal of Mobile Communications	35	2.48	2
Personal Relationships	32	2.27	2
Communications-European Journal of Communication Research	41	2.90	3
Language & Communication	59	4.18	4
Critical Discourse Studies	45	3.19	3
Interaction Studies	10	0.70	1
Visual Communication	45	3.19	3
Javnost-The Public	15	1.06	1
Argumentation	24	1.70	2
Quarterly Journal of Speech	33	2.34	2
African Journalism Studies	14	0.99	1
Critical Studies in Media Communication	28	1.98	2
Journal of African Media Studies	25	1.77	2
Continuum-Journal of Media & Cultural Studies	26	1.84	2
Rhetoric Society Quarterly	27	1.91	2
Narrative Inquiry	15	1.06	1
Translator	22	1.56	2
Text & Talk	63	4.47	4
Signs and Society	17	1.20	1
Technical Communication	13	0.92	1
Journal of Media Economics	10	0.70	1
Tijdschrift Voor Communicatiewetenschap	16	1.13	1
TOTAL NUMBER OF ARTICLES	5,045	357.58	361

Table 2. *Articles published in Communication Methods and Measures in 2022 in the field of Communication according to the JCR*

Numerical ID	Title of the Paper
1	A New Scale for Measuring Identity Insecurity
2	An Empirical Investigation of Inadequate Statistical Reporting Practices in Communication Meta-Analyses and Their Consequences
3	Communication Quality Analysis: A User-friendly Observational Measure of Patient-Clinician Communication
4	Computer Vision and Internet Meme Genealogy: An Evaluation of Image Feature Matching as a Technique for Pattern Detection
5	Correcting Sample Selection Bias of Historical Digital Trace Data: Inverse Probability Weighting (IPW) and Type II Tobit Model
6	Inter-annotator Agreement Using the Conversation Analysis Modelling Schema, for Dialogue
7	Investigating Opinions on Public Policies in Digital Media: Setting up a Supervised Machine Learning Tool for Stance Classification
8	Lifting the Veil on the Use of Big Data News Repositories: A Documentation and Critical Discussion of a Protest Event Analysis
9	Metrics of News Audience Polarization: Same or Different?
10	Promises and Pitfalls of Social Media Data Donations
11	Strong-Form Frequentist Testing in Communication Science: Principles, Opportunities, And Challenges