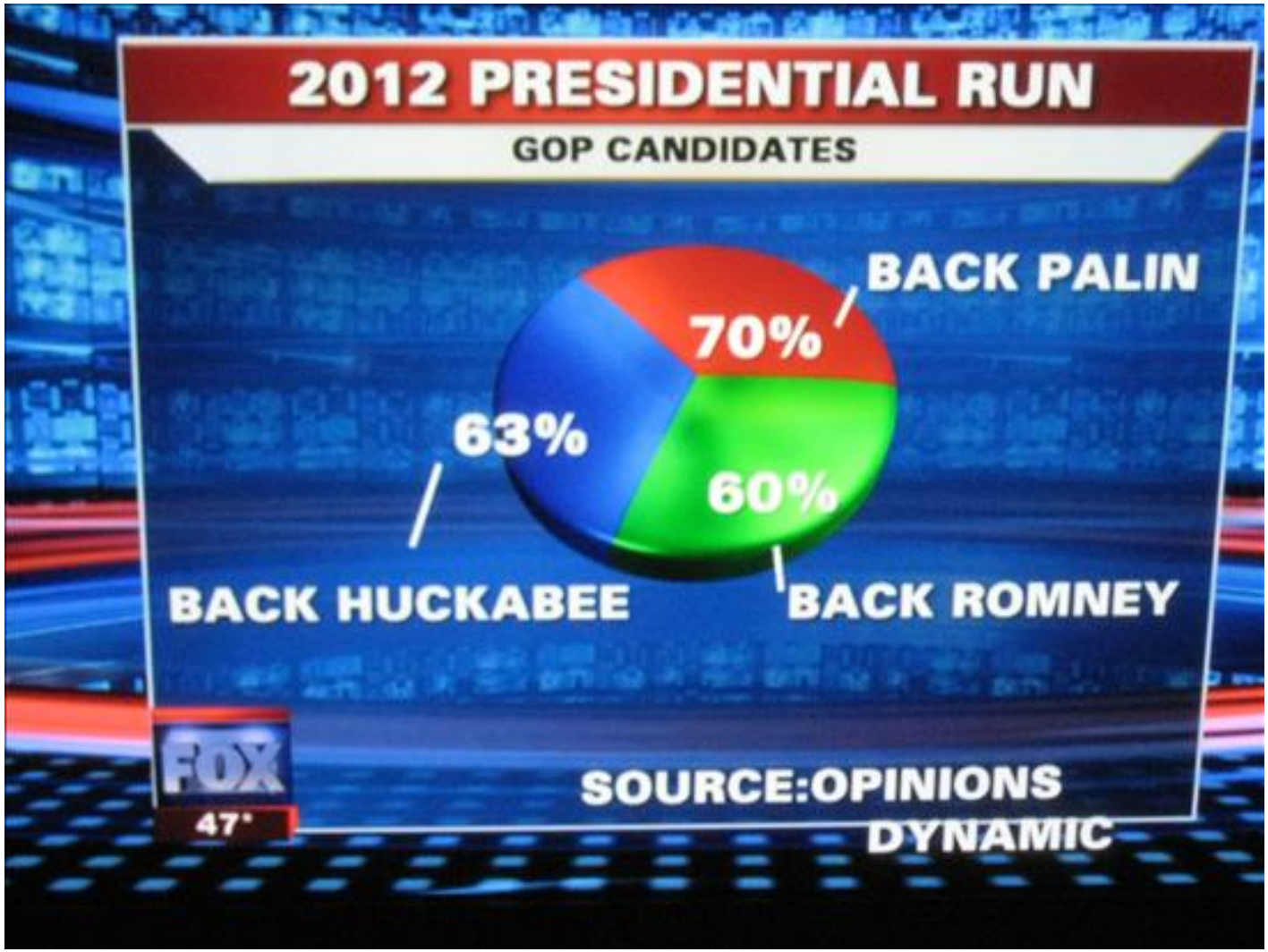
Citizen Data-Driven Journalism
The WikiLeaks Afghanistan War Logs

Drew Conway, Mike Dewar & John Myles White

March 9, 2011

Why is it important to check the data?

Why is it important to check the data?



Source: http://flowingdata.com/

Greater difficulty in the age of "big data"

With crowd-sourced and dubiously disclosed data we have several issues

- Extremely difficult to vet—especially for non-journalists
- Data in "raw" form, high degree of variance
- Motivation of sources unknown

Greater difficulty in the age of "big data"

With crowd-sourced and dubiously disclosed data we have several issues

- Extremely difficult to vet—especially for non-journalists
- Data in "raw" form, high degree of variance
- Motivation of sources unknown

CJR – 'The Challenge of Verifying Crowdsourced Information'

"Even though the information from Twitter is not particularly reliable—and things are being retweeted so its kind of messy—the basic idea is if you crowdsource the information and put it on one map you can really see the clusters of incidents. So even though one particular tweet is not that important, if you have similar reports from the media you can see where the incidents are clustering."

- Jaroslav Valuch, the project manager for Ushahidi Haiti

Greater difficulty in the age of "big data"

With crowd-sourced and dubiously disclosed data we have several issues

- Extremely difficult to vet—especially for non-journalists
- Data in "raw" form, high degree of variance
- Motivation of sources unknown

CJR – 'The Challenge of Verifying Crowdsourced Information'

"Even though the information from Twitter is not particularly reliable—and things are being retweeted so its kind of messy—the basic idea is if you crowdsource the information and put it on one map you can really see the clusters of incidents. So even though one particular tweet is not that important, if you have similar reports from the media you can see where the incidents are clustering."

- Jaroslav Valuch, the project manager for Ushahidi Haiti

But...visual evidence can be deceiving

- Need a real test (statistical) to provide better evidence that data is reliable and not fraudulent

Source: 'The Challenge of Verifying Crowdsourced Information,' see also: 'How WikiLeaks Outsourced the Burden of Verification'

How to verify the data

First, are we trying to verify the information, or data itself?

- Former requires independent information and resources
- Latter we can do statistically for free! ☺

How to verify the data

First, are we trying to verify the information, or data itself?

- Former requires independent information and resources
- Latter we can do statistically for free! ☺

If simply verifying data as it exists, need a **data appropriate test**

- What do our data look like?
- What is the generating process?

How to verify the data

First, are we trying to verify the information, or data itself?

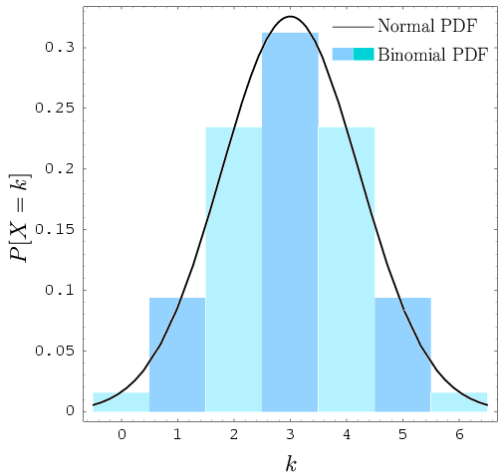
- Former requires independent information and resources
- Latter we can do statistically for free! ☺

If simply verifying data as it exists, need a **data appropriate test**

- What do our data look like?
- What is the generating process?

Journalists often deal with **discrete count data** wherein additional information has been coded into each event

- Election: # people voting for some candidate per-district
- Finance: # securities traded per-day
- Ushahidi: # incidents reported of some type
- WikiLeaks: # SIGACTS per-region or geo-code



How to verify the data

First, are we trying to verify the information, or data itself?

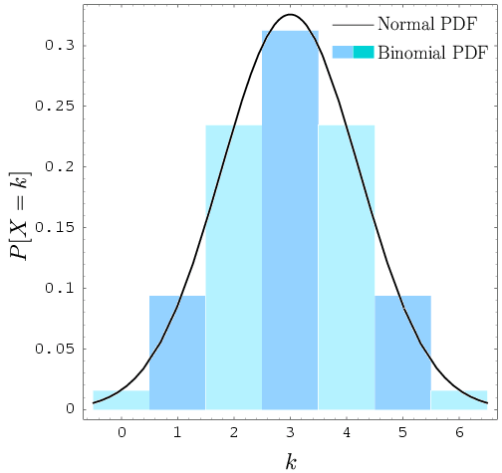
- Former requires independent information and resources
- Latter we can do statistically for free! ☺

If simply verifying data as it exists, need a **data appropriate test**

- What do our data look like?
- What is the generating process?

Journalists often deal with **discrete count data** wherein additional information has been coded into each event

- Election: # people voting for some candidate per-district
- Finance: # securities traded per-day
- Ushahidi: # incidents reported of some type
- WikiLeaks: # SIGACTS per-region or geo-code



We use **Benford's Law** to test that WikiLeaks data fits a *natural data generating process* for count data

Source: `http://en.wikipedia.org/wiki/Binomial_distribution`

What is Benford's Law

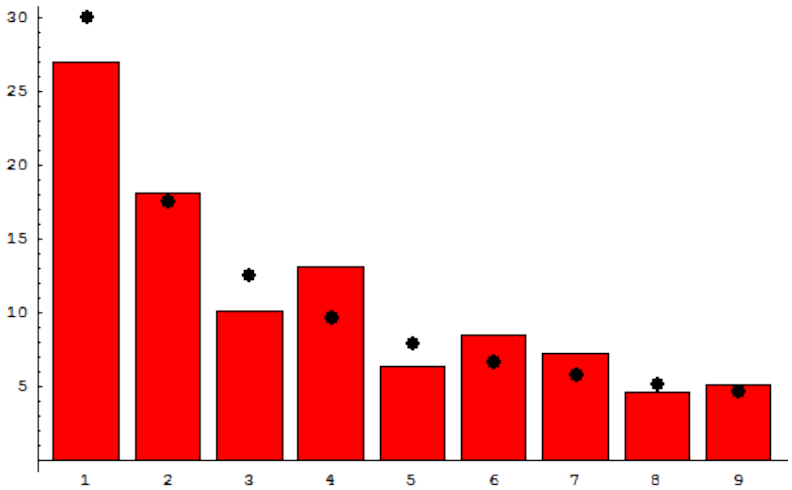
Benford's Law

Benford's law, also called the first-digit law, states that in lists of numbers from many (but not all) real-life sources of data, the leading digit is distributed in a specific, non-uniform way. According to this law, the first digit is 1 about 30% of the time, and larger digits occur as the leading digit with lower and lower frequency, to the point where 9 as a first digit occurs less than 5% of the time.

What is Benford's Law

Benford's Law

Benford's law, also called the first-digit law, states that in lists of numbers from many (but not all) real-life sources of data, the leading digit is distributed in a specific, non-uniform way. According to this law, the first digit is 1 about 30% of the time, and larger digits occur as the leading digit with lower and lower frequency, to the point where 9 as a first digit occurs less than 5% of the time.

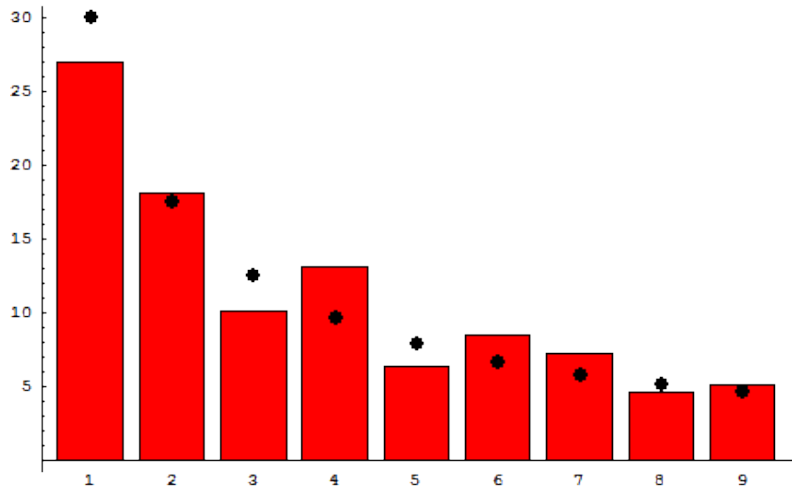


- Distribution of first digits in the population of the 237 countries of the world (at left)

What is Benford's Law

Benford's Law

Benford's law, also called the first-digit law, states that in lists of numbers from many (but not all) real-life sources of data, the leading digit is distributed in a specific, non-uniform way. According to this law, the first digit is 1 about 30% of the time, and larger digits occur as the leading digit with lower and lower frequency, to the point where 9 as a first digit occurs less than 5% of the time.

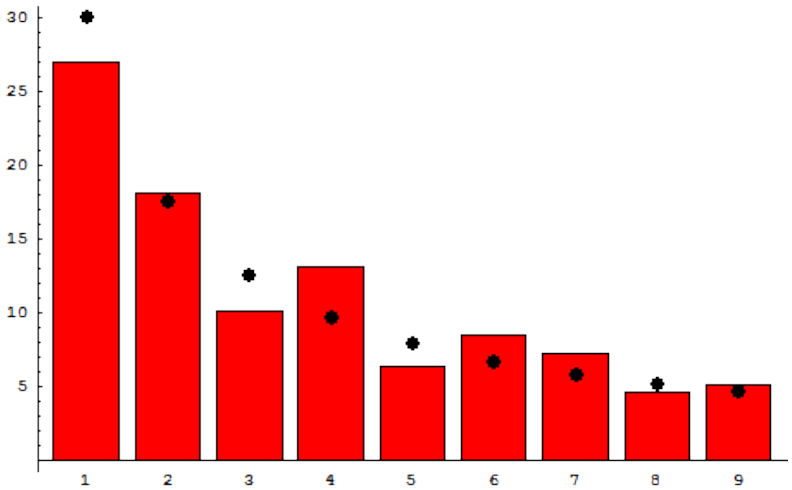


- Distribution of first digits in the population of the 237 countries of the world (at left)
- Detect check fraud (*Journal of Accountancy*)
- Election rigging in Iran (FiveThirtyEight)

What is Benford's Law

Benford's Law

Benford's law, also called the first-digit law, states that in lists of numbers from many (but not all) real-life sources of data, the leading digit is distributed in a specific, non-uniform way. According to this law, the first digit is 1 about 30% of the time, and larger digits occur as the leading digit with lower and lower frequency, to the point where 9 as a first digit occurs less than 5% of the time.



- Distribution of first digits in the population of the 237 countries of the world (at left)
- Detect check fraud (*Journal of Accountancy*)
- Election rigging in Iran (FiveThirtyEight)

For the WikiLeaks, we will analyze the leading digits for **weekly counts of SIGACT reports** in the data

- This includes visual and statistical tests with Benford's Law

Source: `http://en.wikipedia.org/wiki/Benford's_law`

Running the test

```
31  # 2. Benford's test on number of total reports in data per week
32  week.count<-cbind(table(cbind(format.Date(afg$DateOccurred,"%Y %W"))))
33
34  # Function for pulling out leading digit from some integer stored as string
35  leading.dig<-function(x) {
36      as.numeric(strsplit(as.character(x),"")[[1]][1])
37  }
38
39  # Count digits and store as data frame
40  dig.count<-cbind(table(sapply(as.vector(week.count),leading.dig)))
41  dig.count<-as.data.frame(dig.count)
42  colnames(dig.count)<-"DigitCount"
43
44  # Benford's distribution
45  dbenford<-function(d,base=10) {
46      return(log(1+(1/d),base=base))
47  }
```

Running the test

```
31  # 2. Benford's test on number of total reports in data per week
32  week.count<-cbind(table(cbind(format.Date(afg$DateOccurred,"%Y %W"))))
33
34  # Function for pulling out leading digit from some integer stored as string
35  leading.dig<-function(x) {
36      as.numeric(strsplit(as.character(x),"")[[1]][1])
37  }
38
39  # Count digits and store as data frame
40  dig.count<-cbind(table(sapply(as.vector(week.count),leading.dig)))
41  dig.count<-as.data.frame(dig.count)
42  colnames(dig.count)<-"DigitCount"
43
44  # Benford's distribution
45  dbenford<-function(d,base=10) {
46      return(log(1+(1/d),base=base))
47  }
```

```
> head(week.count)
        [,1]
2004 00   12
2004 01   27
2004 02   43
2004 03   31
2004 04   27
2004 05   16
```

```
> tail(week.count)
        [,1]
2009 47  591
2009 48  686
2009 49  525
2009 50  552
2009 51  575
2009 52  305
```

Running the test

```
31  # 2. Benford's test on number of total reports in data per week
32  week.count<-cbind(table(cbind(format.Date(afg$DateOccurred,"%Y %W"))))
33
34  # Function for pulling out leading digit from some integer stored as string
35  leading.dig<-function(x) {
36      as.numeric(strsplit(as.character(x),"")[[1]][1])
37  }
38
39  # Count digits and store as data frame
40  dig.count<-cbind(table(sapply(as.vector(week.count),leading.dig)))
41  dig.count<-as.data.frame(dig.count)
42  colnames(dig.count)<-"DigitCount"
43
44  # Benford's distribution
45  dbenford<-function(d,base=10) {
46      return(log(1+(1/d),base=base))
47  }
```

```
> head(week.count)
        [,1]
2004 00   12
2004 01   27
2004 02   43
2004 03   31
2004 04   27
2004 05   16
```

```
> tail(week.count)
        [,1]
2009 47  591
2009 48  686
2009 49  525
2009 50  552
2009 51  575
2009 52  305
```

Running the test

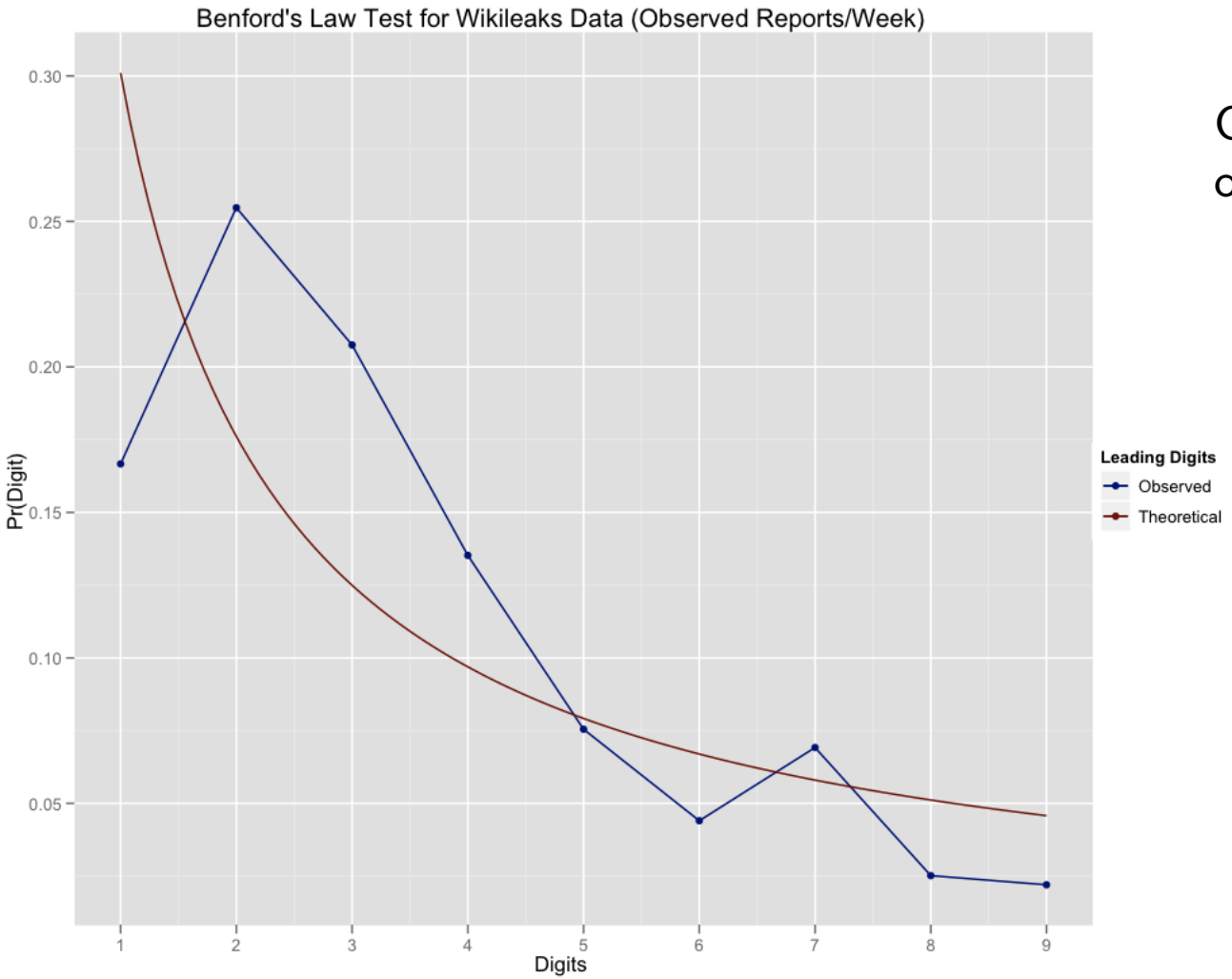
```
31  # 2. Benford's test on number of total reports in data per week
32  week.count<-cbind(table(cbind(format.Date(afg$DateOccurred,"%Y %W"))))
33
34  # Function for pulling out leading digit from some integer stored as string
35  leading.dig<-function(x) {
36      as.numeric(strsplit(as.character(x),"")[[1]][1])
37  }
38
39  # Count digits and store as data frame
40  dig.count<-cbind(table(sapply(as.vector(week.count),leading.dig)))
41  dig.count<-as.data.frame(dig.count)
42  colnames(dig.count)<-"DigitCount"
43
44  # Benford's distribution
45  dbenford<-function(d,base=10) {
46      return(log(1+(1/d),base=base))
47  }
```

```
> head(week.count)
        [,1]
2004 00   12
2004 01   27
2004 02   43
2004 03   31
2004 04   27
2004 05   16
```

```
> tail(week.count)
        [,1]
2009 47  591
2009 48  686
2009 49  525
2009 50  552
2009 51  575
2009 52  305
```

```
> dig.count
  DigitCount
1         53
2         81
3         66
4         43
5         24
6         14
7         22
8          8
9          7
```

Results 1 – all of the data



Chi-squared goodness of fit test

- $\chi^2 = 72$
- p-value=0.23

Cannot reject null that data came from Benford-like data generating process

Results 2 – by region