AN ANALYSIS OF THE DISTRIBUTION OF HUMAN-GENERATED RANDOM NUMBERS

TOM DUNKLEY

ABSTRACT

The project uses a large dataset to look at how humans choose numbers. We use cumulative frequency charts to compare this distribution to others and find that, while we appear to see a logarithmic scale, what we find is a set of uniform distributions. From this, we make a rigid algorithm that represents the way that humans generate random numbers.

INTRODUCTION

Humans have an odd understanding of distributions. While adults believe that their perception of natural numbers is linear, the natural way for humans to count is more similar to a logarithmic scale[1][2], whereby, for example, the perceived difference between 1 and 2 is much greater than the perceived difference between 150 and 151.

We can gain insight into this distribution by looking at the way humans try to generate random numbers that follow a uniform distribution. If a computer were to generate these numbers, each number would be equally likely to be chosen. But in a logarithmic distribution, numbers near the start of the range are more likely to be chosen.

DATA COLLECTION

The human-generated numbers come from a Reddit thread created in 2018 by user *itsrealgood*[3], in which they asked over three hundred people to generate ten numbers from 1 to 1000000, with mixed success. The number 69 occurs 14 times within the data, and 420 occurs 11 times. These numbers have different meanings across the internet, and so might have been given as a joke by respondents. This must always be considered when looking at human-generated datasets.

Another problem with human-generated datasets is that it can be clear that people weren't truthfully answering the questions. Four people only answered 1 as each of their 10 numbers, one person just answered 2 and for with ascending numbers from 1 to 10. By taking these data points into account, and others similar to them, we wipe out the outliers at 1 and 2.

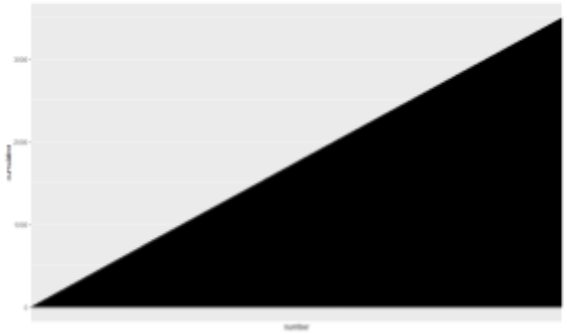
However, this creates an issue in itself - how can we judge what counts as a clear deviance from the question? We would, of course, expect a proportion of repeated numbers in an infinite uniform dataset, so should we remove them? While it would seem that the answer is a resounding yes in this particular scenario, we must consider whether this deviation from uniform is exactly what we are testing for.

---

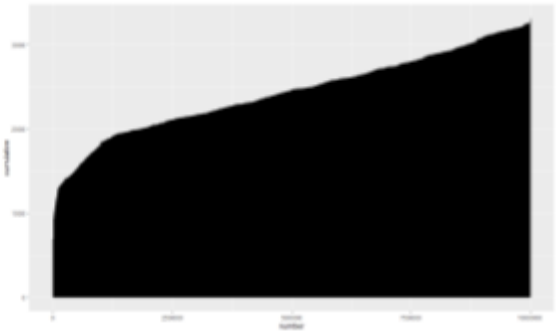
[1] Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures
[2] Why do we perceive logarithmically?
[3] https://www.reddit.com/r/SampleSize/

This is one of the issues with this particular data set. Each individual will have a different way of generating numbers, and asking for ten numbers from each will allow them to have too much influence in the final dataset. However, asking for just one datapoint from each respondent will reduce our data set by a factor of ten, and for any analysis, that is too big to just miss out on.

DATA ANALYSIS



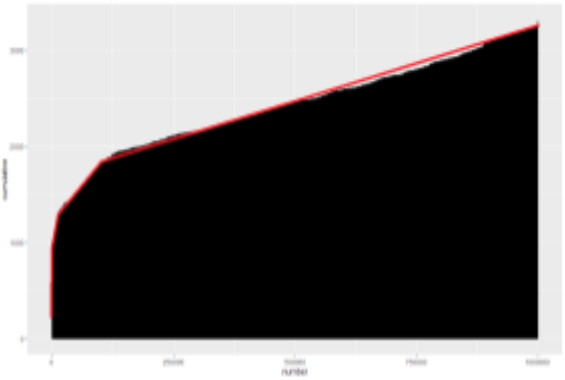
1(a): Cumulative frequency chart showing around 3500 human-generated random numbers in the range 1 to 1000000



1(b): Cumulative frequency chart showing 3500 machine-generated random numbers in the range 1 to 1000000

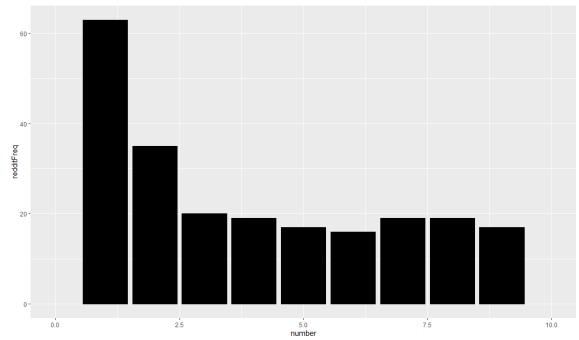
By comparing the two graphs, we can see clearly that the human-generated numbers appear to follow a logarithmic pattern. But really, it's more complicated than that. Looking at the data, you may notice some straight lines that the chart seem to follow, at regular intervals. These have been highlighted in chart 2 to make them more obvious.

These straight lines follow what we would expect from the machine-generated numbers, as in chart 1(b), and so we can conclude that humans used a method of choosing one of these ranges in which we see straight lines, and then using a uniform method to choose the number within the range.

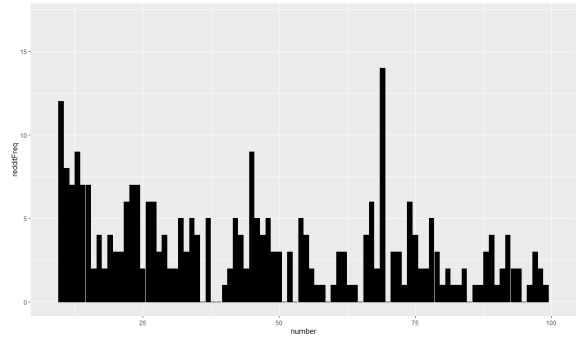


2: Figure 1(a) with approximate straight lines highlighted in red.

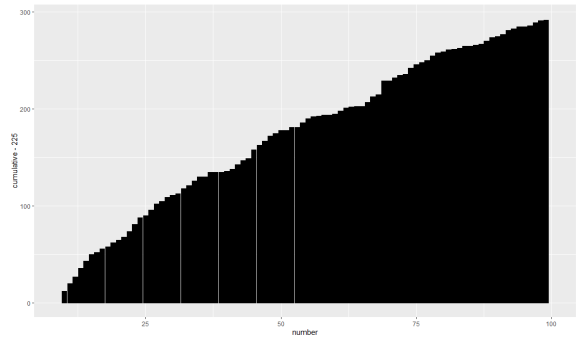
It would be reasonable to suppose that these ranges could correlate with the length of the digits in each number, with people first choosing the number of digits and then randomly choosing their number. We can test this proposal by looking at each range's distributions.



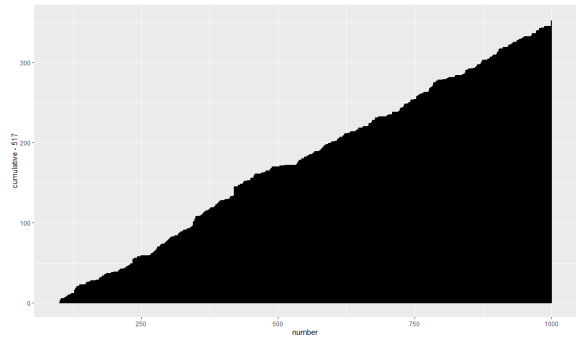
3(a): Frequency chart for the range 1-9 inclusive. While 1 and 2 are noticeably more common than the other numbers, the rest appear with similar frequency.



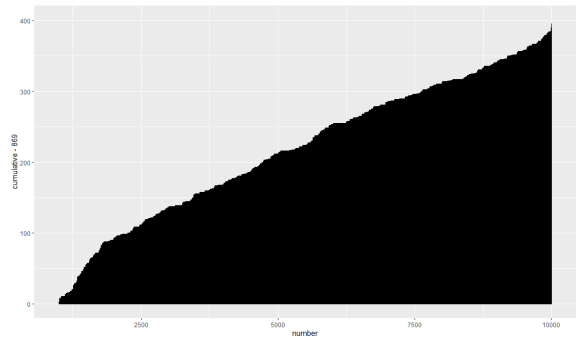
3(b): Frequency chart for the range 10-99 inclusive. Up to about 20, frequencies are higher, and there are multiple peaks and troughs, most noticeably 69. Aside from this, it is hard to tell how the frequencies are distributed.



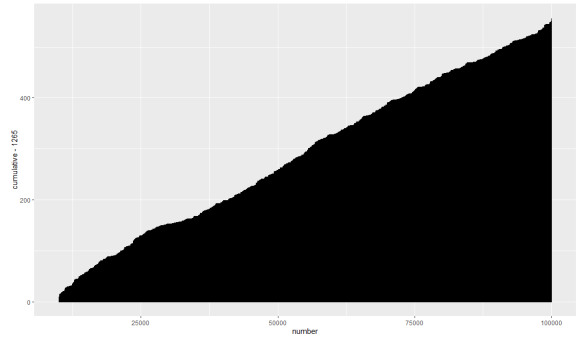
3(c): Cumulative frequency chart for the range 10-99 inclusive. Note the sharp rise at 69.



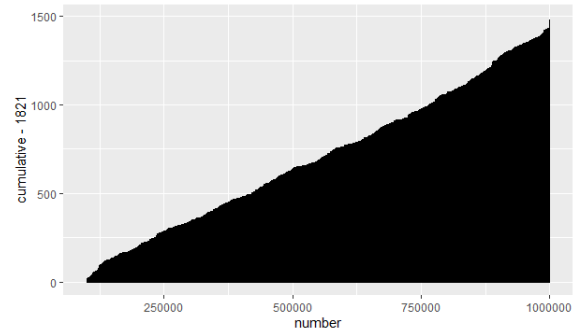
3(d): Cumulative frequency chart for the range 100-999 inclusive. Note the similar sharp rise at 420.



3(e): Cumulative frequency chart for the range 1000-9999 inclusive.



3(f): Cumulative frequency chart for the range 10000-99999 inclusive.



3(g): Cumulative frequency chart for the range 100000-999999 inclusive.

Some of these charts are similar in shape to the one in 1(b), supporting the proposal, but there are some notable exceptions. In 3(c) and 3(e) especially, the graphs appear to be decreasing and so it would seem that the earlier numbers are more likely to occur. Having said that, these distributions are much more uniform than 1(a).

It would seem, then, that people do choose numbers first by choosing the number of digits and then by choosing the number. If we take these ranges into account, human-generated numbers can be uniform. But how is a range chosen?

DEFINING RANGES

LENGTH	COUNT	PROPORTION	EXPECTED PROPORTION	P/E(P)
1	299	0.09279950341	0.000009	10,311.06
2	360	0.1117318436	0.00009	1,241.46
3	405	0.125698324	0.0009	139.66
4	565	0.1753569212	0.009	19.48
5	1496	0.4643078833	0.09	5.16
6	91	0.02824332713	0.9	0.03
7	6	0.001862197393	0.000001	1,862.20

Using the data in table 4, we can see that the frequency of numbers of length six in the human-generated dataset is just 79% of that which is expected if the dataset is uniform.

To put this into context, let us pose the following question: suppose one large dataset, of which 50% of the data points are human-generated, and 50% are generated using a uniform random distribution. If we pluck one data point out at random, and we do not know how it was generated, can we work back and find out the probability that it was generated by a human?

The answer is, of course, yes. We can do this by looking at the ratio of occurrences of this number between each 50%: P/E(P). If we know that the number has six digits, then the human:random ratio is 0.71:0.9. In other words, the probability that the number was generated by a human is 0.71/(0.9+0.71)=0.44. We can do a similar analysis for other attributes, as well as individual numbers.

Our dataset isn't large enough to accurately analyse individual numbers using solely the proportion of that number which occurs. Instead, we can look at a combination of attributes, which define only that number in the range. One example of an attribute is the relationship between consecutive numbers. How often does a 4 follow a 5? This seems like a reasonable thing to analyse, as it could be proposed that people generate a number by first choosing a range, as we discovered, and then choosing each digit one by one.

	1	2	3	4	5	6	7	8	9	0
^	394	291	263	301	284	272	278	315	370	N/A
1	133	177	108	81	86	66	77	82	88	183
2	106	92	182	105	107	90	117	106	105	119
3	71	161	99	149	103	128	136	108	82	76
4	78	137	130	122	176	133	103	127	101	66
5	60	96	126	140	131	171	107	93	61	73
6	75	96	114	102	138	108	155	98	115	67
7	79	114	110	138	118	109	146	143	80	70
8	79	108	103	148	86	99	146	158	127	70
9	105	90	91	96	74	109	87	152	309	116
0	86	63	69	60	52	42	43	39	65	277

5(a): The frequency of each combination of consecutive digits in the human-generated data set. The vertical axis represents the first digit, and the horizontal axis represents the next digit. ^ represents the start of the string.

	1	2	3	4	5	6	7	8	9	0
^	407	393	359	367	374	401	387	385	418	N/A
1	123	127	139	124	152	133	131	130	169	135
2	150	137	125	159	147	151	139	133	147	131
3	138	139	104	136	144	142	149	127	144	149
4	136	124	144	141	162	143	152	147	141	130
5	136	160	153	159	149	179	146	143	128	167
6	146	148	152	169	160	141	137	155	130	139
7	147	150	142	138	162	133	143	154	124	152
8	118	154	143	142	138	133	143	131	154	150
9	136	132	148	144	174	154	158	150	142	138
0	95	98	103	106	113	108	112	104	106	113

5(b): The same table but for a uniformly generated random data set.

digit	frequency
1	2042
2	2063
3	1981
4	2079
5	2169
6	2127
7	2120
8	2065
9	2119
0	1771

5(c): The frequency of each digit within the human-generated data set.

Looking at table 5(a), some patterns seem to emerge. 1 and 9 are the most common starting digits, with 3 and 4 the least common. The most common consecutive digits lie close to each other, with a weak direct proportion between the size of one digit and the one that follows it. The only digit this doesn't apply to is 0, of which the opposite is true, usually preceding or following larger digits. Looking at 5(c), digits in the middle are more frequent, and 0 is noticeably less common, appearing just 8.4% of the time.

Using this data, we can find a supposed probability of every number using the following formula:

$$P/E(P) = P(number\ of\ digits\ is\ as\ given) \times \Pi\, P(digit\ follows\ previous\ digit)$$

The number we can be most sure was generated by a human is 10, appearing 153 times as a human-generated number for every time it appears as a randomly generated number. 316080 is the number that is least likely to be generated by a human using this method, appearing as such just 0.062 times for every time that is randomly generated in a uniform distribution.

Extending this method will always get us caught in a loop for much larger numbers. This can be seen in 316080 - if extended to the least likely 15 digit number, this would be seen as 316080808080808: the formula would put the probability of this being generated by a human as relatively tiny, when it would seem to any observer due to the recursion that this is extremely likely to have been generated by a human.

THE HUMAN DISTRIBUTIONS

We can define this distribution in a much more structured way if we assume that all digits are equally likely. We do so as follows:

$$X \sim Human(p_1, p_2, \ldots, p_R)$$

$$Where\ X \in \mathbb{N},\ 1 \leq X < 10^R$$

$$And\ p_r\ is\ the\ probability\ that\ X\ has\ r\ digits$$

For this distribution, it is clear that

$$P(X = x) = \frac{p_r}{9 \times 10^{r-1}}$$

$$Where\ r\ is\ the\ number\ of\ digits\ in\ x \qquad (1)$$

We can also find E(X) as follows:

$$E(X) = \sum_{all\ x} xP(X = x)$$

$$= \sum_{all\ r} \left(\sum_{x=10^{r-1}}^{10^r - 1} x \frac{p_r}{9 \times 10^{r-1}} \right)$$

$$= \sum_{r=1}^{R} \left(\frac{p_r}{9 \times 10^r} \times \sum_{x=10^{r-1}}^{10^r - 1} x \right)$$

$$= \sum_{r=1}^{R} \left(\frac{p_r}{9 \times 10^r} \times \frac{1}{2} \left((10^r - 1)(10^r) - (10^{r-1})(10^{r-1} + 1) \right) \right)$$

$$= \frac{1}{18} \sum_{r=1}^{R} \left(\frac{p_r}{10^r} (10^{2r} - 10^r - 10^{2r-2} - 10^{r-1}) \right)$$

$$= \frac{1}{18} \sum_{r=1}^{R} \left(\frac{p_r}{10^r} \times 10^{r-1} (10^{r+1} - 10^1 - 10^{r-1} - 10^0) \right)$$

$$= \frac{1}{18} \sum_{r=1}^{R} \left(p_r \times 10^{-1} (10^{r+1} - 10^{r-1} - 11) \right)$$

$$= \frac{1}{180} \sum_{r=1}^{R} \left(p_r (10^{r+1} - 10^{r-1}) - 11 p_r \right)$$

$$= \frac{1}{180}\left(\sum_{r=1}^{R}\left(p_r\left(10^{r+1} - 10^{r-1}\right)\right) - 11\sum_{r=1}^{R} p_r\right)$$

$$E(X) = \frac{1}{180}\sum_{r=1}^{R}\left(p_r\left(10^{r+1} - 10^{r-1}\right)\right) - \frac{11}{180}$$

(2)

Another level of abstraction we can consider is this distribution if p_n is constant - that is to say, all digit lengths are equally likely.

$$X \sim SimpleHuman(R)$$
$$Where\ X \in \mathbb{N},\ 1 \leq X < 10^R$$

We find that

$$P(X = x) = \frac{1}{R}\frac{1}{9\times 10^{r-1}} \qquad (3)$$

When we consider the expected value of the SimpleHuman distribution, the equation simplifies somewhat well, but I will leave that derivation to the reader.

CONCLUSION

There is far further to go when looking at human-generated distributions, but the main issue with this sort of analysis is that it requires huge datasets, the likes of which are extremely difficult to get, let alone with any sort of reliability. The internet makes this slightly easier, but for any sort of in-depth analysis, we would need a good few hundred thousand data points, which could take months to collate.

Having said that, through multiple levels of abstraction we can model these numbers with distributions to create somewhat coherent cumulative distribution functions.

FURTHER READING

- Analysing Humanly Generated Random Number Sequences: A Pattern-Based Approach - *Marc-André Schulz,Barbara Schmalbach,Peter Brugger,Karsten Witt*
- Generation of random sequences by human subjects: A critical survey of literature - *Wagenaar, W. A.*