Benford & RRT

Making Use of "Benford's Law" for the Randomized Response Technique

Andreas Diekmann
ETH-Zurich

1. The Newcomb-Benford Law

► Imagine a little bet. The two betters bet on the first digit of an unknown house number drawn at random. The loser has to pay one euro to the winner. Player A wins if the digit is in the range 1 to 4. Player B wins if the digit is 5 to 9. Is this a fair bet?

1. The Newcomb-Benford Law

► Imagine a little bet. The two betters bet on the first digit of an unknown house number drawn at random. The loser has to pay one euro to the winner. Player A wins if the digit is in the range 1 to 4. Player B wins if the digit is 5 to 9. Is this a fair bet?

► It is not. Paradoxically, the bet is rather unfavourable to player B. The first digits of house numbers follow a logarithmic distribution known as Benford’s law. The betters’ odds are 7:3 in terms of objective probabilities.

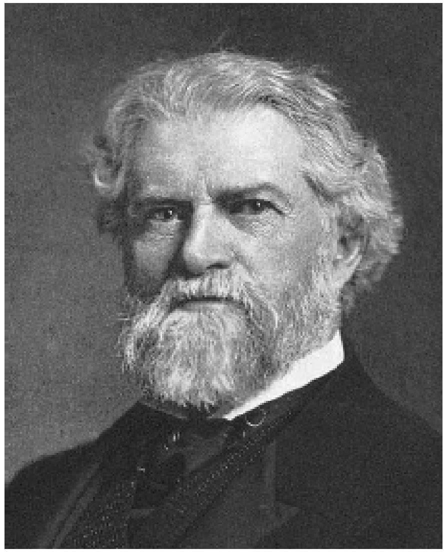


Abbildung 1: Simon Newcomb (1835–1909)

Abbildung 7: Frank Benford (1883–1948)

Die vorderen Seiten einer Logarithmentabelle sind stärker abgegriffen, als die hinteren.

Hungerbühler 2007

Benford's Law

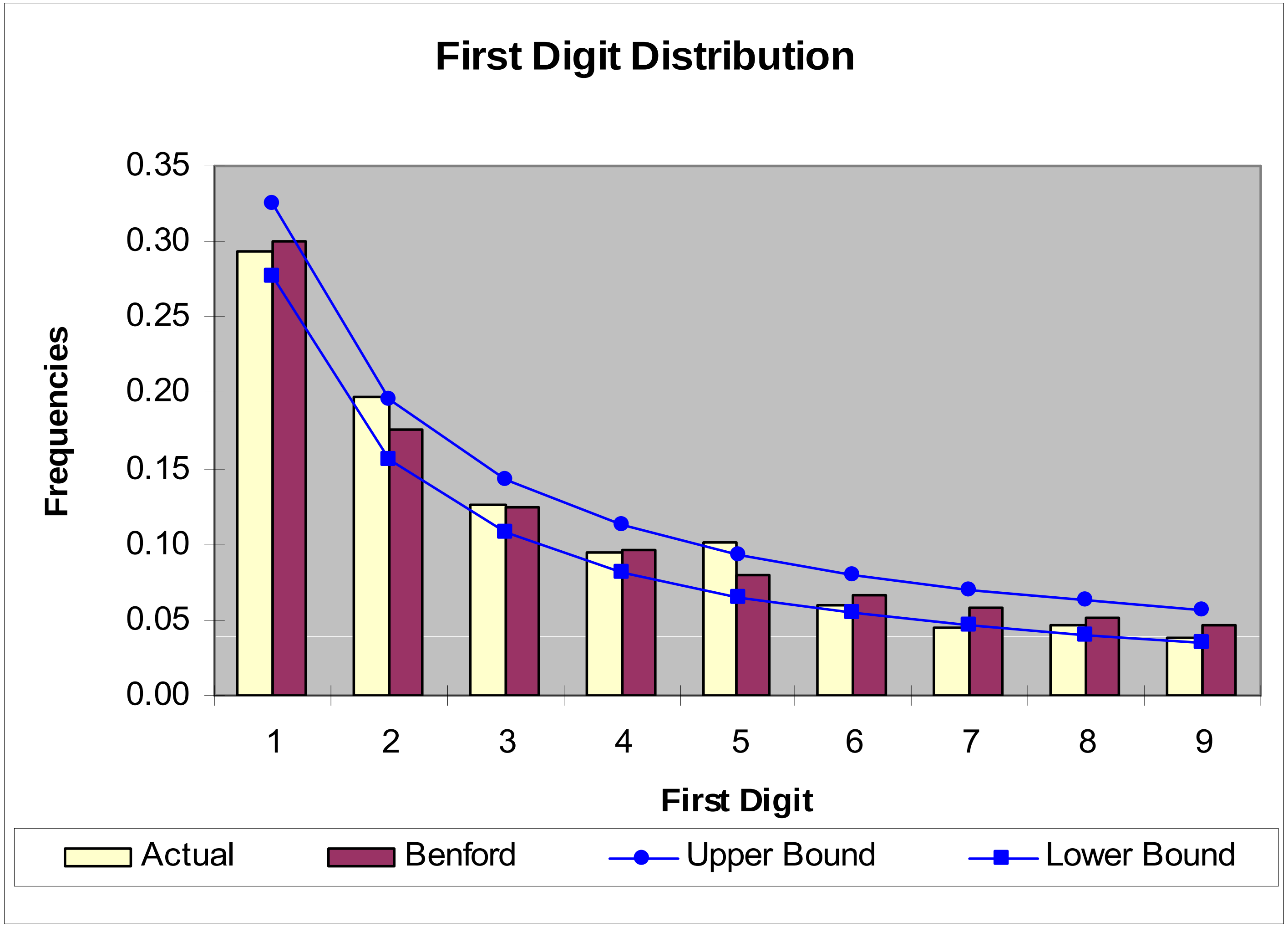
$$P(d_1) = \log_{10} (1 + 1/d_1).$$

1	2	3	4	5	6	7	8	9
0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

$P(D_1 = d_1, \ldots, D_k = d_k) = \log_{10} [\, 1 + (\Sigma d_i\, 10^{k-i})^{-1}]$

with $d_1 = 1, 2, \ldots, 9$ and $d_j = 0, 1, \ldots, 9$ $(j = 2, \ldots, k)$.

Distribution of First Digits of OLS-Regressions Coefficients from Articles Published in the American Journal of Sociology



N = 1457, Tables from AJS 104 / 105.
Deviation from Benford is significant for α=0.05.

Diekmann 2007

Digits in the Bible

Compilation of Digits in the „Elberfelder Konkordanz“

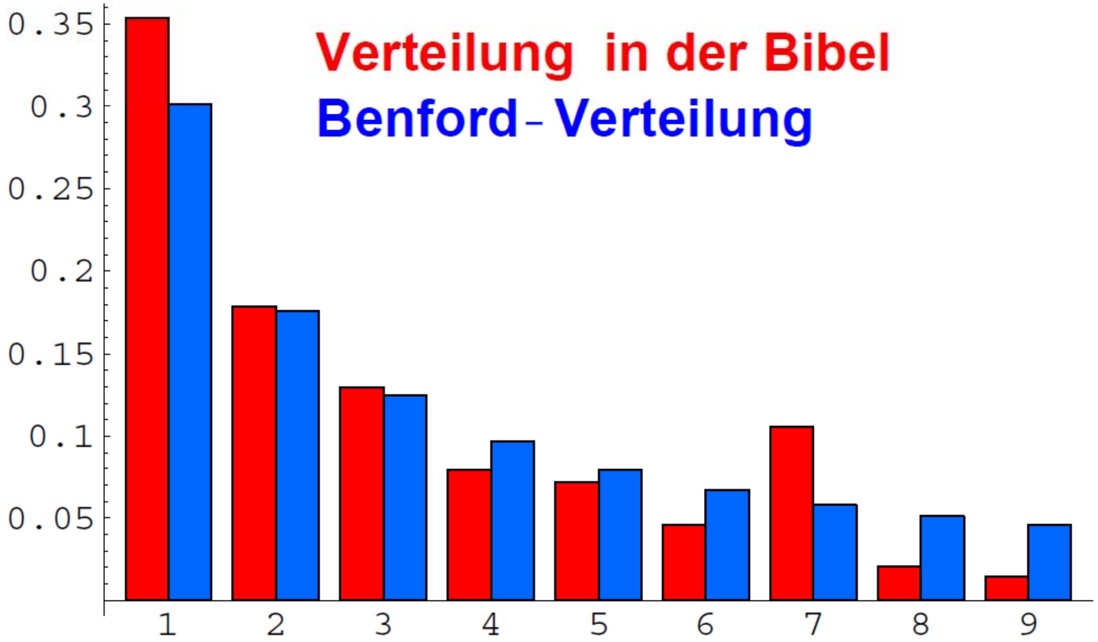
	603 550
2Mo 38,26	der zu den Gemusterten hinüberging, .. 603 550 ⟨Mann⟩
4Mo 1,46	es waren all die Gemusterten 603 550
2,32	Alle Gemusterten der Lager .. waren 603 550
	675 000
4Mo 31,32	das Erbeutete .. war: 675 000 Schafe
	800 000
2Sam 24,9	zwar gab es in Israel 800 000 Wehrfähige
2Chr 13,3	Jerobeam stellte sich gegen ihn .. auf mit 800 000
	1 000 000
1Chr 22,14	für das Haus .. 1 000 000 Talente Silber bereitgestellt
	1 110 000
1Chr 21,5	in ganz Israel 1 110 000 Mann, die das Schwert zogen

Hungerbühler 2007

Digits in the Bible

Compilation of Digits in the „Elberfelder Konkordanz“

	603 550
2Mo 38,26	der zu den Gemusterten hinüberging, .. 603 550 ⟨Mann⟩
4Mo 1,46	es waren all die Gemusterten 603 550
2,32	Alle Gemusterten der Lager .. waren 603 550
	675 000
4Mo 31,32	das Erbeutete .. war: 675 000 Schafe
	800 000
2Sam 24,9	zwar gab es in Israel 800 000 Wehrfähige
2Chr 13,3	Jerobeam stellte sich gegen ihn .. auf mit 800 000
	1 000 000
1Chr 22,14	für das Haus .. 1 000 000 Talente Silber bereitgestellt
	1 110 000
1Chr 21,5	in ganz Israel 1 110 000 Mann, die das Schwert zogen



Hungerbühler 2007

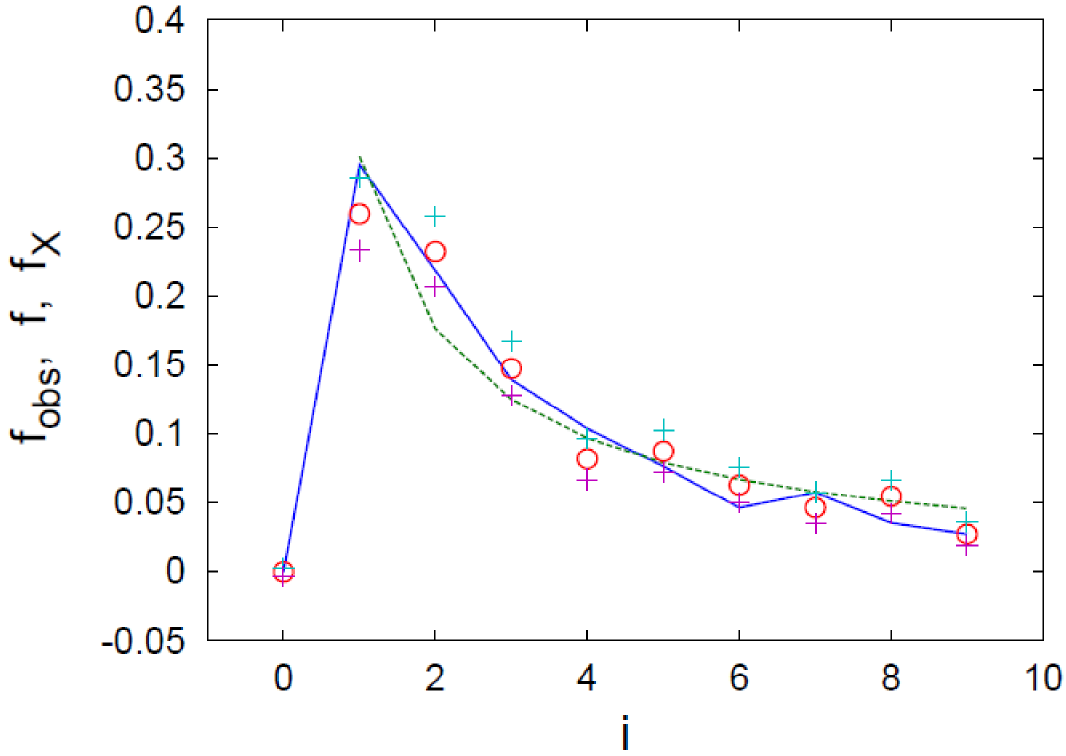


FIG 4. *As for Fig. 3, for candidate A vote counts only.*

Benford's Law and the number of votes for candidate Ahmadinejad (Roukema 2009)

Sensitive Questions

Allen H. Barton, 1958. Asking the Embarrassing Question.

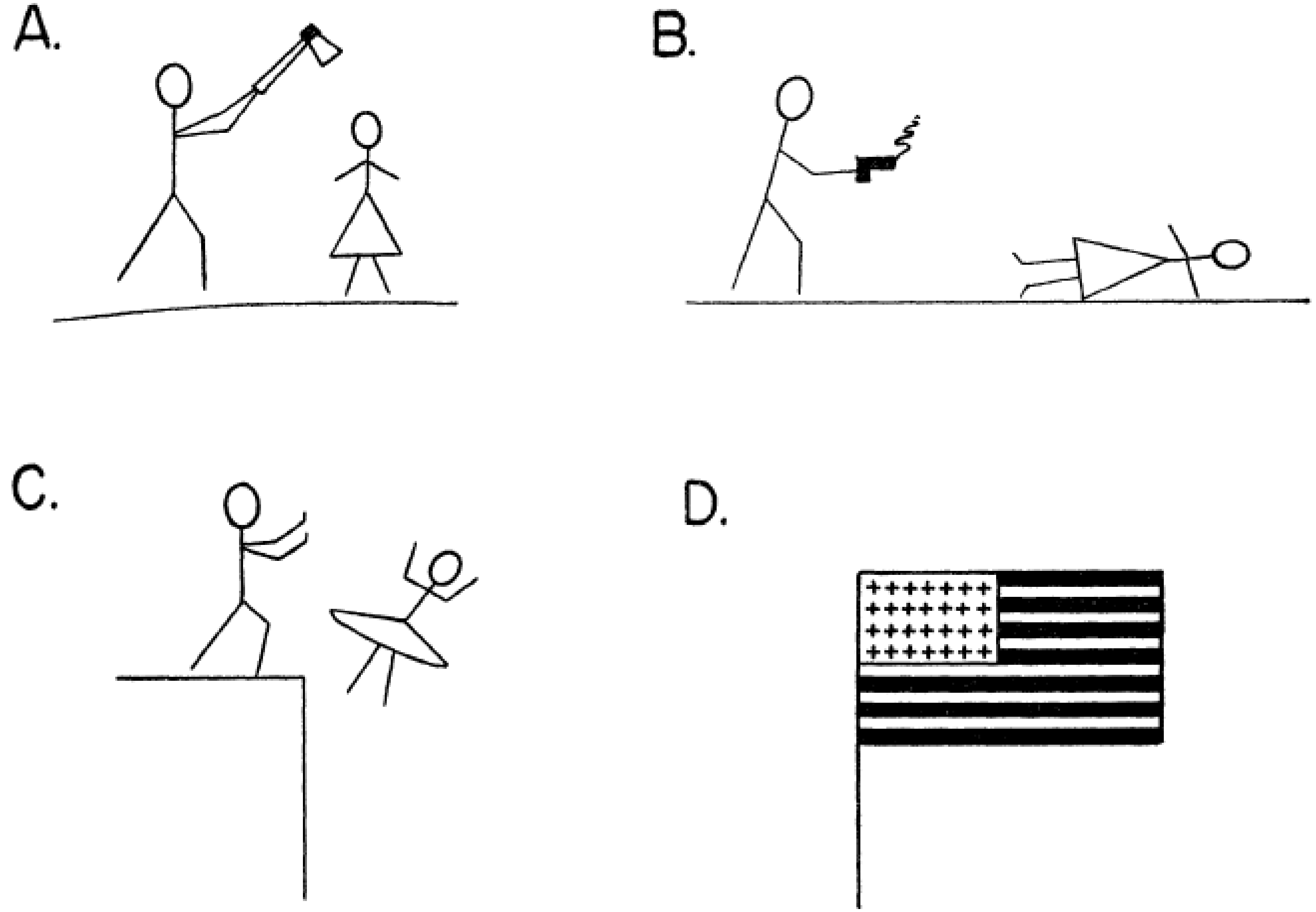
Public Opinion Quarterly 22: 67-68

Barton's (1958) method for a very sensitive question

THE POLLSTER's greatest ingenuity has been devoted to finding ways to ask embarrassing questions in non-embarrassing ways. We give here examples of a number of these techniques, as applied to the question, "Did you kill your wife?"

1. The Casual Approach:
   "Do you happen to have murdered your wife?"
2. The Numbered Card:
   Would you please read off the number on this card which corresponds to what became of your wife?" (HAND CARD TO RESPONDENT)
   1. Natural death
   2. I killed her
   3. Other (What?)

   (GET CARD BACK FROM RESPONDENT BEFORE PROCEEDING!)
3. The Everybody Approach:
   "As you know, many people have been killing their wives these days. Do you happened to have killed yours?"
4. The "Other people" Approach:
   (a) "Do you know any people who have murdered their wives?"
   (b) "How about yourself?"



6. The Projective Technique:

"What thoughts come to mind as you look at the following pictures?"

(Note: The relevant responses will be evinced by picture D.)

8. Putting the question at the end of the interview.

May be RRT is a better method for asking sensitive questions?

2. The Randomized Response Technique (RRT). A Method to Guarantee Full Anonymity for Sensitive Questions

► Subjects had to respond to either a sensitive question A (e.g. shoplifting, tax evasion etc.) or to a random question B (Was your mother‘s birthday in an even month?).

► Assignment to question A or B is by a random device (a dice, a coin etc.)

► The meaning of an individual answer cannot be identified. However, it is possible to estimate the proportion of shoplifting etc. and other statistics on the aggregate level.

- ► Because the random mechanisms are known one can estimate the probability of answering "yes" to the sensitive question by Bayes' formula.

- ► The RRT has the advantage of guaranteeing anonymity, but not without costs. The price is a loss in efficiency. In addition to sampling error, the probabilistic RRT device enlarges the variance of the estimated proportion of positive responses to the sensitive question.

In formal terms:

- p is the probability to answer the question of interest A, q =1-p is the probability to answer the random question B.
- π_y = P("yes"|B) is the probability to response "yes" to the random question.
- Then, we are looking for an estimate of π_x = P("yes"|A), the expected proportion of respondents answering "yes" to the question of interest.
- If we denote the overall proportion of "yes" in the sample by λ we have:

$\lambda = p\,\pi_x + (1-p)\,\pi_y$. ($\lambda$, p, π_y is known)

- Solving for π_x yields:
- $\pi_x = \lambda/p - \pi_y (1-p)/p$.
- p and π_y are determined ex ante by the researcher's RRT-design. A special case is the "forced response" design with $\pi_y = 1$. In this case, a person is "forced" to respond "yes" to the random question.
- With variance: $Var(\pi_x) = \lambda(1- \lambda)/np^2$

3. The Benford distribution as a randomizing device

- ►In face-to-face interviews, a pack of cards, a dice, a coin or some other device may be used to generate randomized outcomes. For example, if a person tosses “head” he or she is instructed to answer the random question, if the result is “tail” the question of interest has to be answered.

- ►This technique has some difficulties in telephone interviews and is particularly problematic in self-administered interviews such as mailed questionnaires or online-surveys.

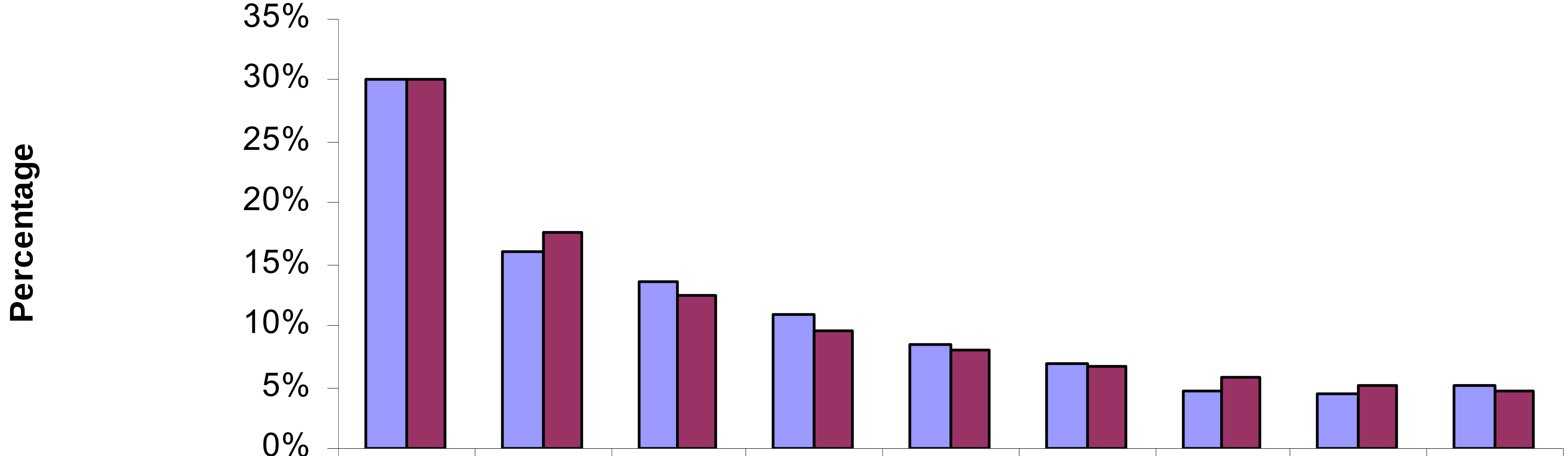
- ►As an alternative, I suggest to make use of the Benford distribution.

House numbers (1st digit) 1,2,3,4 versus 5,6,7,8,9

- The probability that digit 1, 2, 3 or 4 turns out is, therefore, 0.699 or roughly 0.70. The probability to draw a first digit among the set of remaining digits is 0.30.
- The 7:3 rule provides a mechanism to split the sample in a set of respondents answering the question of interest A and respondents answering the random question B. For example, subjects are asked to think of the address of a friend and to keep the house number in mind.
- Depending on the first digit either belonging to the set {1,2,3,4} or belonging to the set {5,6,7,8,9} a person has to answer question A or question B. Other sets may be constructed if a researcher prefers smaller or larger probabilities for the question of interest.

► **However, first we should ask: Do house numbers follow the Benford distribution at all?**

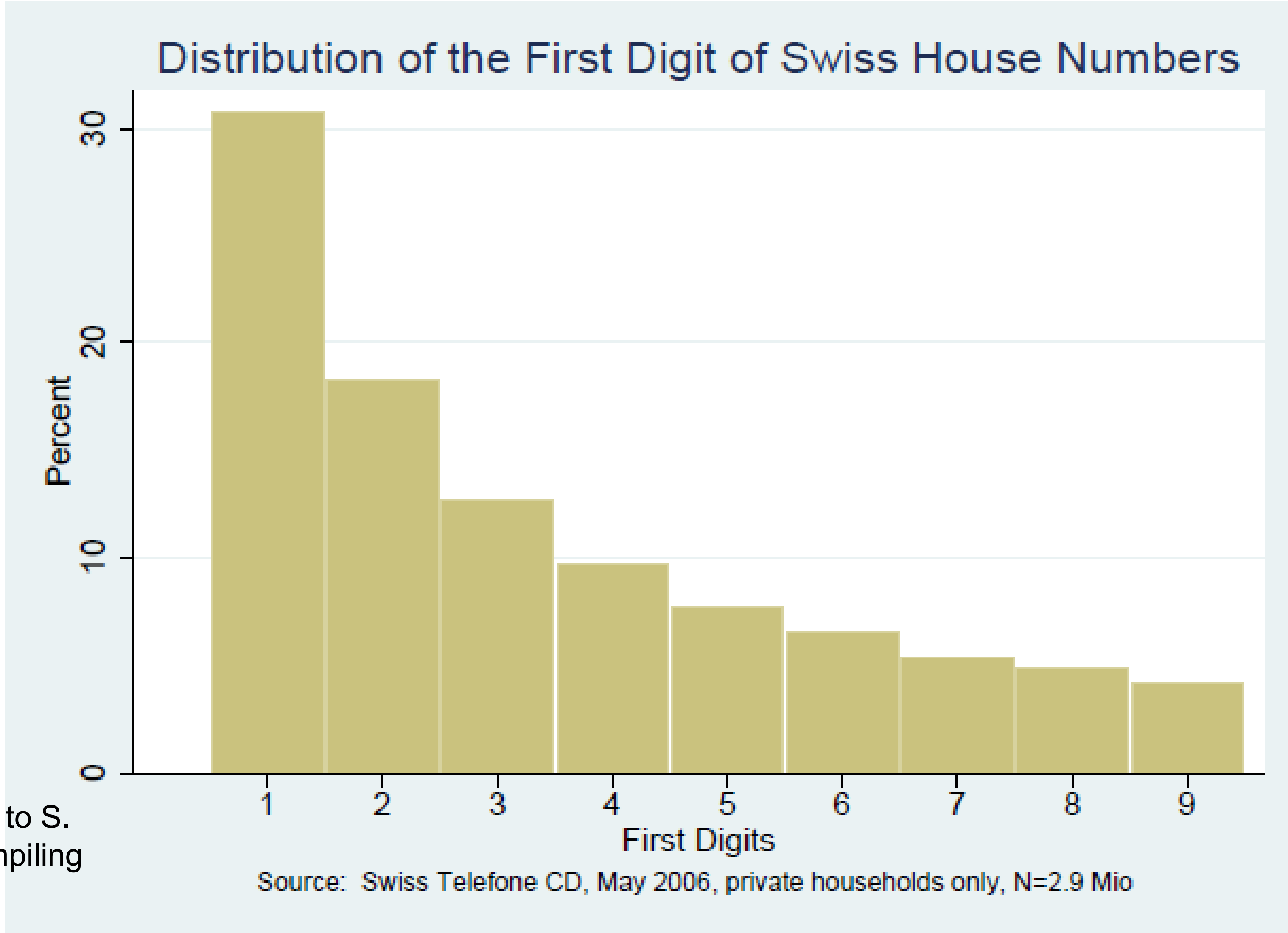
House numbers collected from the telephone directory of Zurich



	1	2	3	4	5	6	7	8	9
House number	29,99%	15,96%	13,51%	10,84%	8,46%	6,90%	4,75%	4,45%	5,12%
Benford	30,10%	17,61%	12,49%	9,69%	7,92%	6,69%	5,80%	5,12%	4,58%

First digit

Benford Distributed House Numbers



I am indebted to S. Wehrli for compiling the data.

4. The “Benford illusion” and other advantages of the method

- The price for the anonymity of the method is an increase in the variance of the estimator for the proportion of yes-responses (π_x) to the question of interest.
- The variance is (Fox and Tracy 1986):

  $Var(\pi_x) = \lambda(1- \lambda)/n(1-q)^2$
- It follows that the variance increases with the probability q = 1-p to arrive at the „random question”.
- On the other hand, the larger q the larger is the degree of anonymity.
- **This is the formal expression for the conflict between efficiency and anonymity.**

„Benford Illusion“

- To use the Benford distribution for the RRT has the advantage to diminish the conflict between efficiency and anonymity.

- The reason is that the perceived probabilities and the objective probabilities differ. Many people believe that the chance to pick a one, two, three or four is much smaller than 70 percent.

- This discrepancy or “Benford illusion” has the positive effect that the perceived q, and, therefore, the perceived anonymity is larger than the objective q. With the little trick of the Benford illusion, the anonymity can be increased without loss in efficiency.

- There are other advantages, too. The method does not require any physical device such as a coin or a dice to generate random numbers.
- In most previous studies, the RRT is applied to sensitive questions in face-to-face interviews.
- However, it is unlikely that most people, asked to fill in online-surveys or mailed questionnaires, follow instructions properly if a coin or dice is required.

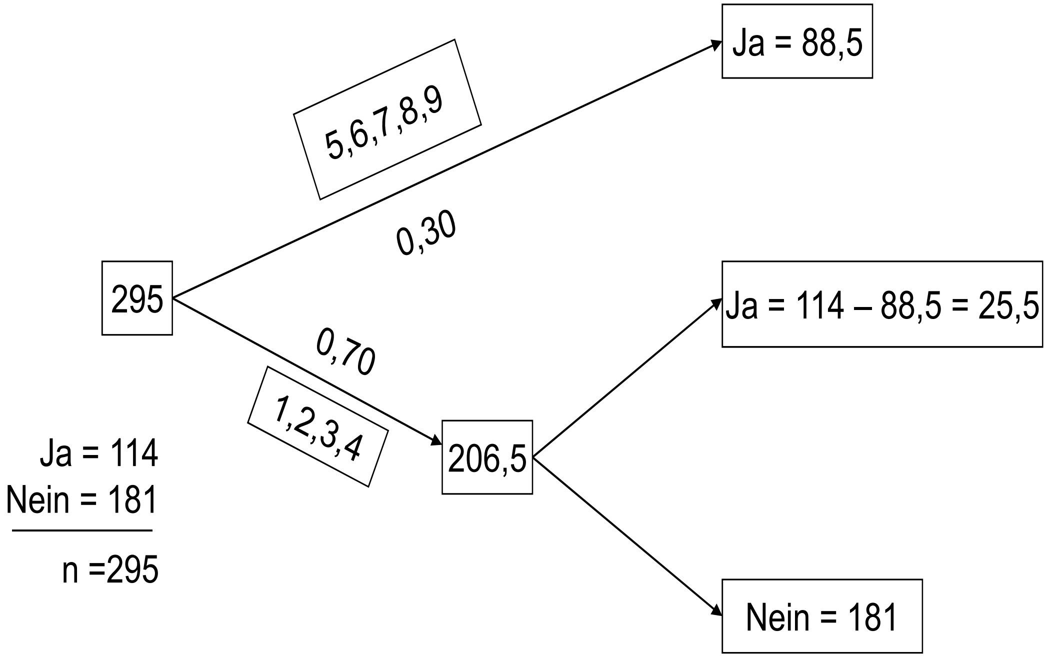
5. Application Shoplifting

Questionnaire

- Imagine a friend or relative who does not live in your house with an address known to you.
- Keep in mind the house number‘s first digit.
- If the digit ist 5,6,7,8 or 9 skip over the next question and mark „yes“
- If the digit is 1,2,3,4, please, answer the following question: „In the last five years, did you ever intentionally pick a shopping item without paying for it?“

Study 1: Shoplifting

RRT Experiment in Vorlesung SS 07

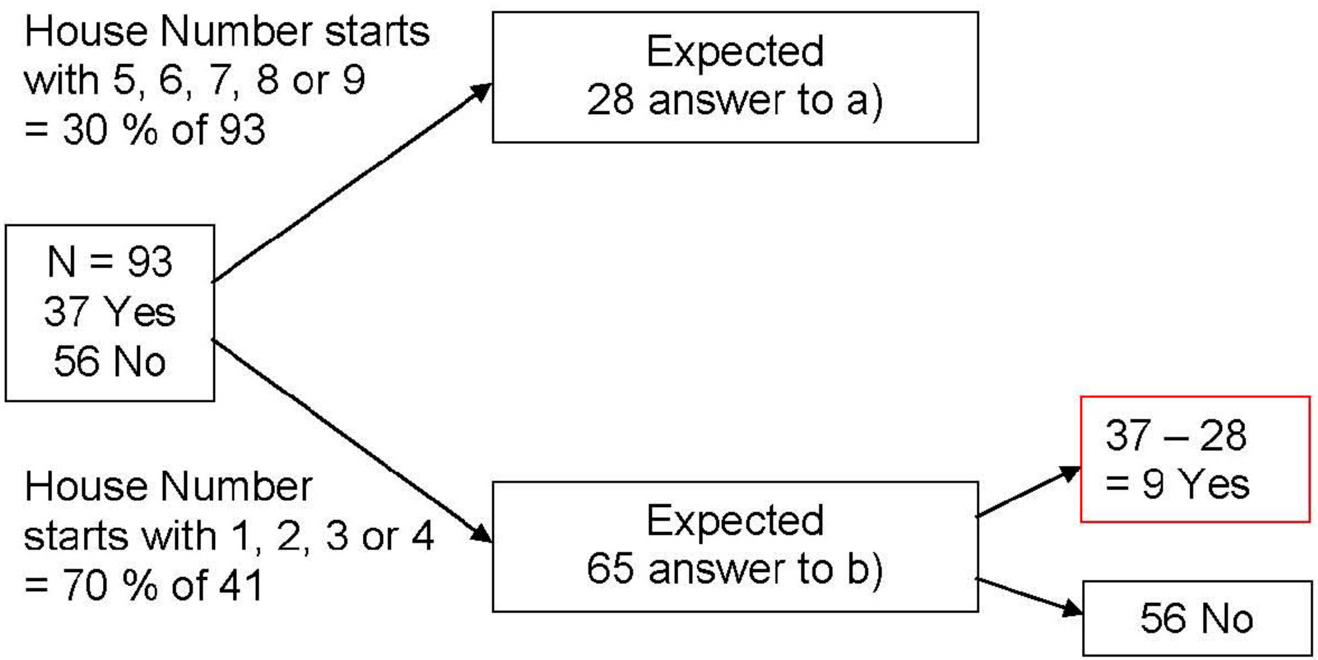


p (Ladendiebstahl) = 25,5/206,5
= 0,12

Questionnaire
in lecture
M. Abraham,
Bern 2007

Result:
n = 295
π_x = 0.12
(SE = 0.04)

Study 2: Shoplifting



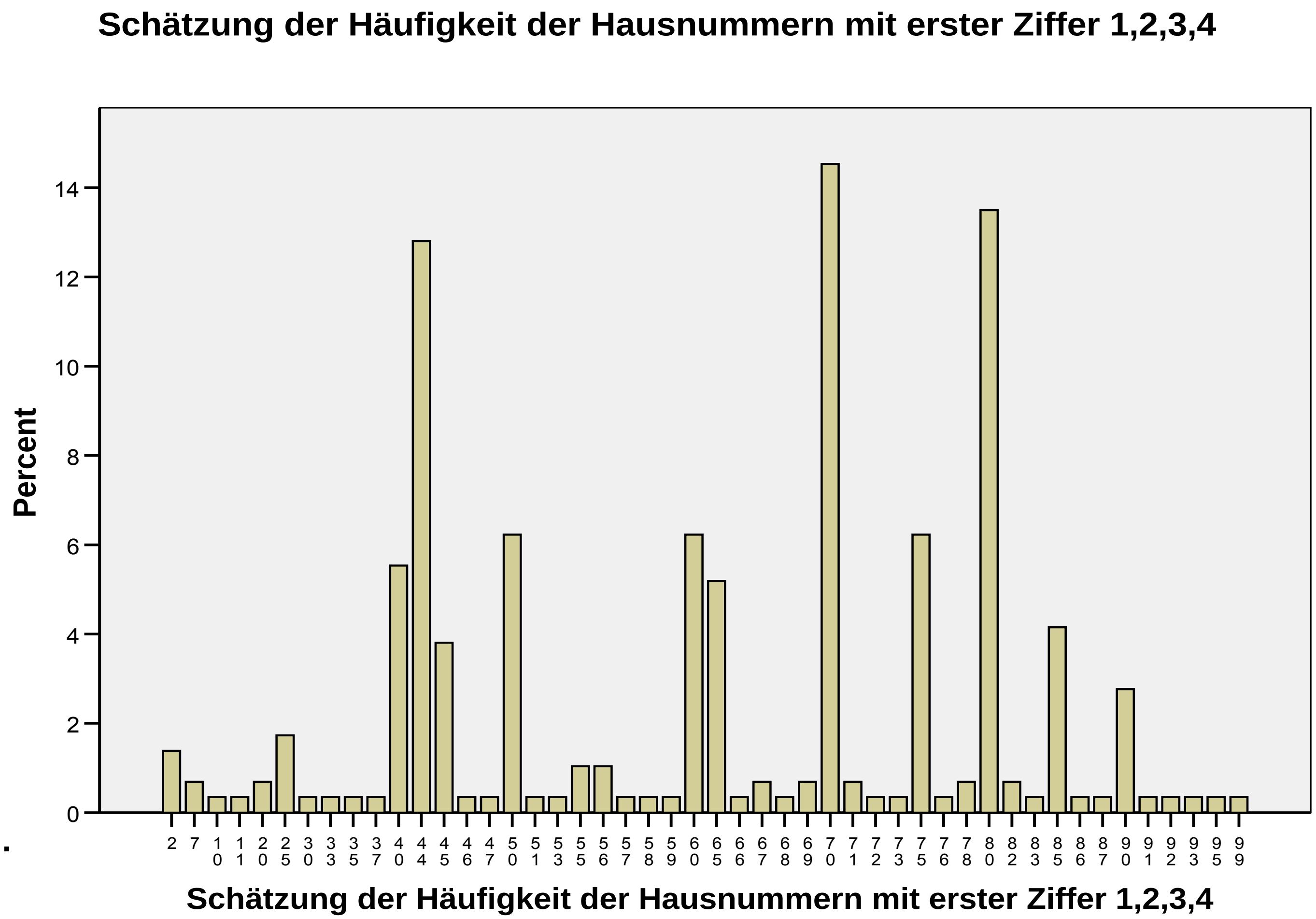
Result: n = 93

$\pi_x = 9/65 = 0.14$

(SE = 0.073)

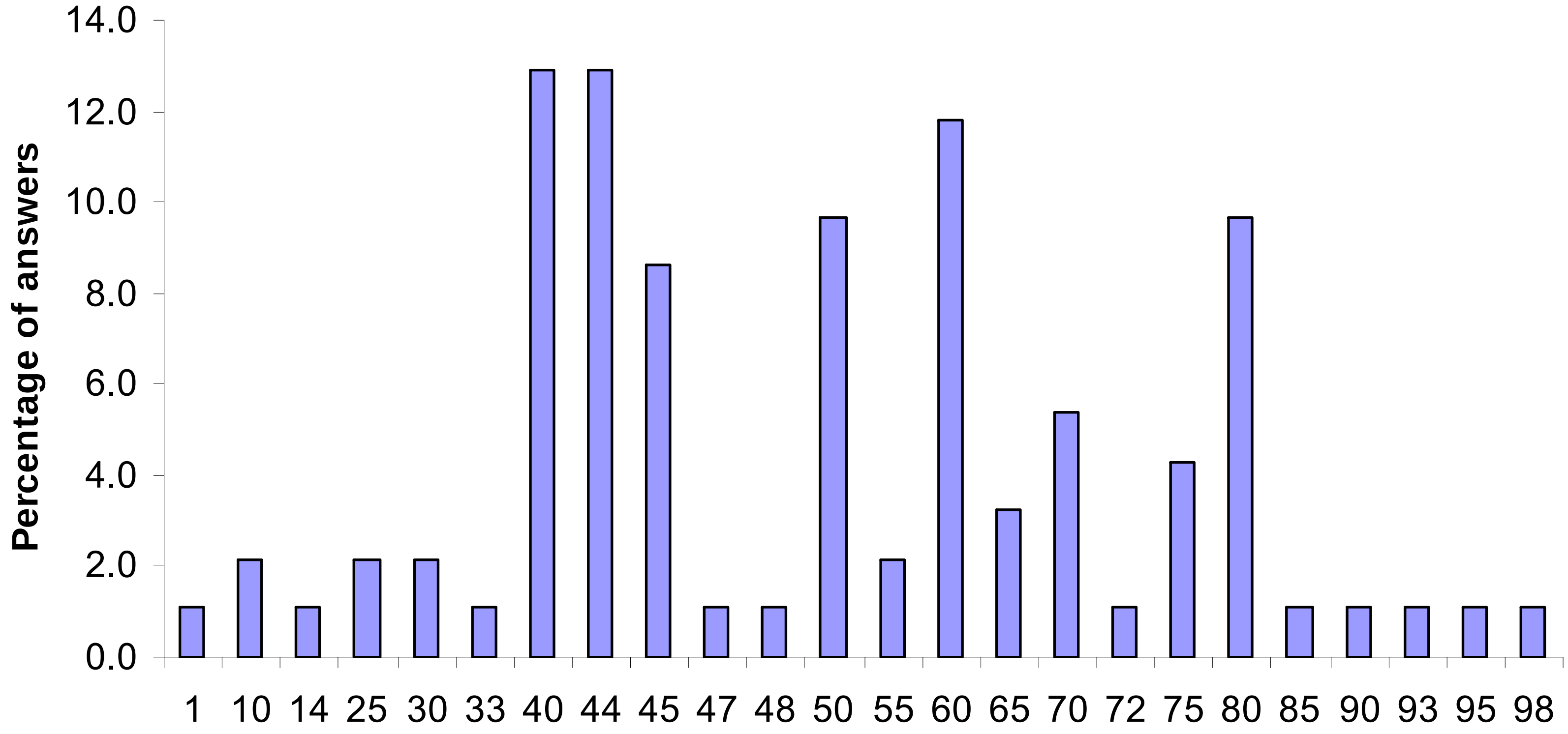
Questionnaire
in lecture Szydlick

6. Do Subjects underestimate the probability of 1,2,3,4? („Benford Illusion“)



N = 289,
Mean =
61.
Lecture M.
Abraham,
Bern 2007

Estimated frequency of house numbers starting with 1, 2, 3 or 4 in per cent



Lecture Szydlik, n = 92, mean = 54

Underestimation of Objective Probability (student population)

	subjective (mean)	objective
Study 1, Bern	61	70
Study 2, Zurch	54	70

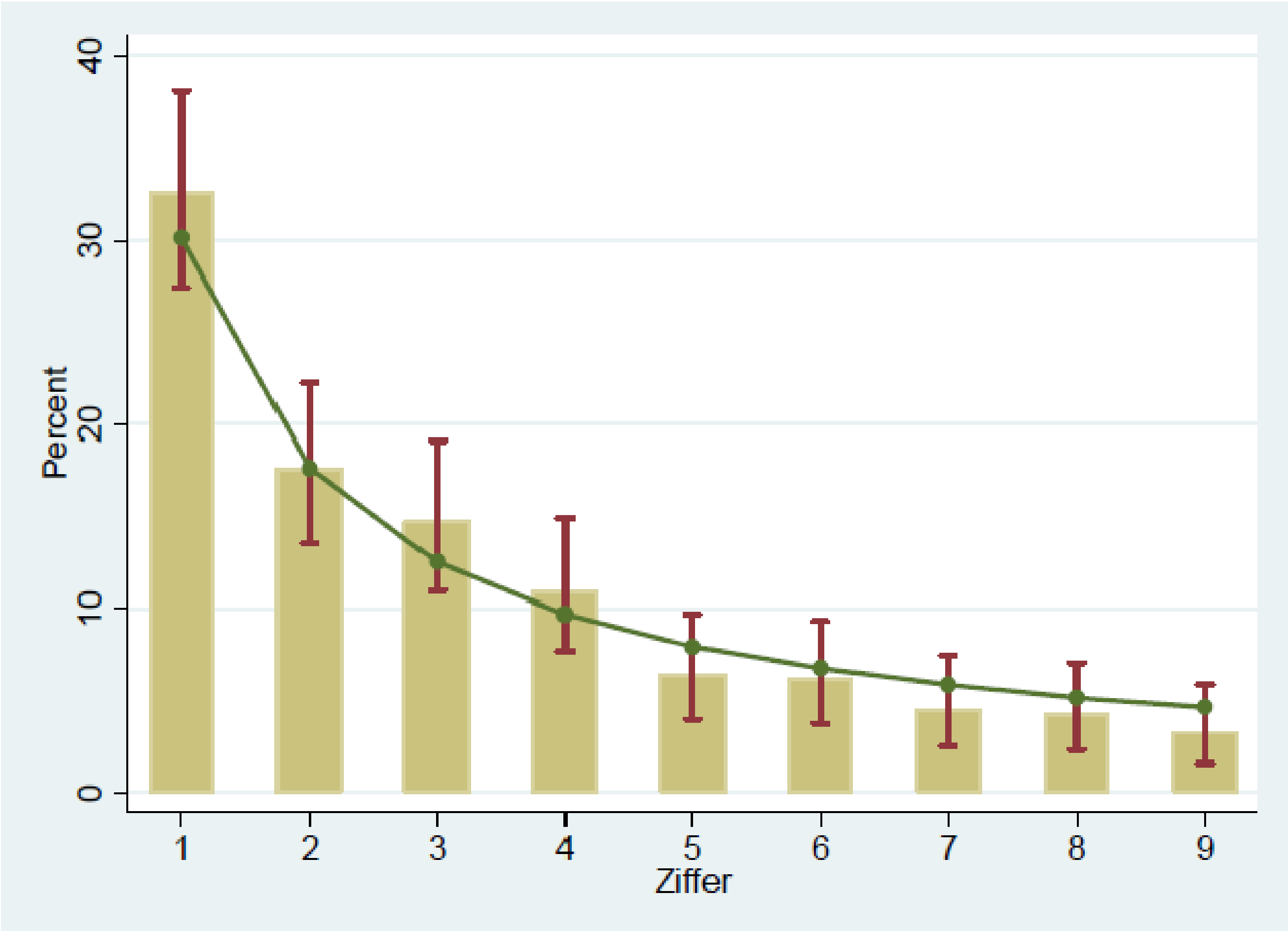
7. Do subjects generate Benford-distributed house numbers?

► As we have seen, objective data follow the Benford distribution.

► However, are the digits produced by the respondents in accordance with Benford as well?

► This is a crucial assumption. Otherwise, the method wouldn‘t work.

7. Do subjects generate Benford-distributed house numbers?



I am indebted to B. Jann for compiling the data.

Survey B. Jann, Wages in Switzerland, 2006/2007, N = 313