Shiny Demonstration Utilizing Bootstrap Approach to the Application of First Digits Analysis

Matthew Corrigan
Statistical Methods Staff (SMS)
Current Employment Statistics (CES)

Disclaimer

- Views expressed are those of the presenter and or author(s) and do not necessarily reflect the views or policies of the Bureau of Labor Statistics.

Problem Space

- 240 data collectors collect employment data from sampled establishments using the CATI system.
- A data collector using CATI may collect data from 50 to 450 establishments each month.
- Data collectors are randomly assigned establishments.
- Randomly selecting an establishment within collector yields a slim possibility of discovering collection errors.

Quality Control

- Supervisors don't use quantitative methods to prioritize which collectors to evaluate.
- Currently selecting which collector to evaluate is done at random.
- There is difficulty in determining if a collector is manufacturing data, also known as curbstoning.

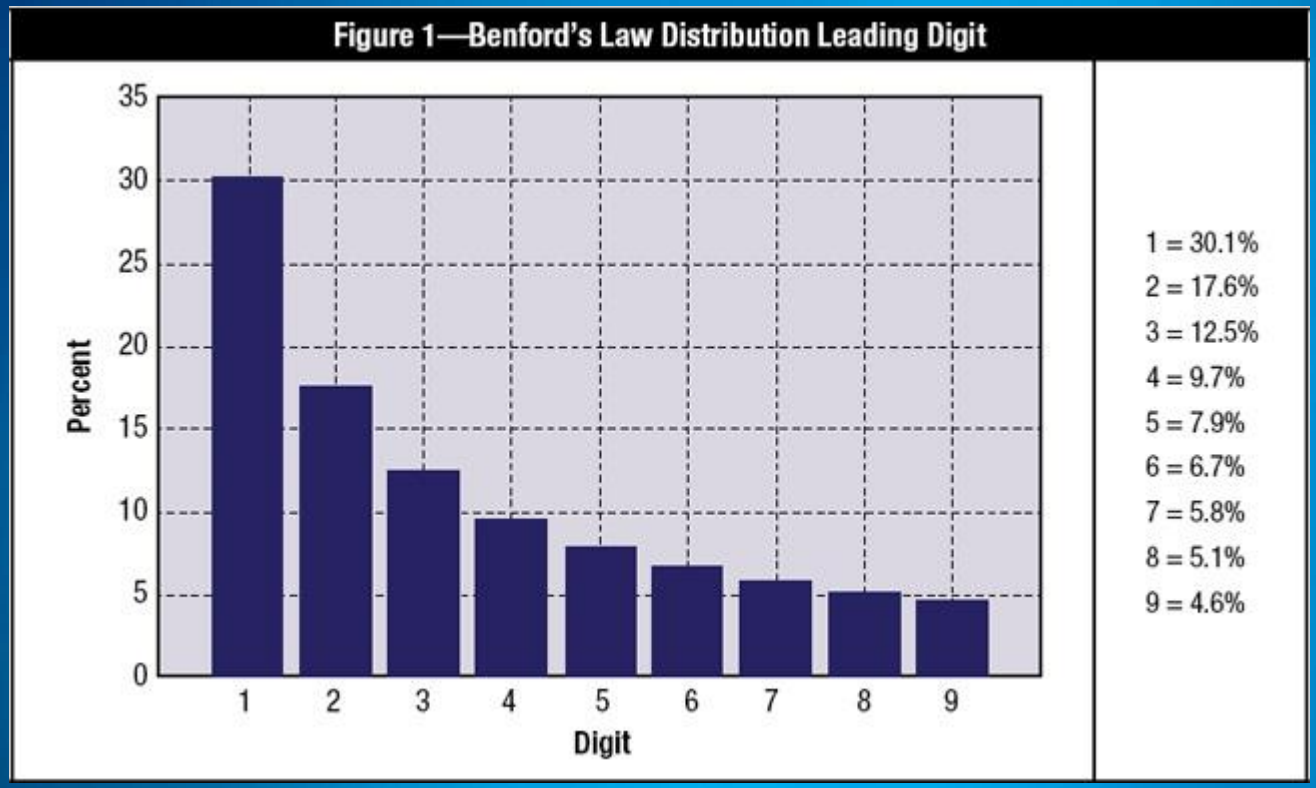
Question

How can we look at all collectors simultaneously and use the collected employment data to increase the probability of discovering systemic collection errors or curbstoning?

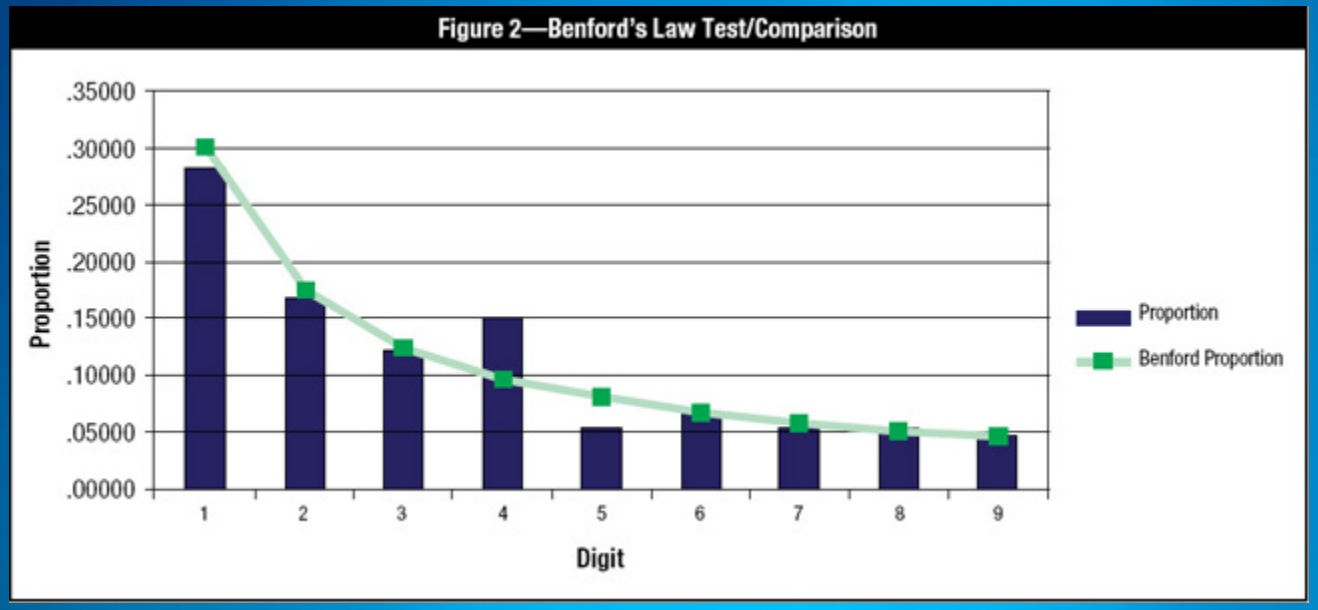
Benford's Law

Benford's Proportion = $log_{10}\left(1+\frac{1}{d}\right)$

First Digit Proportion = $log_{10}\left(1+\frac{1}{1}\right)$ = 0.301



Benford's Law



Benford's Law

- Benford's Law could be ideal in our case.
- A spike for a collector would lead supervisors to recheck the CATI calls for all collected establishments with a first digit the same as starting with the digit of the detected spike.

Question

Not all Collector Data follow Benford's first digit proportions. How do we use the concepts here within the context of our Problem Space?

Modified Benford's

- Modified Benford's uses the same concepts as Benford's Law.
- Uses the proportion of first digits for all data collectors as our Benford's first digit proportions.
- Compares the proportion of first digits of a single data collector to the Modified Benford's first digit proportions.

Goals

- Rank or prioritize data collectors by the likelihood of detecting a systemic problem.
- Be fair to the data collector. We do not want a method that suggests improper collection without conducting further investigation.

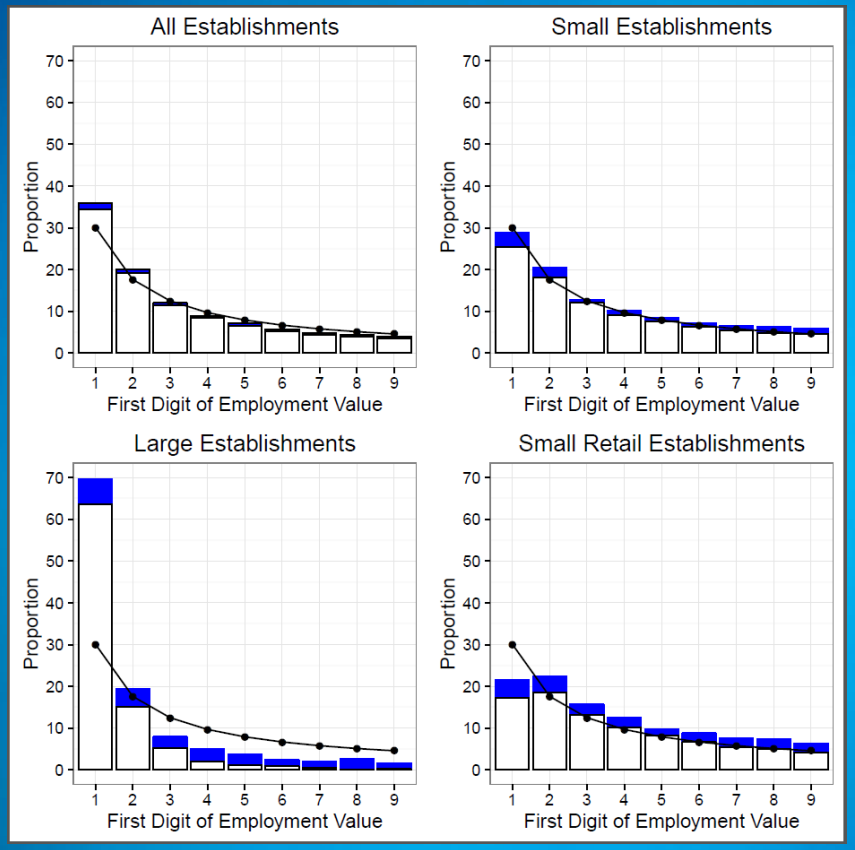
Question

When looking at the proportion of first digits for all data collectors combined, what can we do when the combined proportions follow the Benford's proportions almost exactly?

Stratifying all Collected Data

- Simplify the data used in Modified Benford's proportions to only the categories that a collector has collected in.
- Which industries did the collector report data from?
- What were the reported employment sizes of the establishments collected from? Were they Small, Medium, or Large?

First Digit Proportions of Stratified Collected Data



Stratified Sample

All Collectors	NAICS Super Sector 1	NAICS Super Sector 2	NAICS Super Sector 3
Small	n_{S1}=243	n_{S2}=6273	n_{S3}=1783
Medium	n_{M1}=364	n_{M2}=4969	n_{M3}=392
Large	n_{L1}=110	n_{L2}=2247	n_{L3}=212

Individual Collector	NAICS Super Sector 1	NAICS Super Sector 2	NAICS Super Sector 3
Small		u_{S2}=49	u_{S3}=23
Medium	u_{M1}=2	u_{M2}=46	
Large			u_{L3}=14

Stratified Sample

All Collectors	NAICS Super Sector 1	NAICS Super Sector 2	NAICS Super Sector 3
Small		n_{S2}=6273	n_{S3}=1783
Medium	n_{M1}=364	n_{M2}=4969	
Large			n_{L3}=212

Individual Collector	NAICS Super Sector 1	NAICS Super Sector 2	NAICS Super Sector 3
Small		u_{S2}=49	u_{S3}=23
Medium	u_{M1}=2	u_{M2}=46	
Large			u_{L3}=14

Question

What if I drew a stratified sample from all collectors similar to the Individual Collector I am comparing?

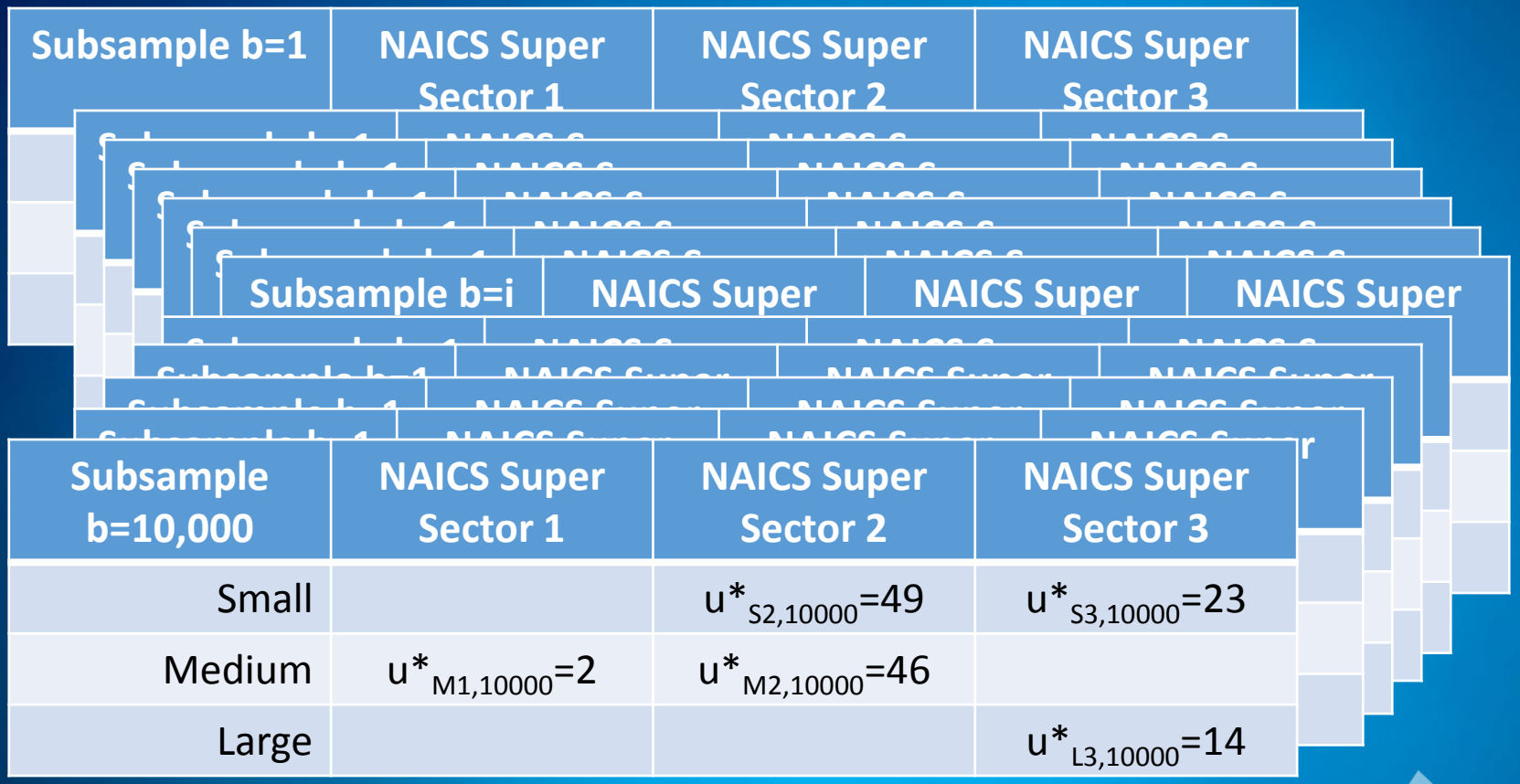
Profile Sample

- Sampled Strata are in the same stratum the collector obtained employment data for.
- Sampled size within the stratum are identical to that of the collector.
- Representative of the workload an individual interviewer has been assigned

Distribution of Profile Samples

- To compare the Individual Collector to the Profile Sample.
- Bootstrap a large number of subsamples.
- Collect the resulting first digit proportions of the subsamples.

Bootstrap Subsamples



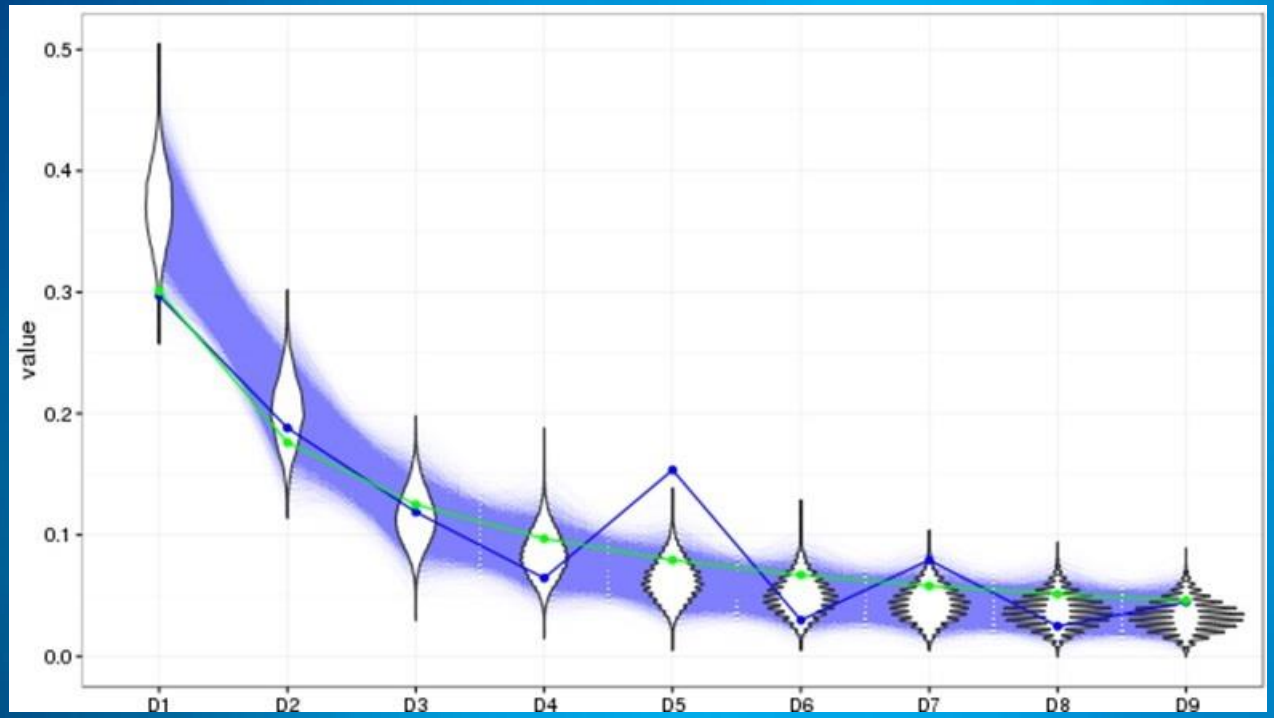
Subsample b=10,000	NAICS Super Sector 1	NAICS Super Sector 2	NAICS Super Sector 3
Small		$u^*_{S2,10000}=49$	$u^*_{S3,10000}=23$
Medium	$u^*_{M1,10000}=2$	$u^*_{M2,10000}=46$	
Large			$u^*_{L3,10000}=14$

First Digit Proportions from Bootstrap

B	{prop(d1), prop(d2), …, prop(d8), prop(d9)}
1	{0.32, 0.16, 0.18, 0.05, 0.07, 0.05, 0.05, 0.06, 0.04 }
2	{0.33, 0.24, 0.12, 0.11, 0.04, 0.06, 0.03, 0.03, 0.04 }
3	{0.36, 0.23, 0.08, 0.10, 0.06, 0.06, 0.05, 0.06, 0.02 }
4	{0.43, 0.15, 0.10, 0.09, 0.05, 0.06, 0.07, 0.02, 0.03 }
…	{…,…,…,…,…,…,…,…,…}
9997	{0.28, 0.21, 0.14, 0.07, 0.08, 0.07, 0.06, 0.05, 0.03}
9998	{0.27, 0.21, 0.10, 0.11, 0.09, 0.07, 0.06, 0.05, 0.03}
9999	{0.26, 0.21, 0.12, 0.09, 0.10, 0.07, 0.05, 0.05, 0.05}
10000	{0.31, 0.20, 0.11, 0.09, 0.07, 0.08, 0.06, 0.04, 0.04}

Violin Plot Display of First Digit Proportions

The Violin plot is a doubled kernel density plot that is used to show multiple densities at once.



Method of Comparison

- To determine how similar the collector's distribution is to the Bootstrap of profile samples, I utilized the distance from the median of the Bootstrapped Samples for each first digit value.
- This is analogous to the Bootstrapped p-value, determined by the percentages of Bootstrapped proportions higher or lower than that of the Collector being evaluated.

Hypothesis Testing

- H_0: the collector's first digit proportions come from the distribution estimated by the profile sample Bootstrap method.
- H_1: the collector's values are not from the estimated distribution.
- Test Statistic: the collector's first digit proportions.
- The test measures the percentages of Bootstrapped samples that have higher and lower first digit proportions than the collector's.

P-values

If the interviewer has a proportion of first digit ones that is in the most extreme 5% of the Bootstrap values, the p-value would be less than or equal to 0.05 for that digit. Since we have 10,000 profile samples, the interviewer's proportion of ones would have to be higher than the top 250 or lower than the bottom 250 profile sample's proportions of ones to fail.

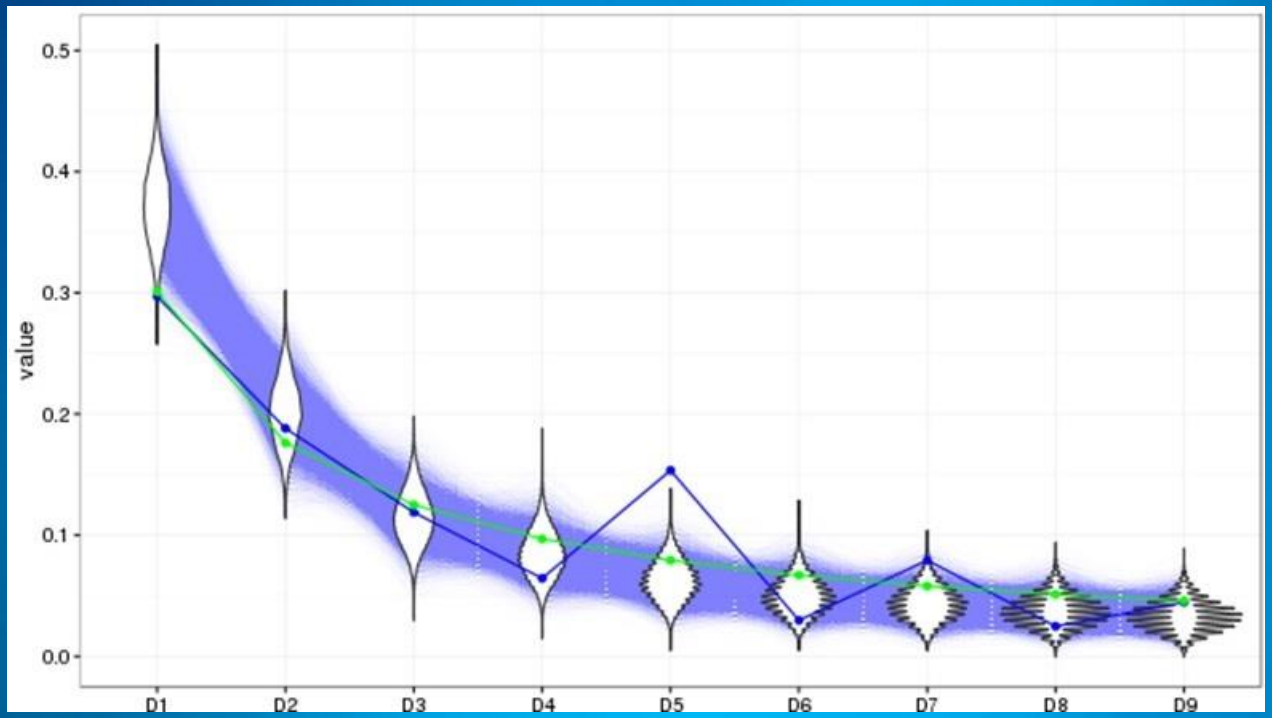
Multiple Testing

- **The minimum of the 9 p-values, detect spikes in proportions.**
- $Pmin = min(for\ all\ p - values_i), where\ i\ is\ the\ i^{th}\ leading\ digit$
- **An interviewer fails the Pmin test with a value smaller than 0.0056 using an alpha of 0.05 and a Bonferroni correction for testing all 9 digits at once (0.05/9 = 0.0056).**

Fairness in Ranking Collectors

- Compared Collector to profile samples with similar workloads.
- Bootstrap with more than the standard 2,000 samples.
- Used Bonferonni correction for the Pmin test instead of less conservative tests.
- Did not account for negative correlation between digits.

Application for Supervisors



Application for Supervisors

- Sort Collectors by their Pmin scores. This increases the probability of discovering systemic collection errors or curbstoning by a specific Collector.
- Review the CATI recordings or perform a call back of establishments that provided a suspect first digit value for a particular Collector.
- Review the month-to-month violin plots of a suspect collector.

Why Not Chi Squared Tests

- In Benford's Distribution analysis, there is a traditional methodology to rely on the Chi-square test to determine if the observed first digit frequencies match the expected Benford's Distribution frequencies.
- The issue with this test is that you will generally not have a significant result with small sample sizes. We had many collectors represented who do not have large enough samples to be detected by a Chi-square test.
- The Chi-Square test was not sensitive enough to be a meaningful discriminator for what we were trying to accomplish.

Sources

Benford's Law Image :
http://www.isaca.org/JOURNAL/ARCHIVES/2011/VOLUME-3/Pages/Understanding-and-Applying-Benfords-Law.aspx

Associated Paper:

https://www.bls.gov/osmr/pdf/st170050.pdf

Contact Information

Matthew Corrigan
SMS CES
202-691-6937
Corrigan.Matthew@bls.gov