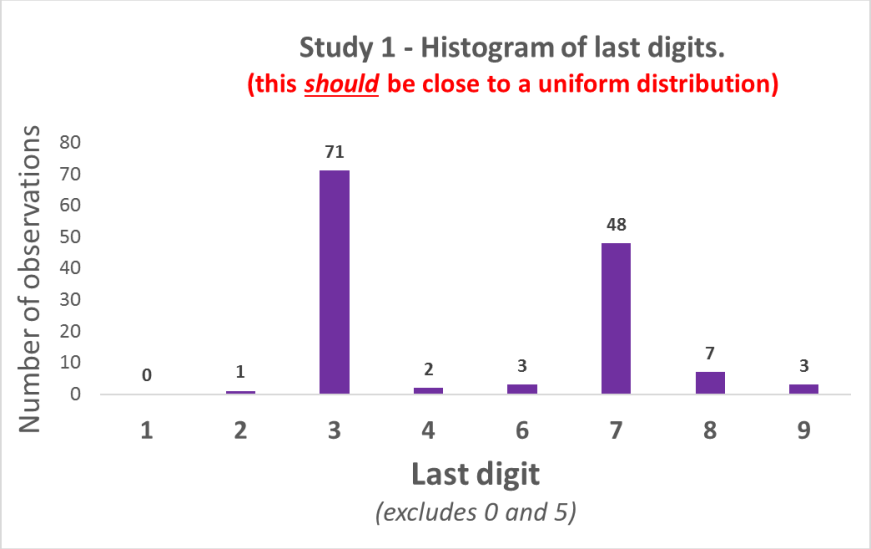
Appendix for DataColada 74 - What distributions of last digits should we expect for Study 1?

Last update: 2018 - 08 - 20

Motivation

Benford's law is technically for 1st digits but it has been generalized for every digit in a number. The expected distribution converges to uniform very quickly as we move to the 2nd, 3rd, 4th, etc digit in a number. But, the speed of convergence depends, to some extent, on the span of observed values. In the post we compare the observed distribution of last digits in Study 1, to the uniform, concluding it looks nothing like it.

As a reminder, the post reads:



A statistical test is not really needed here, but for completeness, the null of a uniform distribution is rejected: $\chi^2(7) = 304.53$, $p<10^{-62}$.

In this appendix we explore how far from the uniform might have we expected the last digit to be in this particular data. We use two standards to answer this question, they both arrive at the same answer: *It should have been extremely close to uniform.*

Standard 1. Normal distribution

In Experiment 1 there are 135 observations that end in a digit that's neither 0 nor 5, and thus purportedly obtained with a scale precise to 1 gram. The mean number of grams for these observations was 52.44, and the SD=14.75.

Drawing a large number of observations (100,000) from the normal distribution with mu=52.44, and sigma=14.75, we rounded to 0 decimals, and tabulated the relative frequency of each last digit. The figure below show that we would in this case expect a distribution of digits that's fairly close to uniform, and very different from what's observed

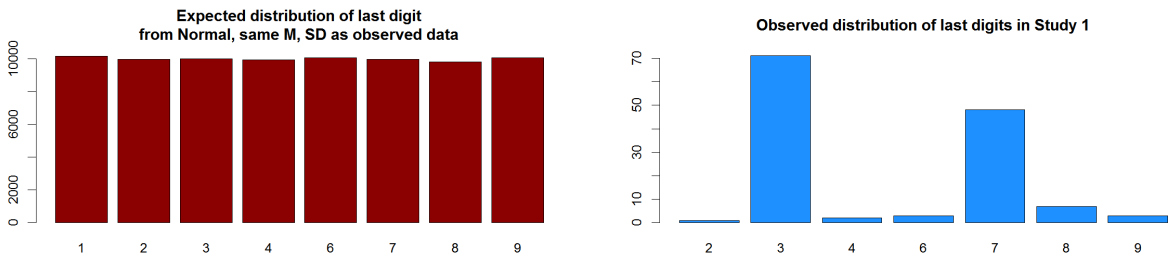


Fig A1. Comparison of expected and observed frequency of last digits, using Normal distribution as standard

Standard 2. Kernell approximation of observed distribution

We now relax the normality assumption.

We begin with a kernel density estimation using all data during the first 20 days (before treatment). The default kernel approach is for continuous data, but we are interested in discrete measures at the 1 gram level. We discretize the kernel by relying on the mgcv() package in R which fits a spline, we predict the density with the values, and then use the fitted values for all discrete values between 0 and 120.

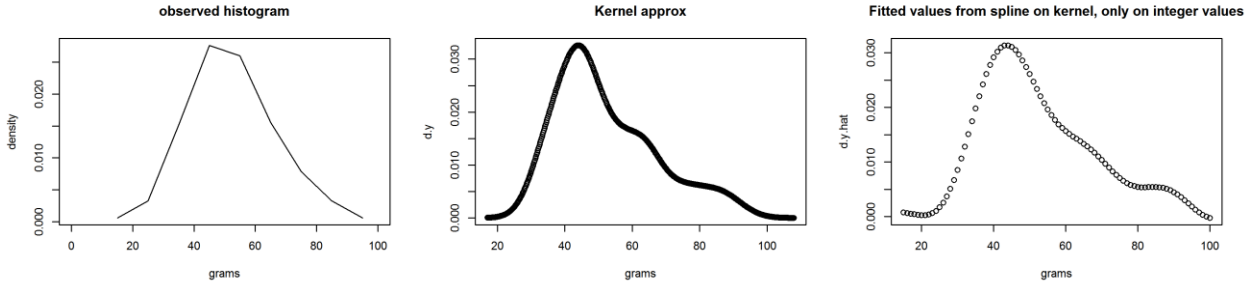


Figure A2. Distribution of values in Study 1, measured with 1 gram precision.

Note: The first panel has the empirical distribution in the data, the 2nd the kernel approximation, and the third panel the discretized version. Only the 135 observations from Study 1, which are not multiples of 5, are included.

We rely on the distribution depicted in the third panel, instead of the normal distribution, to generated an expected distribution of last-digits in the data. In particular, we sum up the density (the y-axis in the 3rd panel of Figure A2) for all discrete values associated with a last digit. For example, for 0 we add up the density for 0, 10, 20… 100. For 1 we add up 1, 11, 21…. The result is an extremely close to uniform distribution, in stark contrast to the posted data.

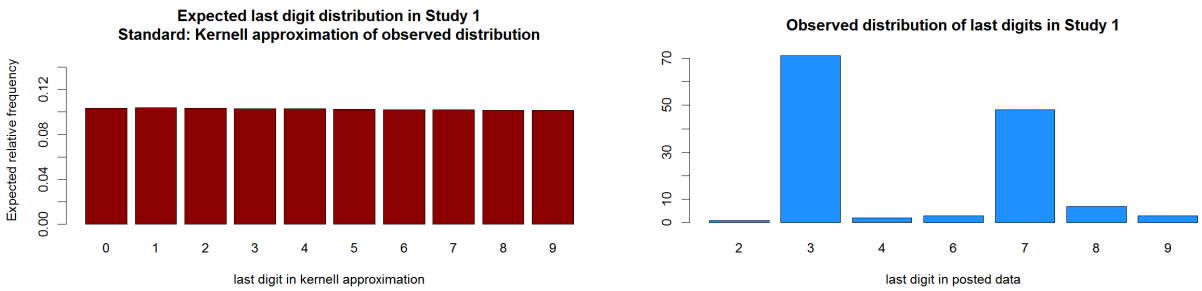


Figure A3. Expected and observed last digits in Study 1, using kernel approximation of observed data as the standard.

In sum.

The uniform distribution is a reasonable standard for the data reported in the paper, and the data reported in the paper look nothing like it.