1 What Is Scientometrics?

R. Heller · Date: 2026-05-09 · Last changed: 2026-10-31

1.1 Learning objectives

After completing this chapter, you will be able to:

- Define scientometrics, bibliometrics, and informetrics and explain how the three fields relate
- Trace the historical development of scientometrics from Garfield's citation indexes through Price's quantitative studies to the present day
- Identify the types of research questions that scientometric methods can address
- Retrieve a small sample of scholarly works from OpenAlex and produce basic descriptive statistics in R
- Create simple visualisations of publication trends and journal distributions

1.2 Setup

```
library(tidyverse)
library(openalexR)
library(glue)

set.seed(20260509)

source(here::here("R", "api_helpers.R"))
source(here::here("R", "utils.R"))
source(here::here("R", "sci_palette.R"))
```

1.3 Conceptual background

Scientometrics is the quantitative study of science itself — its growth, structure, productivity, and impact. Where scientists in most disciplines study the natural or social world, scientometricians turn the lens inward: they study the system that produces scientific knowledge.

The term sits within a family of overlapping labels. **Bibliometrics**, the oldest sibling, focuses on the statistical analysis of publications — books, journal articles, and their metadata such as authors, citations, and keywords. The word was coined by Alan Pritchard in 1969 to replace the earlier "statistical bibliography". **Informetrics** is the broadest umbrella, encompassing any quantitative study of information transfer, including web analytics, patent analysis, and library usage statistics. In practice, the three terms are often used interchangeably, but it helps to remember that bibliometrics emphasises the *document*, scientometrics emphasises the *science*, and informetrics emphasises the *information*.

1.3.1 A brief history

The intellectual roots of the field trace back to at least 1955, when Eugene Garfield (1955) proposed the idea of a citation index for science. Garfield argued that citations create an "association of ideas" — a traceable web of intellectual influence — and that indexing these links would transform how scholars discover literature. His vision was realised in the Science Citation Index (SCI), launched in 1964 by the Institute for Scientific Information (ISI). For the first time, researchers could systematically trace who cited whom across the entire scientific literature.

A few years later, Derek de Solla Price gave the field its quantitative backbone. In *Little Science, Big Science* (Solla Price 1963), Price marshalled data to show that science was growing exponentially — in the number of journals, articles, and scientists — and that this growth followed remarkably regular statistical patterns. He estimated that the volume of scientific literature doubles approximately every 15 years, a regularity so striking that it came to be known as "Price's Law." Price's work established that science could be studied with the same empirical rigour that it applies to other phenomena.

These foundations were reinforced by earlier statistical regularities. Alfred Lotka (1926) had already shown that the distribution of scientific productivity follows a power law: a small number of authors produce the majority of papers. Samuel Bradford (1934) demonstrated a similar pattern for journal coverage: a core set of journals in any field accounts for a disproportionate share of the relevant literature. These laws, which we explore in detail in the next chapter, remain cornerstones of scientometric theory.

From these beginnings, scientometrics expanded rapidly through the latter half of the twentieth century. The h-index (Hirsch 2005), journal impact factors, co-citation networks, and science maps became standard tools in research evaluation and science policy. Landmark declarations such as the Leiden Manifesto (Hicks et al. 2015) and DORA (American Society for Cell Biology 2012) later established principles for responsible use of these tools, warning against mechanical reliance on any single indicator.

1.3.2 The modern era

Today the field is undergoing another transformation. The launch of **OpenAlex** (Priem et al. 2022) — a fully open index of over 250 million scholarly works, built from Crossref, PubMed, institutional repositories, and other open sources — has democratised access to the raw data that powers scientometric research. Researchers no longer need expensive subscriptions to Web of Science or Scopus to conduct large-scale analyses. This book takes advantage of that openness: every example uses data that any reader can reproduce for free.

1.3.3 Scope of this book

This book teaches you to do scientometrics in R. It covers:

- **Data acquisition** — querying OpenAlex, Crossref, and PubMed programmatically.
- **Core bibliometrics** — computing productivity, impact, and citation indicators.

- **Network analysis** — building and analysing co-authorship, co-citation, and keyword co-occurrence networks.
- **Text and topic analysis** — extracting themes and trends from titles, abstracts, and full texts.
- **Advanced topics** — altmetrics, open science, gender equity, retraction analysis, and more.
- **Production** — building reproducible pipelines, dashboards, and reports.

Each chapter follows the same pattern: conceptual background, a fully reproducible worked example, diagnostics, limitations, and exercises.

1.4 Worked example

To make things concrete from the very first chapter, we fetch a small sample of recent works from OpenAlex and compute some basic descriptive statistics. Think of this as a first taste of the workflows that the rest of the book will develop in depth.

1.4.1 Fetching a sample from OpenAlex

We search for works whose title or abstract mentions “scientometrics”, published between 2014 and 2023, and draw a random sample of 500 records.

```
works ← fetch_openalex(
  entity = "works",
  search = "scientometrics",
  from_publication_date = "2014-01-01",
  to_publication_date = "2023-12-31",
  options = list(sample = 500, seed = 42)
)

cat(glue("Fetched {nrow(works)} works\n"))
```

```
#> Fetched 500 works
```

```
glimpse(works)
```

```
#> Rows: 500
#> Columns: 45
#> $ id                              <chr> "https://openalex.org/W4382464140", "htt…
#> $ title                           <chr> "Desvendando a produção científica em to…
#> $ display_name                    <chr> "Desvendando a produção científica em to…
#> $ authorships                     <list> [<tbl_df[1 x 8]>], [<tbl_df[4 x 8]>], […
#> $ abstract                        <chr> "O instrumento econômico \\\"Pagamento p…
#> $ doi                             <chr> "https://doi.org/10.11606/t.18.2023.tde-…
#> $ publication_date                <date> 2023-04-28, 2016-08-23, 2020-12-16, 202…
#> $ publication_year                <int> 2023, 2016, 2020, 2020, 2022, 2019, 2015…
#> $ relevance_score                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
#> $ fwci                            <dbl> NA, 2.667, 0.000, 1.546, NA, 6.047, NA, …
#> $ cited_by_count                  <int> 1, 12, 1, 3, 7, 14, 0, 1, 0, 14, 0, 62, …
#> $ counts_by_year                  <list> [<data.frame[1 x 2]>], [<data.frame[5 x…
#> $ ids                             <list> <"https://openalex.org/W4382464140", "h…
#> $ type                            <chr> "dissertation", "article", "article", "b…
#> $ is_oa                           <lgl> FALSE, TRUE, TRUE, FALSE, TRUE, TRUE, TR…
#> $ is_oa_anywhere                  <lgl> TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRU…
#> $ oa_status                       <chr> "green", "hybrid", "diamond", "closed", …
#> $ oa_url                          <chr> "http://www.teses.usp.br/teses/disponive…
#> $ any_repository_has_fulltext     <lgl> TRUE, FALSE, TRUE, FALSE, FALSE, TRUE, T…
#> $ source_display_name             <chr> "Digital Library of Theses and Dissertat…
#> $ source_id                       <chr> "https://openalex.org/S4377196381", "htt…
#> $ issn_l                          <chr> NA, "0002-9343", "1562-4730", NA, "2542-…
#> $ host_organization               <chr> "https://openalex.org/I17974374", "https…
#> $ host_organization_name          <chr> "Universidade de São Paulo", "Elsevier B…
#> $ landing_page_url                <chr> "http://dx.doi.org/10.11606/t.18.2023.td…
#> $ pdf_url                         <chr> NA, "https://www.amjmed.com/article/S000…
#> $ license                         <chr> NA, "other-oa", "cc-by", NA, "other-oa",…
#> $ version                         <chr> "publishedVersion", "publishedVersion", …
#> $ referenced_works                <list> <"https://openalex.org/W635460351", "ht…
#> $ referenced_works_count          <int> 150, 30, 6, 1, 8, 12, 53, 447, 1, 91, 28…
#> $ related_works                   <list> <"https://openalex.org/W7216826511", "h…
#> $ concepts                        <list> [<data.frame[3 x 5]>], [<data.frame[11 …
#> $ topics                          <list> [<tbl_df[12 x 5]>], [<tbl_df[12 x 5]>],…
#> $ keywords                        <list> [<data.frame[12 x 3]>], [<data.frame[6 …
#> $ is_paratext                     <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE…
#> $ is_retracted                    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE…
#> $ language                        <chr> "pt", "en", "pt", "en", "en", "es", "pt"…
#> $ sustainable_development_goals   <list> NA, [<data.frame[1 x 3]>], NA, NA, NA, …
#> $ awards                          <list> NA, <"https://openalex.org/G1144012484"…
#> $ funders                         <list> NA, [<data.frame[10 x 3]>], NA, NA, NA,…
#> $ apc                             <list> NA, [<data.frame[2 x 5]>], NA, NA, NA, …
#> $ first_page                      <chr> NA, "54", "38", "1", "70", "137", NA, NA…
#> $ last_page                       <chr> NA, "60.e5", "49", "7", "80", "155", NA,…
#> $ volume                          <chr> NA, "130", NA, NA, "7", "13", NA, NA, NA…
#> $ issue                           <chr> NA, "1", "77", NA, "1", "1", NA, NA, NA,…
```

Each row is a scholarly work. Key columns include `display_name` (the title), `publication_date`, `cited_by_count`, `doi`, and `source_display_name` (the source or journal name).

1.4.2 Basic descriptive statistics

How many works did we retrieve, and what does the citation distribution look like?

```r
works_summary ← works ▷
  summarise(
    n_works       = n(),
    doi_coverage  = scales::percent(mean(!is.na(doi))),
    median_cites  = median(cited_by_count, na.rm = TRUE),
    mean_cites    = round(mean(cited_by_count, na.rm = TRUE), 1),
    max_cites     = max(cited_by_count, na.rm = TRUE),
    zero_cites    = sum(cited_by_count == 0)
  )

works_summary
```

```
#> # A tibble: 1 × 6
#>   n_works doi_coverage median_cites mean_cites max_cites zero_cites
#>     <int> <chr>               <dbl>      <dbl>     <int>      <int>
#> 1     500 90%                     7       22.4       446        104
```

The large gap between the mean and median is a hallmark of citation distributions: most papers receive few citations while a small number accumulate many. This skewness is one of the most robust empirical regularities in scientometrics, and we will return to its theoretical basis in 2.3.

1.4.3 Top journals

Which journals appear most often in our sample?

```r
top_journals ← works ▷
  filter(!is.na(source_display_name)) ▷
  count(source_display_name, sort = TRUE) ▷
  head(10)

top_journals
```

```
#> # A tibble: 10 × 2
#>    source_display_name                                                      n
#>    <chr>                                                                <int>
#>  1 Scientometrics                                                          10
#>  2 Sustainability                                                           7
#>  3 International Journal of Environmental Research and Public Health        5
#>  4 Journal of Informetrics                                                  5
#>  5 OSF Preprints (OSF Preprints)                                            5
#>  6 PLoS ONE                                                                 5
#>  7 DOAJ (DOAJ: Directory of Open Access Journals)                           4
#>  8 Environmental Science and Pollution Research                             4
#>  9 HAL                                                                      4
#> 10 LA Referencia (Red Federada de Repositorios Institucionales de Publica…  4
```

1.4.4 Publication trend over time

```
works |>
  mutate(year = year(publication_date)) |>
  count(year) |>
  ggplot(aes(x = year, y = n)) +
  geom_col(fill = palette_sci(1)) +
  labs(
    x = "Publication year",
    y = "Number of works (sampled)"
  ) +
  theme_sci()
```

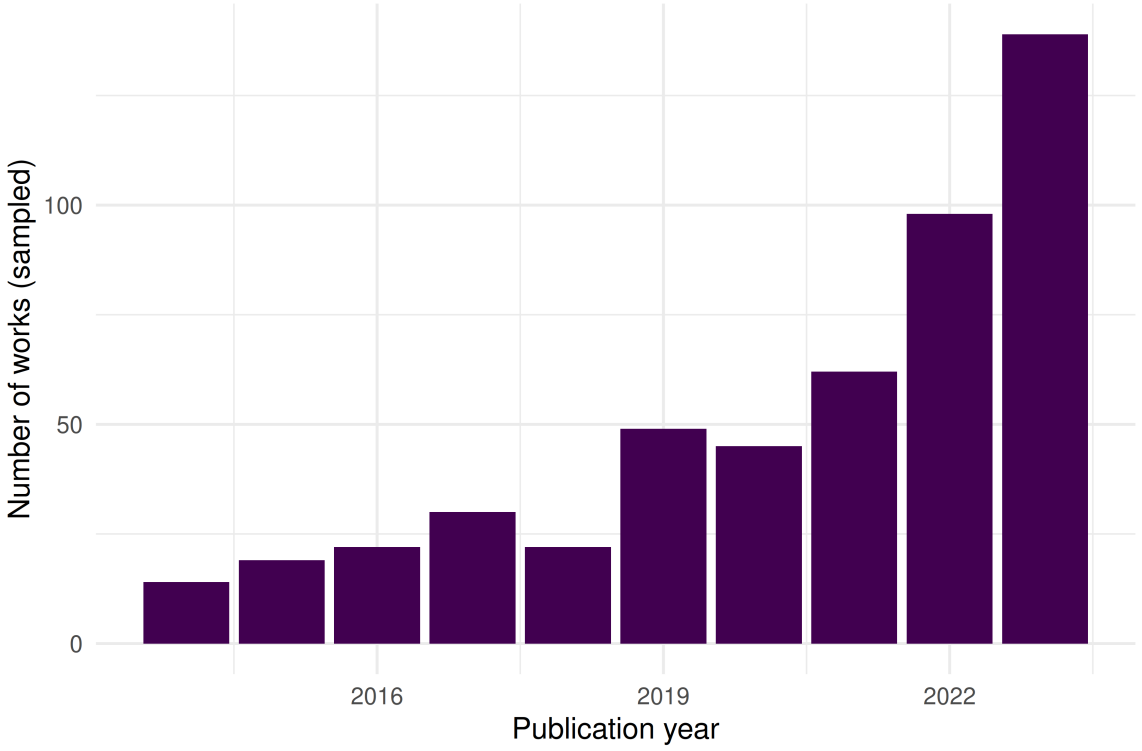


Figure 1.1: Annual publication counts for sampled works related to scientometrics (2014–2023).

The trend shows a growing volume of scientometrics-related publications over the past decade — consistent with the broader exponential growth of science documented by Solla Price (1963).

1.4.5 Top journals visualised

```
top_journals |>
  mutate(source_display_name = fct_reorder(source_display_name, n)) |>
  ggplot(aes(x = n, y = source_display_name)) +
  geom_col(fill = palette_sci(1)) +
  labs(
    x = "Number of works",
    y = NULL
  ) +
  theme_sci()
```

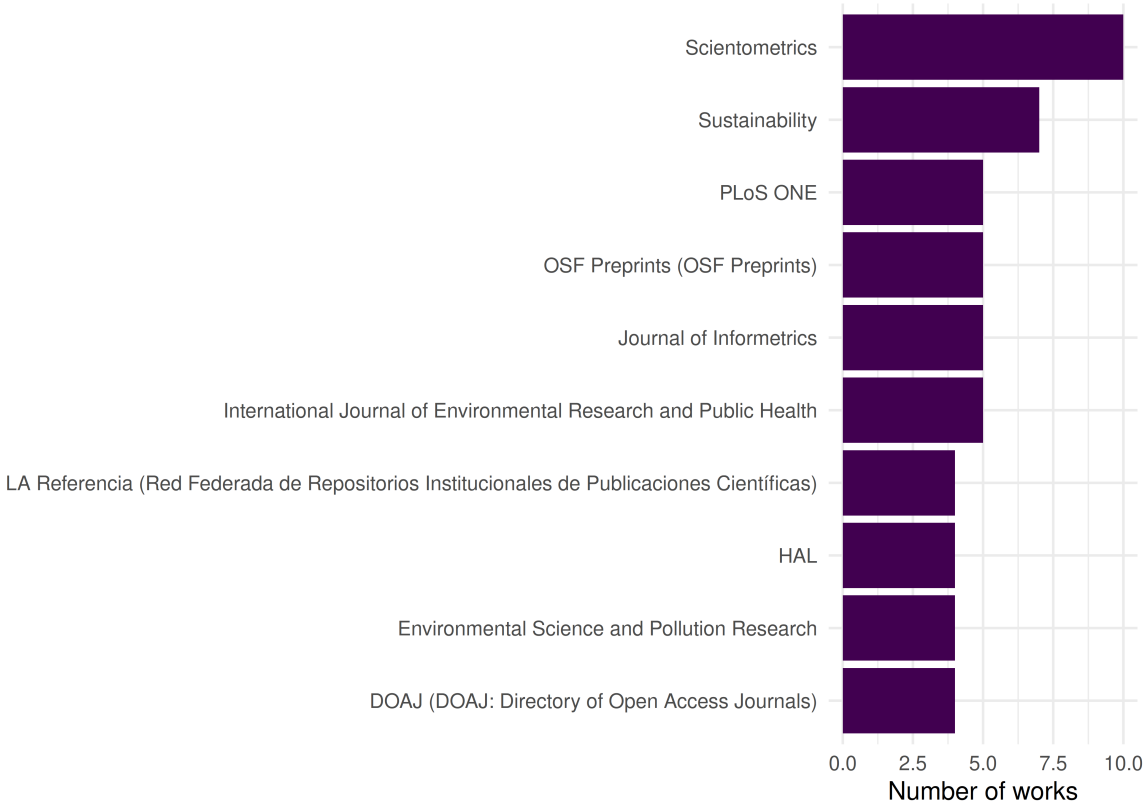


Figure 1.2: Top 10 journals by number of works in the sampled scientometrics corpus.

1.4.6 Citation distribution

```
works ▷
  filter(cited_by_count > 0) ▷
  ggplot(aes(x = cited_by_count)) +
  geom_histogram(bins = 40, fill = palette_sci(1), colour = "white") +
  scale_x_log10() +
  labs(
    x = "Citation count (log scale)",
    y = "Number of works"
  ) +
  theme_sci()
```

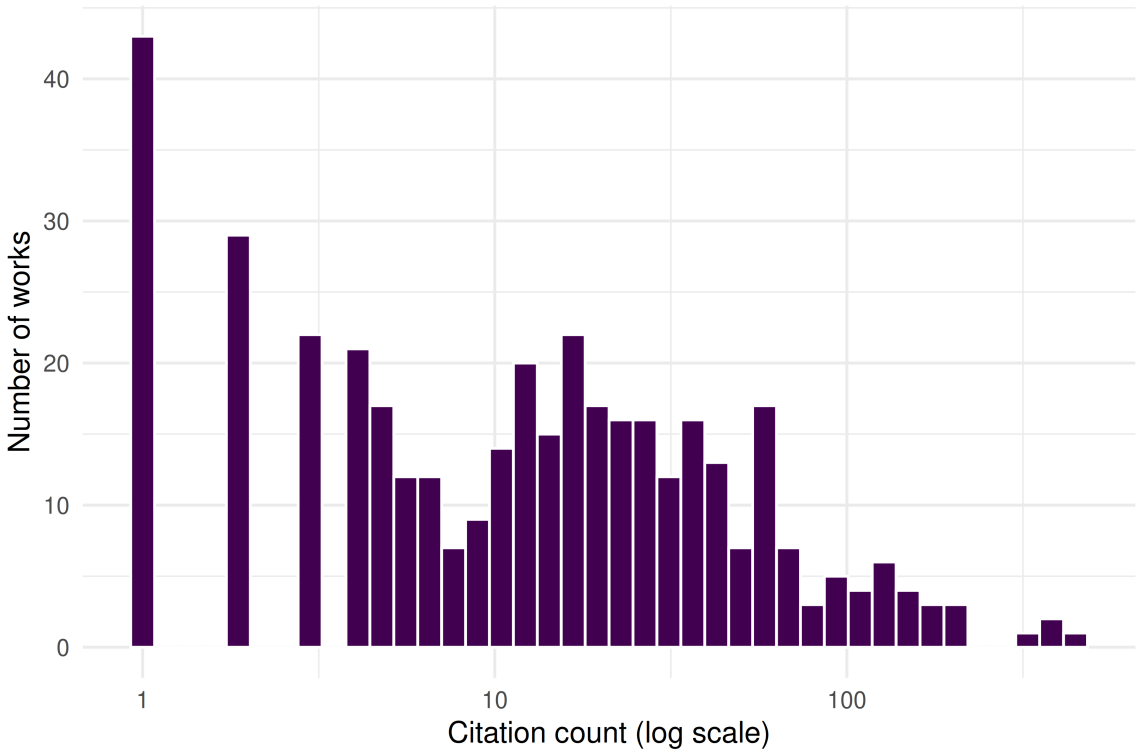


Figure 1.3: Citation distribution of sampled scientometrics works (log-scaled x-axis).

This right-skewed citation distribution is a universal feature of scholarly communication. In Chapter 2, we will see that it connects to deeper theoretical regularities: Lotka's Law, Bradford's Law, and the Matthew Effect.

1.5 Diagnostics and interpretation

Even for a simple descriptive analysis, a few diagnostic checks are essential:

- **Sample vs. population.** We drew a random sample of 500 works. The publication trend reflects the underlying population only if the sampling was truly random within the date window. OpenAlex's `sample` parameter does this, but always verify the date range of returned results.
- **Source coverage.** OpenAlex is strongest in English-language, peer-reviewed journal articles. Conference proceedings, book chapters, and grey literature may be underrepresented. Our "top journals" list reflects what OpenAlex indexes, not the full universe of scientometrics publishing.
- **Citation counts are snapshots.** The `cited_by_count` field reflects the current citation total at the time of the API query. Older papers have had more time to accumulate citations, so direct comparisons across years are misleading without normalization.

```
cat(glue("Date range: {min(works$publication_date)} to
{max(works$publication_date)}\n"))
```

```
#> Date range: 2014-01-01 to 2023-12-31
```

```
cat(glue("Works with zero citations: {sum(works$cited_by_count == 0)}\n"))
```

```
#> Works with zero citations: 104
```

```
cat(glue("Unique journals: {n_distinct(works$source_display_name, na.rm = TRUE)}\n"))
```

```
#> Unique journals: 406
```

1.6 Limitations and responsible use

1.7 Limitations and responsible use

- **Descriptive statistics are not evaluations.** Counting publications and citations describes *activity*, not *quality*. A high publication count may reflect salami slicing; a high citation count may reflect controversy rather than excellence.
- **Data source bias.** OpenAlex, like all databases, has systematic coverage gaps. English-language STEM journals are overrepresented; social sciences, humanities, and non-English scholarship are underrepresented. Always report the data source and its known limitations.
- **Ecological fallacy.** Aggregate statistics about a field do not apply to individual researchers. A growing field does not mean every researcher in it is productive.
- **The Leiden Manifesto** (Hicks et al. 2015) and **DORA** (American Society for Cell Biology 2012) provide essential guardrails. Quantitative data should support, never replace, expert qualitative judgement.

1.8 Common pitfalls

1.9 Common pitfalls

- **Confusing scientometrics with research evaluation.** Scientometrics provides tools for studying science; it does not provide verdicts on the value of individual researchers or papers.
- **Treating OpenAlex counts as ground truth.** OpenAlex counts may differ from Web of Science or Scopus due to different coverage, deduplication, and

metadata parsing. Always triangulate when precision matters.

- **Ignoring the citation window.** Comparing raw citation counts across publication years is misleading. A 2023 paper with 5 citations may be performing better than a 2015 paper with 20. Field- and time-normalised indicators (covered in Chapter 10) address this.
- **Overgeneralising from a sample.** A random sample of 500 works is useful for learning and prototyping. Drawing conclusions about an entire field requires a comprehensive corpus.
- **Forgetting to set a seed.** Random samples are only reproducible if you set the seed both in R (`set.seed()`) and in the API call (`seed` parameter). Without this, every render produces different results.

1.10 Exercises

1. **Explore a different field.** Replace the search term "scientometrics" with a topic of your choice (e.g., "machine learning", "climate change"). Fetch 500 works and compare the publication trend and top journals to those we found above. What differences do you notice?

2. **Uncited works.** What fraction of the sampled works have zero citations? Filter to uncited works and examine their publication years. Is there a pattern? *(Hint: very recent works are more likely to be uncited.)*

3. **Open Access status.** OpenAlex records an `is_oa` field. Compute the proportion of open-access works in your sample and plot how it has changed over time. What trend do you observe?

4. **Growth curve.** Extend the publication trend analysis to 2000–2023. Does the annual publication count follow an exponential growth pattern, as Solla Price (1963) predicted? *(Hint: plot on a log-scaled y-axis and check for linearity.)*

1.11 Solutions

Solutions are provided in E.

1.12 Further reading

- Garfield (1955) — The foundational paper proposing citation indexes for science; the intellectual starting point for the entire field.
- Solla Price (1963) — *Little Science, Big Science*; the first systematic quantitative study of the growth of science.

- Priem et al. (2022) — The OpenAlex paper; describes the open data infrastructure that powers the examples in this book.
- Hicks et al. (2015) — The Leiden Manifesto; ten principles for the responsible use of research metrics.
- American Society for Cell Biology (2012) — The San Francisco Declaration on Research Assessment; a landmark call to improve research evaluation.
- Aria and Cuccurullo (2017) — The `bibliometrix` R package; a comprehensive tool for science mapping and bibliometric analysis.
- Hirsch (2005) — The h-index paper; illustrates how a single indicator reshaped research evaluation discourse.
- Lotka (1926) — The foundational study of scientific productivity distributions; a bridge to Chapter 2.

1.13 Session info

```
sessionInfo()
```

```
#> R version 4.6.1 (2026-06-24)
#> Platform: x86_64-pc-linux-gnu
#> Running under: Ubuntu 24.04.5 LTS
#>
#> Matrix products: default
#> BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
#> LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.26.so;  LAPACK
version 3.12.0
#>
#> locale:
#>  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
#>  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
#>  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
#> [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
#>
#> time zone: UTC
#> tzcode source: system (glibc)
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#>  [1] glue_1.8.1      openalexR_3.1.2 lubridate_1.9.5 forcats_1.0.1
#>  [5] stringr_1.6.0   dplyr_1.2.1     purrr_1.2.2     readr_2.2.0
#>  [9] tidyr_1.3.2     tibble_3.3.1    ggplot2_4.0.4   tidyverse_2.0.0
#>
#> loaded via a namespace (and not attached):
#>  [1] viridis_0.6.5      utf8_1.2.6         sass_0.5.0         generics_0.1.4
#>  [5] xml2_1.6.0         stringi_1.8.9      hms_1.1.4          digest_0.6.39
#>  [9] magrittr_2.0.5     evaluate_1.0.5     grid_4.6.1         timechange_0.4.0
#> [13] RColorBrewer_1.1-3 bookdown_0.48      fastmap_1.2.0      rprojroot_2.1.1
#> [17] jsonlite_2.0.0     gridExtra_2.3.1    httr_1.4.9         viridisLite_0.4.3
#> [21] scales_1.4.0       jquerylib_0.1.4    cli_3.6.6          rlang_1.3.0
#> [25] withr_3.0.3        cachem_1.1.0       yaml_2.3.12        otel_0.3.0
#> [29] tools_4.6.1        tzdb_0.5.0         memoise_2.0.1      here_1.0.2
#> [33] curl_8.1.0         vctrs_0.7.3        R6_2.6.1           lifecycle_1.0.5
#> [37] fs_2.1.0           pkgconfig_2.0.3    pillar_1.11.1      bslib_0.12.0
#> [41] gtable_0.3.6       xfun_0.62          tidyselect_1.2.1   rstudioapi_0.19.0
#> [45] knitr_1.52         farver_2.1.2       htmltools_0.5.9    labeling_0.4.3
#> [49] rmarkdown_2.32     compiler_4.6.1     S7_1.0.0           downlit_0.4.5
```