Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data*

Andreas Diekmann
Swiss Federal Institute of Technology Zurich

October 2004
diekmann@soz.gess.ethz.ch

*For data collection I would like to thank Regula Bieri (section III), Ben Jann (Experiments 1 and 3), and Kurt Schmidheiny (Experiment 2). I also like to thank Niels Lepperhoff and Peter Preisendoerfer for valuable hints. I am very much indebted to Elisabeth Coutts for helpful comments and assistence.

Abstract

Digits in statistical data produced by natural or social processes are often distributed in a manner described by "Benford's law". Recently, a test against this distribution was used to identify fraudulent accounting data. This test is based on the supposition that real data follow the Benford distribution while fabricated data do not. Is it possible to apply Benford tests to detect fabricated or falsified scientific data as well as fraudulent financial data? We approached this question in two ways. First, we examined the use of the Benford distribution as a standard by checking digit frequencies in published statistical estimates. Second, we conducted experiments in which subjects were asked to fabricate statistical estimates (regression coefficients). These experimental data were scrutinized for possible deviations from the Benford distribution. There were two main findings. First, the digits of the published regression coefficients were approximately Benford distributed. Second, the experimental results yielded new insights into the strengths and weaknesses of Benford tests. At least in the case of regression coefficients, there were indications that checks for digit-preference anomalies should focus less on the first and more on the second and higher-digits.

I. Introduction

The digits of numerical data produced by a large number of very different natural and social processes take the form of a logarithmic distribution described by Benford's law. Given the number and variety of processes that produce Benford-distributed data, it is often assumed that many kinds of real data adhere to Benford's law. The further assumption that fabricated or falsified data are detectable through the deviation of their digits from the Benford distribution has been tested recently in several contexts. For example, some studies have reported success in identifying fraudulent information with a check of digital frequencies in tax or other financial data against the Benford distribution (Carslow 1988, Berton 1995, Nigrini 1996, Quick and Wolz 2003). Similar results have been reported for fabricated survey interviews (Schraepel and Wagner 2003, Schäfer et al. 2004). It may well be that "Benford tests" can also be used to identify fraudulent scientific data or results.

In the empirical sciences, publications often report large tables with statistical estimates (such as regression coefficients) whose digits might fruitfully be compared with the Benford distribution. In this article, we will empirically investigate the application of the Benford test to regression coefficients and other statistics. Regression coefficients were chosen as an object of study because of their ubiquity in the scientific literature, and not only in fields such as sociology or psychology. Estimates for regression coefficients are, for example, frequently reported in econometrics. Biomedical researchers also use regression analysis or related techniques such as logistic regression.

However, before we can apply Benford tests to these data, it must be demonstrated that the digits of regression coefficients or other statistical estimates are generally distributed in accordance with Benford's law. And, even

if there is evidence for the use of this standard, employing the Benford test to identify fraudulent data means that the deviation of fraudulent data from the standard set by Benford's law must also be demonstrated. Good evidence is required for both of these hypotheses before the Benford test can be accepted as a valid procedure for detecting anomalies in scientific publications. The first of the above hypotheses (that real data are Benford distributed) is tested in Section III of this paper. In an effort to learn more about the distributional properties of the digits from estimated regression coefficients, we collected a large sample of regression coefficients from the published literature. In Section IV, we report on the results of three experiments designed to test the second hypothesis (that the digits of fraudulent data deviate from the Benford distribution). In these experiments, students attending university-level statistics courses were asked to construct a table of regression coefficients in support of a certain hypothesis. The second hypothesis predicts that the first and second digits of the fabricated data will deviate from Benford's law.

II. Benford's Law

The logarithmic distribution of the first digit d_1 of various naturally occurring quantities is described by "Benford's law" or the "first digit phenomenon" (Hill 1998, Raimi 1969, 1976):

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

According to the formula, the probability that a number's first digit is "1" is 0.301, while a "9" is expected with a much lower probability of 0.046 (see Table 1).

This phenomenon was discovered by Newcomb (1881), who observed that tables of logarithms were used more often for smaller digits than for larger ones. Half a century later, Benford (1938) happened upon this regularity through the same observation (Hill 1995a). However, Benford went further in computing frequency distributions for the first digits of a variety of data such as the area of riverbeds, figures from newspaper articles, population figures and other data. The digits of these data could be closely approximated by the logarithmic distribution.

A generalized distribution describes the data's other digits. The joint distribution of first and higher-order significant digits takes the form (Hill 1995a):

$$P(D_1 = d_1, ..., D_k = d_k) = \log_{10} [\ 1 + (\Sigma d_i\ 10^{k-i})^{-1}] \qquad (2)$$

whith $d_1 = 1, 2, ... ,9$ and $d_j = 0, 1, ... ,9$ $(j = 2, ..., k)$. For example, if digits are Benford distributed the combination of significant digits 1028 (e.g. 0.001028) is expected with probability $\log_{10} [1 + 1/1028]$. This “general significant-digit law” (Hill 1995a) permits the derivation of the marginal distributions of second-order and higher-order digits. Table 1 displays the probabilities for the first three significant digits.

[Table 1]

It follows from the joint distribution described above that the distribution of higher-order digits increasingly approximates the uniform distribution.

Since Benford’s publication, substantial progress has been made in explaining the mechanism behind the generation of Benford-distributed digits. Hill (1995a, 1998) proved a “random samples from random distributions theorem”. If one

first chooses a sample of distributions at random and then samples digits from those distributions, the resulting distribution will – under certain conditions – approximate Benford's law. Also, Hill (1995a) was able to prove the base and scale invariance of Benford's law rigorously. Hence, if Benford's law, for example, applies to the distribution of the digits of data on the area of lakes in units of acres it will (on average) also apply to the same data in units of square meters. Moreover, Hill (1995b) has shown that Benford's logarithmic distribution is the only scale-invariant distribution.

III. Digit Distribution of Statistical Estimates

A necessary prerequisite for the application of Benford tests for the accuracy of any kind of data is that the real (i.e. not fabricated or falsified) data should be Benford distributed. Little information exists on the Benford conformity of raw data, and even less exists on whether statistical estimates generally take the form of the Benford distribution. To our knowledge, with the exception of Becker's (1982) analysis of failure rates, there have been no published investigations of the typical distribution of digits for statistical estimates such as standard deviations or regression coefficients. To examine the use of the Benford distribution as a standard, we created a dataset of first digits from means, standard deviations, correlation coefficients, and standardized and unstandardized regression coefficients (including those from ordinary least squares and logistic regression models), including about one thousand digits for each statistic. These data were collected from tables published in two volumes of the "American Journal of Sociology" from January 1996 (Vol. 101) to May 1997 (Vol. 102).

The relative frequencies of the first digits of unstandardized regression coefficients closely approximate the Benford distribution. For example, a "one"

has a relative frequency of 0.307 in our sample, while the value predicted by Benford's law is 0.310 (Figure 1). For a significance level of $\alpha = 0.05$, a comparison with the Benford distribution supports the null hypothesis of no difference between the predicted and observed distributions ($\chi^2 = 7.115$, df = 8, $p = 0.524$).

[Figure 1]

On the other hand, the fit between the distribution of the statistical estimates' first digits and the Benford distribution is much worse for means, standard deviations, correlations, and standardized coefficients (results not shown).
To explore the robustness of the above result and to gather information on the Benford conformity of the estimates' second digits, we inspected an additional sample of regression coefficients. The second sample was drawn from the same journal as the first and contains 1,457 first and second digits from all the tables of (OLS) regression coefficients published in Volume 104, Issues 1-6 (1999) and Volume 105, Issues 1-5 (2000) of the same journal.

Although the χ^2 test results in the rejection of the null hypothesis that the first digits of the second set of regression coefficients are drawn from a Benford distribution ($\chi^2 = 21.072$, df = 8, $p = 0.007$), the approximation is not all that poor in descriptive terms. The significant deviation is caused mostly by the higher-than-expected occurrence of the digit "5", which has a relative frequency 0.101 in the sample of regression coefficients, as compared to an expected frequency of 0.079. Moreover, the second digits are distributed largely in accordance with the monotonic decline of digit frequencies predicted by Benford's law (Figure 2).

The observed distribution of second digits yields a better approximation of the Benford-predicted distribution ($\chi^2 = 7.115$, df = 9, p = 0.524). Note that the observed values exhibit the typical pattern of a monotonic decline and therefore deviate systematically from a uniform distribution.

[Figure 2]

In summary, the largest discrepancy between the predicted and observed digit frequencies is 0.022 for a first digit of "5" in this second sample. Further, all of the above tests on regression coefficients reveal the pattern of a monotonic decline in the digital frequencies. Hence, the conclusion that the digits of regression coefficients closely approximate Benford's law is justified.

IV. Experiments with Fabricated Regression Coefficients

In three separate experiments, students participating in statistics courses at the University of Berne in Switzerland were asked to fabricate regression coefficients in accordance with a given hypothesis. Students were mainly from the sociology (experiment 1 in January 2001, and experiment 3 in January 2004) and economics (experiment 2 in October 2001) departments. Subjects were asked to construct "plausible values" of regression coefficients that would support a controversial hypothesis from neoclassical economics, and then record these values on a form provided by the researchers. The hypothesis was, "The higher the unemployment benefits, the longer the duration of unemployment". They were asked to generate four-digit coefficients for the unemployment benefit variable and nine other independent variables or "controls", such as education in years, job experience, gender, and so on. In experiments 1 and 2, each subject produced the ten coefficients detailed above. The task in experiment 3 was the same, except that subjects were asked to fabricate those

ten coefficients for ten separate samples, in other words to fabricate 100 four-digit regression coefficients.

A few students produced data that indicated they had either not understood the task or not followed instructions to any meaningful extent; their questionnaires were excluded from the analysis. Data from a total of 10 questionnaires were used from experiment 1 (n=100 coefficients), 13 questionnaires from experiment 2 (n = 130), and 14 questionnaires from experiment 3 (n = 882). Only four subjects completed the entire experiment 3 questionnaire within the time allotted (about 35 minutes), while the other ten filled in the questionnaire at least partially. Data were aggregated for analysis across subjects in experiments 1 and 2, while the large number of fabricated coefficients collected in experiment 3 allowed for a separate analysis of the data for every individual.

The distribution of first digits produced in both experiments 1 and 2 exhibits a pattern similar to the one predicted by Benford's law. In both experiments, χ^2 tests for the equivalence of the expected and the observed distributions do not permit the rejection of the null hypothesis for $\alpha = 0.05$ (experiment 1: $\chi^2 = 10.644$, df = 8, p = 0.223; experiment 2: $\chi^2 = 15.295$, df = 8, p = 0.054), although the test statistic for experiment 2 just failed to reach the level of statistical significance. More importantly, the shape of the frequency distribution mirrors the monotonic decline of the Benford distribution for both experiments. Thus, data from these experiments do not support the idea that the first digits of fabricated data deviate from Benford's law.

[Figure 3]

What about the second digit? In both experiments, the observed distributions of the second digits deviate significantly from the Benford distribution (experiment

1: $\chi^2 = 27.000$, df = 9, p = 0.001; experiment 2: $\chi^2 = 23.570$, df = 9, p = 0.005). The hypothesis that "true" regression coefficients follow Benford's law while fabricated data do not is supported by the analysis of the second digits although it was not supported by the analysis of the first.

Of course, a weakness of the experiments is that they permit only the analysis of aggregated data. Assuming that there is individual variance in the falsification patterns, an individual level analysis might be more informative. The third experiment was conducted to collect enough data from each subject to permit an individual-level analysis. This procedure allows for the separate analysis of individual data.

[Table 2]

In principle, the results from the third experiment are very much in line with those from aggregate-level experiments. Most subjects exhibit fabrication patterns that conform to Benford's law for the first digit, but not for the second or higher-order digits. With the Benford distribution as the null-hypothesis, the pattern of the failure to reject the null-hypothesis for the first digit and of the rejection of the null-hypothesis for the second and higher-order digits is supported by most of the individual-level significance tests conducted for these data: Out of 14 tests, three are significant ($\alpha = 0.05$) for the first digit, while ten tests are significant for the second digit, 12 for the third digit, and 13 for the fourth digit (Table 2).

It is not the first digit that matters! This result fits well with the finding by Mosimann et al. (1995) that the inspection of the higher-order digits of fabricated data provides better clues to errors or data fabrication than does the inspection of the first digit. Quite interestingly, subjects favour smaller first

digits in fabricating regression coefficients, resulting in a Benford-like pattern for the distribution of first-digits in fabricated data. So, a test for the fabrication of regression coefficients might most fruitfully focus on the second, third or higher-order digits. If second and higher-order digits deviate from the Benford distribution, this deviation may yield an indication that the data have been fabricated. At least for regression coefficients, it appears that using a Benford test of first digits for data fabrication would provide misleading results.

References

Becker, P, 1982: Patterns in Listings of Failure-Rate and MTTF values and listings of other data. IEEE Transactions on Reliability R-31: 132-134.

Benford, Frank, 1938: The Law of Anomalous Numbers. Proceedings of the American Philosophical Society 78: 551-572.

Berton, L., 1995: He's Got their Number. Scholar Uses Math to Foil Financial Fraud. Wall Street Journal, July 10.

Carslow, C., 1988: Anomalies in Income Numbers. Evidence of Goal Oriented Behavior. The Accounting Review 63: 321-327.

Drake, Philip D. and Nigrini, Mark J., 2000: Computer Assisted Analytic Procedures Using Benford's Law. Journal of Accounting Education 18: 127-146.

Hill, Theodore P., 1995a: A Statistical Derivation of the Significant-Digit Law. Statistical Science 10: 354-363.

Hill, Theodore P., 1995b: Base Invariance Implies Benford's Law. Proceedings of the American Mathematical Society 123: 887-895.

Hill, Theodore P., 1998: The First Digit Phenomenon. American Scientist 86: 358-363.

Mosimann, James E., Wiseman, Claire V., Edelman, Ruth, E., 1995: Data Fabrication: Can People Generate Random Digits? Accountability in Research 4: 31-55.

Newcomb, Simon, 1881: Note on the Frequency of Use of the Different Digits in Natural Numbers. American Journal of Mathematics 4: 39-40.

Nigrini, Mark J., 1996. A Taxpayer Compliance Application of Benford's Law. The Journal of the American Taxpayer Association 18: 72-91.

Quick, Reiner and Wolz, Matthias, 2003: Benford's Law in deutschen Rechnungslegungsdaten. Betriebswirtschaftliche Forschung und Praxis: 208-224.

Raimi, Ralph A., 1969: The Peculiar Distribution of First Digits. Scientific American: 118-120.

Raimi, Ralph A., 1976: The First Digit Problem. American Mathematical Monthly 83: 521-538.

Schäfer, Christian, Schräpler, Jörg-Peter, Müller, Klaus-Robert and Wagner, Gert G., 2004. Identification of Faked and Fraudulent Interviews in Surveys. To be published in Schmollers Jahrbuch.

Schraepler, Joerg-Peter and Wagner, Gert G., 2003. Identification and Characteristics of Faked Interviews in Surveys. Unpublished manuscript Ruhr University Bochum and DIW Berlin.

Table 1: Probabilities Predicted by Benford's Law for the First and Higher-Order Digits*

d_i	$P(d_1)$	$P(d_2)$	$P(d_3)$	$P(d_4)$
0		0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.10882	0.10097	0.10010
3	0.12494	0.10433	0.10057	0.10006
4	0.09691	0.10031	0.10018	0.10002
5	0.07918	0.09668	0.09979	0.09998
6	0.06695	0.09337	0.09940	0.09994
7	0.05799	0.09035	0.09902	0.09990
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.08500	0.09827	0.09982

*Figures are adapted from Nigrini 1996.

Table 2: Analysis of Fabricated Data for Individual Subjects (Experiment 3)

	1st Digit			2nd Digit			3rd Digit			4th Digit		
Subject	*χ2*	*p value*	*n*	*χ2*	*p value*	*n*	*χ2*	*p value*	*n*	*χ2*	*p value*	*n*
1	**18.49**	**0.018**	100	**30.11**	**0.000**	100	**32.35**	**0.000**	99	**28.28**	**0.001**	98
2	14.14	0.078	100	**23.88**	**0.004**	100	**25.49**	**0.002**	100	**19.29**	**0.023**	100
3	9.08	0.336	100	12.58	0.182	100	**19.59**	**0.021**	100	**33.04**	**0.000**	100
4	7.90	0.443	100	**30.15**	**0.000**	99	**33.70**	**0.000**	93	**35.83**	**0.000**	85
5	5.60	0.692	26	14.75	0.098	26	11.72	0.229	26	12.44	0.190	26
6	9.19	0.326	20	9.44	0.398	20	**24.61**	**0.003**	20	**18.05**	**0.035**	20
7	3.12	0.926	24	**17.16**	**0.046**	24	**57.31**	**0.000**	24	**22.60**	**0.007**	23
8	**34.03**	**0.000**	45	**17.09**	**0.047**	45	**17.69**	**0.039**	45	**38.68**	**0.000**	45
9	7.85	0.448	68	**42.69**	**0.000**	68	**40.91**	**0.000**	68	**25.36**	**0.003**	67
10	6.42	0.601	60	**17.26**	**0.045**	60	**44.47**	**0.000**	60	**103.07**	**0.000**	56
11	9.13	0.331	63	**40.88**	**0.000**	63	**113.22**	**0.000**	62	**162.69**	**0.000**	52
12	13.64	0.092	46	**22.91**	**0.006**	46	**19.64**	**0.020**	46	**40.46**	**0.000**	44
13	**19.39**	**0.013**	50	**23.79**	**0.005**	49	8.33	0.502	47	**29.32**	**0.001**	42
14	5.49	0.705	80	13.40	0.145	80	**22.48**	**0.007**	79	**27.34**	**0.001**	75
All	12.26	0.140	882	**31.83**	**0.000**	880	**59.90**	**0.000**	869	**112.74**	**0.000**	833

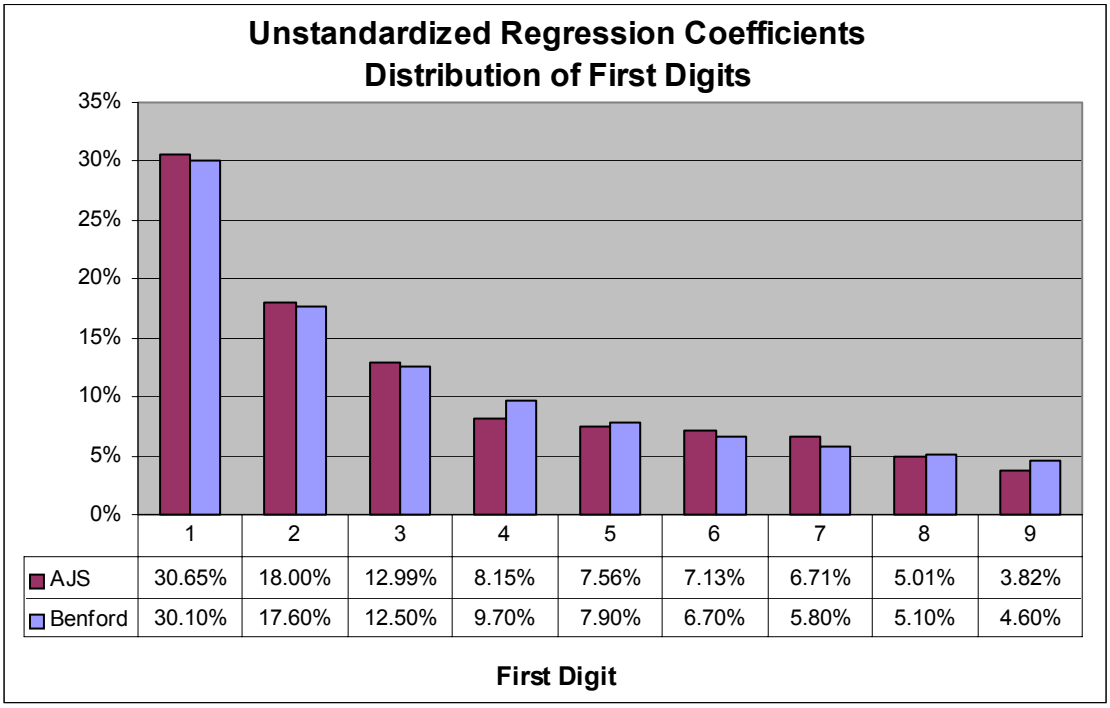
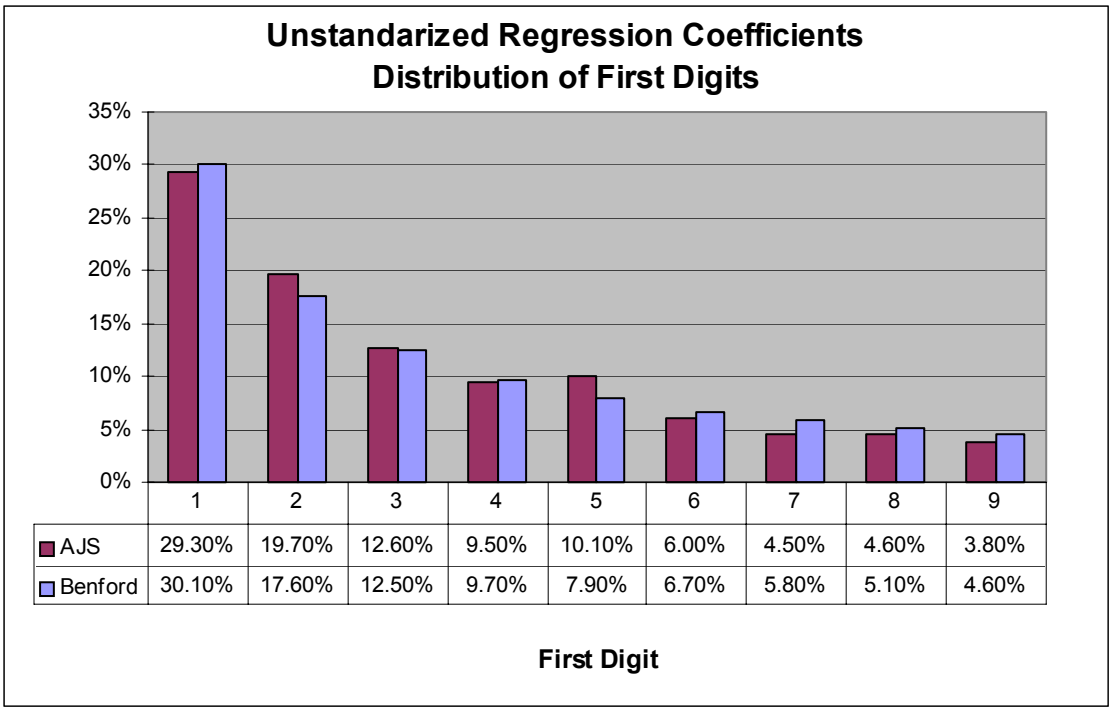


Figure 1: Relative frequencies of first digits of regression coefficients from articles published in the American Journal of Sociology (Sample 1, Volumes 101 and 102).



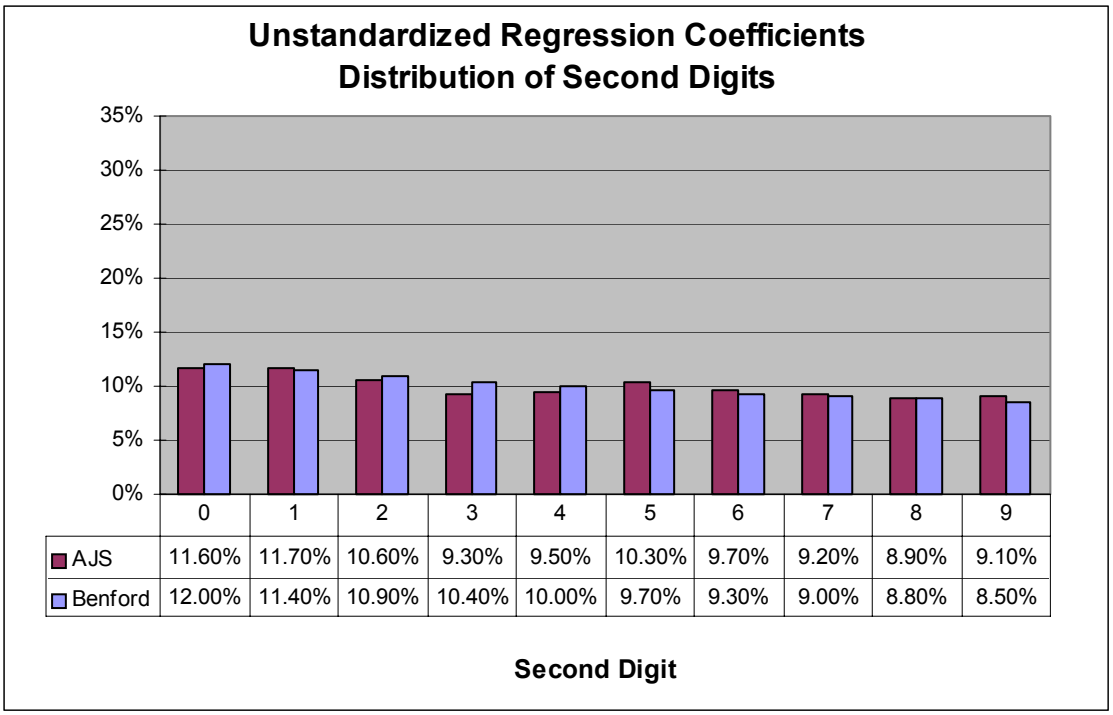
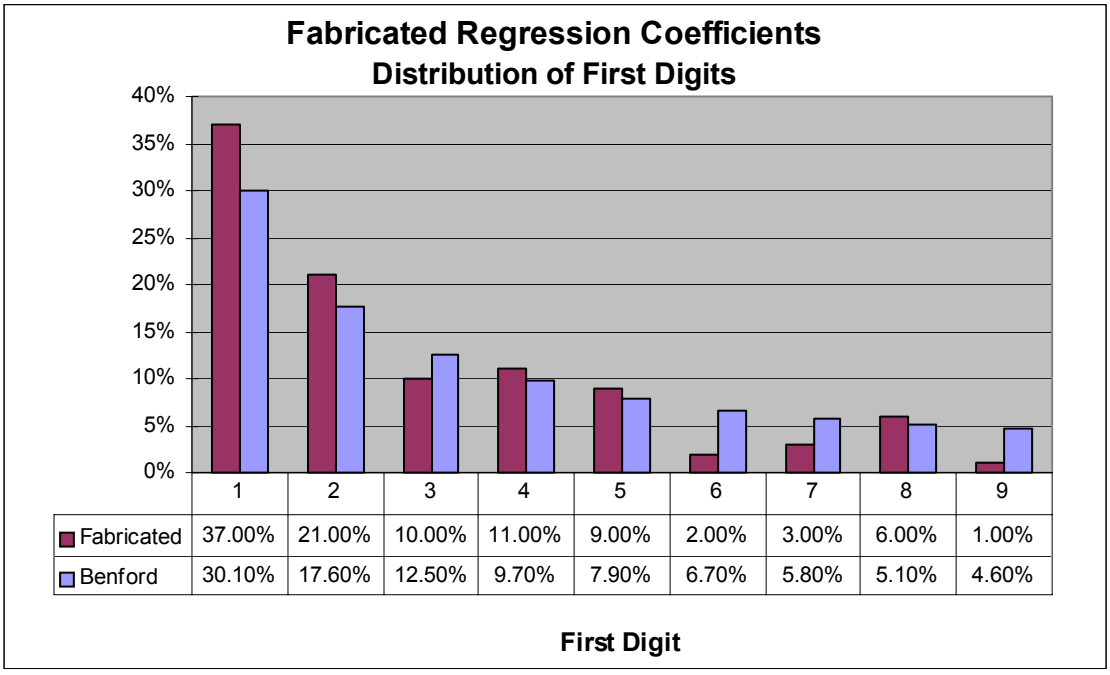


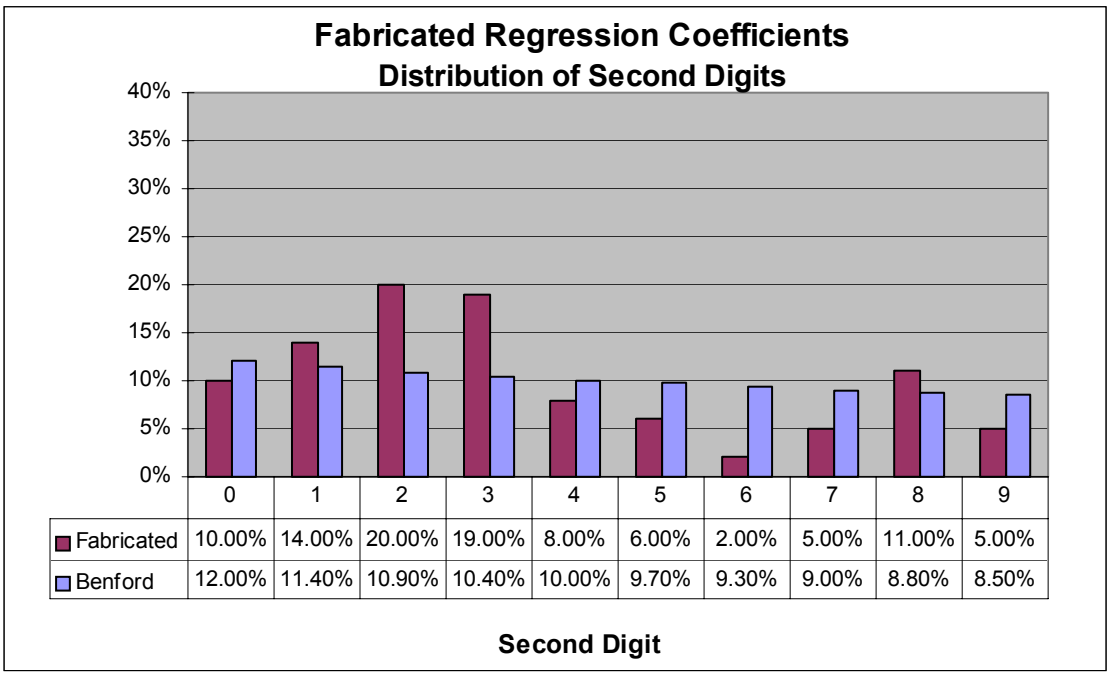
Figure 2: Relative frequencies of first and second digits of regression coefficients from articles published in the American Journal of Sociology (Sample 2, Volumes 104 and 105)

Figure 3: Relative frequencies of first and second digits of fabricated regression coefficients

(a) Experiment 1, n=100

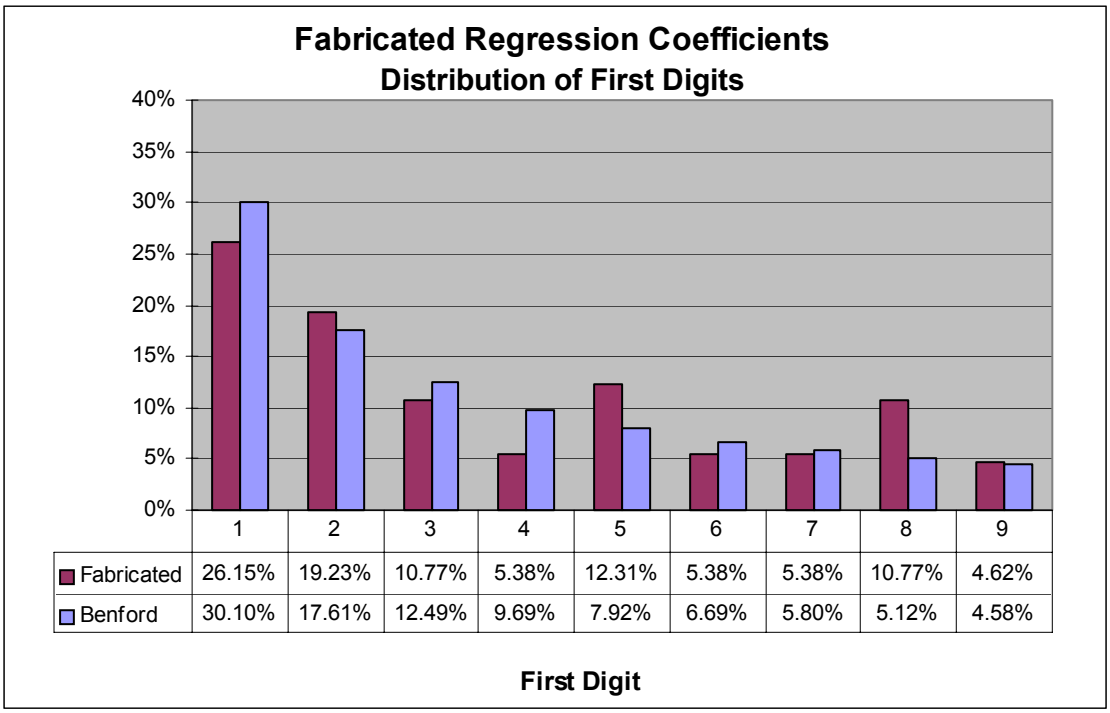


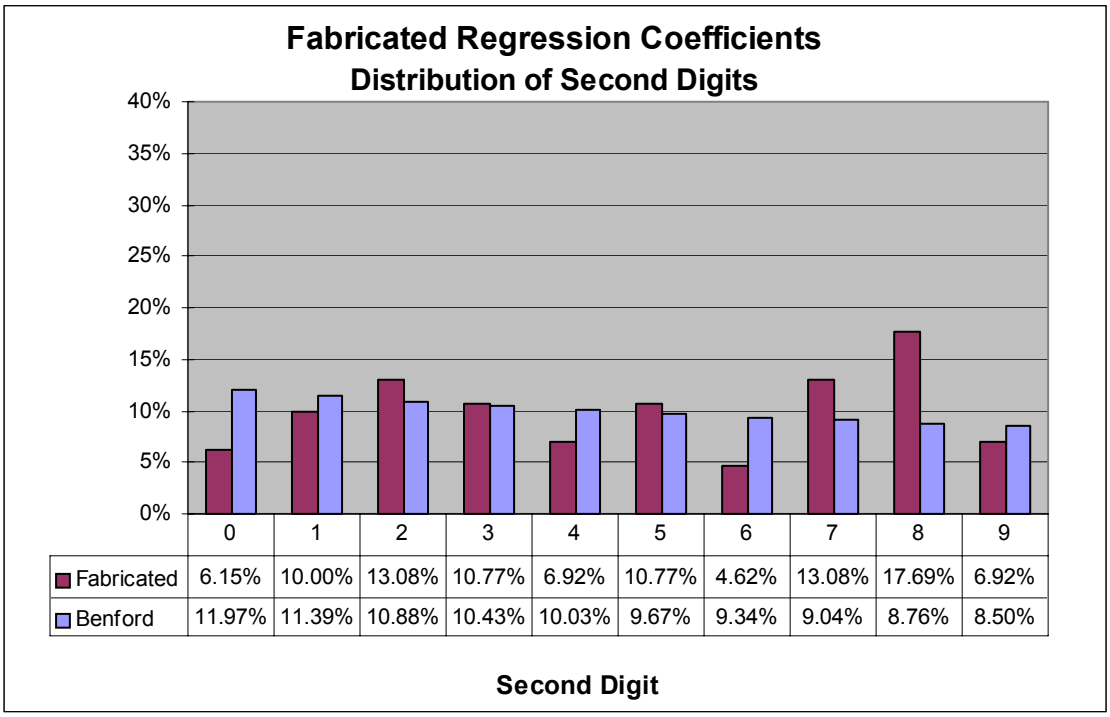
	1	2	3	4	5	6	7	8	9
Fabricated	37.00%	21.00%	10.00%	11.00%	9.00%	2.00%	3.00%	6.00%	1.00%
Benford	30.10%	17.60%	12.50%	9.70%	7.90%	6.70%	5.80%	5.10%	4.60%



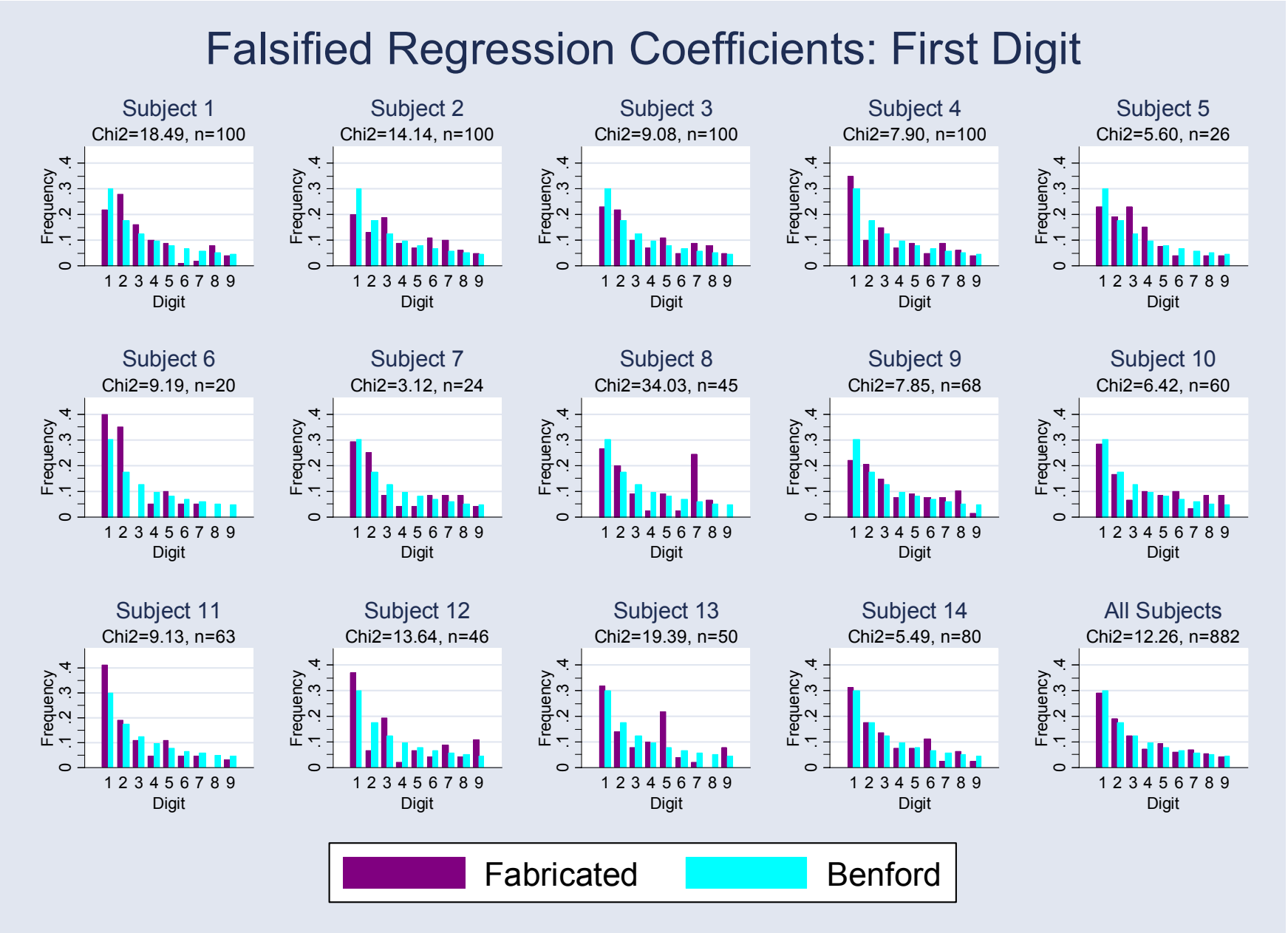
	0	1	2	3	4	5	6	7	8	9
Fabricated	10.00%	14.00%	20.00%	19.00%	8.00%	6.00%	2.00%	5.00%	11.00%	5.00%
Benford	12.00%	11.40%	10.90%	10.40%	10.00%	9.70%	9.30%	9.00%	8.80%	8.50%

(b) Experiment 2, n=130

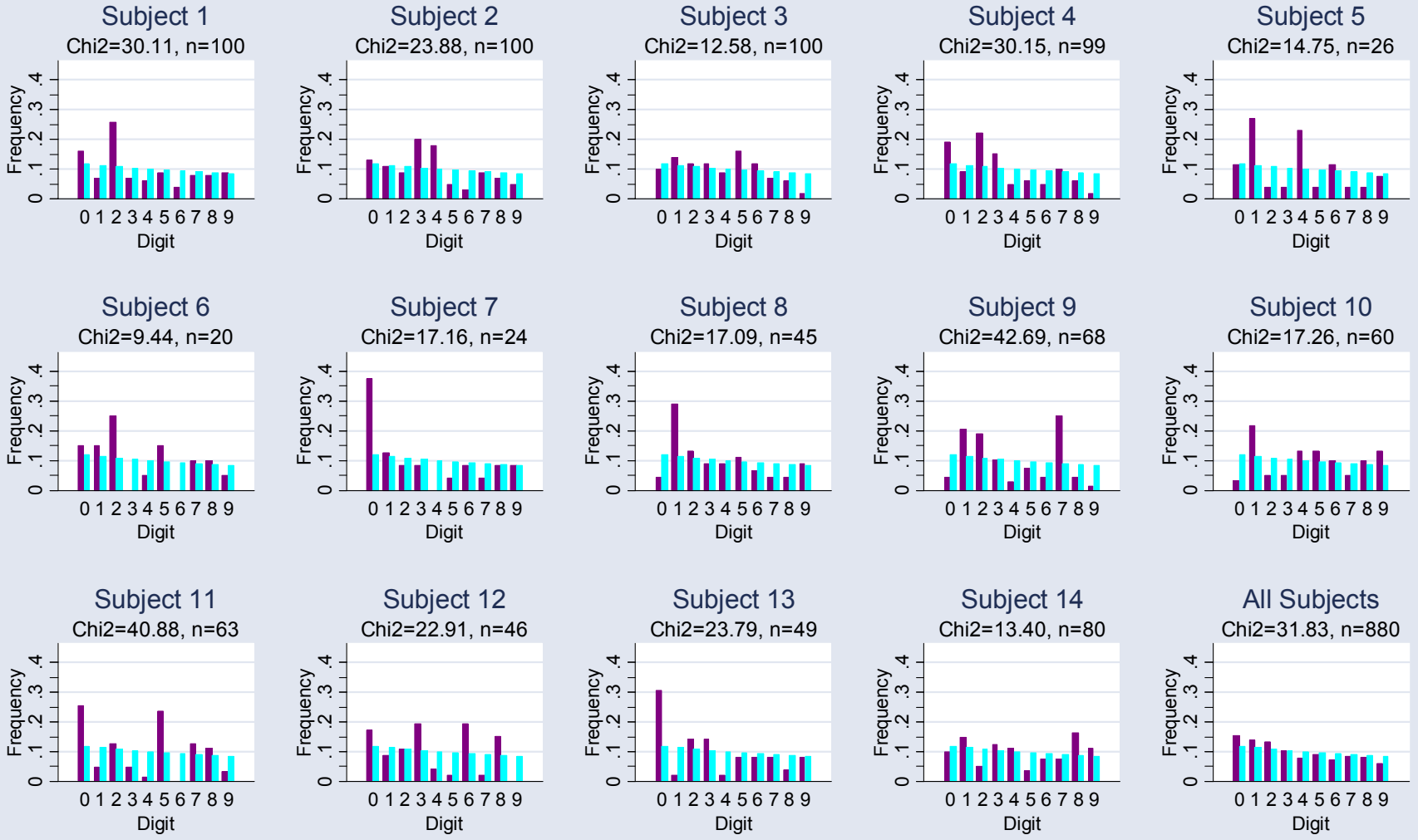




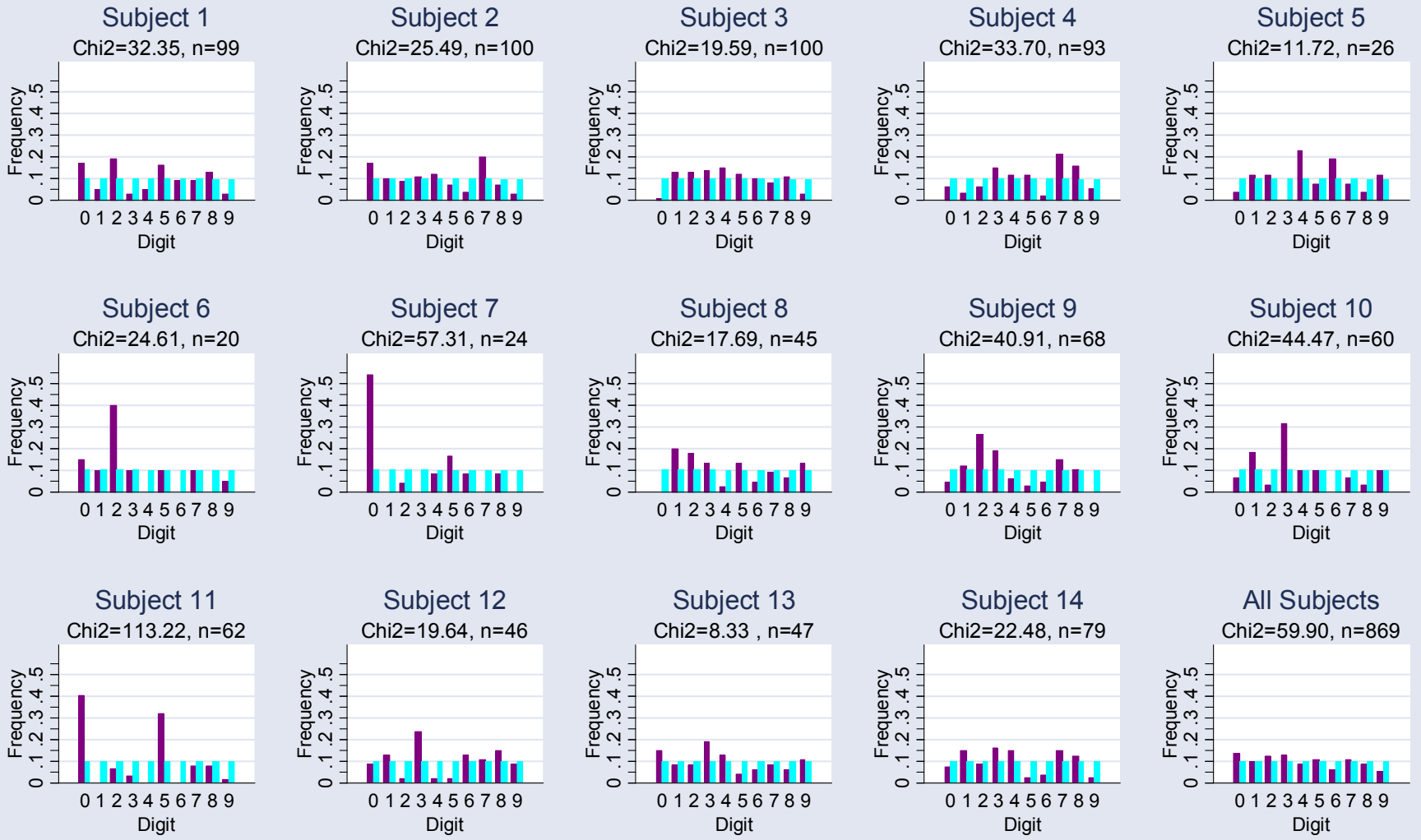
Appendix 1: Relative Frequencies of First and Higher Order Digits for 14 Subjects (Experiment 3)



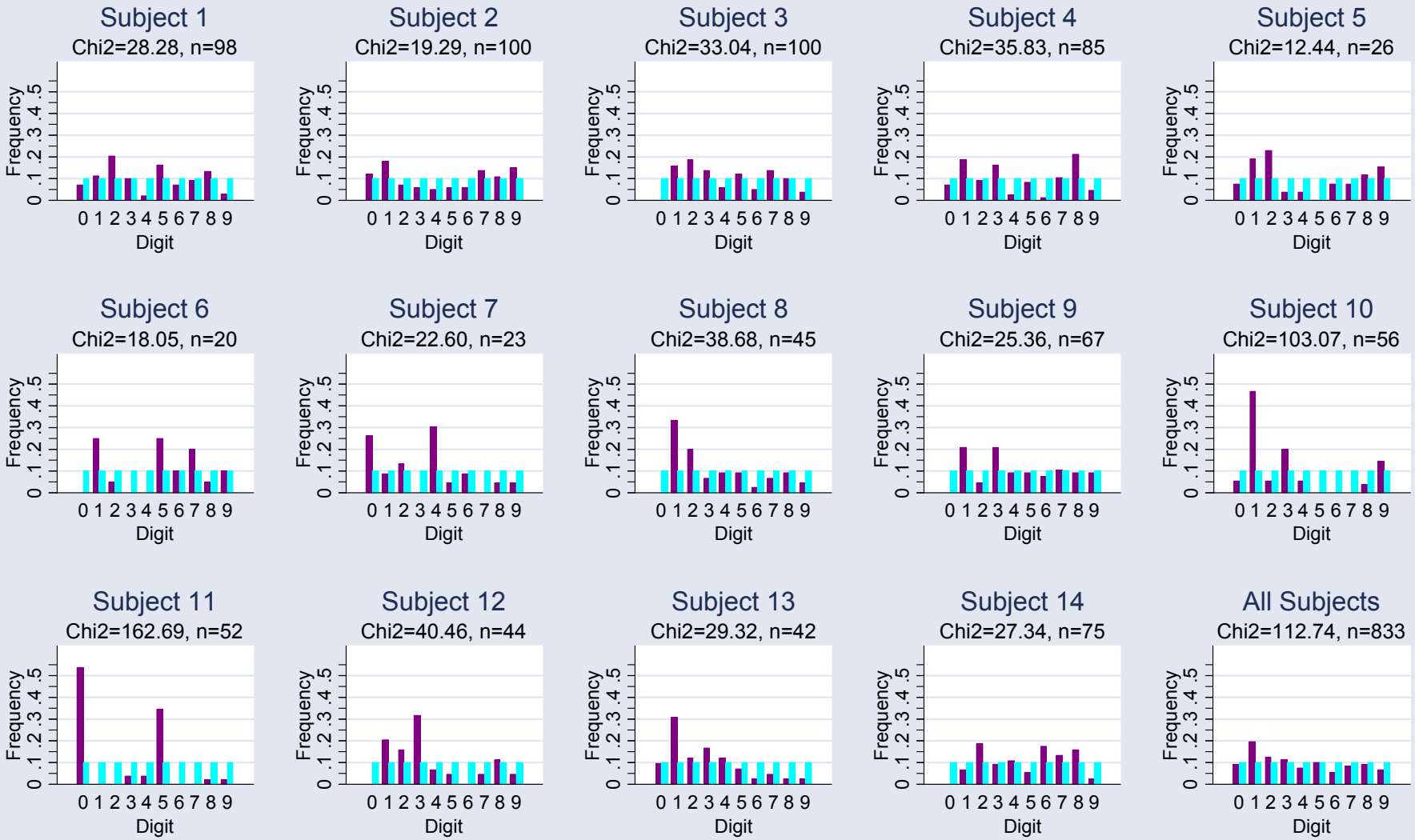
Falsified Regression Coefficients: Second Digit



Falsified Regression Coefficients: Third Digit



Falsified Regression Coefficients: Fourth Digit



Appendix 2: Questionnaire for the Fabrication Experiments 1 and 2*

Name:
Major:
Semester:

Your task is to construct a table of (unstandardized) regression coefficients (for a multiple linear regression) that support the following hypothesis:

"The higher the unemployment benefits, the longer unemployment will last."
The values should be plausible and they should seem to you to have been produced by actual data analysis.

A few more things to consider:

1. Keep in mind that a coefficient can be meaningfully interpreted only for a certain scale. If, for example, unemployment benefits are measured in Swiss francs, then you will have to select different coefficients depending on whether one unit of the unemployment benefits variable is equal to 100 francs or 1,000 francs. You should take the units of all the other variables into account in a similar way. First select a scale (by placing an x next to the option you choose) and then fill in the table with coefficients that you think would produce realistic results.
2. Be sure to put down a standard error as well as a coefficient. As you know, a coefficient with a probability of error of alpha = .05 is significant if the value of the coefficient is more than twice as large as the value of the standard error. **Please denote significant coefficients with an asterix.**
3. As you also know, the regression coefficient for a dichtotomous- 0/1 coded- variable denotes the amount by which the dependent variable changes when the independent variable is equal to 1 versus when it is equal to 0. For example, the coefficient for a variable that takes on the values of 1 for a city and 2 for a town or a rural area might be -3.642. If the length of the unemployment spell is measured in weeks, then the length of the unemployment spell in a city is 3.642 weeks shorter in a city than in a town or a rural area.
4. Be sure to note the coefficients and standard errors to **four digits, not including the zeroes before the first digit.** For example, the numbers 0. 001438 or 91.24 would both fulfill this condition.

*A slightly modified version of this questionnaire was used in experiment 3.

So, let's get started:

First, select a scale for the length of the unemployment spell:

Days: .....
Weeks: .....
Months: .....

Table: Determinants of the length of unemployment: Estimates from a multiple regression (standard errors in parentheses)

Independent Variables	Regression Coefficients (Standard Errors)
Unemployment benefits In units of CHF 1 ..... CHF 100 ..... CHF 1000	..... (.....)
Years of education	..... (.....)
Years of job experience	..... (.....)
Mother's years of education	..... (.....)
Father's years of education	..... (.....)
Sex (Female = 1)	..... (.....)
Marital status (married = 1 , otherwise 0)	..... (.....)
Last position was in the service sector (service sector = 1, otherwise 0)	..... (.....)

Monthly income for the last job held, in units of CHF 1 ..... CHF 100 ..... CHF 1000	..... (.....)
Distance between residence and place of business in units of: 1 km ..... 10 km	..... (.....)
Adjusted multiple R-squared Number of cases (N)	....