Is lottery fair?

Michal Horáček
Matrikelnummer 6373382
michal.horacek@student.uni-tuebingen.de

Carson Zhang
Matrikelnummer 6384481
carson.zhang@student.uni-tuebingen.de

Abstract

We investigate lottery.

1 Introduction

Games such as lottery and dice are likely one of the first application of randomness in human culture. But for the same time organizers have cheated in lotteries, skewing the uniform distribution of drawn numbers expected by common sense for their own monetary gain and misleading customers.

We begin by describing the data-gathering process in Section 2. Given the enormous demands placed on data volume by the second part of this paper, a sizable portion of our work involved preparing the input data.

In Section 3, we investigate the distribution of answers for the German Lotto lottery drawn from 1955 to the present day.

We continue by reformulating lottery as a process generating random numbers and explore their quality via the Diehard battery of tests [1] in Section 4. Through their means we attempt to prove whether lottery holds properties such as mutual uncorrelatedness or an absence of a period of repetition.

This paper is concluded in Section 5 by a brief discussion of the limitations of our work.

2 Dataset

The chief component of our dataset is formed by numbers drawn in German lottery Lotto from 1955 onwards. We have focused on these dataset because of geographical locality and historical depth of 70 years. Since its beginning, the rules have undergone slight changes, such as the introduction of "super numbers" in December of 1991. Nonetheless, the core principle of the lottery, six numbers between 1 and 49 has not changed once in almost 70 years and thus provides a consistent basis for our work.

Is there a dataset for a rigged lottery? This looks sketchy as hell.

However even 70 years of Lotto numbers is not sufficient to produce enough data for the diehard tests. These require 10 to 12 MiB of random bits, which is substantially more than 28.4 KiB of Lotto numbers. Therefore we downloaded other lottery datasets and combined them together. In total, our dataset reached more than 750 000 numbers drawn in 18 different lotteries. These majority of these lotteries come from various english-speaking countries such as USA, Australia or UK because we have been looking for them with english search queries. Most of the complement is formed by other european nations like Italy, Czech Republic or Germany.

Project Report for *Data Literacy* 2022/23

Merge datasets?

The numbers are drawn individually, but their order within a single lottery draw does not matter - but maybe it does for some of our tests?

Investigate this

3 Distribution testing

In the first part we test whether the drawn numbers come from a uniform distribution $\mathcal{U}(1, 50)$. To investigate this, we use Pearson's χ^2 test and the Kolmogorov-Smirnov test.

For more information about the Diehard battery of tests we refer the reader to the original paper [2].

4 Diehard tests

Fairness entails more than the question whether are lottery numbers from the expected distribution. For instance, the Kolmogorov-Smirnov test used in the first part of this paper does not concern itself with the order the numbers are drawn. However if we saw a lottery whose numbers were always drawn in a descending sequence, for example, we would become suspicious.

Thus a more comprehensive test is clearly required to establish a more detailed answer to our question. We approach this problem by reformulating lottery as a process producing a stream of (supposedly) random numbers, which themselves are simply bit sequences. Under this formulation, we can deploy standard statistical tests developed for testing random number generators: we have a file of one and zero bits and wish to investigate if its bits are correlated, repeating with a period or other quantities undesirable for randomness.

A number of these test suites has been developed over time. Donald Knuth presented an initial set of empirical tests in the second volume of his computer science bible The Art of Computer Programming in 1969. Many general cryptography textbooks such as Handbook of Applied Cryptography or Foundations of Cryptography contain multiple tests of their own. The american National Institute of Standards & Technology has published a guideline discussing this matter too.

We decided to use the Diehard battery of tests, which was developed by the american statistician George Marsaglia in the nineties. While this package used to be quite popular in its day, it has been superceded today by other suites, including its derivatives such as Dieharder or TestU01. **Why did we pick this one?**

1. About Diehard.
2. The following part is divided into several section. In section 1, we discuss data gathering, processing and general creation of input files for Diehard. Special attention is given to the problem of attaining input file of sufficient size and the asymetric requirements of individual Diehard tests.
3. Part 2 highlights result on Diehard suite, interprets them and compares them against commonly used PRNGs.
4. Section 3 clarifies the shortcomings of above approach.

5 Conclusion

1. Lottery draws without replacement
2. Some datasets are sorted in draw order
3. Reminder: too few numbers for comfort
4. The Diehard tests are flawed (Linear Feedback Shift Registers)

References

[1] George Marsaglia. The marsaglia random number cdrom including the diehard battery of tests of randomness. https://web.archive.org/web/20160125103112/http://stat.fsu.edu/pub/diehard/, 1995.

[2] George Marsaglia. A current view of random number generators. In Elsevier Science Publishers, editor, *Computer Science and Statistics, Sixteenth Symposium on the Interface*, 1985.