Validating Survey Data Using Benford's Law

Kevin Foster

Federal Reserve Bank of Boston

2015 Joint Statistical Meetings
Seattle, WA
August 9, 2015

FEDERAL RESERVE BANK OF BOSTON™

Introduction

Certain kinds of survey data are prone to measurement error. In the worst case, these errors indicate fraud. In more benign cases, it might be recall error or recording error.

1. What types of data are susceptible to these errors?
2. One method of detecting these errors: Benford's Law
3. Application: 2012 Diary of Consumer Payment Choice

FEDERAL RESERVE BANK OF BOSTON™

Categorical or Continuous data?

The range of values in categorical data are bound by the constraints of the survey design.

- Yes/No questions are typically coded as (1,2) or (1, 0).

However, continuous data can be misrepresented in many ways.

- Fraud
- Curbstoning
- Rounding
- Recording error

FEDERAL RESERVE BANK OF BOSTON™

Potential errors

- *Fraud*: Tax records, election results
- *Curbstoning*: Field representitives who make up data without ever visiting or contacting the respondent.
- *Rounding*: The Consumer Expenditure Quarterly staff believe that respondents round certain dollar values and show evidence to support their belief.
- *Recording error*: mis-typing responses
- Other errors??

FEDERAL RESERVE BANK OF BOSTON™

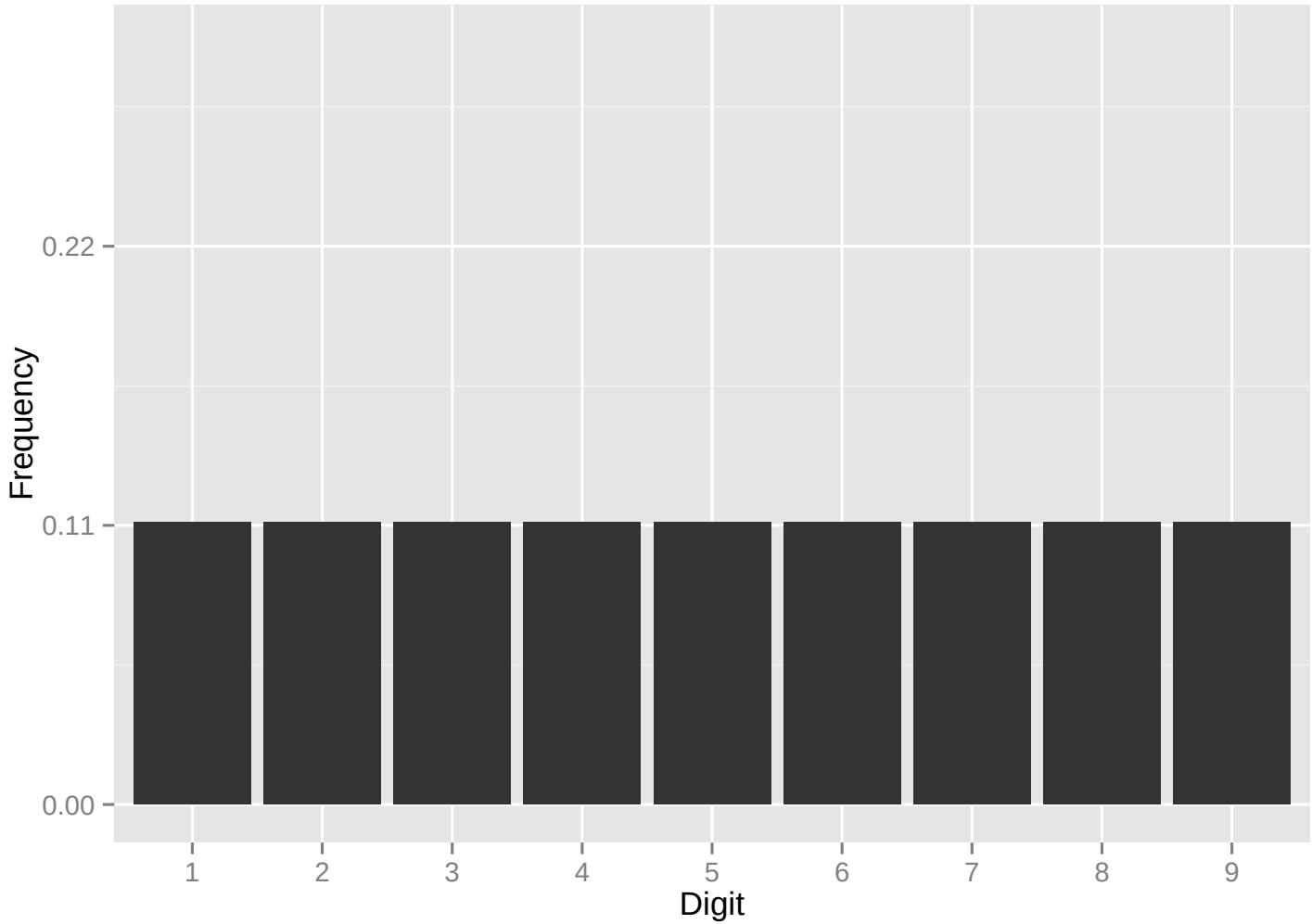
First digits

How are the first significant digits of large collections of numbers distributed?

FEDERAL RESERVE BANK OF BOSTON™

First digits

How are the first significant digits of large collections of numbers distributed?

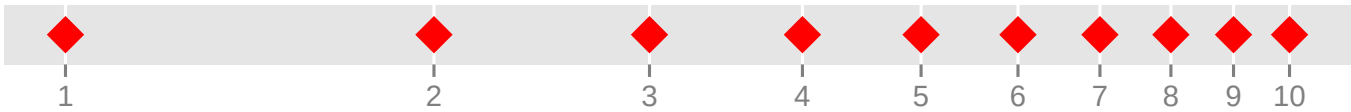


FEDERAL RESERVE BANK OF BOSTON™

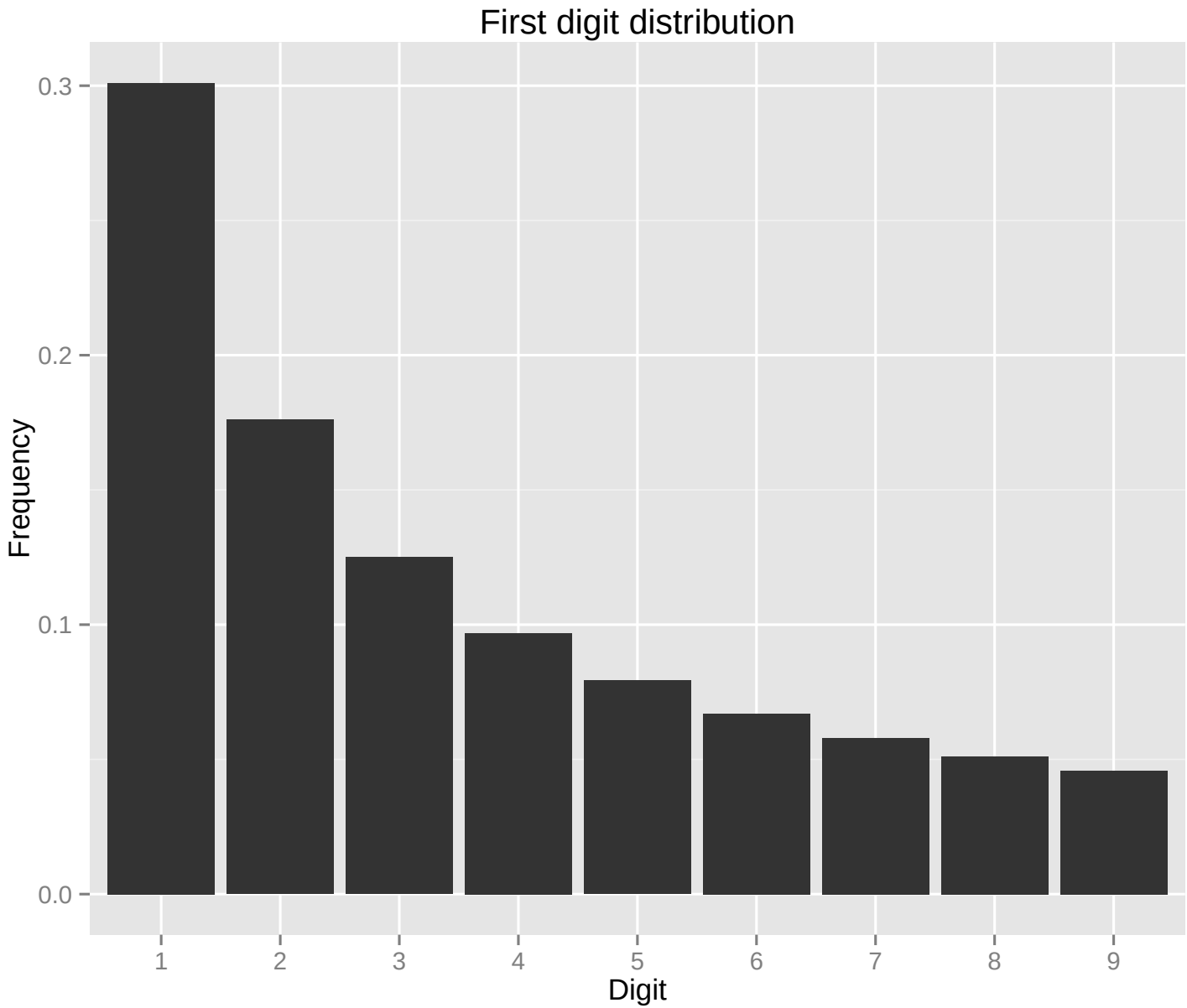
Benford's Law

Benford (1938) says that the first or leading digits of many large collections of numbers have the following distribution:

$$P(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}\left(1+\frac{1}{d}\right), d = 1, \ldots, 9$$



FEDERAL RESERVE BANK OF BOSTON™

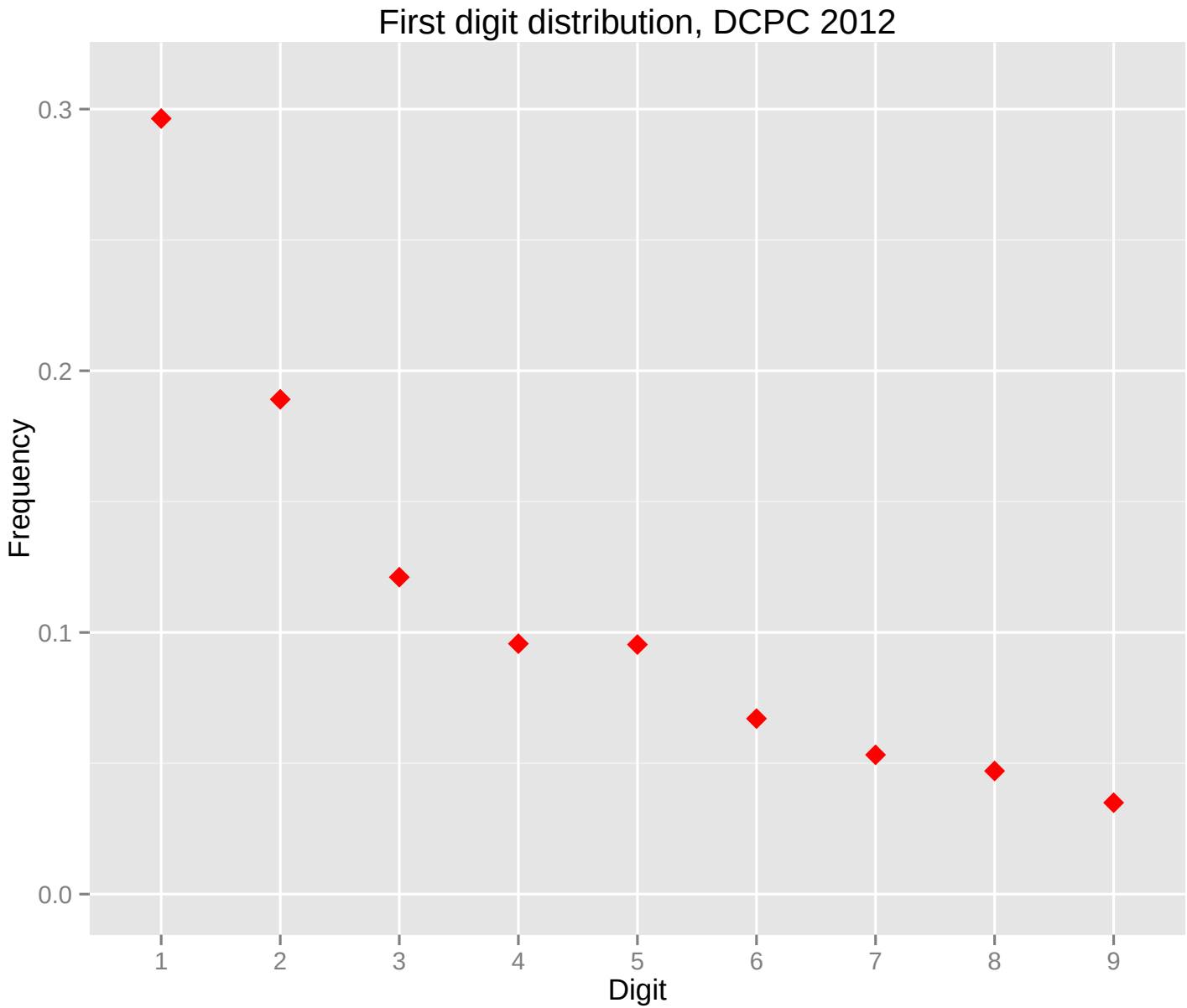


FEDERAL RESERVE BANK OF BOSTON™

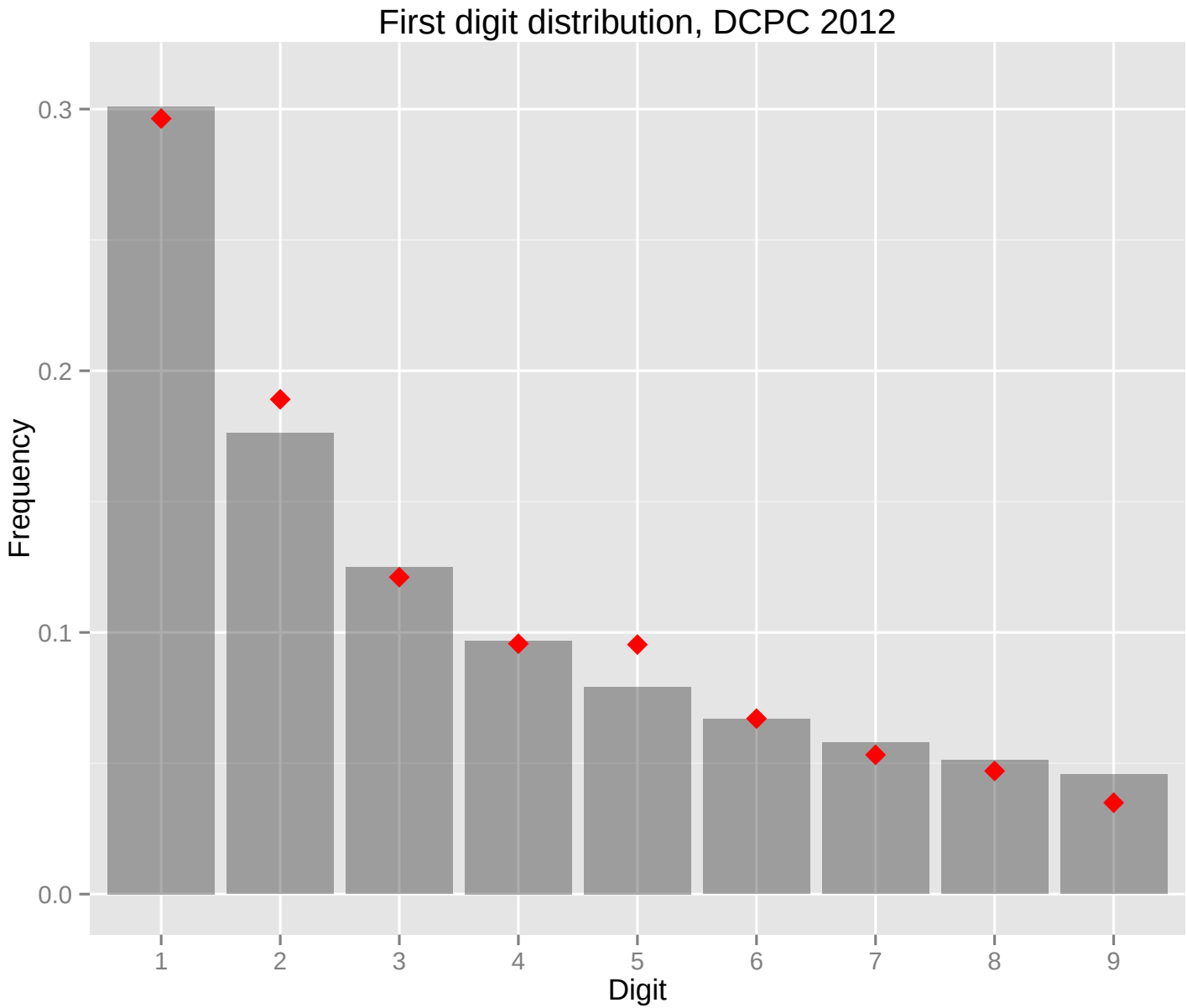
Diary of Consumer Payment Choice, 2012

- The DCPC has 2,468 diarists who each participate in the diary study for three days.
- The diarists made around 14,596 payments spanning day-to-day purchases, bills, and automatic bill payments.
  - In order to test the first three digits, we deleted 176 observations that are less than one dollar.
- We'll look at the first significant digits of the dollar amounts of all these payments.

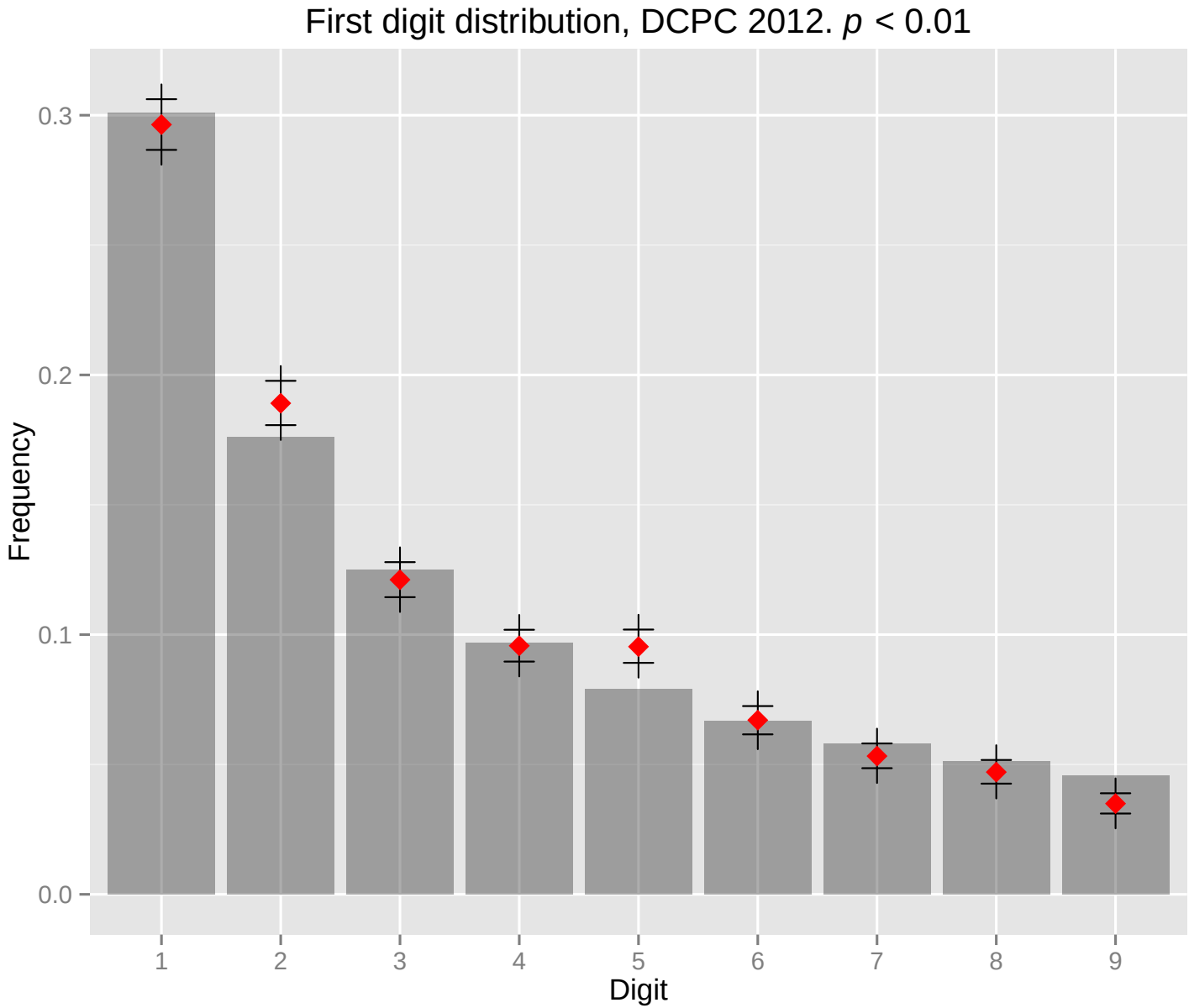
FEDERAL RESERVE BANK OF BOSTON™



FEDERAL RESERVE BANK OF BOSTON™



FEDERAL RESERVE BANK OF BOSTON™



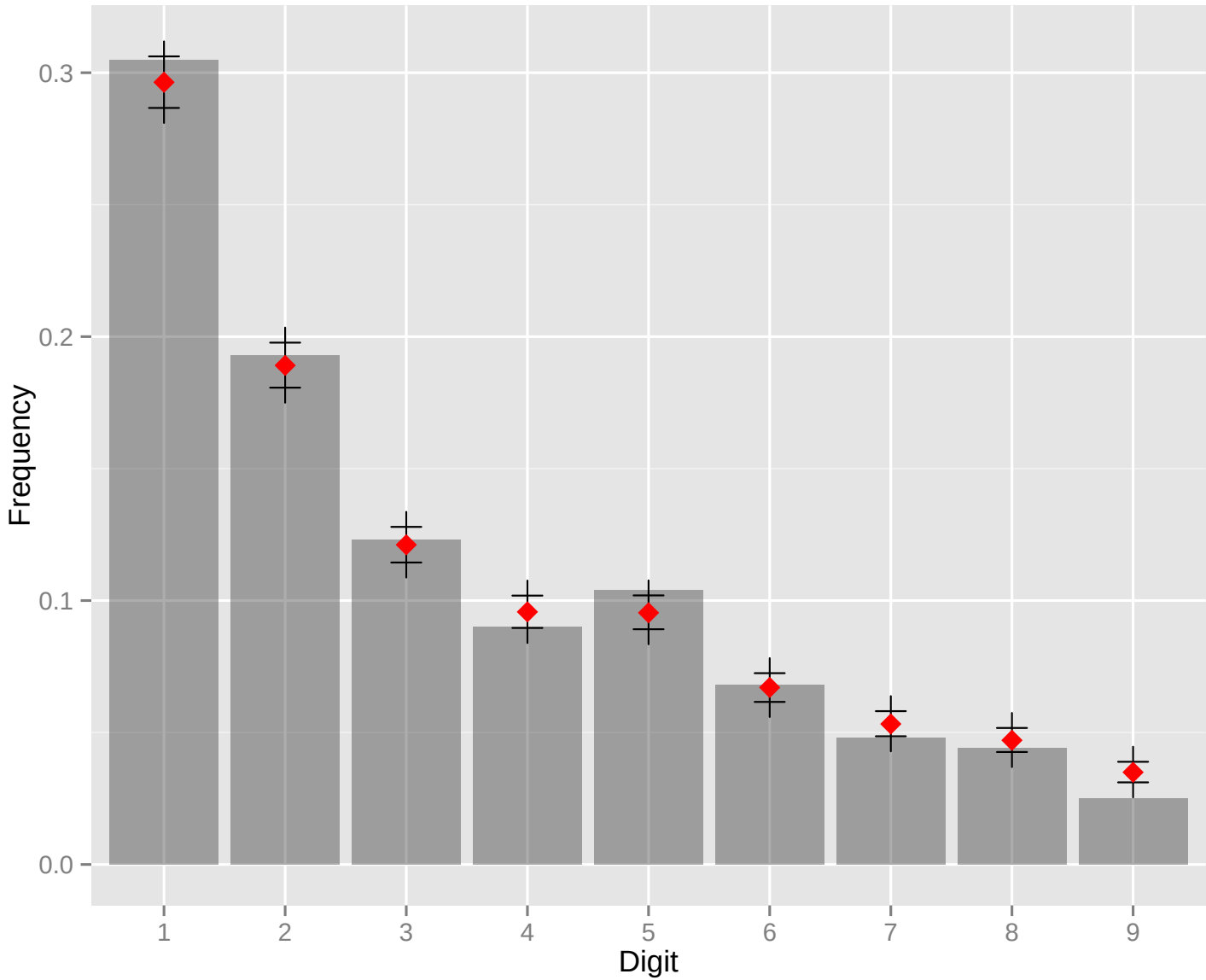
FEDERAL RESERVE BANK OF BOSTON™

Compare DCPC to BLS CE

- In Swanson et al (2003), the authors apply Benford's Law to 734,684 dollar values from the Consumer Expenditure Survey in the year 2000.
- They see the same overrepresentation of 2's and 5's, and a slight shortage of 9's.
- In the next figure, the bars represent the BLS numbers, and DCPC numbers the red dots.

FEDERAL RESERVE BANK OF BOSTON™

First digit distributions, DCPC 2012 versus CE 2000. $p < 0.01$



FEDERAL RESERVE BANK OF BOSTON™

Generalization to 2nd digits and beyond

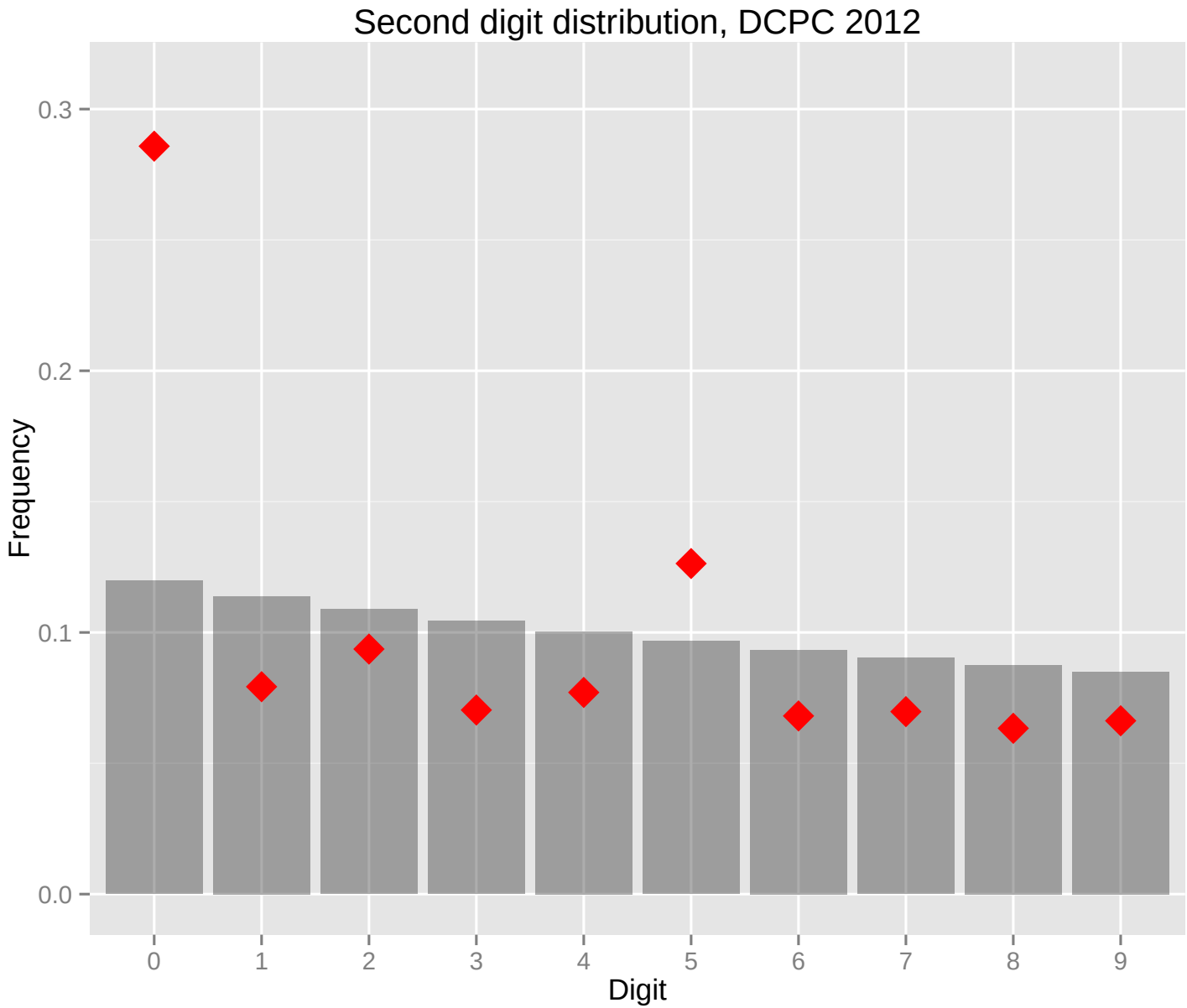
In the BLS paper, they only analyzed the first digits. We can generalize Benford's Law to the nth digit d.

$$P(n\text{th digit} = d) = \sum_{k=10^{n-2}}^{10^{n-1}-1} \log_{10}\left(1 + \frac{1}{10k+d}\right)$$

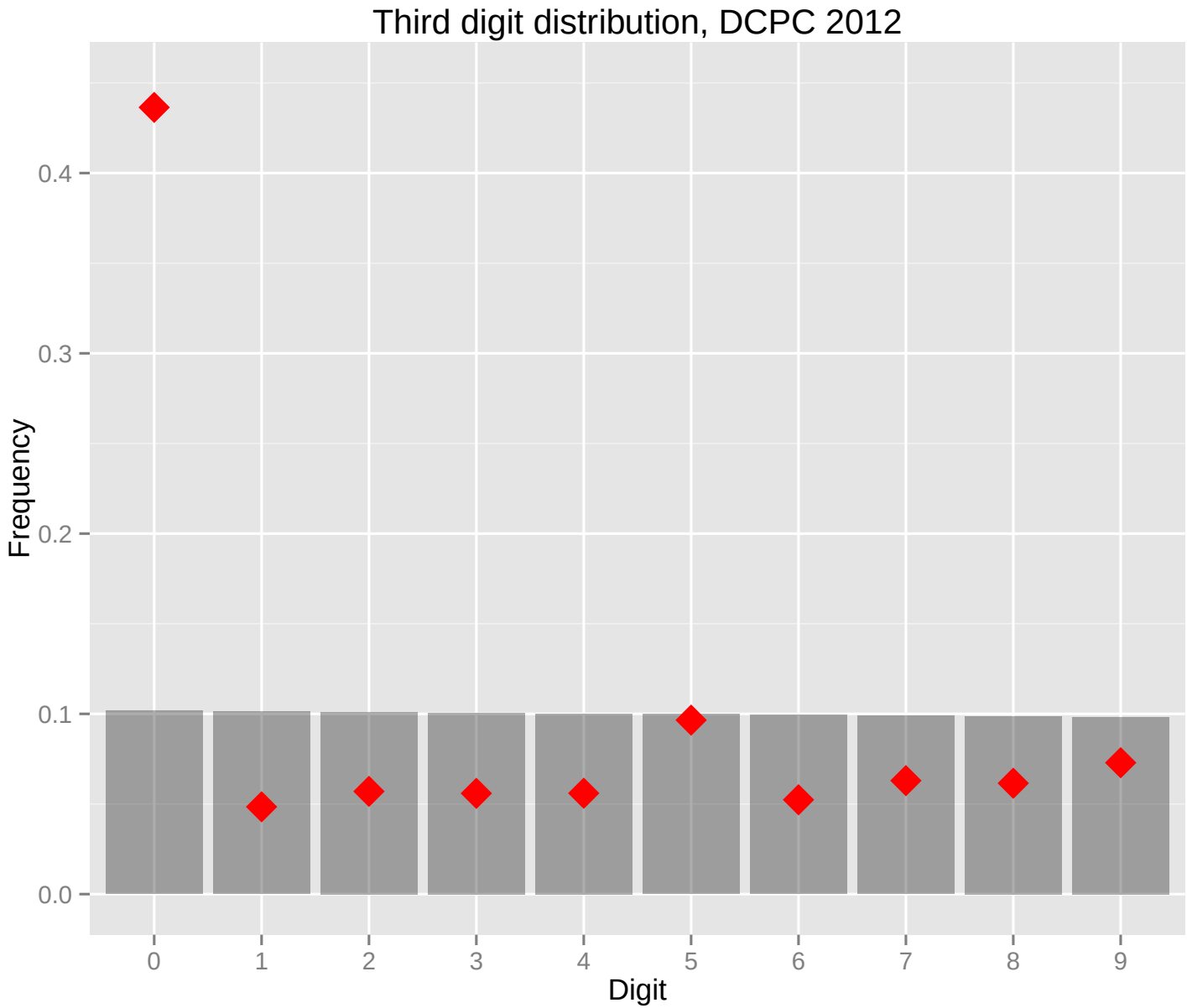
Example using $n = 2$ and $d = 2$.

$$\log_{10}\left(1 + \frac{1}{12}\right) + \cdots + \log_{10}\left(1 + \frac{1}{92}\right) \approx 0.109$$

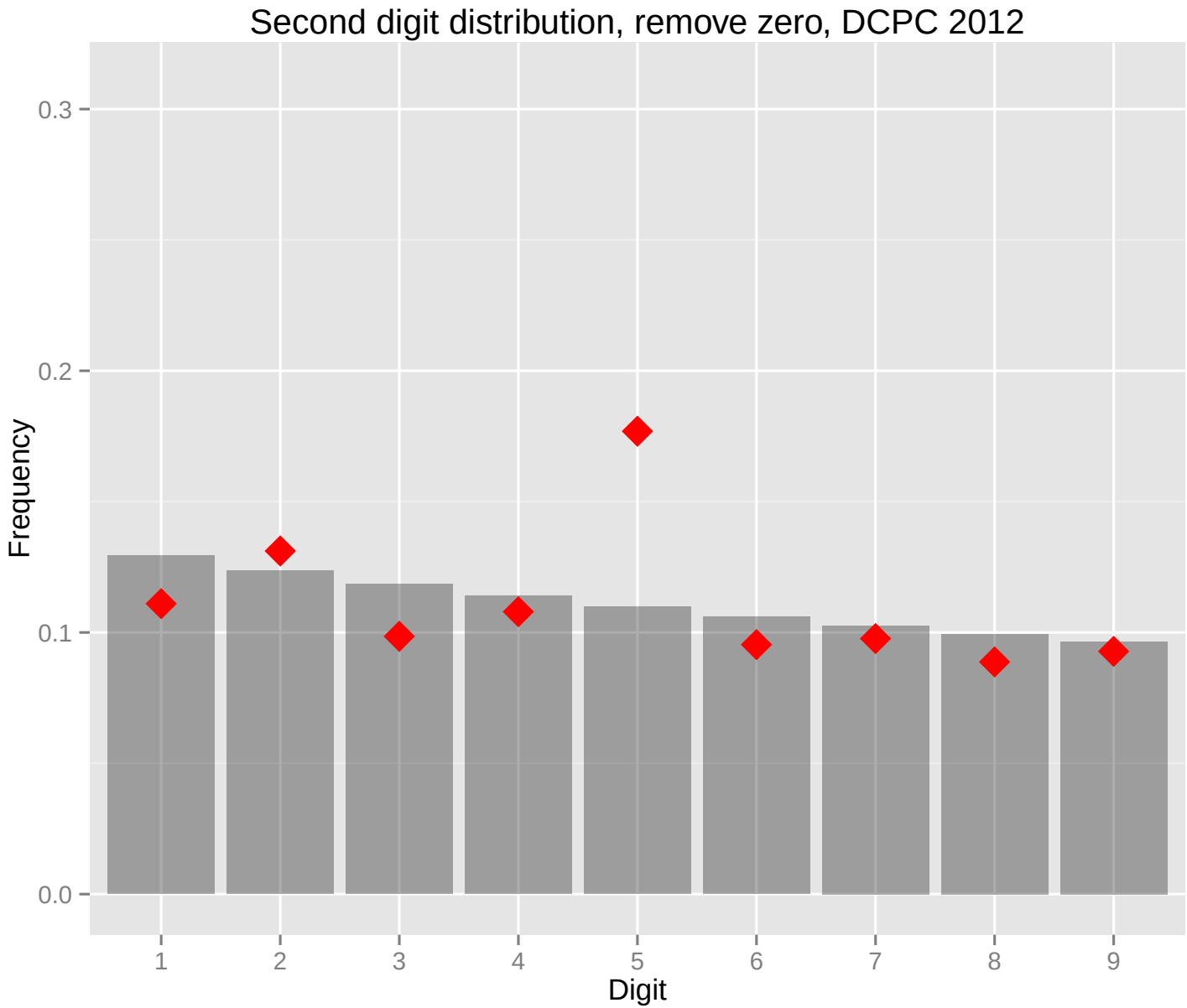
FEDERAL RESERVE BANK OF BOSTON™



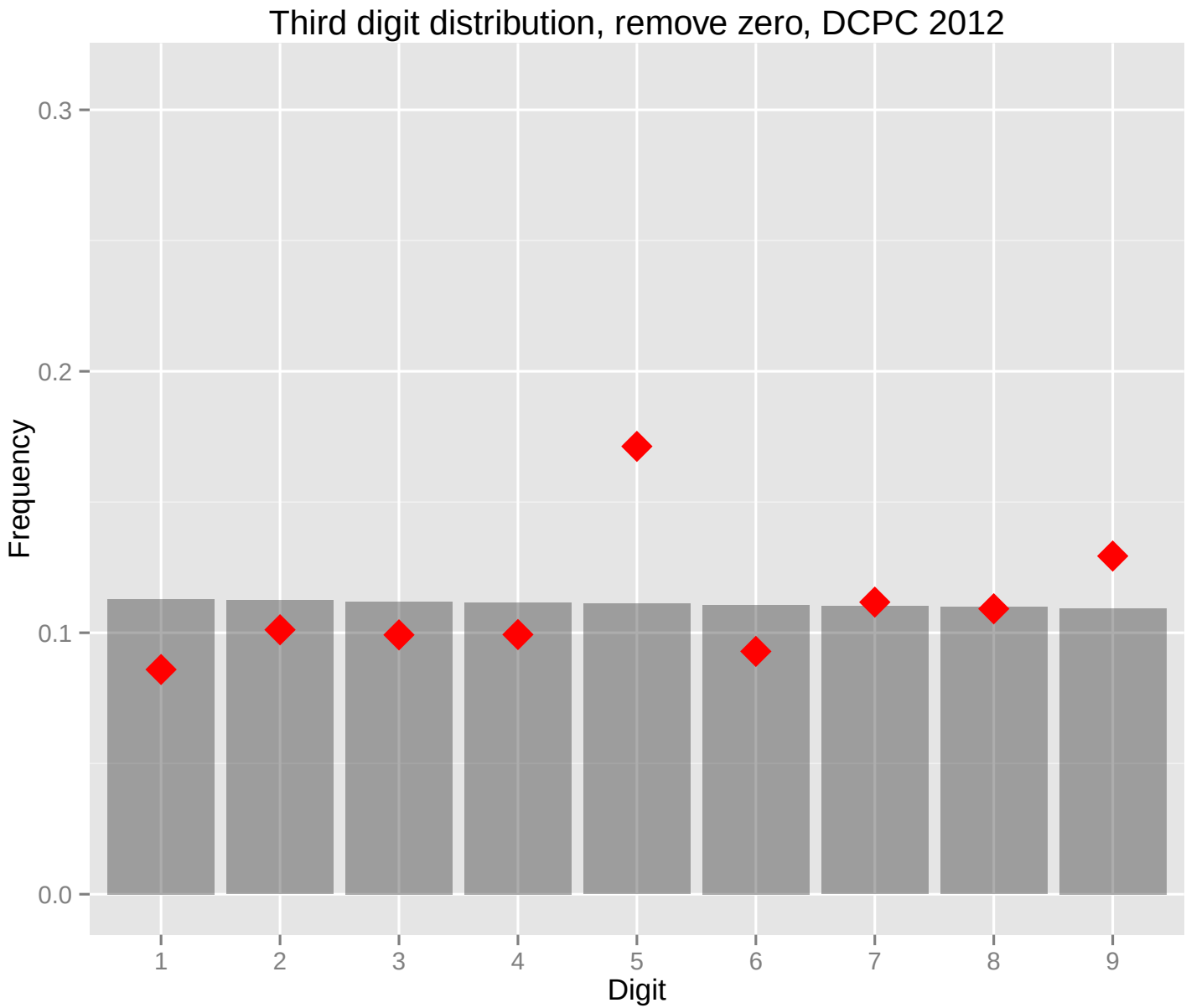
FEDERAL RESERVE BANK OF BOSTON™



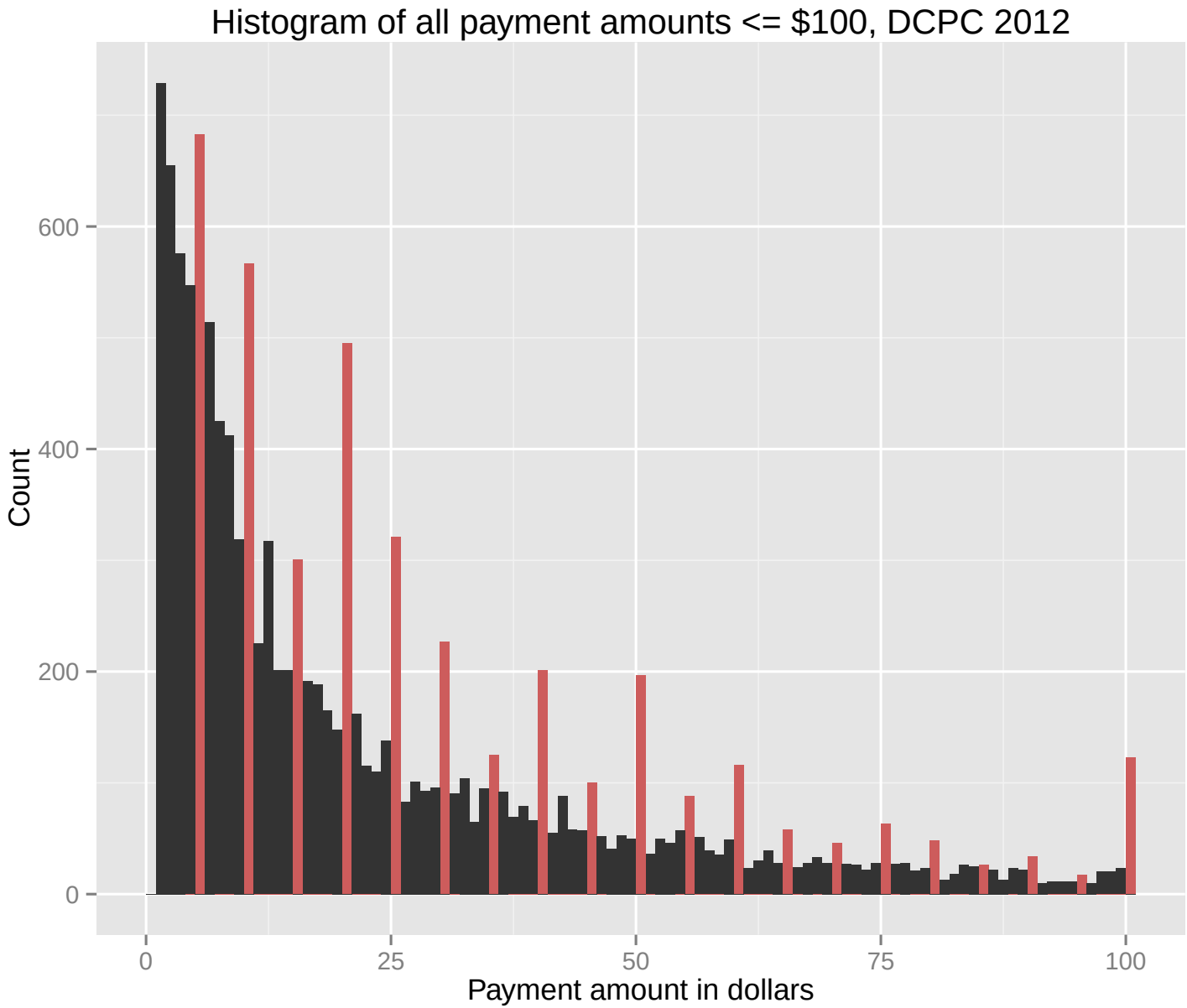
FEDERAL RESERVE BANK OF BOSTON™



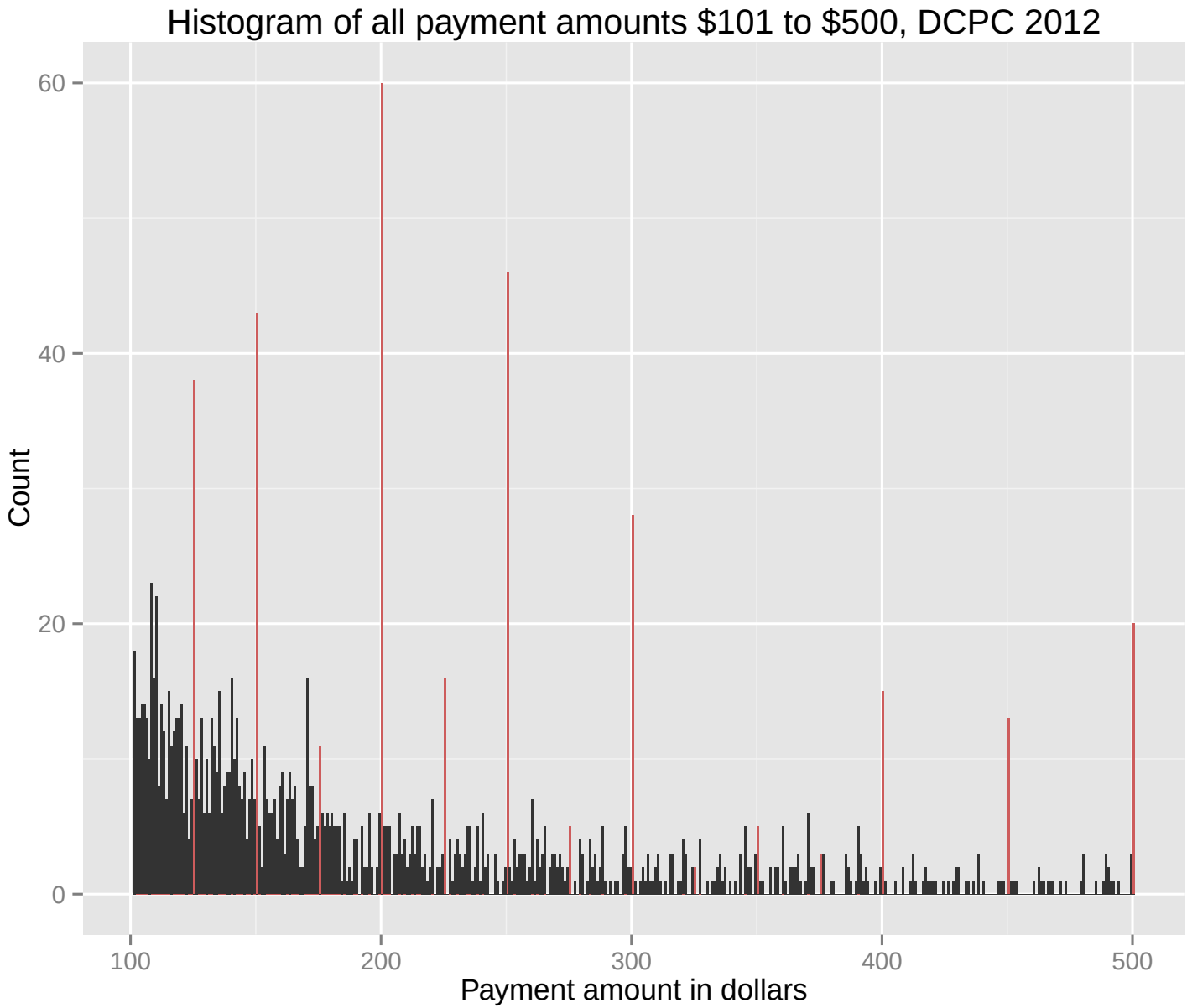
FEDERAL RESERVE BANK OF BOSTON™



FEDERAL RESERVE BANK OF BOSTON™



FEDERAL RESERVE BANK OF BOSTON™



FEDERAL RESERVE BANK OF BOSTON™

Rounded vs Non-rounded amounts

- The set is all dollar amounts ≤ 1000.
- Identify rounded dollar amounts as those ending in ".00"
- Confidence intervals are bootstrap estimates.

	N	mean	95% c.i.
Rounded	5766	66.22	(62.91, 69.45)
Non-rounded	8679	42.84	(41.05, 44.68)

	25P	50P	75P	90P	95P	99P	max
Rounded	9	23	60	160	296	725	1000
Non-rounded	6.0	15.0	42.1	96.8	165	479.6	983.2

FEDERAL RESERVE BANK OF BOSTON™

Conclusions

Our data is close to the distribution described by Benford's Law for $n = 1$.

- DCPC 2012 shows similar differences that BLS discovered in the CE 2000 data for $d = 2, 5, 9$.
- Zero is heavily overrepresented in the 2nd and 3rd digits of the DCPC 2012 data.
- The data suggests that rounded dollar values are larger than those that are not rounded.

Does any of this suggest fraud or data manipulation? Probably not, but we could improve data accuracy by emphasizing that it is important to know the exact dollar value of every payment.

FEDERAL RESERVE BANK OF BOSTON™