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. First, we test the null hypothesis that, for the German Lotto, the distribution of the minimum distance between lottery numbers, d, is as it would be in a uniformly sampled lottery. The χ^2_6 goodness-of-fit test resulted in a p-value of 0.4989. 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

Lottery is one of the most obvious instances of randomness in everyday life. Many forms of lottery exist around the world, though they are all based on the same fundamental idea: integers are being drawn from a certain interval without replacement, with each integer being equally as likely as the others remaining. This project searches for evidence of this principle.

This report consists of four sections. Section 2 describes the data gathering process. In section 3, we apply a statistic test from [1] to several lotteries. Subsequently we explore the randomness of lottery through the Diehard battery of tests in Section 4. Finally we briefly discuss the limitations of our work in Section 5.

2 Dataset

We began with the winning numbers of the German lottery Lotto, which are tracked at [2] from 1955 onwards. It became apparent early on, that much more data will be required, so we added numbers from other lotteries. Ultimately we have collected 37 different lotteries, which in total provide almost 30 million numbers.

By volume, our data is dominated by the New York Quick Draw lottery, because it has been drawn every 4 minutes for the last decade, adding up to to 25,663,100 numbers at our collection of the dataset. Washington DC's Keno functions similarly, reaching 2,299,563 numbers over 3 years of existence. Our dataset is completed by the "other" category, consisting of 1,917,890 winning numbers combined from lotteries drawn less often worldwide.

Our entire dataset originates from West-aligned countries. The reason is twofold - we have been looking for them with English search queries and these nations are more likely to subscribe to ideas like open data and thus offer such dataset as CSV files.

3 Methods

3.1 Testing the distribution of d

We conducted the following hypothesis test.

H_0: the distribution of d is what it we expect when the lottery numbers are sampled uniformly.

H_A: the distribution of d is different.

We used Pearson's χ^2 test, a commonly recommended test for the probabilities of observing categorical data. This is a popular test for performing exactly the type of hypothesis test we intend to perform. Furthermore, it is the same type of test performed by Drakakis et al., which allows us to check our (and their) results. We can also satisfy the assumptions of the χ^2 test, and follow the common rules-of-thumb for using the test. **Motivate the choice of this particular test.**

3.1.1 Description of minimum-distance statistic.

To create the frequency tables for the $\chi^2 test$, we used the minimum distance statistic d used by Drakakis et al.

write the definition from Drakakis et al.

This statistic is useful for detecting human tampering because it is known that humans usually do a poor job of imitating the random choices that occur under the null hypothesis. Boland and Pawitan showed that when humans are asked to sample m integers from the set $1, ..., n$, the value of this statistic is greater on average than it is under uniform sampling. In particular, they observed that the true probability of $d = 1$ is greater than 0.5, a result that is highly unintuitive to humans: humans tend to underestimate how likely it is that two consecutive numbers are picked. In particular, Boland and Pawitan computed a χ^2 goodness-of-fit statistic of the human-produced d against the uniformly produced d , which if they had conducted a hypothesis test, would have yielded a p-value of < 0.0000001.

Therefore, a low p-value could be consistent with human tampering, especially if we observe an unusually small number of lottery drawings where $d = 1$, as Drakakis et al. did in the French lottery.

3.1.2 Description of the test statistic.

The χ^2 goodness-of-fit test tests how well an expected discrete distribution fits an observed discrete distribution.

We computed the expected distribution using the following formula proved by Drakakis (2007).

Let the lottery game be a sample of m integers drawn from the integers $1, ..., n$. (In the German Lotto, $n = 49$ and $m = 6$.) Let $r_1, ..., r_m$ be the numbers drawn in the sample of size m.

The minimum distance $d = min_{1 \le i < j \le m} |r_j - r_i|$ has the following distribution. For $k = 1, ..., \lfloor \frac{n-1}{m-1} \rfloor$,

$$P(d < k) = 1 - \frac{\binom{n-(k-1)(m-1)}{m}}{\binom{n}{m}}$$

We define $\alpha = 0.05$ as our significance level because it is popular and Krakakis used the same significance level.

To prepare the German Lotto dataset for the hypothesis test, we computed d, the minimum distance between winning numbers, for each lottery day. We counted the frequencies of each value of d, and combined the counts of the two largest possible values, 7 and 8. This is the same type of data preparation that was performed in the paper, and we did it to satisfy the common rule of thumb for usage of the χ^2 goodness-of-fit test: the expected frequency in each bin must be ≥ 5.

This prepration yielded the dataset in Table 1.

We estimated no parameters from the data, so we use $p - 1 = 7 - 1 = 6$ degrees of freedom for the χ^2 statistic.

Table 1: Frequencies of d statistic for German Lotto

d	Expected frequency	Actual frequency
1	2310.595965	2383
2	1266.759926	1247
3	639.888057	615
4	290.255446	273
5	113.589766	105
6	35.857667	35
7 and 8	9.025145	8

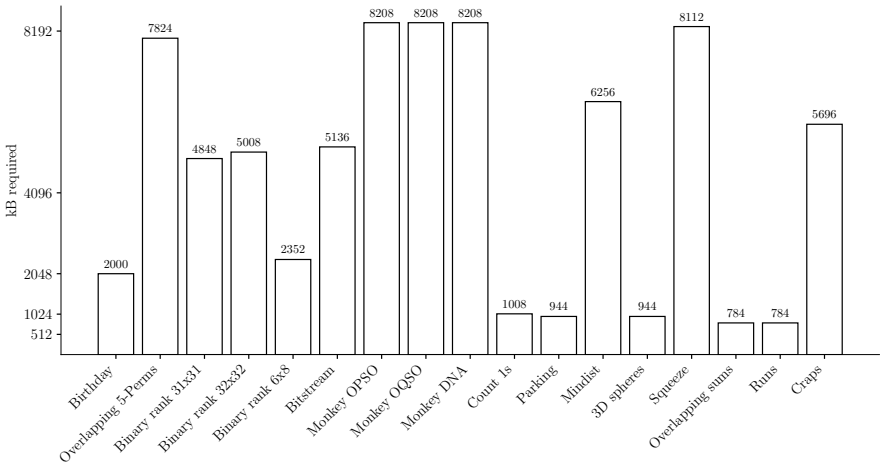


Figure 1: Amount of data required to run tests at default settings.

3.2 Diehard tests

For more information about the Diehard battery of tests we refer the reader to the original paper [3].

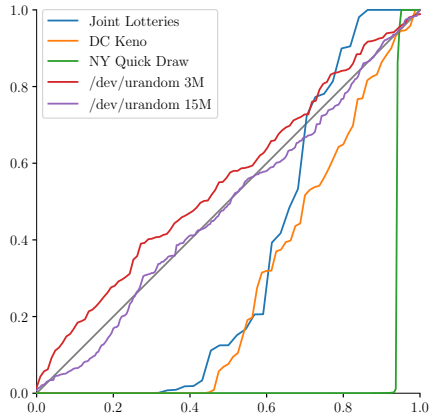
4 Diehard tests

Fairness entails more than the question whether are lottery numbers from the expected distribution. We need to search for deeper patterns in the data. Thus a more comprehensive test is required to establish a more detailed answer to our question. We reformulate winning numbers of a lottery as a stream of (supposedly) random bits. This enables us to deploy standard statistical suites for testing pseudo random number generators.

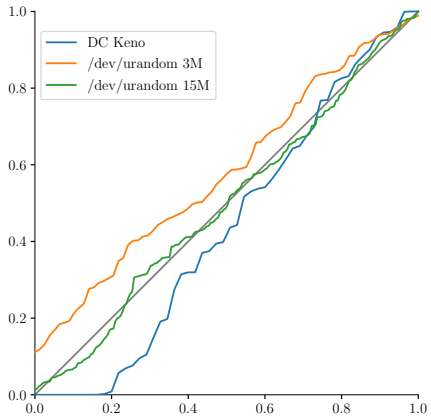
Donald Knuth presented an initial set of empirical tests in the second volume of his computer science bible The Art of Computer Programming in 1969. Since then numerous new test suites have been developed [4, 5, 6, 7]. We decided to use George Marsaglia's Diehard battery of tests [8]. Diehard has been superceded by newer tests for its original purpose [6], however it remains the best choice for us because it requires the least amount of data to run as illustrated in Figure 1 and [9]. Nonetheless even less data greedy test still requires quite a lot of data. We are thus limited in which tests we can run given our dataset size. **Transform figure into number of standard draws 1-49**

4.1 Data augmentation

Diehard battery expects a random stream of zero and one bits as input. We therefore require a function which produces uniformly distributed bit sequences from lottery numbers. Because it is impossible to directly transform a discrete $\mathcal{U}(1, 49)$ of lottery numbers into $\mathcal{U}(0, 31)$ (the uniform distribution for



(a) Randomness quality as measured by Diehard. Gray line is perfect randomness.



(b) Removed p-values from Binary Rank 6x8 test.

Figure 2: p-value distribution

5-tuples of bits) we simply reject all numbers greater than 32 and substract one from the rest. This causes a fair amount of data loss, but we do not see a better way. (If a lottery draws from interval e.g. 1-80, we can throw away everything greater than 64 and get uniform distribution of bit 6-tuples.)

4.2 Results

The Diehard battery of tests provides 15 statistical tests. These tests operate with the null hypothesis that the data is truly random. Each of them outputs one or more p-values, which are uniformly distributed if the null hypothesis is correct [5]. In some cases, a Kolmogorov-Smirnov test is used to test for uniformity of p-values provide a single final, composite p-value. For more information on the nature of these tests, consult either the Diehard source code appended to this report or one of George Marsaglia various papers [10, 11].

We compare our lottery datasets against randomness collected from /dev/urandom, the preferred source of randomness in Linux systems [12, 13]. These datasets ensure that tests are working as intended and represent the best digital randomness available to the common man.

We immediately notice that the NY Quick Draw catastrophically failed to pass the Diehard tests. We believe this is caused by the structure of NY Quick Draw csv files. Each row contains 20 winning numbers of an individual draw, all sorted in ascending order. It is our opinion that Diehard identifies this ordering. This hypothesis is supported by the behaviour of Joint Lotteries on the Runs test. This test performs two runs and provides two p-value in each run. The first run is provided a 50 % of drawn and 50 % of ascending-order numbers, reports p-values of 0.981149 and 0.904974. The second run uses solely ascending-order numbers and both p-values become 0.

The DC Keno dataset which is entirely created from drawn-order CSV files displays a remarkable even slope after it starts rising. Curiously, the 25 of the 32 zero p-values for DC Keno come from the Binary Rank 6x8 test. If we disregard this test, the resulting p-values are fairly uniformly distributed, although the dataset still struggles at some of them, corresponding to 20 % of zero p-values. This is reasonable considering some enduring faults of the dataset (lottery drawn without replacement).

The Joint Lotteries mixed dataset achieves somewhat worse performance in comparison with the DC Keno dataset. We interpret this evidence for the importance of using drawn-order input files as later runs of the same test visible degrade in performance since they work with ascending-order numbers.

4.3 Discussion

There is a great number of pitfalls to take into account. Some of them have already been addressed, such as data size, drawn/ascending order or lottery being drawn without replacement and thus not providing a truly uniformed distribution.

Furthermore, even the Diehard battery of tests possesses flaws of its own. Its critique can be found at [6, 14]. For instance, Linear Feedback Shift Registers (a type of pseudorandom number generator) passes the Diehard battery of tests [15], despite being predictable and cryptographically weak.

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)

Some other choices of hypothesis tests were possible. We discuss why we did not choose some alternatives.

We initially wanted to perform a hypothesis test of the distribution of the numbers. In particular, we considered performing a test of the goodness-of-fit of the discrete uniform distribution from 1 to 49 for the observed frequencies of each Lotto number. However, we discarded this approach because the common goodness-of-fit tests (χ^2, Kolmogorov-Smirnov) require that the events are i.i.d., and the Lotto numbers are not i.i.d.: in each Lotto drawing, the numbers are sampled without replacement, so each number drawn on a particular day depends on the numbers that were drawn before on that day.

Each lottery day is i.i.d. from a multivariate hypergeometric distribution, so we can apply one of these common goodness-of-fit tests when each day contributes a single event. However, another rule-of-thumb for the χ^2 goodness-of-fit test is that each category should have an expected count ≥ 5. Directly counting the frequencies of each unique combination would not satisfy this rule-of-thumb. To handle this, we could compute a function of lottery draws and use that to group lottery draws together: for example, we could compute the sum of the 6 Lotto numbers from a single drawing, and count the frequencies of the sums being in specific bins. Then, we could safely apply the χ^2 goodness-of-fit test. However, it is unknown how well this approach may detect human tampering. Prior research has shown that humans do a poor job of reproducing the distribution of d, so we felt that it was more appropriate for tampering detection. However, this approach may still be useful for a player of the lottery: a deviation from the expected distribution of sums may guide a player to choose a combination of numbers that is more likely than would be expected under a "fair" null hypothesis. We provide an implementation of this apporach in R.

References

[1] Konstantinos Drakakis, Ken Taylor, and Scott Rickard. A statistical test to detect tampering with lottery results. 01 2009.

[2] Johannes Friedrich. Lotto number archive. https://github.com/JohannesFriedrich/LottoNumberArchive, 2023.

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

[4] Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone. *Handbook of Applied Cryptography*. CRC Press, 2001.

[5] Lawrence Bassham, Andrew Rukhin, Juan Soto, James Nechvatal, Miles Smid, Stefan Leigh, M Levenson, M Vangel, Nathanael Heckert, and D Banks. A statistical test suite for random and pseudorandom number generators for cryptographic applications, 2010-09-16 2010.

[6] Robert G. Brown. Dieharder: A random number test suite. https://webhome.phy.duke.edu/ rgb/General/dieharder.php, 2004.

[7] Pierre L'Ecuyer and Richard Simard. Testu01: A c library for empirical testing of random number generators. *ACM Trans. Math. Softw.*, 33(4), aug 2007.

[8] 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.

[9] Crypto Stackexchange. How to compute the dataset size required by dieharder tests? https://crypto.stackexchange.com/questions/90076/how-to-compute-the-dataset-size-required-by-dieharder-tests, 2021.

[10] G. Marsaglia and Arif Zaman. Monkey tests for random number generators. *Computers & Mathematics with Applications*, 26:1–10, 11 1993.

[11] George Marsaglia, B. Narasimhan, and Arif Zaman. The distance between random points in rectangles. *Communications in Statistics - Theory and Methods*, 19:4199–4212, 01 1990.

[12] Linux manual. https://man7.org/linux/man-pages/man4/random.4.html.

[13] Thomas' Digital Garden. Myths about /dev/urandom. https://www.2uo.de/myths-about-urandom/.

[14] Dirk Eddelbuettel. Sts critique. https://github.com/eddelbuettel/dieharder/blob/master/NOTES, 2019.

[15] Nidhi Gupta and G. P. Biswas. Wep implementation using linear feedback shift register (lfsr) and dynamic key. In *2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011)*, pages 422–427, 2011.