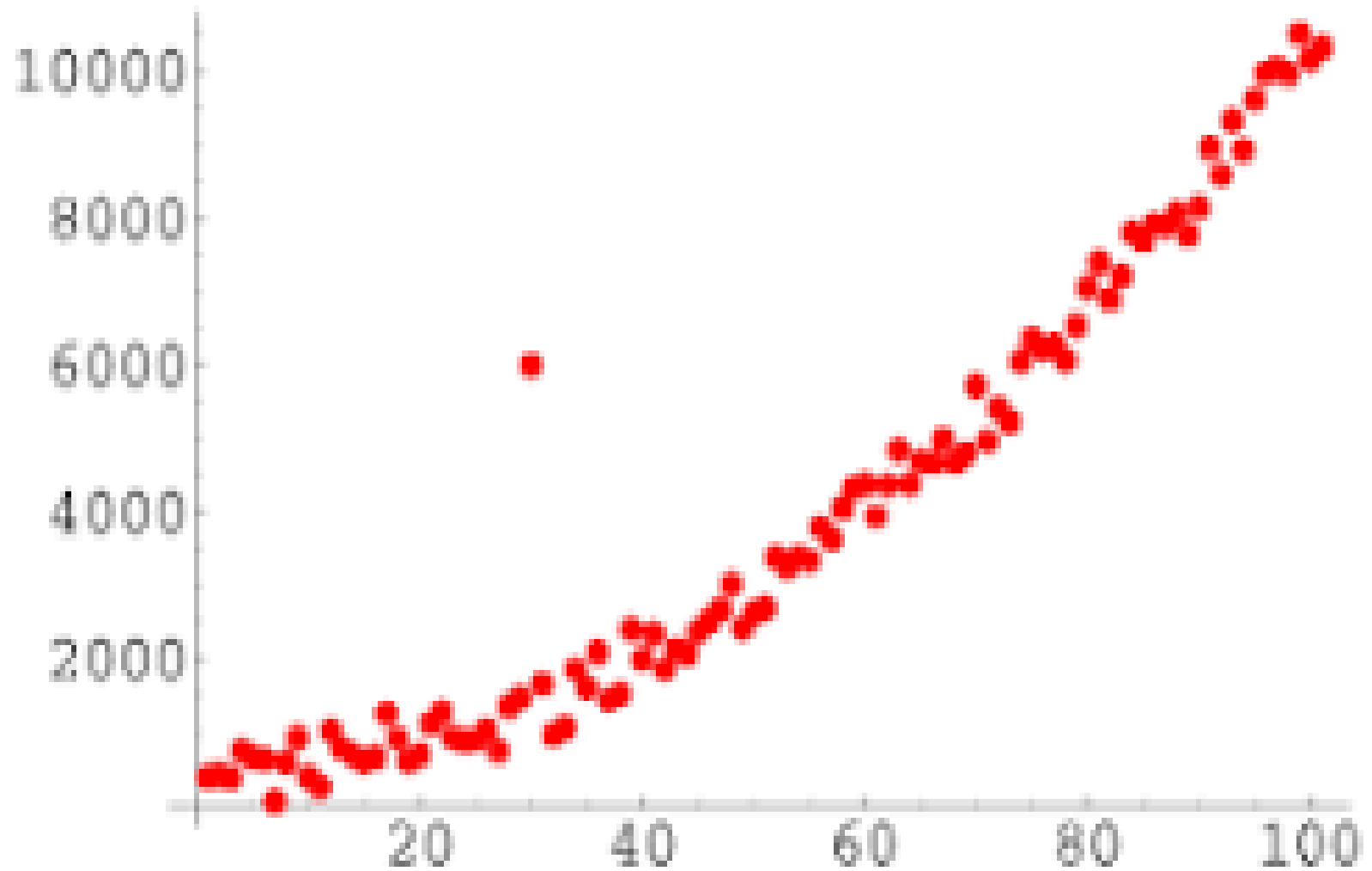


Outliers

Andrew McCarren

What's going on here?



Suspicious Data

- ▶ Consider the data points
3, 4, 7, 4, 8, 3, 9, 5, 7, 6, 92
- ▶ “92” is suspicious – an *outlier*
- ▶ Outliers are potentially legitimate
- ▶ Often, they are data or model glitches
- ▶ Or, they could be a data miner’s dream, e.g. highly profitable customers

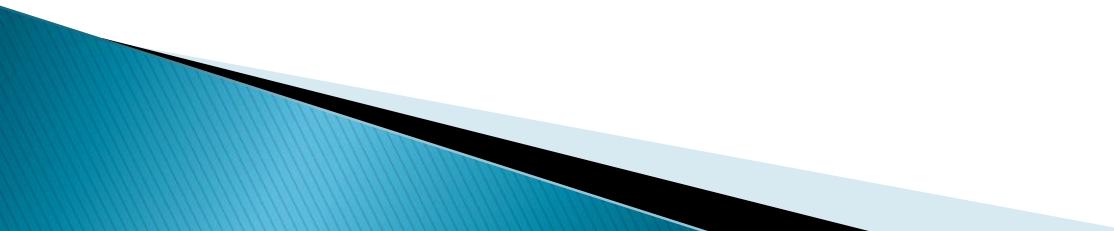
Outliers

- ▶ Outlier – “departure from the expected”
- ▶ Types of outliers – defining “expected”
- ▶ Many approaches
 - Error bounds, tolerance limits – control charts
 - Model based – regression depth, analysis of residuals
 - Geometric
 - Distributional
 - Time Series outliers

Distributional Outliers

- ▶ For each point, compute the maximum distance to its k nearest neighbors.
 - *DB(p, D)-outlier* : at least fraction p of the points in the database lie at distance greater than D .
- ▶ *Local Outliers* : adjust definition of outlier based on density of nearest data clusters.

Set Comparison and Outlier Detection

- ▶ “Model” consists of partition based summaries
 - ▶ Perform nonparametric statistical tests for a rapid section-wise comparison of two or more massive data sets
 - ▶ If there exists a baseline “good” data set, this technique can detect potentially corrupt sections in the test data set
- 

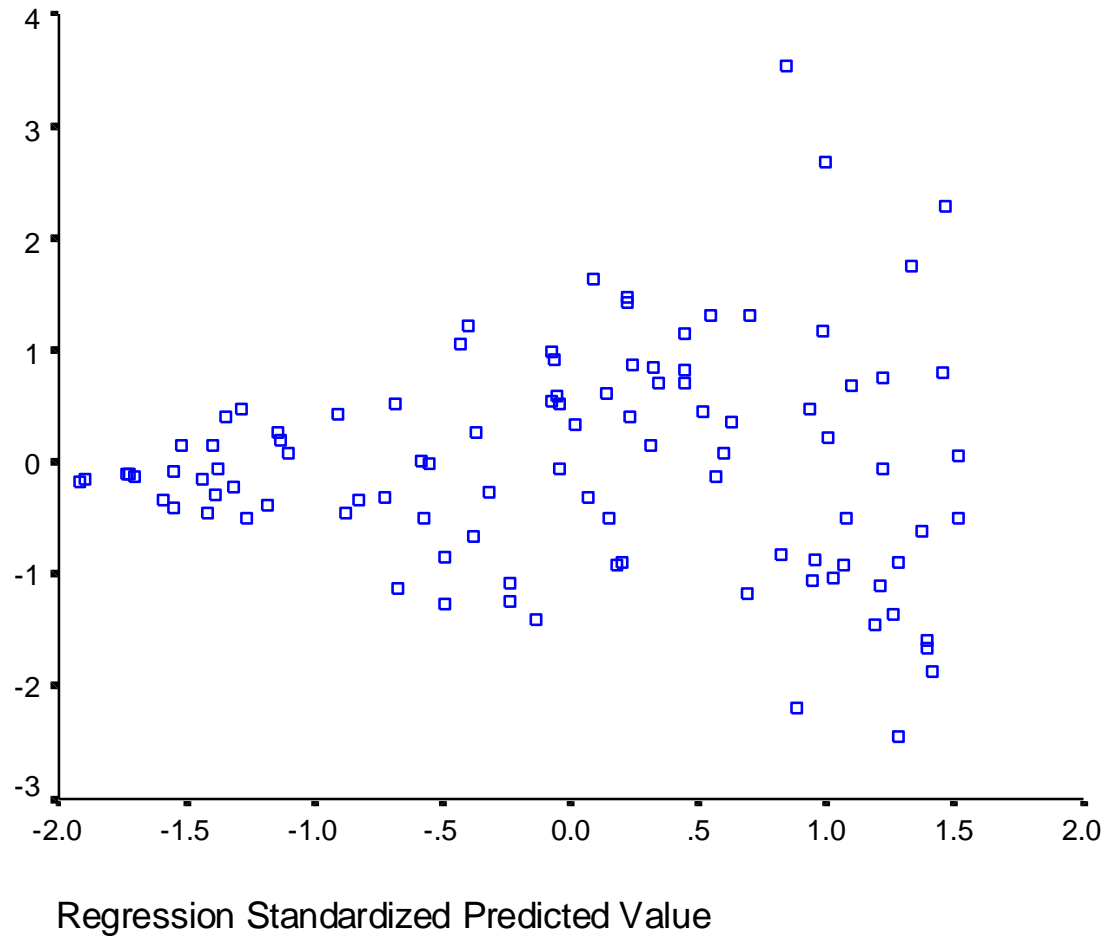
Goodness of Fit – 1

- ▶ Chi-square test
 - Are the attributes independent?
 - Does the observed (discrete) distribution match the assumed distribution?
- ▶ Tests for Normality (Sharpiro Wilks)
- ▶ Q-Q plots (visual)
- ▶ Kolmogorov-Smirnov test (used to compare 2 samples)
- ▶ Kullback-Liebler divergence (information gain)

Goodness of Fit – 2

- ▶ Analysis of residuals
 - Departure of individual points from model
 - Patterns in residuals reveal inadequacies of model or violations of assumptions
 - Reveals bias (data are non-linear) and peculiarities in data (variance of one attribute is a function of other attributes)
 - Residual plots
 - Regression depth is basically measures impact of outliers on the model (see <https://www.jstor.org/stable/pdf/2670155.pdf?refreqid=excelsior%3A82bf94c45b1e373345937b035a2f7266>)

Detecting heteroscedasticity



Time Series Outliers

- ▶ Data is a time series of measurements of a large collection of entities (e.g. customer usage).
- ▶ Vector of measurements define a trajectory for an entity.
- ▶ A trajectory can be glitched, and it can make radical but valid changes.
- ▶ Approach: develop models based on entity's past behavior (*within*) and all entity behavior (*relative*).
- ▶ Find potential glitches:
 - Common glitch trajectories
 - Deviations from within and relative behavior.

Benford's Law

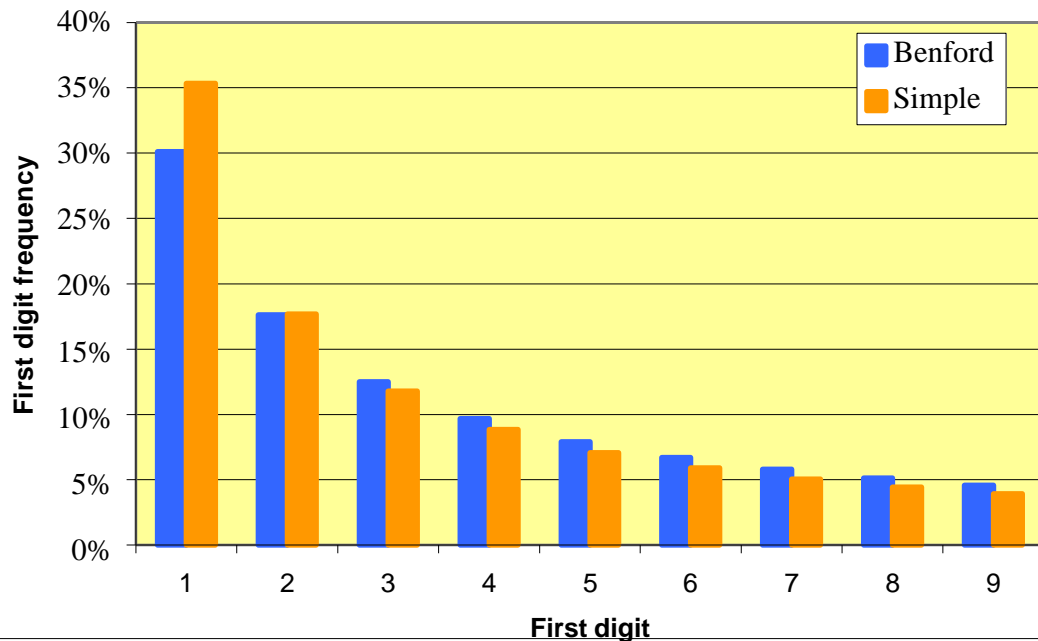
The Law of First Digits

- ▶ Examines the leading digit in real life sets of data.
- ▶ A set of numbers is said to follow Benford's law if the leading digit
 - $d \in \{0, 1 \dots, 9\}$
 - $P(d) = \log_{10}(d+1) - \log_{10}(d)$
- ▶ Examples
 - Population figures
 - Social security volumes

... a simple explanation...

The general principle is that there are more smaller observations vs larger ones. There are probably nearly twice as many 1s as there are 2s and three times as many 1s as there are 3s, etc... Using such a principle throughout gives us a frequency that is close to Benford's Law.

First Digit frequency
Benford's Law vs Simple rule



Digit	Benford $\log(1+1/d)$	Simple rule	Simple rule proportion $1/d$
1	30.1%	35.3%	1.00
2	17.6%	17.7%	0.50
3	12.5%	11.8%	0.33
4	9.7%	8.8%	0.25
5	7.9%	7.1%	0.20
6	6.7%	5.9%	0.17
7	5.8%	5.0%	0.14
8	5.1%	4.4%	0.13
9	4.6%	3.9%	0.11
			2.83

We would need a sample $> 1,000$ to reach statistical significance at the 0.05 level that those two distributions are different.

A few Benford's Law applications...

- Checking tax returns for fraud;
- Uncovering accounting fraud;
- Detecting false insurance claims.