Data Science in Action: Forensic Data Science

professor Evelina Gavrilova
NHH Norwegian School of Economics

Lectures in Data Science 2023-09-28

OSINT

- Open Source Intelligence (OSINT) - analysis of publicly available data to produce knowledge
- Forensic Data Science - looking at data to find evidence
- Everyone has access to publicly available data
- What matters are skills:
  - Applying statistical tests
  - Creating your own tests
  - Programming in Stata, R, Python, Matlab or others
- You get these skills in a PhD | advanced Master degree &| a lot of perseverance

Me

My website-> https://sites.google.com/site/evegavrilova/research

- PhD in Turin, Italy

My research:

- Economics of Crime (peer effects, police militarization, medical marijuana)
- Public Economics (payroll tax, cum ex, divident-witholding tax)
- Empirical applications

Over the last 10 years I developed a course on *Detecting Corporate Crime* at NHH.

OSINT

2 types of OSINT:

- Detecting patterns at the aggregate level, e.g. presence of a numbers fraud
  - academic interest
  - financial investors and short-sellers
- Detecting the guilty party - requires unit-level data, e.g. firm, individual, purchase, etc
  - journalists
  - investigators & lawyers
  - financial investors and short-sellers

What happens next?

- Data evidence is almost always circumstantial - evidence that relies on inference to connect it to a conclusion or fact, e.g. fingerprint at the scene of the crime
- Very rare to observe direct evidence linking one entity to crime (with public data!)
- If you have evidence and you have access -> Interview people
  - open questions
  - subject can reflect as much as they want
  - get more information

Circumstantial evidence

Data is:

- Noisy
- Data errors
- Get false positives - you obtain evidence that supports your suspicions, when there is no wrongdoing
- There will be false negatives - you will miss the small time frauds

Detection

Detection Strategy Checklist:

1. What is the cheating incentive in the context?
2. Define treatment and control groups
3. See what is the available data
4. Choose a method
5. Find the sufficient statistic
6. Do robustness checks
7. View the crime observations and verify that the data corresponds to cheating behavior and not to data error

1.Incentive

Incentive - the reason to commit a crime.

Becker Crime Model (simplified):

$$E(U) = Y(1-p) + p(-J) <> W$$

- Y are the criminal earnings - p is the probability of detection

- If the criminal is not detected, they get Y
- $-J$ is the punishment
  - if the criminal gets detected she will be punished
- W is the legal wage

The crime will be committed if Y are high, if p is low, if punishment J is low or if opportunity cost W is low.

Things not captured by the model - many, e.g. gray areas in law like cryptocurrency regulation

2. Treatment and Control

Define the groups:

- Treatment - the group where you expect to observe crime
- Control - the group where there is no crime, which gives you information on the behavior of data in the absence of crime (can be data or a distribution), calles also counterfactual

Optimally, apples to apples.

Apples to Pears - difference in data could be driven by crime or by difference between the groups.

3. Data

- Use your google
- Read academic articles to see how they source their data
- Replication packages
- Github data packages
- Others

Search terms: Libor historical data

4 & 5. The comparison

Sufficient Statistic describes the distance between treated and control on some characteristic

In the following example - treated is the libor data, control is a theoretical distribution

Sufficient statistic is going to be the chi-squared

Another example

Gap= Export of Salmon from Norway to China - Import of Salmon from Norway to China

6. Robustness

- Switch control group
- Use a different source for the main variables of interest
- Trim the data
- Cut the data, etc

7. View the observations

Look at the data in the browser window

"Does this make sense?"

Benford's Law and LIBOR cartel

Libor Cartel

- British Bankers Association surveys 18 banks on the question "At what rate could you borrow funds, were you to do so by asking for and then accepting inter-bank offers in a reasonable market size just prior to 11 am?"
- Throws out highest 4 and lowest 4, averages the middle 10 - LIBOR
- What are the opportunities for crime here?

Libor Cartel

`Predisposition:`

- Profit incentive
- Low punishments
- Peers
- Coalition

Libor Cartel

- Allegations of coordinated cartel behavior since 1991
- Cartel participants were Royal Bank of Scotland, HSBC, Deutsche Bank, JP Morgan Bank, Citibank, Bank of America, Barclays and more. There were several cartels.
- Cartel participants would also coordinate on their bids, sometimes in chats

Financial Crisis

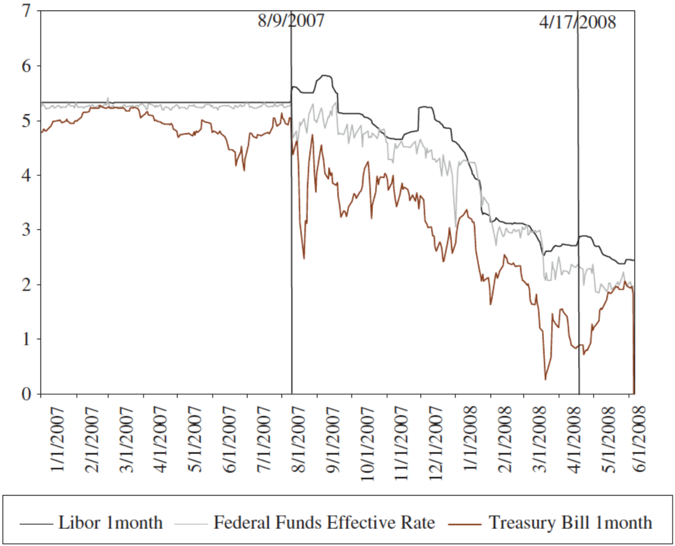


Fig. 4. Libor 1 month, Federal Funds Effective Rate and Treasury Bill 1 month

Benford's Law

- In many naturally occurring sets of data the leading digits and the last digits follow a known frequency distribution
- E.g. electricity bills, stock price, house prices, death rates, population numbers, physical and mathematical constants
- Not always fitting, important to check with comparable dataset

The best way to observe that is here:

```r
a <-round(100*rnorm(20),0)
a
```

```
##  [1]  -34   67  -29   63   31   55  132  -57  -99  -95
## [16]  -37  -47   12   -2   36
```

```r
b <-round(a %% 5,0)
b
```

```
##  [1] 1 2 1 3 1 0 2 3 1 0 3 0 3 1 0 3 3 2 3 1
```

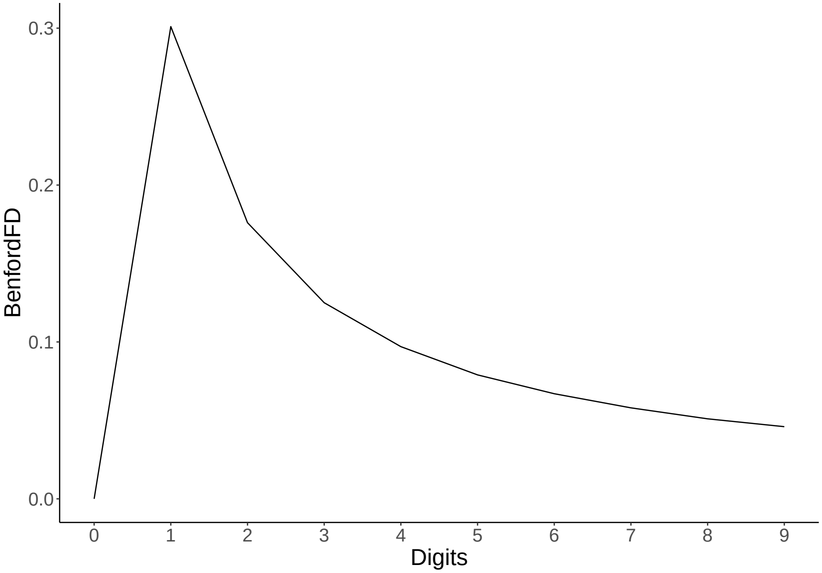
The first series ends in all types of numbers

The second series ends only in 0,1,2,3,4

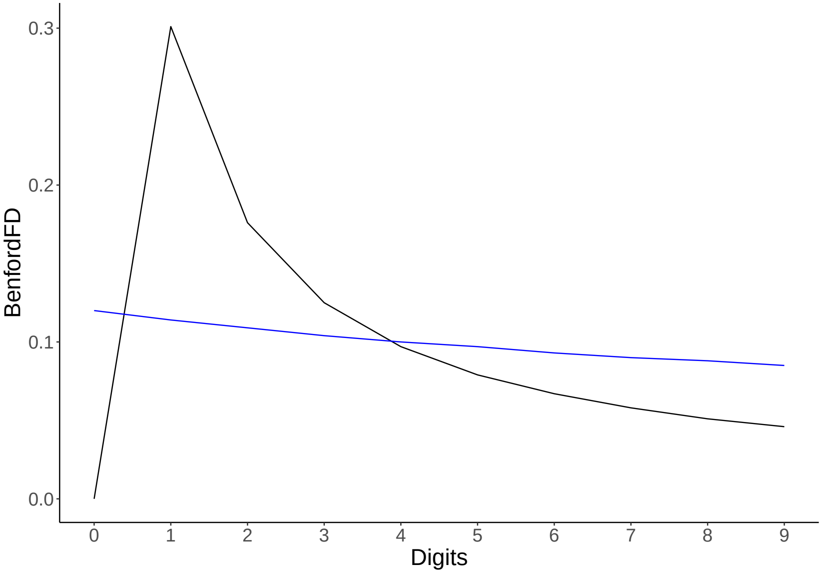
The Law

```
Digits<-seq(0,9,1)
BenfordFD<-c(30.1,17.6,12.5,9.7,7.9,6.7,5.8,5.1,4.6)
BenfordSD<-c(12.0,11.4,10.9,10.4,10.0,9.7,9.3,9.0,8.8,8.5)
Benford3D<-c(10.2,10.1,10.1,10.1,10.0,10.0,9.9,9.9,9.9,9.8
BenfordnD<-rep(10,10)
```

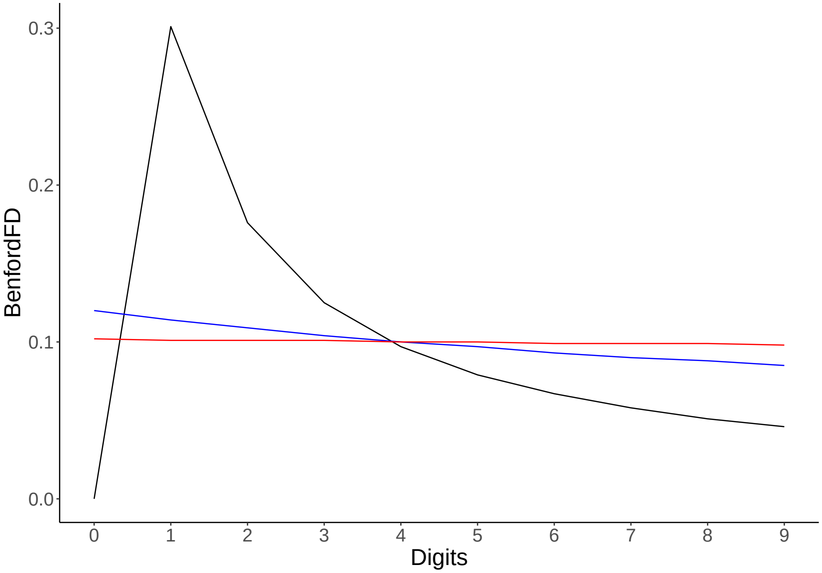
The First Digit



The Second Digit



The Third Digit



Chi-squared Test for Categorical Data

$$\chi^2 = \sum_i \frac{(e_i - p_i)^2}{p_i}$$

- e_i are the observed frequencies
- p_i are the theoretical frequencies - Benford's Law
- High value of χ^2 is going to be **sufficient** to convince us that there is *something*
- χ^2 is the sufficient statistic

Switch to R