Improving the accuracy of the Benford algorithm via Monte Carlo simulation for sample size determination

By Franco Arda

Doctoral (DBA) Thesis Defense

Supervisor: Prof. Dr. George Iatridis

Date: 25 February 2021

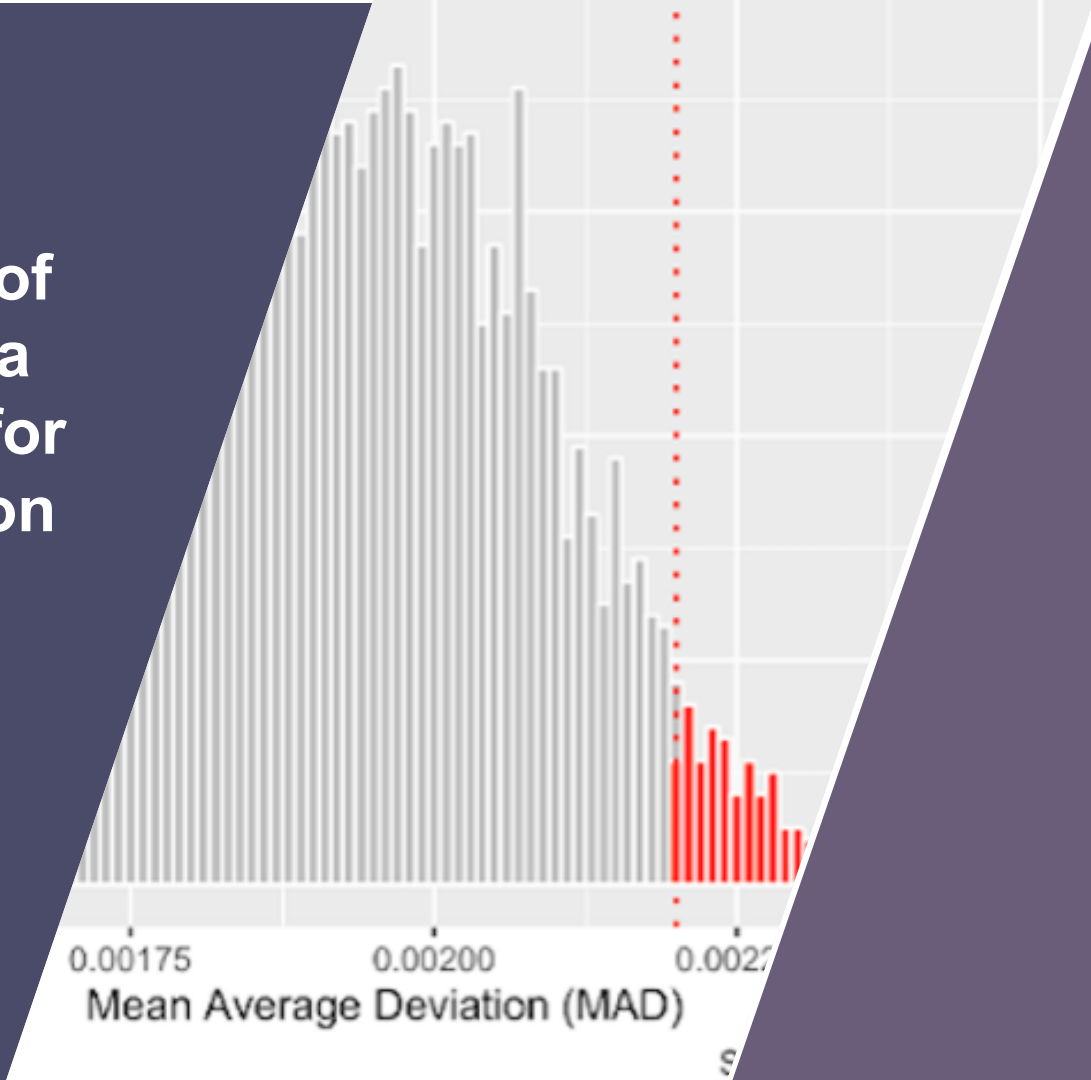
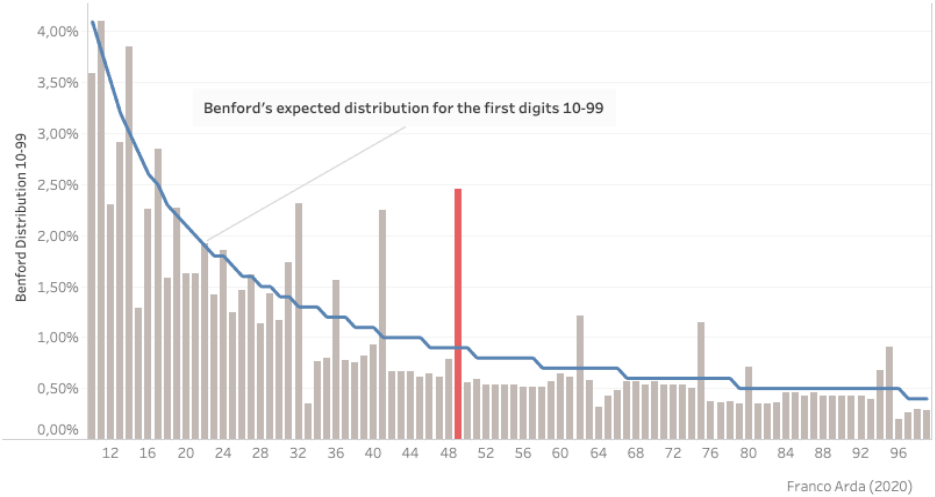


Table of Content – Proposal Defense Outline

Benford’s Law to detect fraud



From the chapter “A short introduction to Benford’s Law” page 23.

*"… for auditors, the **false positives** have become a strong deterrent to the use of Benford's Law …"*

(Miller, 2015)

From the chapter "Introduction" page 16.

Empirical false positive rate at 9%?

	Predicted	
	Positive	Negative
Actual True	1	0
Actual False	0.09	1

From the chapter ”Classification accuracy: benchmark II vs. alternative” page 112.

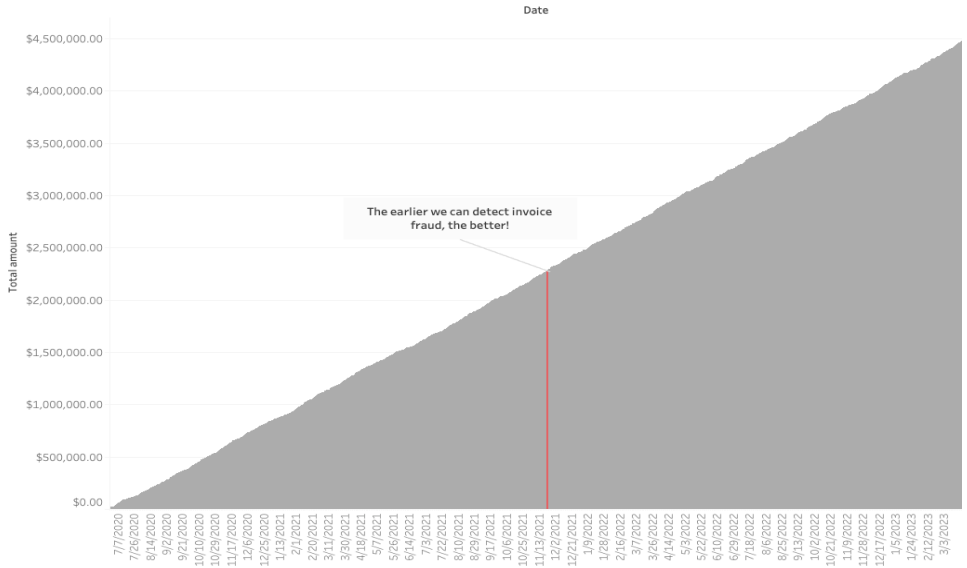
Research Question

"… it is not clear, how large our numbers (sample size) have to be…"

(Nigirini, 2012 and 2020)

From the chapter "Literature Review" page 16.

With fraud detection, the false positives AND the sample size must be *as small as possible.*



From the chapter “The importance of sample size in fraud detection” page 33.

Research Hypotheses

01

Research Hypothesis #1

For Benford's Law, can we determine the sample size required at a given confidence interval?

02

Research Hypothesis #2

With the required sample size, can we improve the accuracy of the Benford algorithm at a statistically significant level with a p-value 0.05?

From the pages 37 and 79.

Review of Literature

Research on Bedford's Law:

1. **Forensic Analytics: Methods and Techniques for Forensic Accounting,** 2020
   Prof. Nigrini is probably the world's foremost expert on Benford's Law.

2. **An Introduction to Bedford's Law**, 2015
   Highly mathematical, but includes problems with Bedford's Law in practice, in particular the high false positive rate.

3. **The Use of Bedford's Law as a Tool for Detecting Fraud in Accounting Data**, 2016
   Offers a deeper analysis of the first-two digit algorithm.

Research on computational statistics in R:

4. **Mathematical Statistics with Resampling**, 2019
   Advanced techniques with Monte Carlo simulations, bootstrap, and challenges with nonlinear distributions.

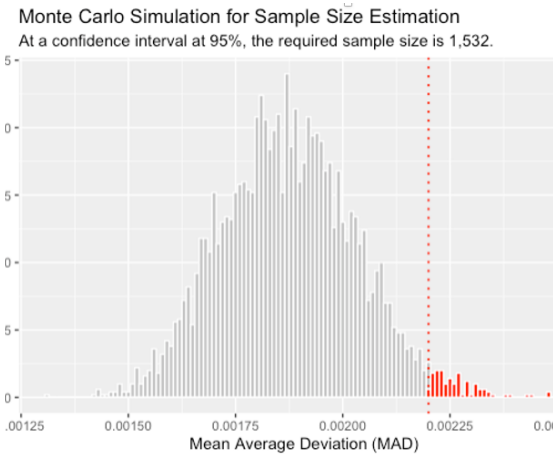
5. **Data Science: Data Analysis and Prediction Algorithms with R**, 2019
   Probability and random variables for running plain-vanilla Monte Carlo simulations in R.

6. **Statistical Inference via Data Science**, 2019
   Offers a library in R to run permutation based hypothesis test with beautiful visualizations.

From the chapter "Literature Review" pages 16.

Research Process

endor_3	Vendor_4	Vendor_5	Vendor_6	Vendor_7	Vendor_8	Vendor_9	Vendor_
7153.52	9418.90	463.45	385.83	20950.77	59.21	157.04	894.54
952.22	24.46	1793.08	61.72	17234.55	16248.00	5243.24	2068.2
18.20	35.38	1487.30	153.89	53.02	575.44	36.04	6474.4
16.35	28131.96	5105.05	32151.41	4830.59	11376.27	651.63	39120.1
9476.62	15.95	2307.81	26.79	9585.17	745.42	1301.37	172.82
0956.55	27874.05	94885.53	21897.78	281.06	1102.55	114.39	5385.1
238.23	78.78	1803.02	13.43	55.41	39336.89	35.03	105.25
30.42	18297.85	3218.10	3922.83	9061.50	554.12	34.67	6263.2
775.64	125.20	258.46	4317.18	19.41	36610.02	2639.98	88470.8
306.40	63620.93	124.62	27593.07	12.27	3735.94	54600.92	16323.0
24.52	98.36	3897.62	83.64	1332.91	681.08	35.78	20155.8
978.42	792.87	20.72	11.58	20.95	33.48	6492.32	21837.3
5377.37	4070.05	140.86	1440.12	49249.29	42.74	24.25	142.65
145.34	11.25	639.73	3760.11	137.66	7284.50	48.13	12.59
12.75	13.33	380.89	2255.28	129.30	15289.74	2000.78	9727.4
133.44	1964.26	833.30	21.14	2434.45	472.93	29.40	6338.7
19.02	46687.43	10.13	23615.66	13.89	98.08	12953.88	973.64
5318.32	81.36	6956.65	2823.58	114.92	333.27	1741.00	4458.6
414.95	83791.51	676.08	377.05	10.06	53161.84	401.05	43411.0
4859.36	247.74	24479.36	190.90	19.23	7085.98	10.38	2284.5
9796.64	3535.09	36.81	84.26	169.04	74.27	207.40	31739.4
313.91	347.70	18382.31	14401.24	97.90	133.78	44.87	1875.8
1713.03	10.11	782.71	5425.00	19.73	63.15	35.97	17.52
813.00	26.55	2696.50	10.89	1561.71	62058.32	27.87	2525.8
7948.97	146.82	98.27	544.50	60701.58	14.60	19.23	10794.4
427.73	62.06	41304.75	29.32	24524.49	10.70	99815.96	5039.6



Vendor:	Benchmark II:	Ground truth:	Alternative:	Ground truth:
Vendor 14	0.000239	✔	0.000239	✔
Vendor 15	0.000239	✔	0.000239	✔
Vendor 16	0.000239	✔	0.000239	✔
Vendor 17	0.000121	✔	0.000121	✔
Vendor 18	0.000121	✔	0.000121	✔
Vendor 19	0.000121	✔	0.000121	✔
Vendor 20	0.000121	✔	0.000121	✔
Vendor 21	0.000121	✔	0.000121	✔
Vendor 22	0.000121	✔	0.000121	✔
Vendor 23	0.000121	✔	0.000121	✔
Vendor 24	0.000121	✔	0.000121	✔

Step 01

Create a dataset of 100 vendors with different invoice amounts.

Step 02

Simulate the required sample size based a confidence interval (e.g., 5%).

Step 03

Test the accuracy of the algorithm and run a hypothesis test at a p-value of 0.05.

The synthetic dataset of almost 1 million invoices.

In order to test our hypothesis, we created a synthetic dataset of 100 vendors with up to 10,000 invoices each.

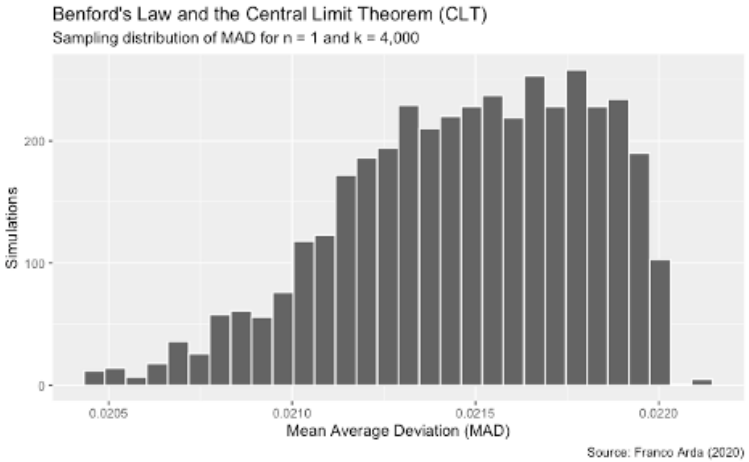
- 85 vendors with Benford conform invoice amounts.
- 10 vendors with only a few invoices.
- 5 vendors with randomized invoices (i.e., fraudulent).

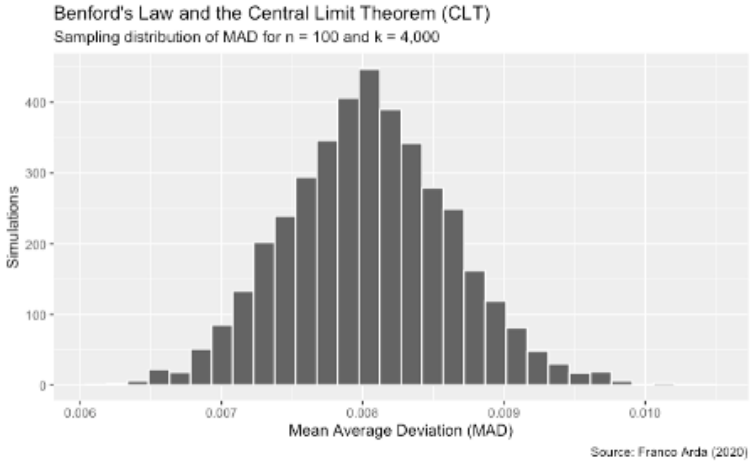
ndor_3	Vendor_4	Vendor_5	Vendor_6	Vendor_7	Vendor_8	Vendor_9	Vendor_
153.52	9418.90	463.45	385.83	20950.77	59.21	157.04	894.5
952.22	24.46	1793.08	61.72	17234.55	16248.00	5243.24	2068.2
18.20	35.38	1487.30	153.89	53.02	575.44	36.04	6474.4
16.35	28131.96	5105.05	32151.41	4830.59	11376.27	651.63	39120.
476.62	15.95	2307.81	26.79	9585.17	745.42	1301.37	172.8
956.55	27874.05	94885.53	21897.78	281.06	1102.55	114.39	5385.1
238.23	78.78	1803.02	13.43	55.41	39336.89	35.03	105.2
30.42	18297.85	3218.10	3922.83	9061.50	554.12	34.67	6263.2
775.64	125.20	258.46	4317.18	19.41	36610.02	2639.98	88470.
306.40	63620.93	124.62	27593.07	12.27	3735.94	54600.92	16323.
24.52	98.36	3897.62	83.64	1332.91	681.08	35.78	20155.
978.42	792.87	20.72	11.58	20.95	33.48	6492.32	21837.
377.37	4070.05	140.86	1440.12	49249.29	42.74	24.25	142.6
145.34	11.25	639.73	3760.11	137.66	7284.50	48.13	12.59
12.75	13.33	380.89	2255.28	129.30	15289.74	2000.78	9727.4
133.44	1964.26	833.30	21.14	2434.45	472.93	29.40	6338.7
19.02	46687.43	10.13	23615.66	13.89	98.08	12953.88	973.6
318.32	81.36	6956.65	2823.58	114.92	333.27	1741.00	4458.6
14.95	83791.51	676.08	377.05	10.06	53161.84	401.05	43411.
859.36	247.74	24479.36	190.90	19.23	7085.98	10.38	2284.5
796.64	3535.09	36.81	84.26	169.04	74.27	207.40	31739.
13.91	347.70	18382.31	14401.24	97.90	133.78	44.87	1875.8
713.03	10.11	782.71	5425.00	19.73	63.15	35.97	17.52
313.00	26.55	2696.50	10.89	1561.71	62058.32	27.87	2525.8
948.97	146.82	98.27	544.50	60701.58	14.60	19.23	10794.
427.73	62.06	41304.75	29.32	24524.49	10.70	99815.96	5039.6

From the chapter “Creating a dataset of 100 vendors” on page 27.

Unexpected

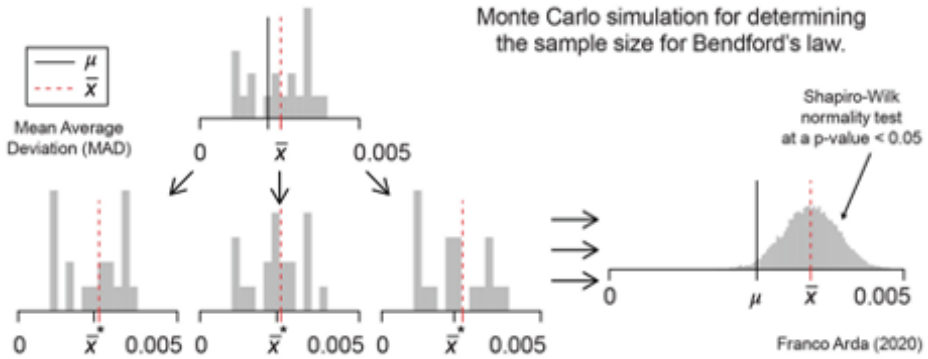
Benford’s distribution and the Central Limit Theorem (CLT)





From the chapter “Benford’s distribution and the Central Limit Theorem” on page 58.

Breakthrough approach in sample size determination

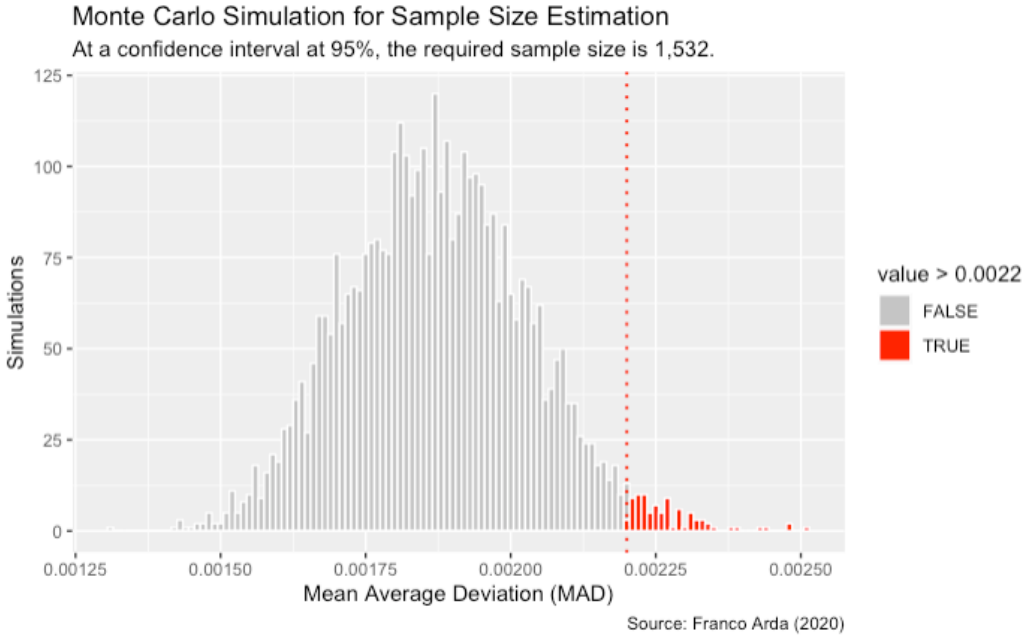


Research Methodology: An iterative process

Sample size	Number of simulations	Shapiro-Wilk normality test p-value	Mean Average Deviation	Lower bound confidence level	Upper bound confidence level	Confidence Interval
10	1,000	0.048	[illegible]	[illegible]	[illegible]	0%
100	1,000	0.001	[illegible]	[illegible]	[illegible]	0%
1,000	1,000	0.321	[illegible]	[illegible]	[illegible]	17.1%
1,700	3,000	0.044	[illegible]	[illegible]	[illegible]	99.4%
1650	4000	0.065	[illegible]	[illegible]	[illegible]	99.3%
1600	4,000	0.021	[illegible]	[illegible]	[illegible]	97.6%
1,550	4,000	0.388	[illegible]	[illegible]	[illegible]	96.3%
1,540	4,000	0.000	[illegible]	[illegible]	[illegible]	95.4%
1,525	4,000	0.000	[illegible]	[illegible]	[illegible]	92.4%
1,530	4,000	0.000	[illegible]	[illegible]	[illegible]	94.2%
1,535	4,000	0.000	[illegible]	[illegible]	[illegible]	95.4%
1,532	4,000	0.000	[illegible]	[illegible]	[illegible]	95%

From the chapter “Monte Carlo sample size determination at a 95% confidence interval” on page 72.

Probably for the first time ever …



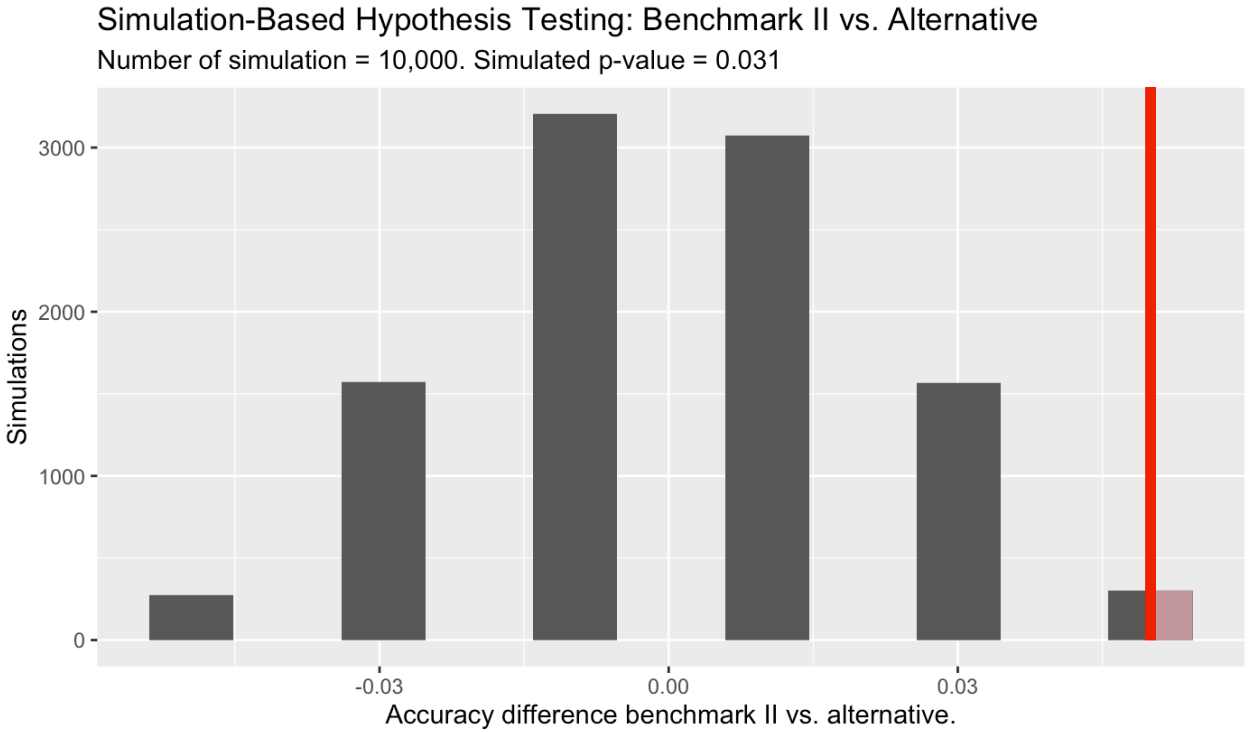
From the chapter “Monte Carlo sample size determination at a 95% confidence interval” on page 72.

Research Results

	Benchmark I	Benchmark II	Alternative
Accuracy	95.23%	95.50%	100%
Recall	0	0.38	1

From the chapter “Empirical findings” on page 119.

Hypothesis test



Source: Franco Arda (2020).

From the chapter “Accuracy” on page 99.

Importance of the study

Our empirical study revealed for the first time ever the sample size required for the Benford algorithm:

- At a 90% confidence interval, we need a sample size of 1,480.
- At a 95% confidence interval, we need a sample size of 1,532.
- At a 99% confidence interval, we require a sample size of 1,670.

From the chapter “Conclusion” on page 125.

Thank you!

Special thanks to:

- Prof. Dr. George Iatridis
- Prof. Dr. Mark J. Nigrini