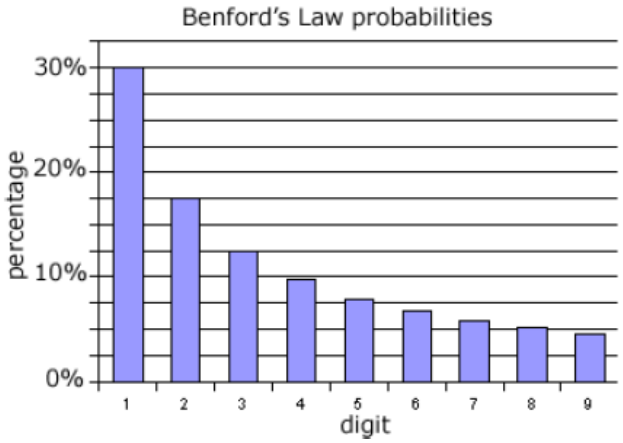
Benford's Law:

Is naturally occurring data the same as randomly generated data?

To spot real data, we look at the first digit of each number. In a randomly generated sequence of numbers, one would expect each of the digits 1, 2, 3, 4, 5, 6, 7, 8 and 9 to occur with equal frequency of $\frac{1}{9}$. However, Benford's law tells us that in naturally-occurring data this equal frequency does not actually occur. In such data, the first digit is 1 much more often than the other digits, and the higher the digit, the less likely it is to occur as the first digit. In fact, the probability of each digit being the first is given in the following table and graph. (The proportions of the digits i = 1, 2, …, 9 are given by $Log_{10}\left(1+\frac{1}{i}\right)$.) This was first discovered in 1881 by Simon Newcomb who noticed that books of log tables had more wear and tear on the first pages than on others.

Leading digit	Probability
1	30.10%
2	17.60%
3	12.50%
4	9.70%
5	7.90%
6	6.70%
7	5.80%
8	5.10%
9	4.60%



It was rediscovered by Benford in 1938 and has been shown by many other investigations, such as Giles (2007) who looked at prices in eBay auctions to hold true. It is commonly used to determine if any fraud is taking place as if the naturally occurring data has been tampered with it will not fit Benford's law.

Investigations to do:

1) We offer some data on the next page from CensusAtSchool for you to investigate to see if you can use Benford's law to determine which is the real dataset and which is the false data that has been generated.

2) You can also investigate with other datasets such as those found in the following places:

The number of employees in various occupations in New York State 2006 found at http://www.bls.gov/oes/current/oes_ny.htm#b00-000
Data from the Population Reference Bureau www.prb.org such as population figures or the number of threatened plant species in 104 countries http://www.prb.org/Datafinder/Topic/Bar.aspx?sort=v&order=d&variable=104

3) Investigate whether the numbers in the Fibonacci sequence follow Benford's law or not. Why do you think this happens?

Exercise 1.

Can you tell which is the real dataset?

One set of data is real data representing the number of secondary school pupils that took part in CensusAtSchool Phase 5 in a number of different postcode areas, the other is randomly generated.

Dataset 1

1279	257	161	60
1249	257	159	58
1126	256	156	58
893	251	156	53
782	250	153	52
744	247	146	49
676	240	135	48
672	237	119	47
579	223	119	46
542	218	112	43
537	215	112	41
532	211	107	39
481	211	105	37
467	202	94	36
414	201	91	34
329	201	85	31
326	193	79	29
325	193	78	23
320	192	77	15
313	172	73	15
299	170	71	15
286	168	69	8
277	166	67	8
276	165	66	7
259	164	62	5

Dataset 2

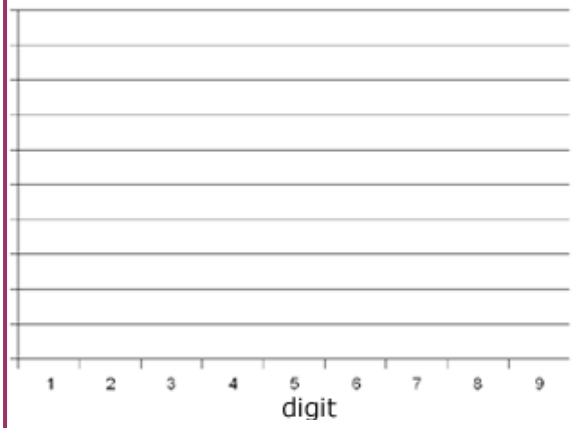
1340	569	292	68
1214	526	218	65
1064	593	251	66
998	560	211	68
861	592	214	60
814	597	261	60
832	522	271	64
828	540	230	66
747	577	155	53
724	520	132	56
778	591	123	59
727	482	152	52
775	443	145	32
748	420	99	39
765	430	94	35
732	457	88	28
625	489	81	23
641	447	81	16
690	333	86	19
648	337	77	13
675	312	75	11
615	355	72	9
619	353	79	8
695	343	67	6
520	345	67	8

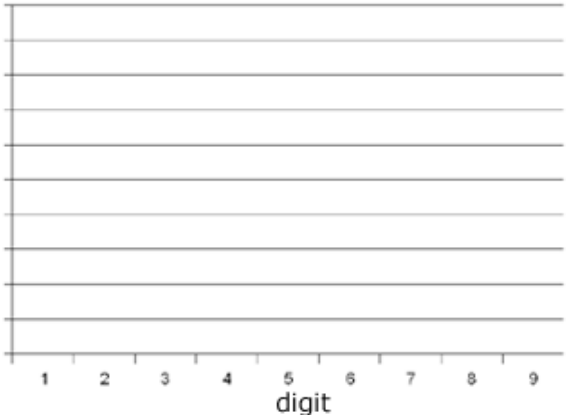
Go through the data given above and record how many numbers in each group begin with each digit in the table below, and plot these on the blank graphs. Can you tell which dataset is the real data?

(Hint: Have a look which graph is closest to the Benford's law probabilities above.)

Dataset 1	
Leading digit	Number
1	
2	
3	
4	
5	
6	
7	
8	
9	

Dataset 2	
Leading digit	Number
1	
2	
3	
4	
5	
6	
7	
8	
9	





Conclusion:

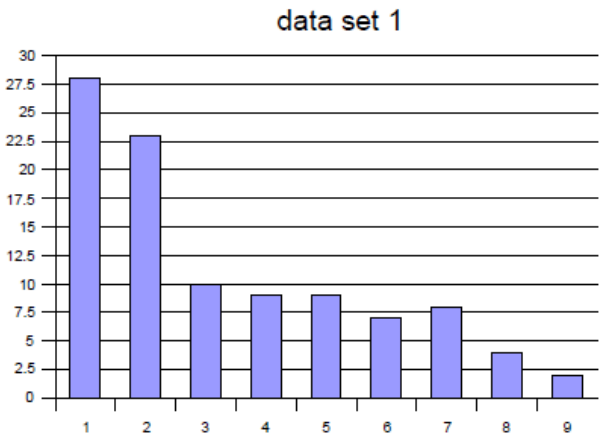
Solution:

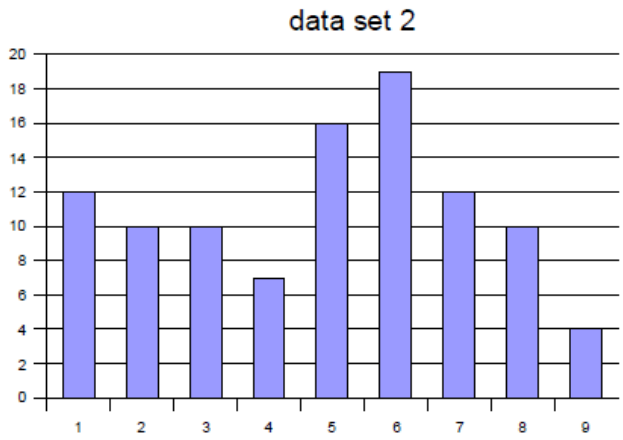
Dataset 1

Leading digit	Number
1	28
2	23
3	10
4	9
5	9
6	7
7	8
8	4
9	2

Dataset 2

Leading digit	Number
1	12
2	10
3	10
4	7
5	16
6	19
7	12
8	10
9	4





It should be obvious that the first set of data has a similarity with the Benford's Law probabilities where the second one doesn't! Thus we conclude dataset 1 is the real data.

Dataset 1 contains population data from 100 postcode areas from which children took part in CensusAtSchool Phase 5.
Dataset 2 contains a sequence of random numbers with first digits determined by rolling a die with digits 0-9 on it and filled in to simulate the population data using a computer pseudo-randomised number generator.

Included pupils from: LE, NG, CM, PO, B, ST, TN, SG, SO, MK, SK, BB, PE, GL, CV, PL, TR, HG, DT, DE, HA, DL. RG, L, CH, BS, WR, WS, NN, DA, IP, CW, WF, SE, BN, LU, CT, DY, WD, S, HR, RH, EN, HD, EX, SR, RM, NR, WV, DH, NW, N, CB, LA, NE, OX, E, BL, HP, SN, AL, SP, LN, DN, M, IM, IG, W, TA, PR, BH, BR, UB, CO, WA, CF, SM, YO, BD, SW, GU, PH, WN, ME, KT, LS, TQ, CR, SL, GY, DD, EH, FY, NP, KY, SS, BA, WC, OL

Zero pupils took part from: BT, HX, DG, KA, HU, ML, TF, MR, LD, FK, PA, KW, ZE, A
Less than five pupils took part in each of: TS, CA, JE, TW, TD, G, IV, SY, SA, LL, HS, EC