Softtek®

TSP Symposium

Using Benford's Law to Monitor Process Fidelity

Data Veracity

- **Process measurements are critical in any process improvement initiative.**

- **Gathered data must be:**
  - **Clearly defined,**
  - **Precise,**
  - **Invariant,**
  - **Reliable.**

- **Benfords's Law helps to detect possible frauds or invalid data.**

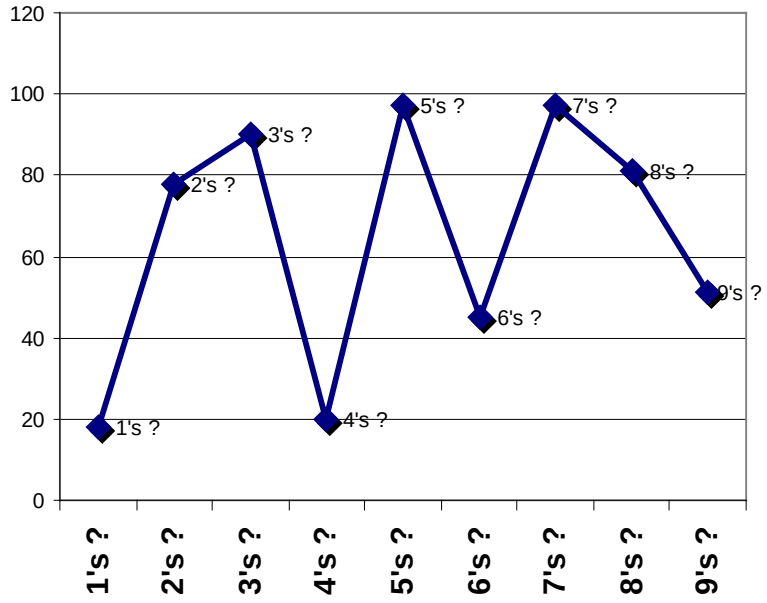
Softtek®

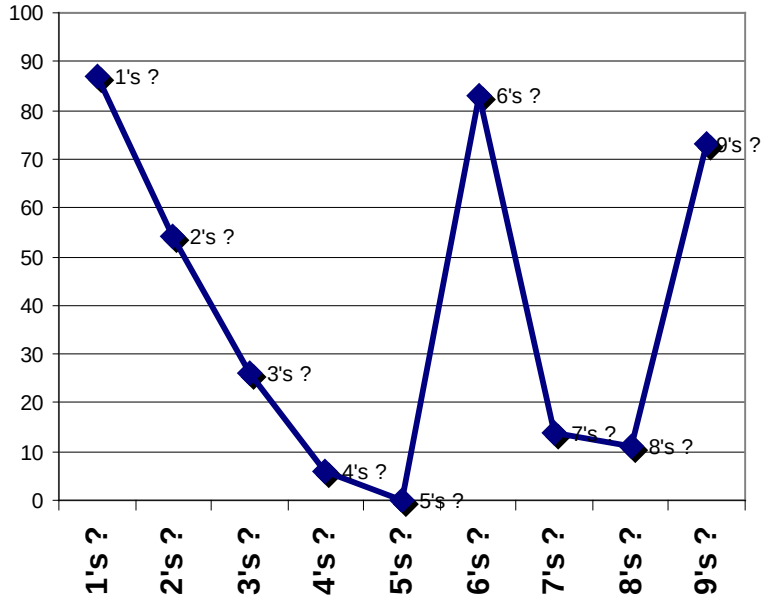
How is first digit of each value in these data sets distributed?

- For different sets of data:
  - “Votes in the Mexican Presidential Election”.
  - “Economic Activity in the Different Mexican States”.
  - Minutes in PSP Time Logs.
  - IMDB Top 250 movies user votes.

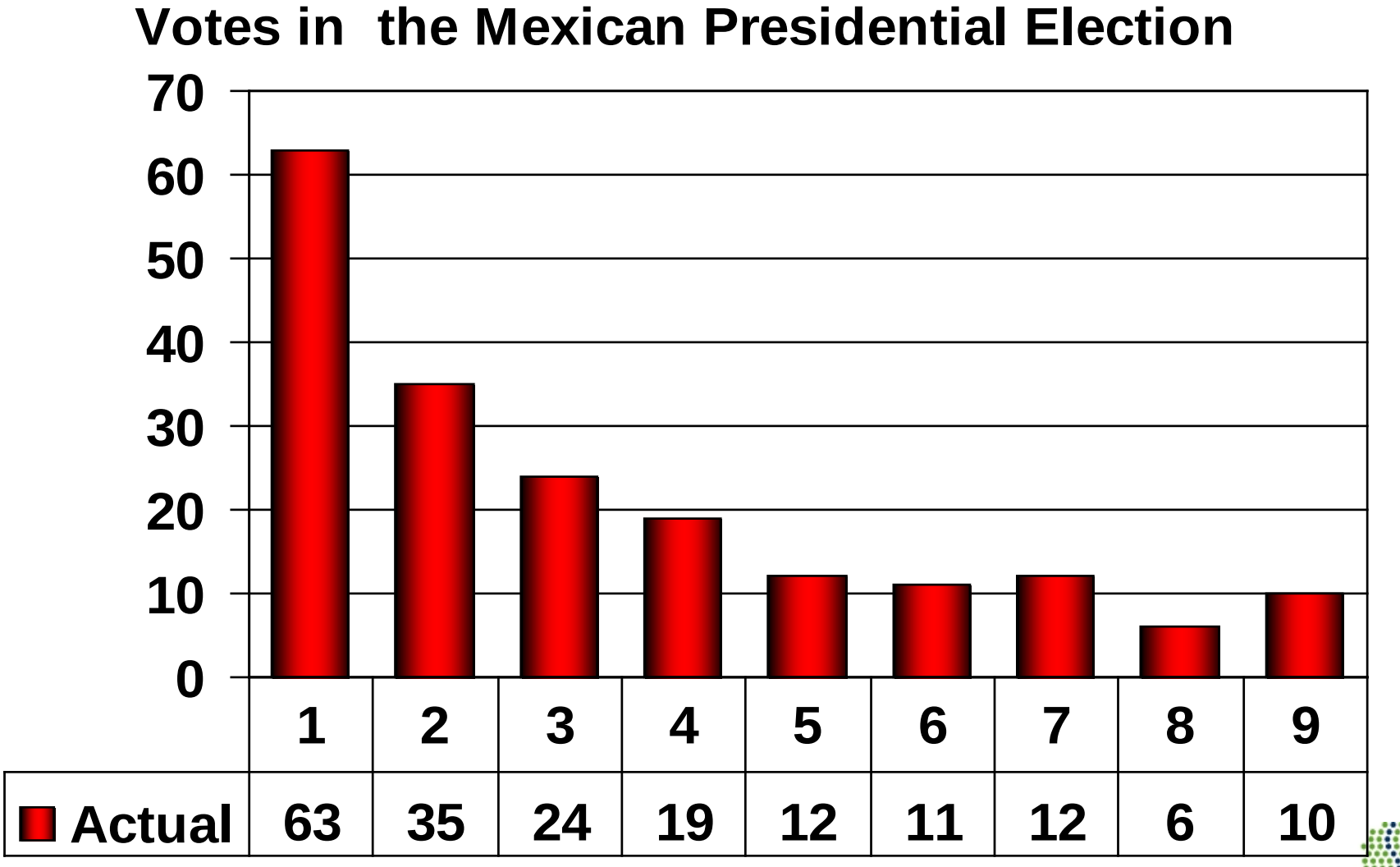
Softtek®

What distribution did you expect for the data you analyzed?



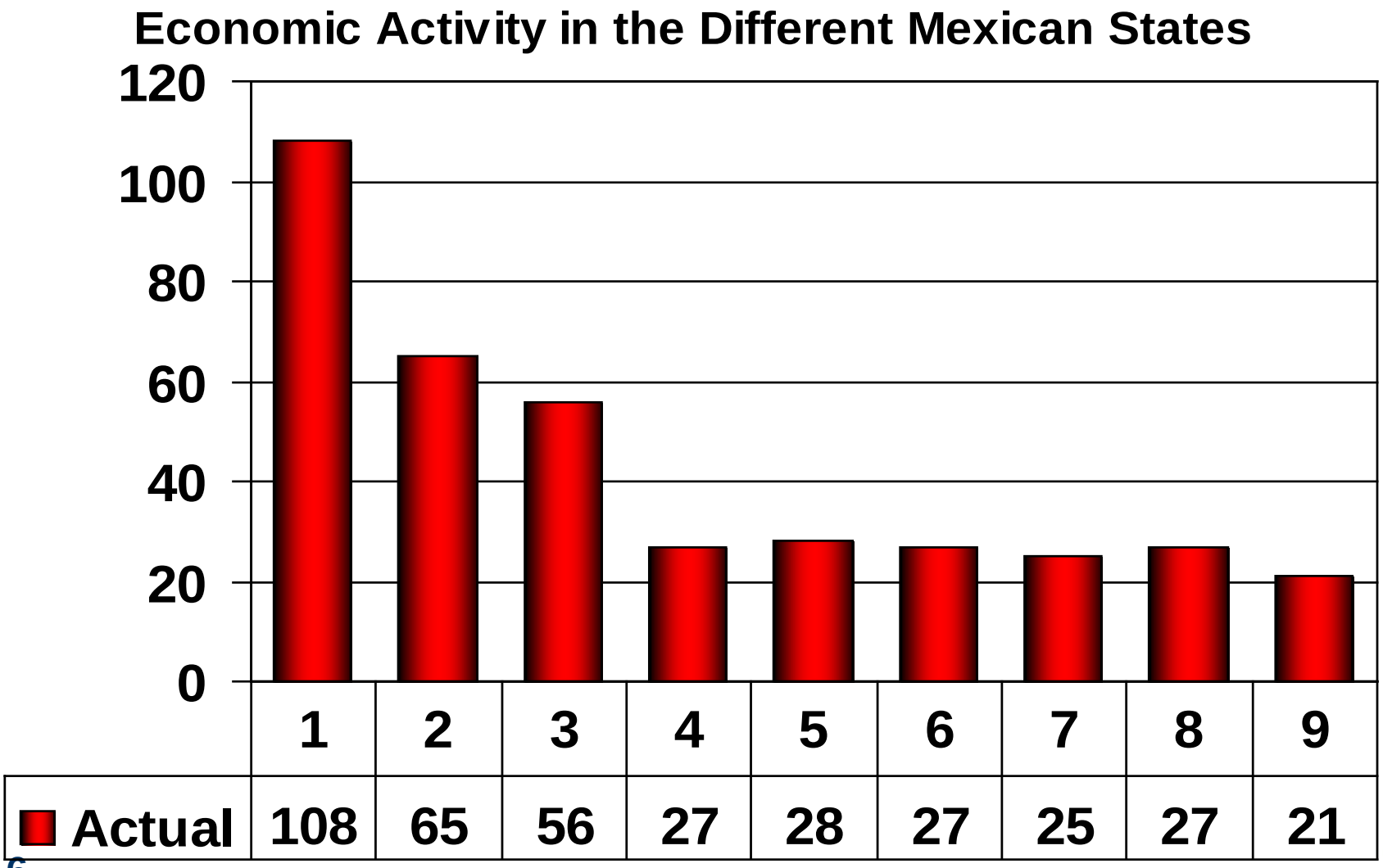


Actual Results - 1



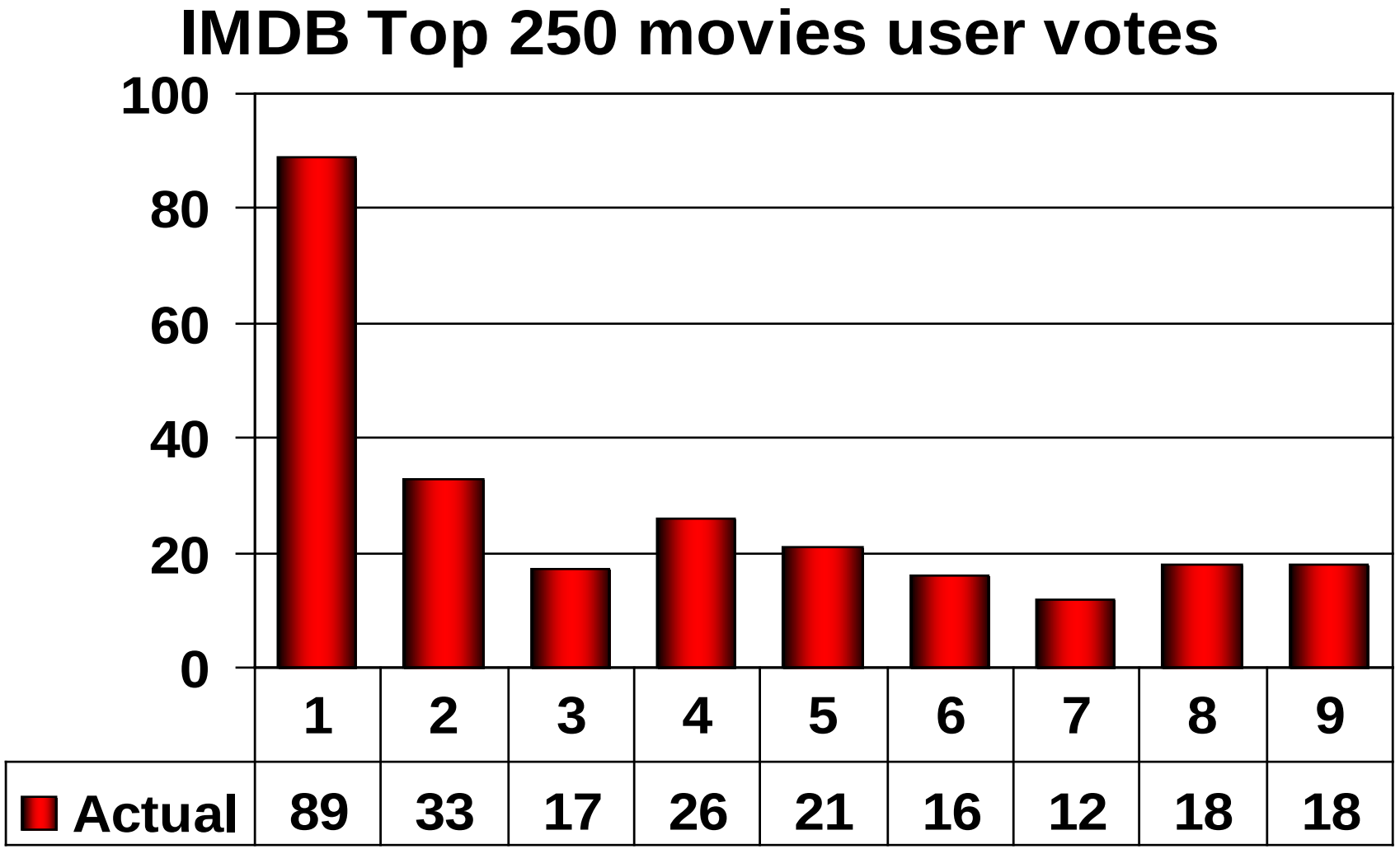
Softtek®

Actual Results - 3



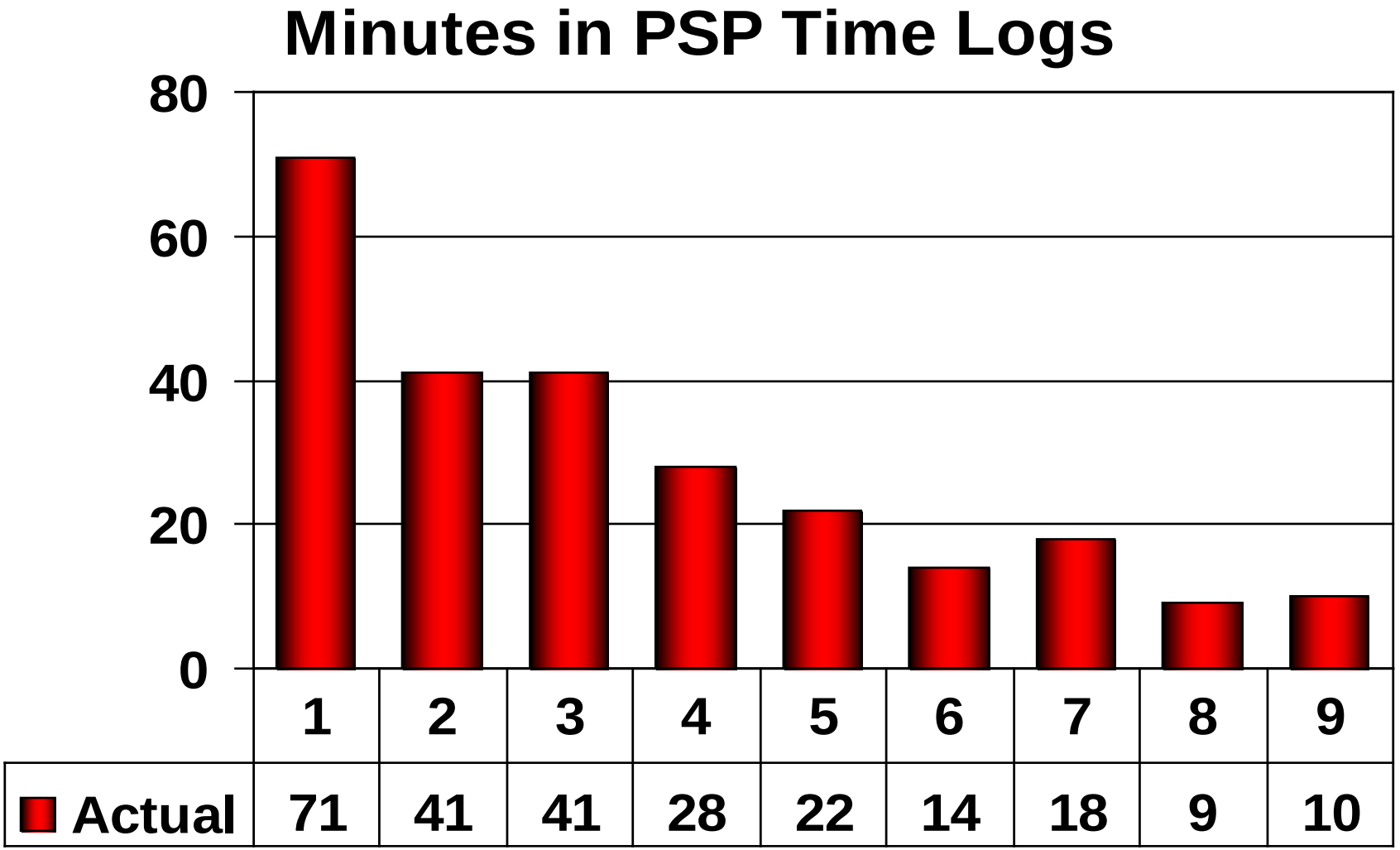
Softtek®

Actual Results - 4



Softtek®

Actual Results - 4



Softtek®

Common Points

- What do these results have in common?
- Is there a common distribution?
- How would you state a rule describing first digit distribution?

Softtek®

Benford's Law

The leading digit *d* (*d* ∈ {1, 2, ..., 9}) occurs with probability p

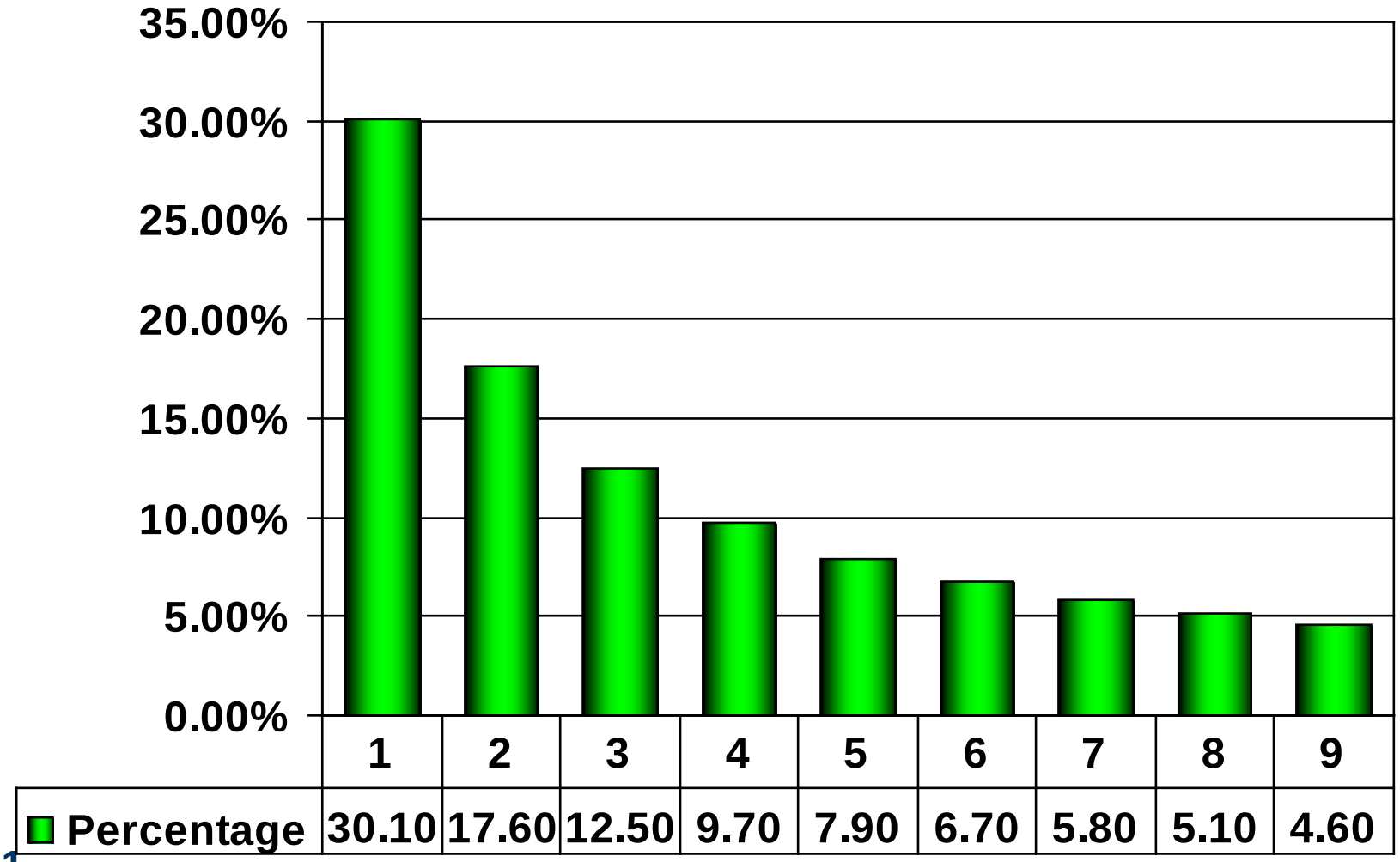
$$P(d) = \frac{\ln\left(1 + \frac{1}{d}\right)}{\ln(10)}$$

- **The most frequent leading number is 1, then 2 and so on.**

- **Data not following this distribution is probably invalid.**

Softtek®

1st digit distribution



	1	2	3	4	5	6	7	8	9
Percentage	30.10	17.60	12.50	9.70	7.90	6.70	5.80	5.10	4.60

Softtek®

Benford's Law applications

- **Benford's Law apply for data with the following characteristics:**
  - **Without natural limits.**
    - **There is no natural or imposed upper or lower limit to data.**
  - **Invariant.**
    - **Data does not depend on a measuring system.**
  - **Not assignable.**
    - **Values are not artificially generated.**

Softtek®

Uses of Benford's Law

- This counter-intuitive result applies to a wide variety of figures, including:
  - electricity bills,
  - street addresses,
  - stock prices,
  - population numbers,
  - death rates,
  - lengths of rivers,
  - physical and mathematical constants,
  - and processes described by power laws (which are very common in nature).

Softtek®

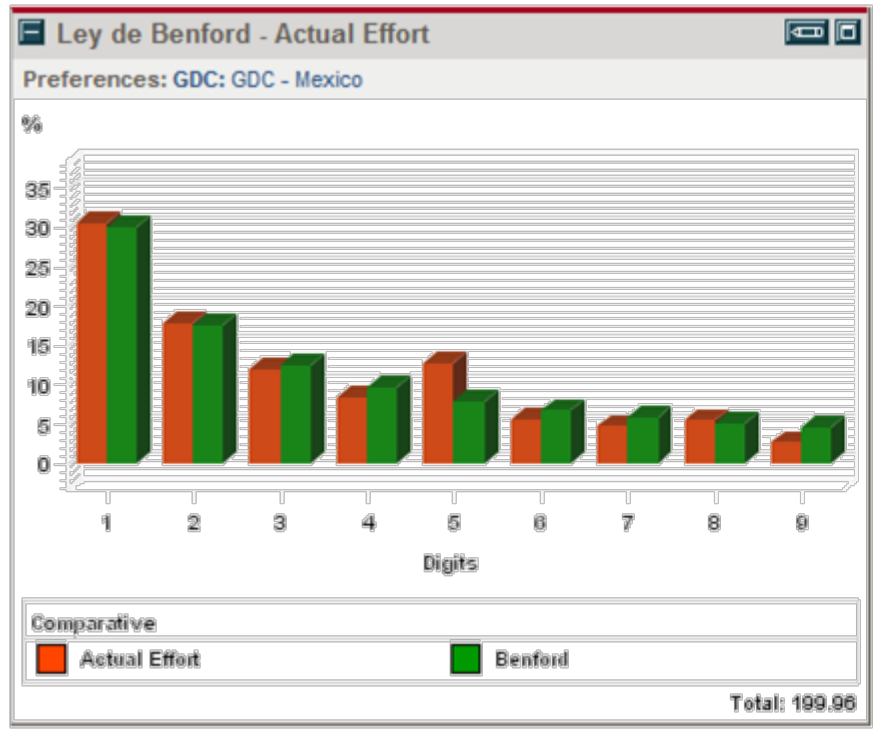
Actual Use in Softtek

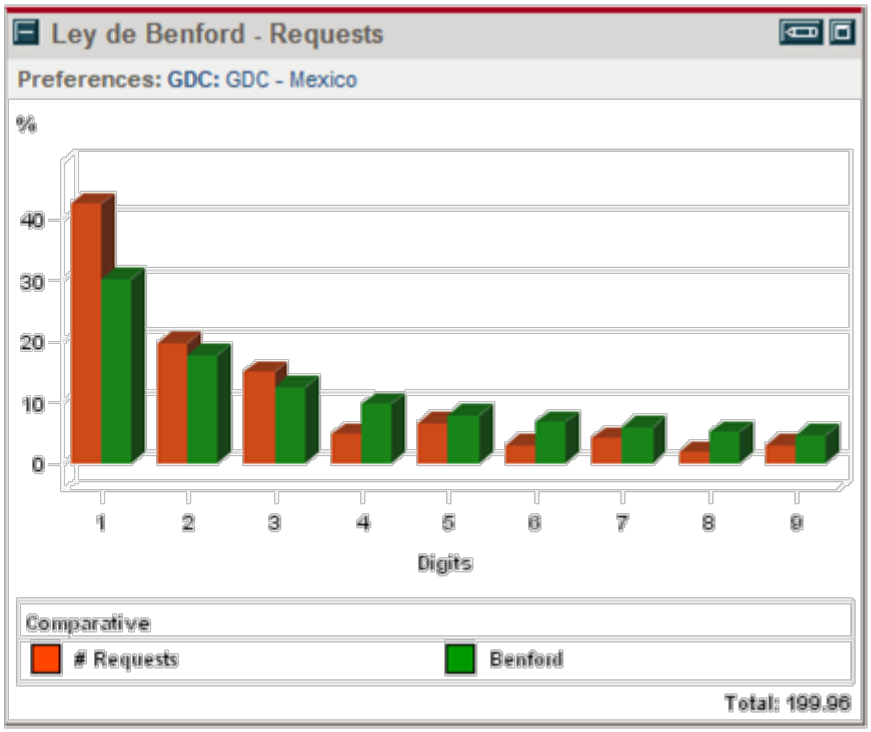
- **ITG Portlets:**
  - **To validate ‘Actual effort’ applied to requests.**
  - **To validate ‘Number of Requirements’.**

- **Management tracking:**
  - **PL’s, OL’s and GDC Managers must track periodically Benford’s Law charts in ITG.**

Softtek®

Actual ITG portlets





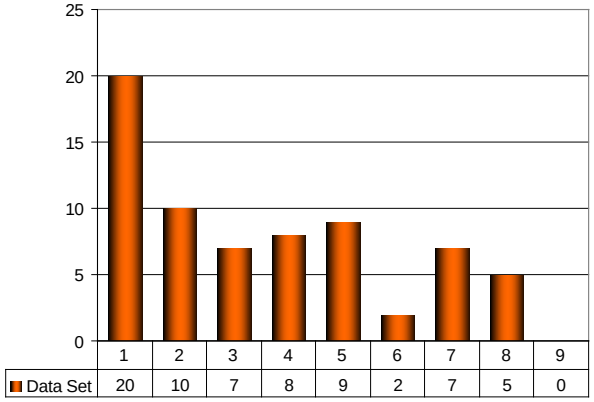
- **Benford's Law is a competitive advantage:**

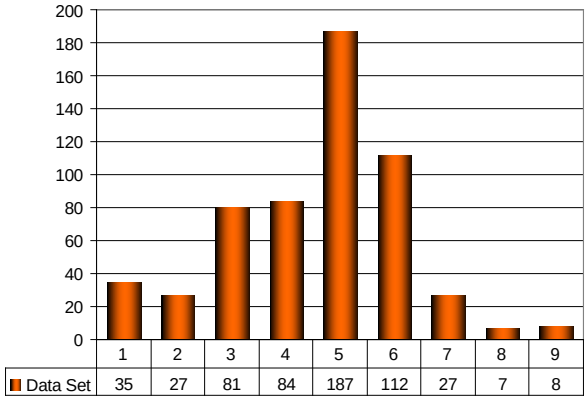
  Understand better with less effort

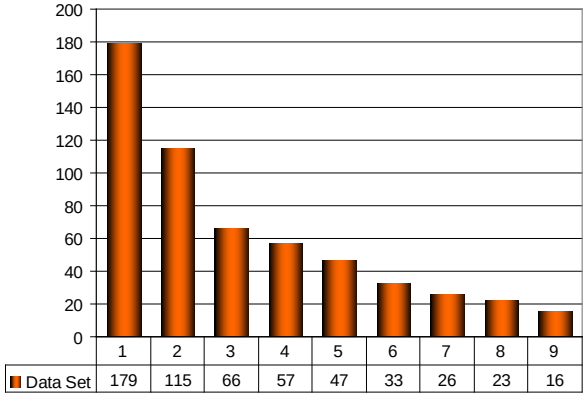
Softtek®

Actual Softtek's projects data

- **Data provided in your handouts was extracted from actual data entered on ITG for actual Softtek's projects.**





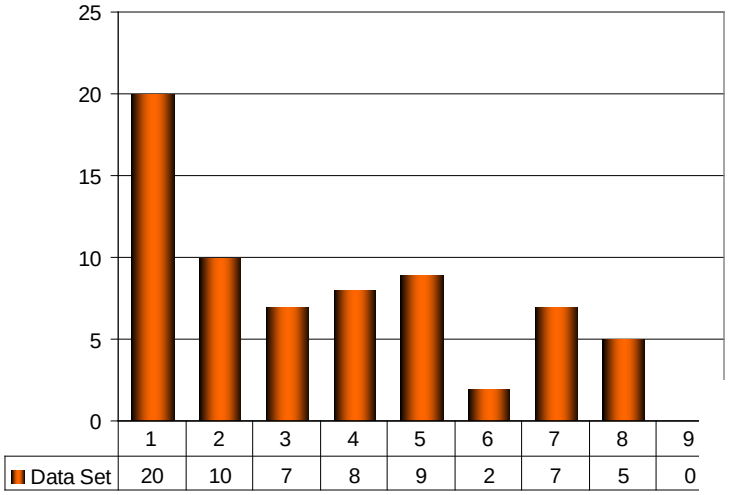


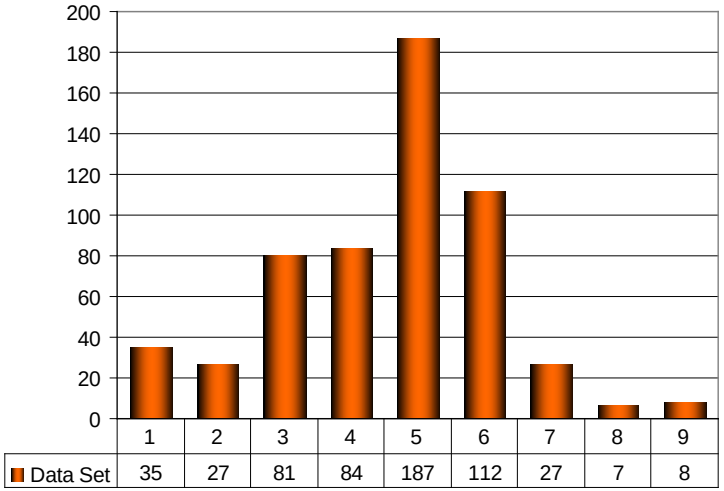
Data Analysis

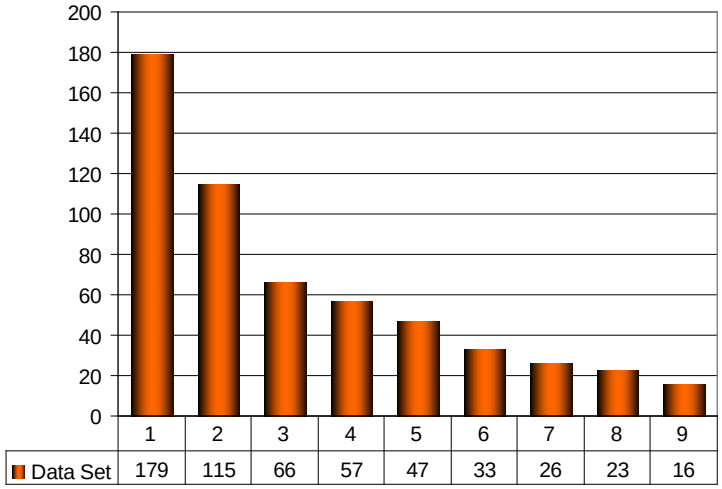
- In case you were in charge of these projects:
  - Would you trust the provided data?
  - Which charts are reliable and which are not?
  - What would you do to improve data reliability?

Softtek®

Reliability Discussion.







Softtek®

Reliability Rules

- **Rules of Thumb to be used to make quick decisions on sample data:**
  - **Absolute Difference:**
    - **Consider data is a defect if the actual frequency is greater than expected frequency plus 10% of total digit data count.**
  - **Relative Difference:**
    - **Consider data is a defect if actual percentage is greater than expected frequency plus 35%.**

Softtek®

Data Set 1 Analysis

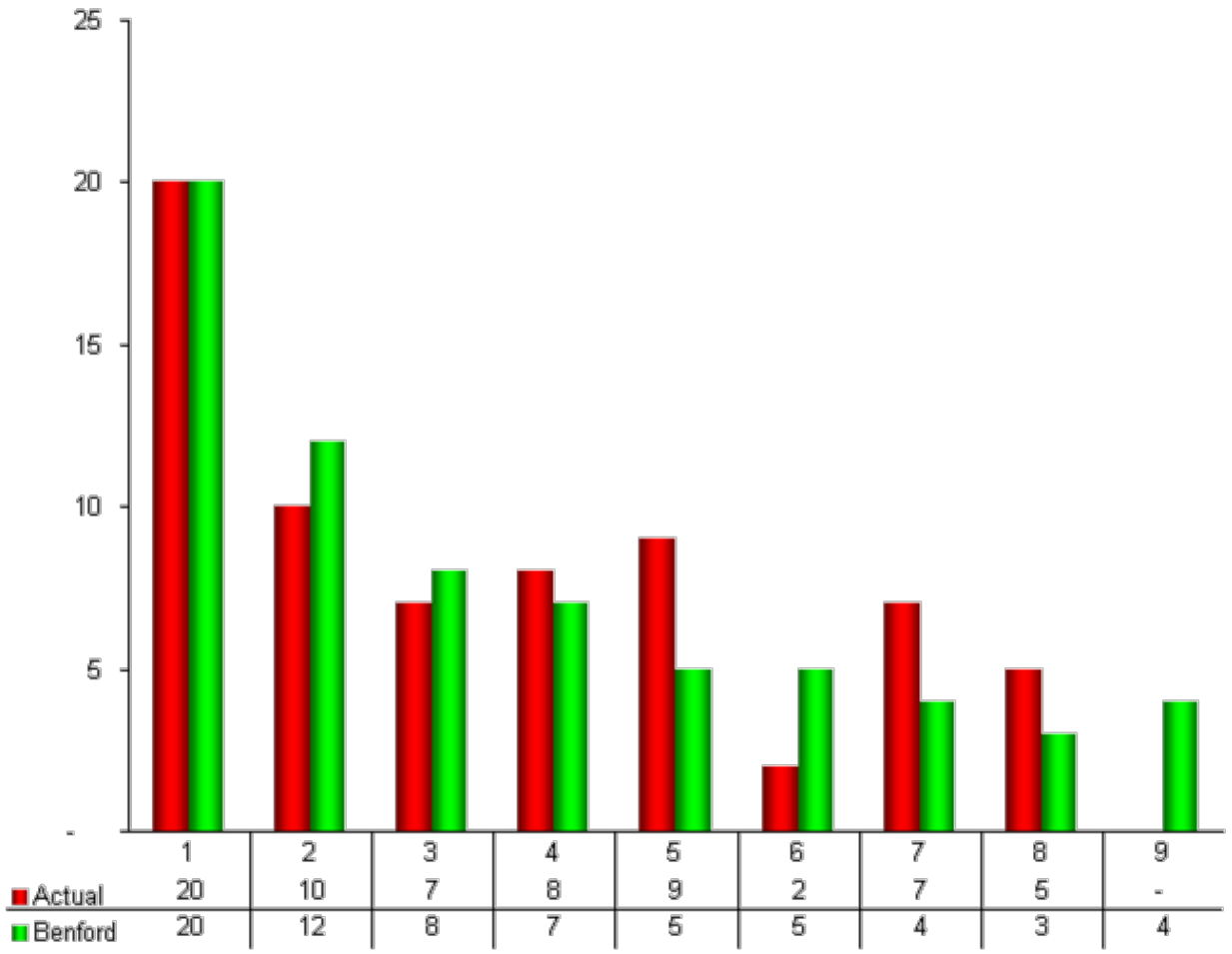
Digit	Actual	Benford	Chi2	10% Abs	35% rel
1	20	20	-	TRUE	TRUE
2	10	12	0.3333	TRUE	TRUE
3	7	8	0.1250	TRUE	TRUE
4	8	7	0.1429	TRUE	TRUE
5	9	5	3.2000	TRUE	FALSE
6	2	5	1.8000	TRUE	TRUE
7	7	4	2.2500	TRUE	FALSE
8	5	3	1.3333	TRUE	FALSE
9	-	4	4.0000	TRUE	TRUE
	68	68	13.18		

Data Volume

Not enough data

p-value	0.106

< 0.01	Strong evidence against
< 0.10	Moderate evidence against
>= 0.10	Little or no evidence against



Softtek®

Data Set 2 Analysis

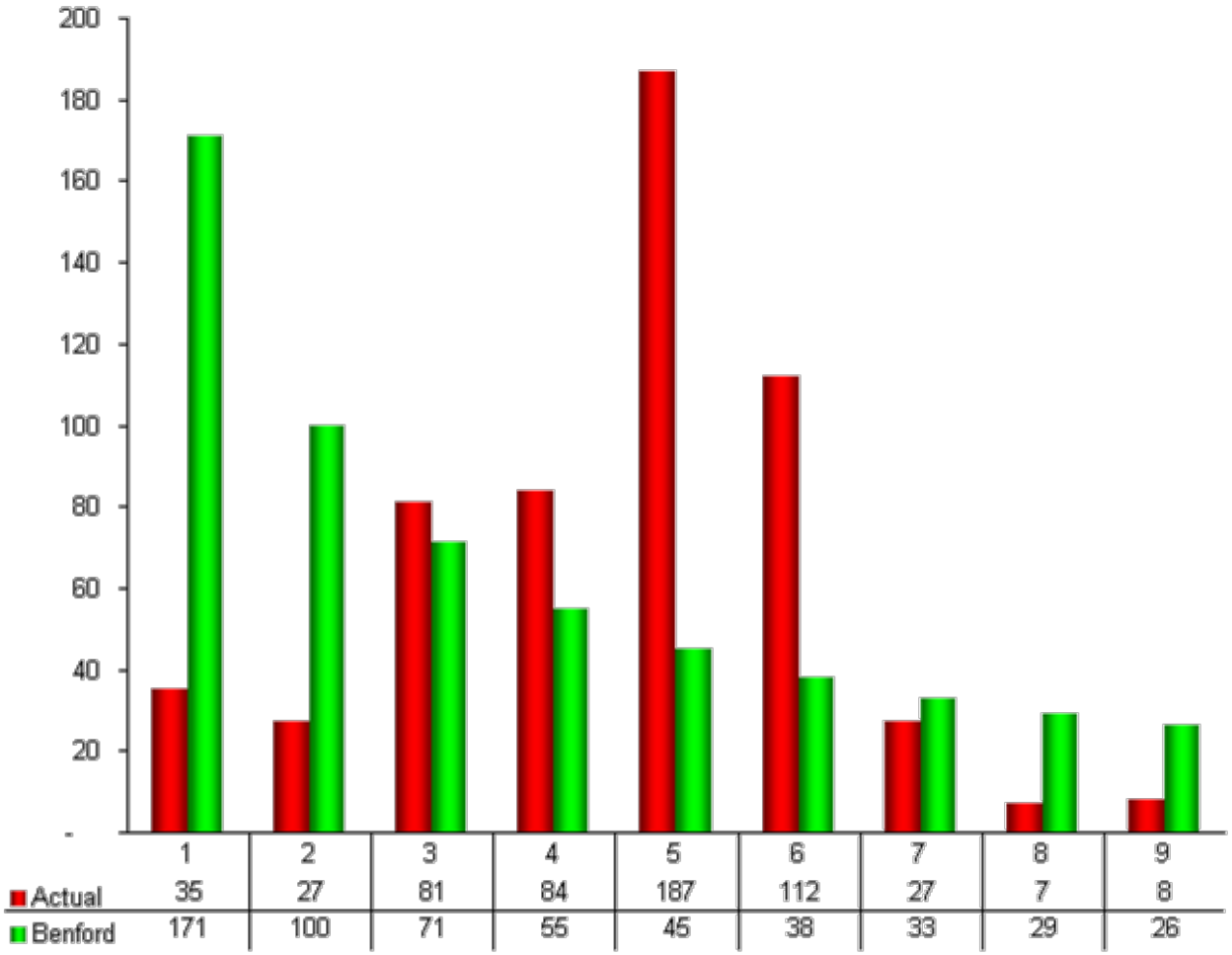
Digit	Actual	Benford	Chi2	10% Abs	35% rel
1	35	171	108.1637	TRUE	TRUE
2	27	100	53.2900	TRUE	TRUE
3	81	71	1.4085	TRUE	TRUE
4	84	55	15.2909	TRUE	FALSE
5	187	45	448.0889	FALSE	FALSE
6	112	38	144.1053	FALSE	FALSE
7	27	33	1.0909	TRUE	TRUE
8	7	29	16.6897	TRUE	TRUE
9	8	26	12.4615	TRUE	TRUE
	568	568	800.59		

Data Volume

Enough data

p-value 0.000

< 0.01	Strong evidence against
< 0.10	Moderate evidence against
>= 0.10	Little or no evidence against



Softtek®

Data Set 3 Analysis

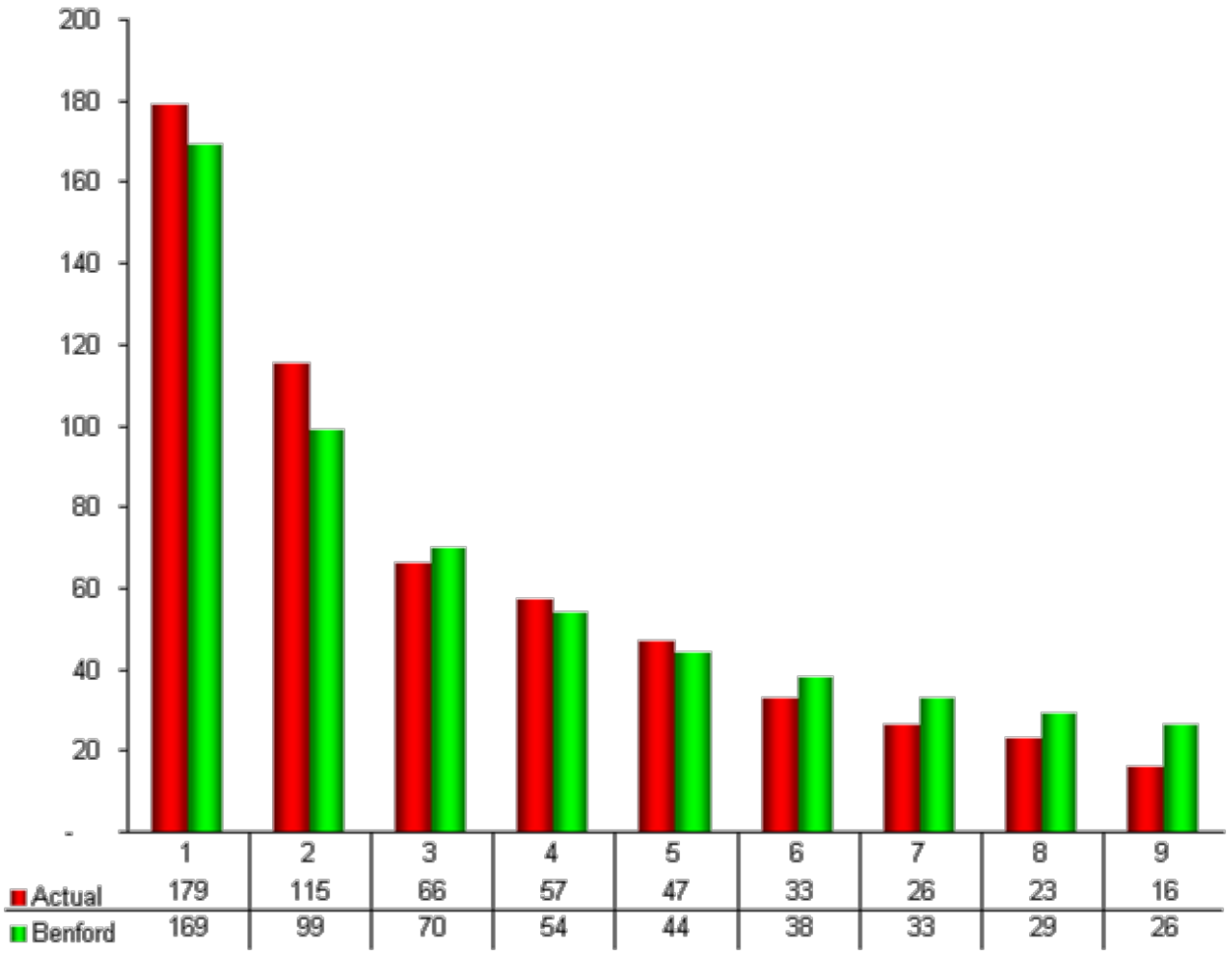
Digit	Actual	Benford	Chi2	10% Abs	35% rel
1	179	169	0.5917	TRUE	TRUE
2	115	99	2.5859	TRUE	TRUE
3	66	70	0.2286	TRUE	TRUE
4	57	54	0.1667	TRUE	TRUE
5	47	44	0.2045	TRUE	TRUE
6	33	38	0.6579	TRUE	TRUE
7	26	33	1.4848	TRUE	TRUE
8	23	29	1.2414	TRUE	TRUE
9	16	26	3.8462	TRUE	TRUE
	562	562	11.01		

Data Volume

Enough data

p-value 0.201

< 0.01	Strong evidence against
< 0.10	Moderate evidence against
>= 0.10	Little or no evidence against



Softtek®

Next Steps

- **Use Benford's Law to validate data from other sources:**
  - **Defect Log Fix times**
  - **Defect Counts????**
- **What else?**

Understand better with less effort

Softtek®

Questions and Comments

Ricardo Garza
Global Process Change Manager
Softtek Near Shore® Services

ricardo.garza@softtek.com
+52 (55) 2626-5247

http://www.softtek.com

Softtek®