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 the uniformity of various lotteries. We perform a hypothesis test of the null hypothesis that the German Lotto numbers are sampled with equal probability, using the χ^2 test statistic, with a p-value of **the p-value**. Furthermore, to evaluate the randomness of the random number generators underlying different lotteries, we apply the Diehard battery of tests **with the following results**.

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 main dataset that we investigate is the numbers drawn from the German lottery Lotto, from 1955 onwards. We have chosen this dataset because it is relevant to us and readers of this analysis as current residents of Germany, and it contains almost 70 years of data, which we believ is enough data to perform a meaningful analysis.

The dataset has been compiled by Johannes Friedrich, a software developer who has made the data publicly available via his Github repository.

Perform verification of the correctness of the dataset by sampling from it and manually inspecting the data.

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.

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

3.1 Testing uniformity

DECIDE WHICH TEST TO USE, BETWEEN: MINIMUM DISTANCE, BINNED SUMS, SIMPLE NUMBER FREQUENCIES

Decide whether we should specify the multivariate hypergeometric assumption. Decide whether this assumption would change the degrees of freedom (at the moment, I think it wouldn't).

We are interested in whether each of the lottery numbers appear with equal probability. However, each sample of 6 balls from the lottery bowl is sampled without replacement. Therefore, within a single lottery drawing, the balls are *not* independent.

Each lottery day is independent. So we can

We conducted the following hypothesis test.

H_0: each lottery number is drawn with equal probability.

H_A: the lottery numbers are drawn with unequal probabilities: some numbers are more likely to appear than others.

We used Pearson's χ^2 test, a commonly recommended test for the probabilities of observing categorical data. **Motivate the choice of this particular test.**

Under H_0, we compute the following expected frequencies of each lottery number: **Create a table with the expected frequencies. They will all be the same, so perhaps this can just be stated in a sentence.**

The χ^2 statistic is a function of how much our actual observed frequencies deviate from these expected frequencies. Larger deviations result in higher values of the χ^2 statistic and therefore lower p-values.

Justify the choice of degrees of freedom.

Decide whether we should spell out the computation of the p-value, as if we computed it manually.

Report the value of the χ^2 statistic and the p-value.

3.2 Diehard tests

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.