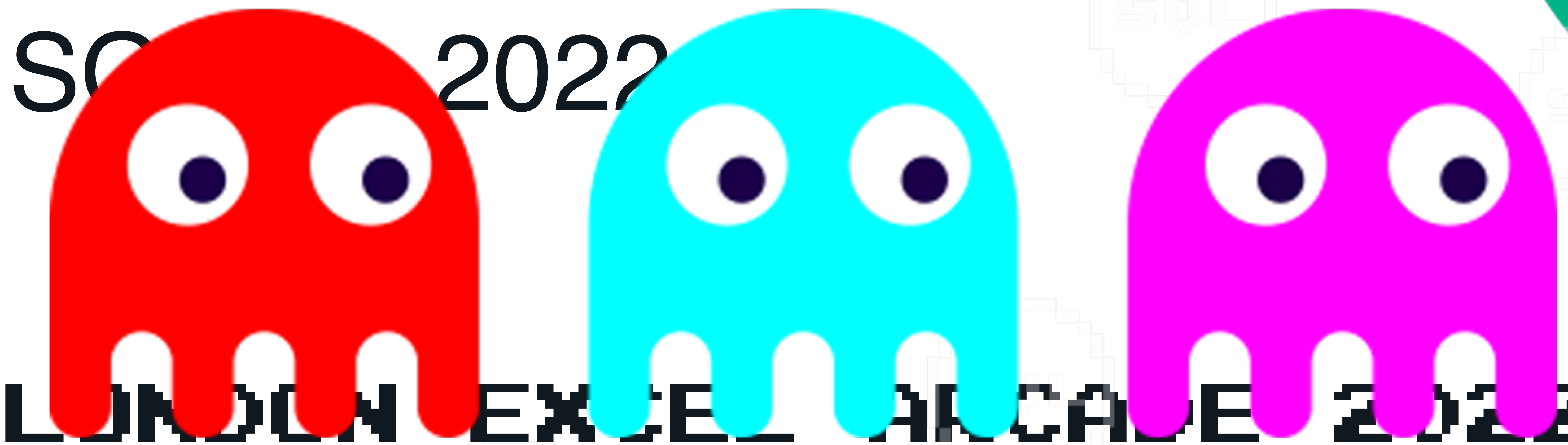
Hunting for fraud with Bedford's law in R

Thomas Hütter

S[illegible] 2022



LONDON EXCEL ARCADE 2022

Hunting for fraud with Bedford’s law in R

Thomas Hütter, Diplom-Betriebswirt

- Application/BI developer, consultant, accidental DBA
- Worked at consultancies, ISVs, end user companies
- SQL Server > 6.5, former „Navision“ > 3.0, R > 3.1.2
- Speaker at SQL events around Europe

@DerFredo https://twitter.com/DerFredo

de.linkedin.com/in/derfredo

www.xing.com/profile/Thomas_Huetter

sqlbits

Developer! DeveloperDeveloper! UK Community Events

SQLdays konferenz

DATA INNOVATION SUMMIT 2021

Agenda

- Introducing: R, the language
- Benford’s law: basics
- Benford’s law: maths
- Applying Benford’s law
- Wrap-up, resources, feedback

Introducing: R the language

- Programming language for statistical computing and visualization, widely used by statisticians, data miners, analysts, data scientists
- Created by **R**oss Ihaka and **R**obert Gentleman, Uni Auckland, in 1993 as an open source implementation of the (1970s) S language
- GNU project, maintained by the R Foundation for Statistical Computing, compiled builds für Mac OS, Linux, Windows, supported by R Consortium
- Extensible through user-created packages, ≈ 19000 available on CRAN
- Commercial support, e.g. since 2007 by Revolution Analytics, acquired by Microsoft in 2015, now provide Microsoft R Open, R Server
- IDEs: R.App, RStudio, MS R Tools for VisualStudio (< Version 2019)
- Support for R now in SQL Server (R services), Power BI, Azure ML

Bedford’s law: basics

- Aka „Newcomb-Benford-Law“ or „Law of the first digit“
- Discovered by astronomer and maths prof Simon Newcomb (published 1881 as „Note on the frequency of use of the different digits in natural numbers“), credited to Frank Benford 1938 („The law of anomalous numbers“)
(a fact which follows Stigler’s law, discovered by Merton ;-))
- An observation about the distribution of leading digits in naturally occurring collections of numerical data
- Intuition: all digits are evenly distributed
- Observation: In logarithm tables, the front pages were more worn
- Conclusion: leading digits are more likely to be small

Bedford’s law: maths

- Evenly distributed digits:
  $P = 1/9 \approx 0.1111$

- First digit D_1 according to Benford:
  $P(D_1=d)$
  $= \log_{10}(d+1) - \log_{10}(d)$
  $= \log_{10}(1 + 1/d)$

d	evenly	Benford
1	0.1111	0.3010
2	0.1111	0.1761
3	0.1111	0.1249
4	0.1111	0.0969
5	0.1111	0.0792
6	0.1111	0.0669
7	0.1111	0.0580
8	0.1111	0.0512
9	0.1111	0.0458

Applying Bedford’s law

- Determine the data / measure to examine
- extract first *significant* digits, regardless of magnitude
- calculate the table of relative density
- compare to Benford’s table
- visualize

DEMO TIME!

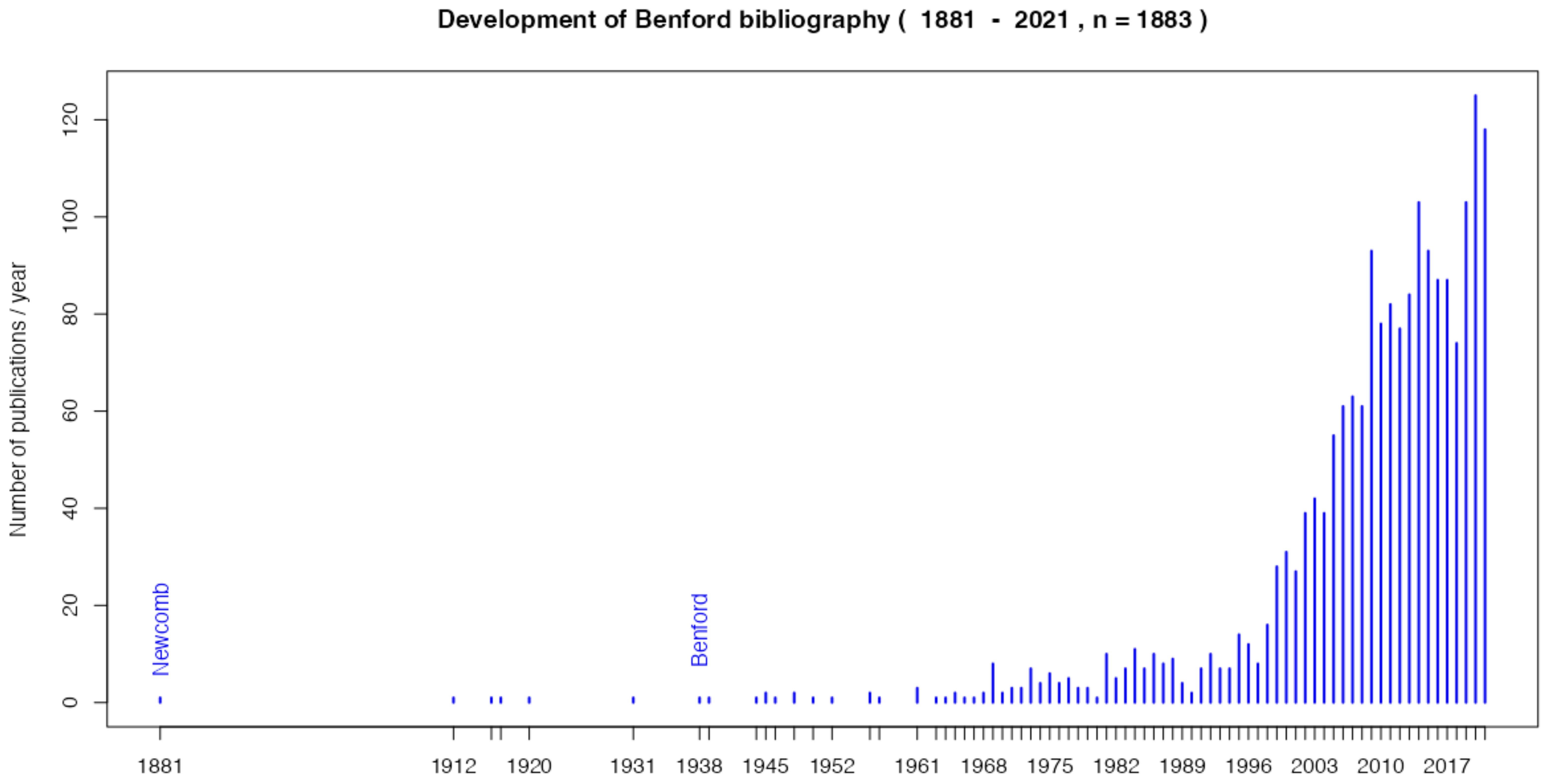
Wrap-up

Benford’s law:

- helps finding deviations from the ‘natural’ distribution of digits
- works for natural or transactional data, the bigger the sample the better
- works best when the values spread over several orders of magnitude
- does not work for numbers influenced by human rules or actions
- has applications in accounting/economics (fraud detection), engineering, environment sciences, medicine, social sciences, election forensics, statistics…

Resources

- Explanations and applications of Benford, incl. 70+ references: en.wikipedia.org/wiki/Benford's_law
- Benford online bibliography: https://www.benfordonline.net/



Hunting for fraud with Bedford’s law in R

Please do the organizers and myself one favour:

(and yourself… and mother nature…)

Submit your feedback for this session now!

https://sqlb.it/?6951

(There’s a text box at the bottom 😉)

Hunting for fraud with Bedford’s law in R

Thank you for your time and interest & keep in touch:

- @DerFredo https://twitter.com/DerFredo
- de.linkedin.com/in/derfredo
- www.xing.com/profile/Thomas_Huetter

This file and the demo script can be found at:

http://j.mp/DerFredoBits22