7-2023

Dice: Blessed or Cursed?

Warren Campbell
Civil Engineering, warren.campbell@wku.edu

Cameron Miller
Western Kentucky University, cameron.miller014@topper.wku.edu

Xavier Simpson
Western Kentucky University, xavier.simpson77@gmail.com

Dice: Blessed or Cursed?

C. Warren Campbell
School of Engineering and Applied Sciences
Western Kentucky University

Cameron Miller
The Gatton Academy of Mathematics and Science
Western Kentucky University

Xavier Simpson
The Gatton Academy of Mathematics and Science
Western Kentucky University

Abstract:

Every year, hundreds of millions of dice are manufactured and sold. Because of the impossibility of precise dimensional control and nonuniform density, none of these dice are fair. Polyhedral dice manufactured for role-playing games like Dungeons and Dragons typically contain 4-sided, 6-sided, 8-sided, 10-sided, 12-sided, and 20-sided dice (D4, D6, D8, D10, D12, and D20). D20s are especially problematic. In 3000-roll tests of several D20s, only about one-quarter tested fair. In light of the inherent unfairness of most dice, we explored the possibility of using dice mechanics involving multiple dice to obtain fairer results. For D20s, summing three dice gave promising results. Even using dice that tested highly unfair individually, sums of three dice tested fair. We also considered Fate or Fudge dice mechanics which effectively use the sum of 4 D3s. With one exception, these dice tested fair. In our tests, three D20s tested fairer than four Fate dice.

Background:

According to Hasbro, 50M copies of Yahtzee are sold each year. Each game includes at least 5 dice. One game uses one-quarter of a billion dice each year. Millions more polyhedral dice are sold for use in role-playing games like Dungeons and Dragons, and Pathfinder. None of these dice are fair because they cannot be manufactured with precise dimensions and uniform density. Casino dice which are controlled by Federal and state gaming laws come the closest. The cubes of these D6 dice are machined to within a few ten-thousandths of an inch, and the material of the pips (spots) has the same density as the body of the die. Six-sided dice are probably the easiest to manufacture precisely.

Every year gamers in tens of thousands gather at gaming conferences across the world. In 2019, Gen Con, the largest gaming conference in North America, had 70,000 attendees. At this conference at least 18 kiosks in the exhibit hall primarily sold dice. One topic consistently discussed at these conferences and on numerous Internet sites is the fairness of dice. Are dice cursed, roll consistently low, or blessed, roll consistently high? At least one company, Precision Play Dice, is going to extreme measures in an attempt to manufacture fairer dice.

This study addresses two questions: 1) What is the best method of testing dice and identifying the most unfair in the fewest number of rolls? 2) What dice mechanics using unfair dice can produce fair results?

Best Single Dice Test Methods:

We tried three different methods and applied them to testing dice. They were the running chi-square test, the modified Kolmogorov-Smirnov (KS) test, and a double binary test. The KS test is strictly for continuously distributed random variables, which dice rolls are not. So the KS test must be modified for use with discretely distributed variables like dice rolls. It uses the maximum absolute value of the difference between the fair die cumulative probability distribution function, and the unfair die cumulative probability distribution function. This problem has been solved mathematically, but the results are difficult to apply. For this reason, we simulated billions of D20 dice rolls to determine the 95 percent value of the KS statistic for different numbers of rolls. These are given in Table 1. If this data is fitted with a power law, the result is given by Equation 1. The fit has an r^2 value of 1.

$$KS_{95} = 1.1608\, rolls^{0.493} \quad (1)$$

Table 1. Kolmogorov-Smirnov 95 percent statistic for D20 dice

Rolls	ks95
100	0.12
500	0.054
1000	0.039
2000	0.0275
3000	0.022333

Of the three methods tested, the running chi-square method was the best. The chi-square statistic is given by Equation 2.

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i} \qquad (2)$$

O_i = number of observations of outcome i

E_i = expected number of observations of outcome $i = p_i N$

p_i = the fair probability of outcome i

N = the total number of observations (rolls)

k = total number of possible outcomes

We refer to outcomes rather than the number of sides of the dice because when summing dice rolls, there are more possible outcomes than sides of the dice. For example, when rolling two D6 dice, there are eleven possible outcomes (2 through 12) though each die has only six sides.

The difference between a normal chi-square test and a running chi-square test is that we calculate the statistic after each roll, typically from 1 roll to 3000 rolls. Campbell and Wimsatt (2022) showed that for an unfair die, the chi-square statistic has a linear trend. See Figure 1 for a typical example for a D20. For multiple dice, it is easy to show that the slope of the trend line is given by Equation 3.

$$m = \sum_{i=1}^{k} \frac{\delta_i^2}{p_i} \qquad (3)$$

m = the slope of the running χ^2 trend line

δ_i = the deviation of the unfair dice probabilities from p_i

This is a slight generalization of the Campbell and Wimsatt result in which all of the $p_i = 1/s$ where s is the number of dice sides. Equation 3 applies to single rolls or multiple dice roll sums. It is easy to demonstrate its correctness by taking 3000 rolls of a die or dice and calculating the best estimate of the outcome probabilities by dividing the number of observations of each outcome by 3000. If the results are then multiplied by different numbers of rolls the calculated χ^2 values for each falls perfectly on a straight line. In Excel, you can add a trendline and show the equation of the line. The slope of the trendline will match the slope calculated with Equation 3. See Figure 2. In the figure, the actual slope is 0.0001803421, and if you use this and calculate the χ^2 values they will match to six decimal places. The figure corresponds to four Fate dice that were rolled 3000 times.

Figure 1. A typical example of a running χ^2 graph.

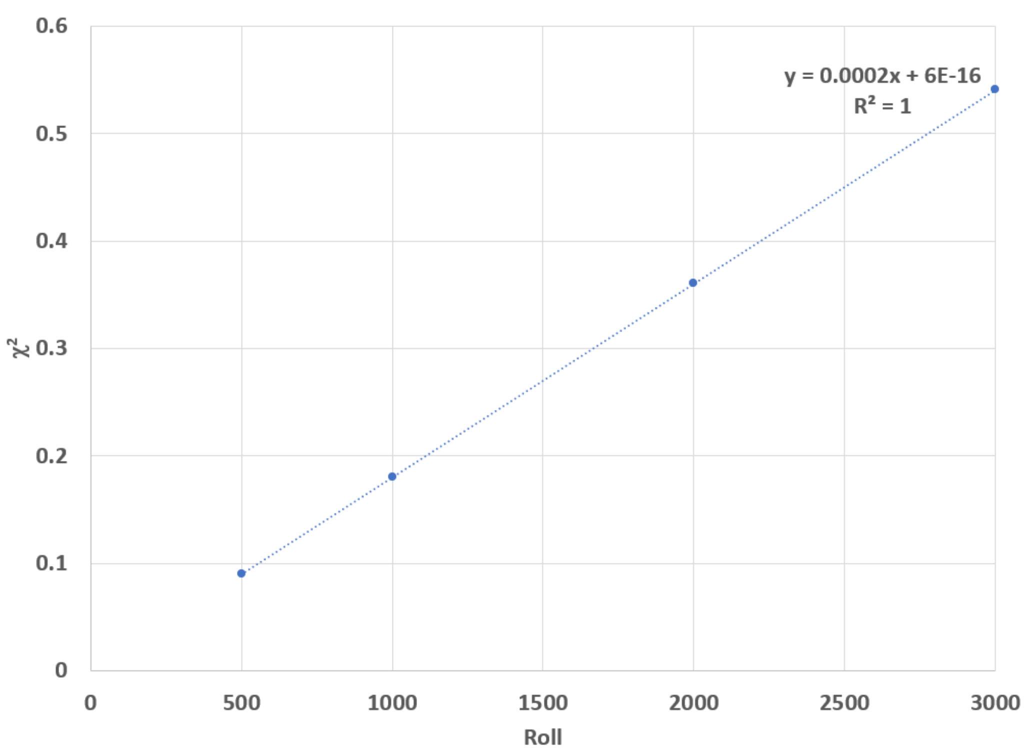

Figure 2. Comparison of the calculated slope of the χ^2 trendline with the fitted trendline.

Multiple Dice Mechanics

Summing dice rolls were considered because of the central limit theorem. No matter how unfair the dice are, the sum of enough of them always has a probability distribution that approaches a normal distribution. We have tested loaded D6s that rolled one number more than 80 percent of the time. To illustrate the approach of loaded dice like this and how their sum approaches the normal distribution, we took a hypothetical D6 and calculated probability distributions of sums of 2, 4, 8, 16, 32, and 64 of them. The closer the original single-die distribution is to the normal distribution, the faster the approach. This is illustrated by Figure 3 which shows a loaded D6 distribution, and distributions for sums of these dice. The figure shows the distributions for the loaded dice in blue and for the fair dice in orange. While both distributions approach normal, their means and standard deviations are significantly different. For the loaded die the side probabilities were 1 – 0.0, 2 – 0.0, 3 – 0.0, 4 – 0.1, 5 – 0.1, and 6 – 0.8. The mean roll for a fair die is 3.5 and for the loaded die is 5.7. The mean for 64 fair D6s is 224 and for the loaded D6s is 364.8. Similarly, the standard deviation for a single fair die is 1.708, while for the loaded die is 0.64. For the 64 loaded dice, the χ^2 value was 1,630,000 as compared to the 95-percent value of 91.67. Determination of these values was obtained by the usual procedure of making sure the expected value was no less than 5 for any one bin and does not include any randomness. We simply multiplied the probability of each roll by 3000 rolls.

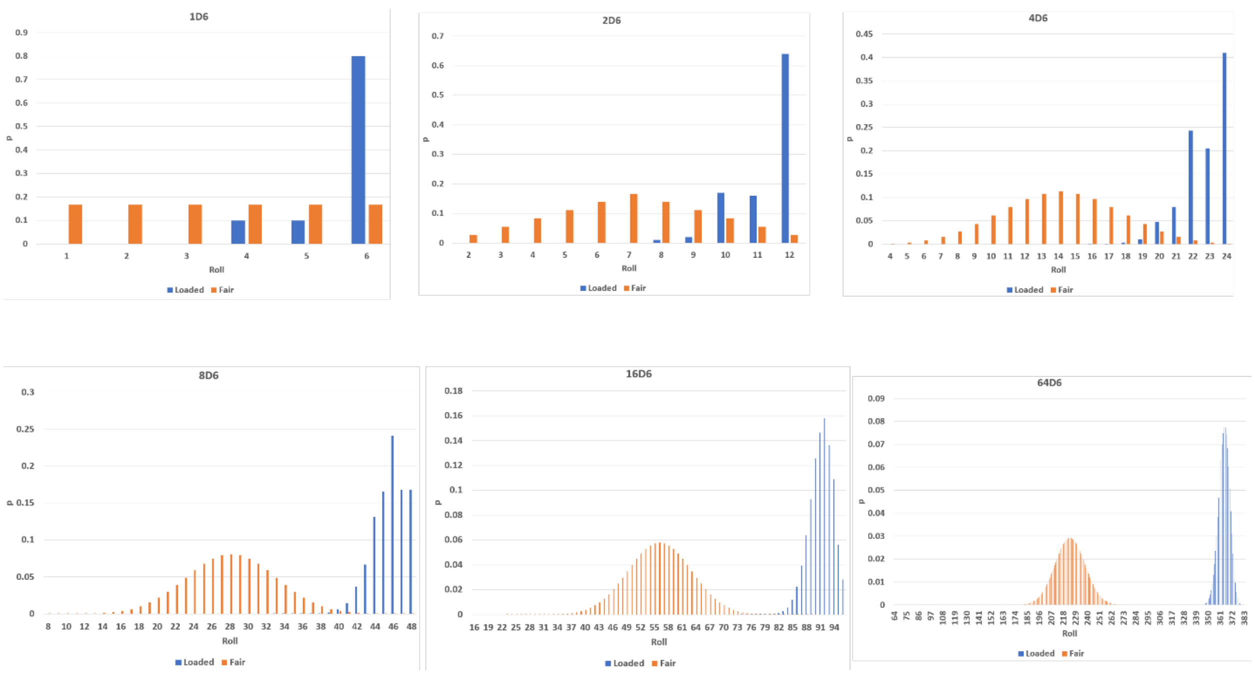

Figure 3. The approach of dice sum-distributions to normal for fair D6s and for loaded D6s.

We also calculated probabilities for a centered loaded die with a mean of 3.5 for the roll of a single die. The side probabilities were rearranged so that the non-zero probabilities were 3 – 0.8, 5 – 0.1, and 6 – 0.1. This makes quite a difference in the χ^2 value. The 64 dice sum had a χ^2 value of approximately 919.

While it is not something most people would want as a dice mechanic, rolling three D20s and summing results offers some promise for fairness. We had several D20s that had been rolled 3000 times each. Table 2 summarizes the running chi-square values for each set of 3. In the table, the first 2 dice identification letters identify the manufacturer, e.g. Ch means Chessex. The second set of two letters identifies the color of the dice body usually using the first and last letter of the color. For example, Be represents blue and Mr multicolor. The next set of two letters identifies the color of the numbering using the same color system as the body colors. The next letter identifies the material of the die, X for plastic translucent, O for plastic opaque, and M for metal. The number identifies the number of sides of the die. The last two letters identify the method of rolling, HR for the High Roller automated roller, and Hand for hand rolling with a dice tower.

The 95 percent value of χ^2 for a single D20 is 30.14 and for the sum of 3 D20s is 75.62. In each of the six sets tested, at least one die rolled unfair after 3 rolls. The three dice of set 6 had the highest χ^2 values after 3000 rolls of any dice we tested. The value of p for the sum (last column in Table 2) would have to be less than 5 percent for us to reject the null hypothesis that the three dice rolled unfair. Sets 1 and 3 tested the closest to unfair. The other 4 sets had p-values greater than 50 percent, including set 6.

The set 6 running χ^2 curve for the sum is given in Figure 4. The curve flattens out and has no discernable trend which is characteristic of fair single-dice curves. Considering the unfairness of the three dice, this result is surprising.

Table 2. Summary of 3D20 dice sum χ^2 tests

Set	Dice	χ^2 Value	Sum χ^2	p
1	ChBeWeX20 HR	21.73	65.89	19.6%
	ChRdWeX20 HR	10.88		
	NFGyBkM20 HR	35.56		
2	ChPeWeO20 HR	22.39	53.12	62.1%
	ChBeWeX20 HR	27.69		
	WzBkGdO20 HR	34.03		
3	BnOeBkX20 HR	43.41	61.25	32.6%
	BnBkWeO20 HR	20.13		
	GSGyWeX20 HR	24.29		
4	GSBeWeX20 HR	15.21	52.60	64.1%
	MPRdGdX20	37.64		
	MPWeMrO20	34.25		
5	MPRdGdX20 Hand	21.04	51.12	69.4%
	MPRdGdX20 HR	37.67		
	MPRdBeX20	78.08		
6	DHBkGdM20 Hand	107.43	54.53	56.8%
	DHBkGdM20 HR	371.72		
	WzBkGdO20 Hand	90.84		

Figure 4. Running χ^2 statistic for three D20 dice each with χ^2 values greater than 90.

Summing the rolls for three D20s is not convenient and would require a calculator for most people to ensure an accurate sum. However, the Fate dice mechanic uses 4 D6s and is easy to sum. This led us to consider the Fate dice mechanic as a possible fair approach. To that end, we rolled 4 sets of Fate dice 3000 times each and recorded the roll and the sums. Figure 5 shows a set of 4 Fate dice. Each die has two sides with a +, two blank sides, and two sides with a minus. A plus counts as +1, a blank as 0, and a minus as -1. Of the 16 dice rolled, only one tested unfair after 3000 rolls.

We calculated the χ^2 value of the sum for each of the four sets of dice. They were 12.31, 5.16, 5.01, and 19.62 as compared to the 95 percent value of 15.51. The last result was surprising since all four dice tested very fair individually. See Figures 6 and 7.

Figure 5. Fate or fudge dice.

Figure 6. Individual Fate dice χ^2 curves.

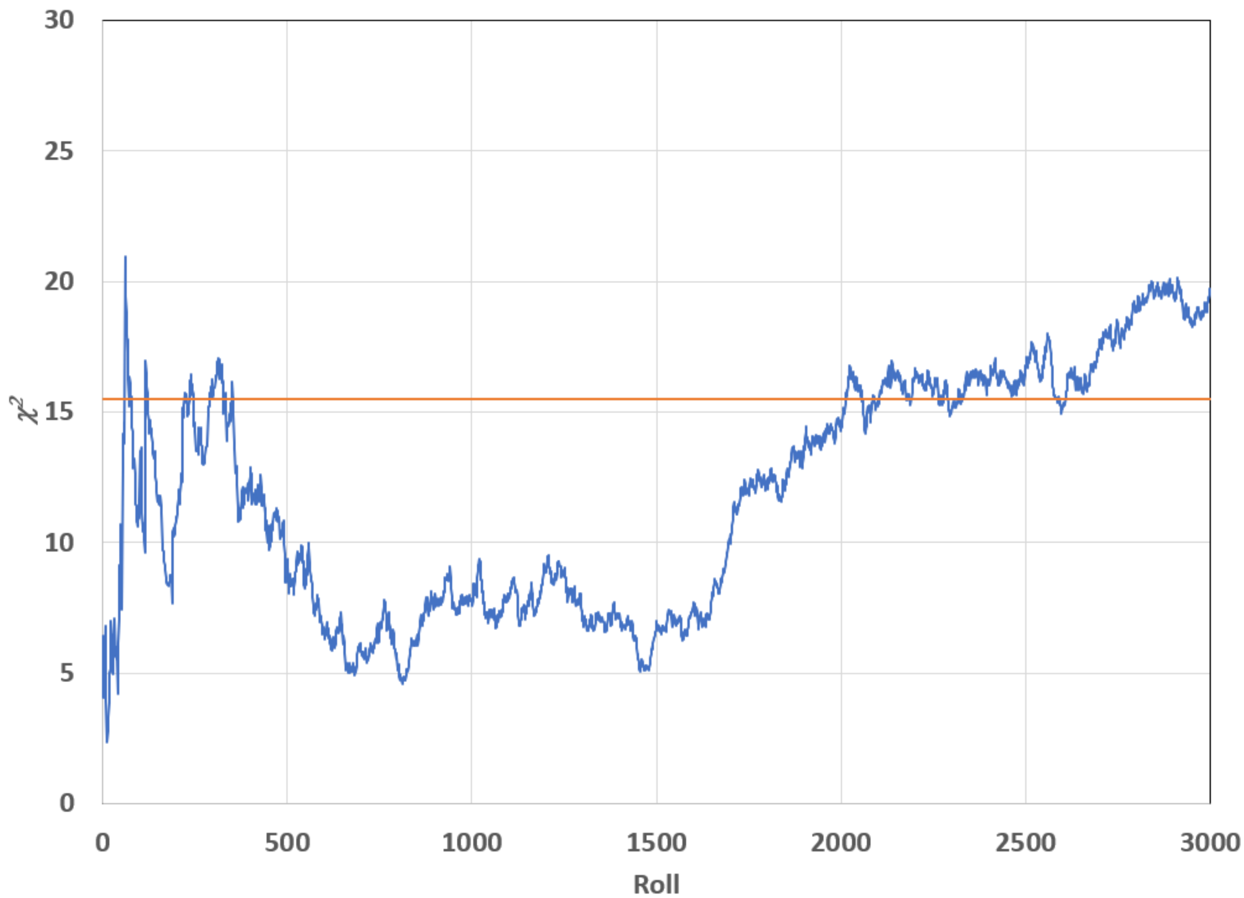

Figure 7. Fate dice sum running χ^2 curve

The results were surprising, so we considered whether having fair Fate dice rolls test collectively unfair was possible. We checked and rechecked the calculations. It is possible. For example, suppose we let die 1 (D1) roll +1 for a thousand rolls, D2 and D3 roll 0, and D4 roll -1. For the next thousand rolls, let D1 and D4 roll zeros, D2 roll +1s, and D3 roll -1s. For the last thousand rolls, let D1 and D2 roll -1s, and D3 and D4 roll +1s. Each of the four dice has 1000 +1s, 1000 0s, and 1000 -1s for individual χ^2 values of zero. Meanwhile, for all 3000 rolls, the sum is zero giving the sum χ^2 value of 9789.

It is also instructive to look at the individual running χ^2 curves (Figure 8). The individual running χ^2 curves for the 4 dice plot as exactly the same curve. They rise to $\chi^2 = 2000$ in the first 1000 rolls, then with two loops work down to zero at 3000 rolls. Meanwhile, the sum χ^2 curve goes linearly to the final value of 9789 (Figure 8). The critical value of 15.51 is indistinguishable from zero on the chart.

Returning to Figure 7, the final χ^2 value is 19.62 which corresponds to a *p*-value of 1.19%. It doesn't fall into the category of impossible, but it does call into question the hypothesis that Fate dice sums are more fair. To explore this result in greater depth, we took the 3000-roll probability estimates for the four dice, and simulated 3000 rolls several times (Figure 10).

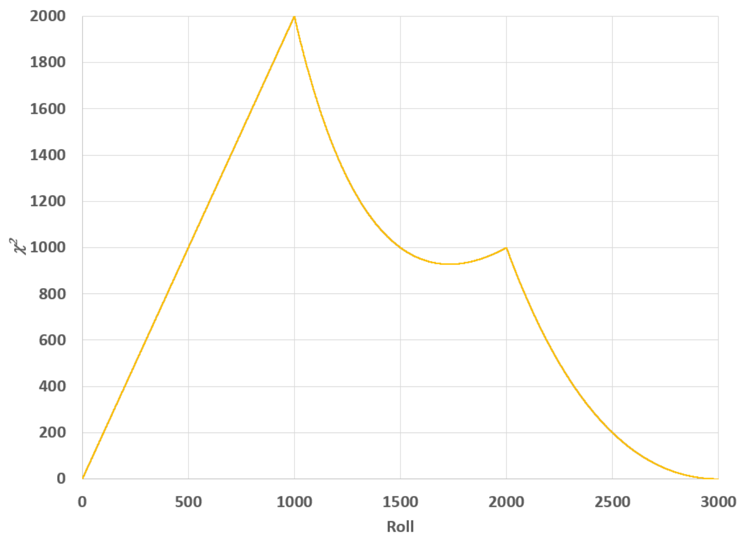

Figure 8. Hypothetical Fate dice rolls plot as a single χ^2 line.

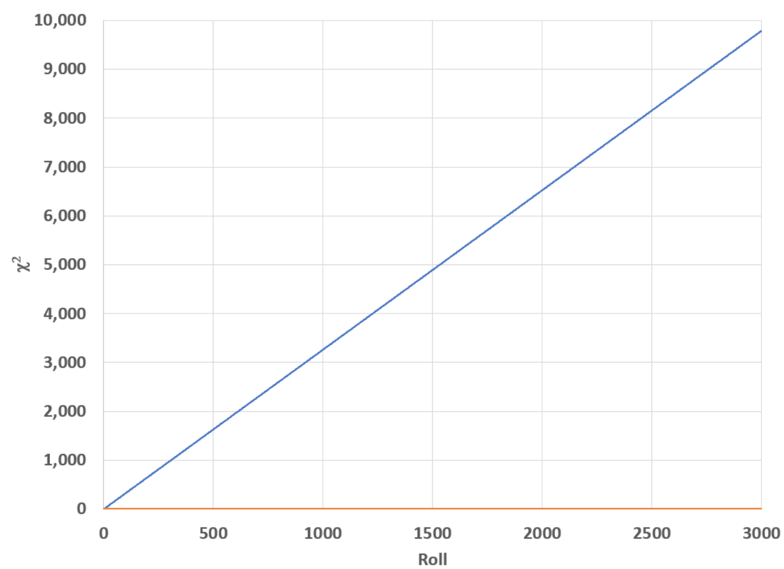

Figure 9. Hypothetical 4 Fate dice sum χ^2 curve.

The four Fate dice that had fair individual dice, but unfair sum was simulated using the individual dice 3000-roll probability estimates. All of the sums were fair after 3000 four-dice rolls as shown. However, the figure shows significant random excursions even as the number of rolls approached 3000. One curve climbed to twice the critical value, then barely dipped under the critical χ^2 value right at the end. From a gamer's perspective, it would be easy to conclude that the dice were not fair.

From this testing, it appears that summing 3D20s is a fairer dice mechanic than summing 4 Fate dice.

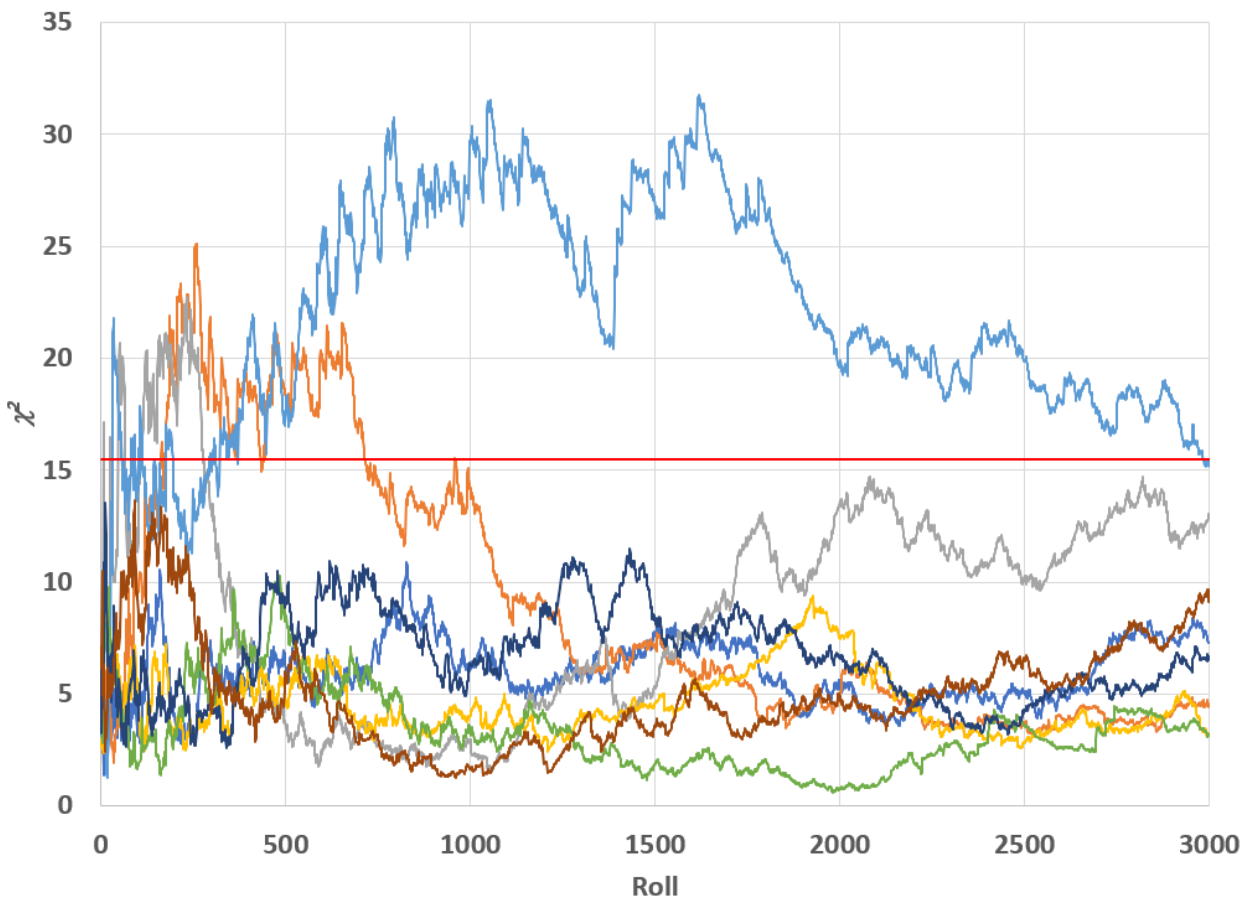

Figure 10. Simulated sum rolls of 4 Fate dice.

Power of Dice Testing

One of the purposes here was to find an easier way of testing dice. Unfortunately, we found that nothing was better than large numbers of rolls of dice testing with the chi-square test. Campbell and Dolan (2019 and 2020) showed that float testing of dice is not effective because the dice side that comes up in a float test is uncorrelated with the side that comes up most often in a roll test. They concluded that the dynamics of a float test are radically different from that of rolling dice.

We also tried a modified Kolmogorov-Smirnov test for dice, but the running χ^2 test consistently outperformed the KS test. We also tried a double binomial test, but again the running chi-square test had the better power. Power is defined for dice as the probability that if dice are unfair, they are identified as unfair. That is, 1 – the probability of a type II error. This is an important question because of the widespread misinformation available on the Internet where a D20 is rolled 100 times and conclusions drawn about its fairness or unfairness. Figure 11 illustrates the uselessness of rolling only 100 times, calculating the chi-square statistic, and drawing conclusions from that test. The dice in question had a chi-square value of 35.56 after 3000 rolls. The die tested slightly unfair. The power of the test is estimated at 8%. This means that only 8 times out of a hundred would you conclude this die is unfair after rolling it only 100 times. If increased to 500 rolls, Figure 12 is the result. There is some separation in the fair and unfair probability distributions. The fair die χ^2 distribution is also ragged because the rolls of fair dice are being simulated as well. However, the power is only 24%. You have about a 1 in 4 chance of correctly identifying this die as unfair.

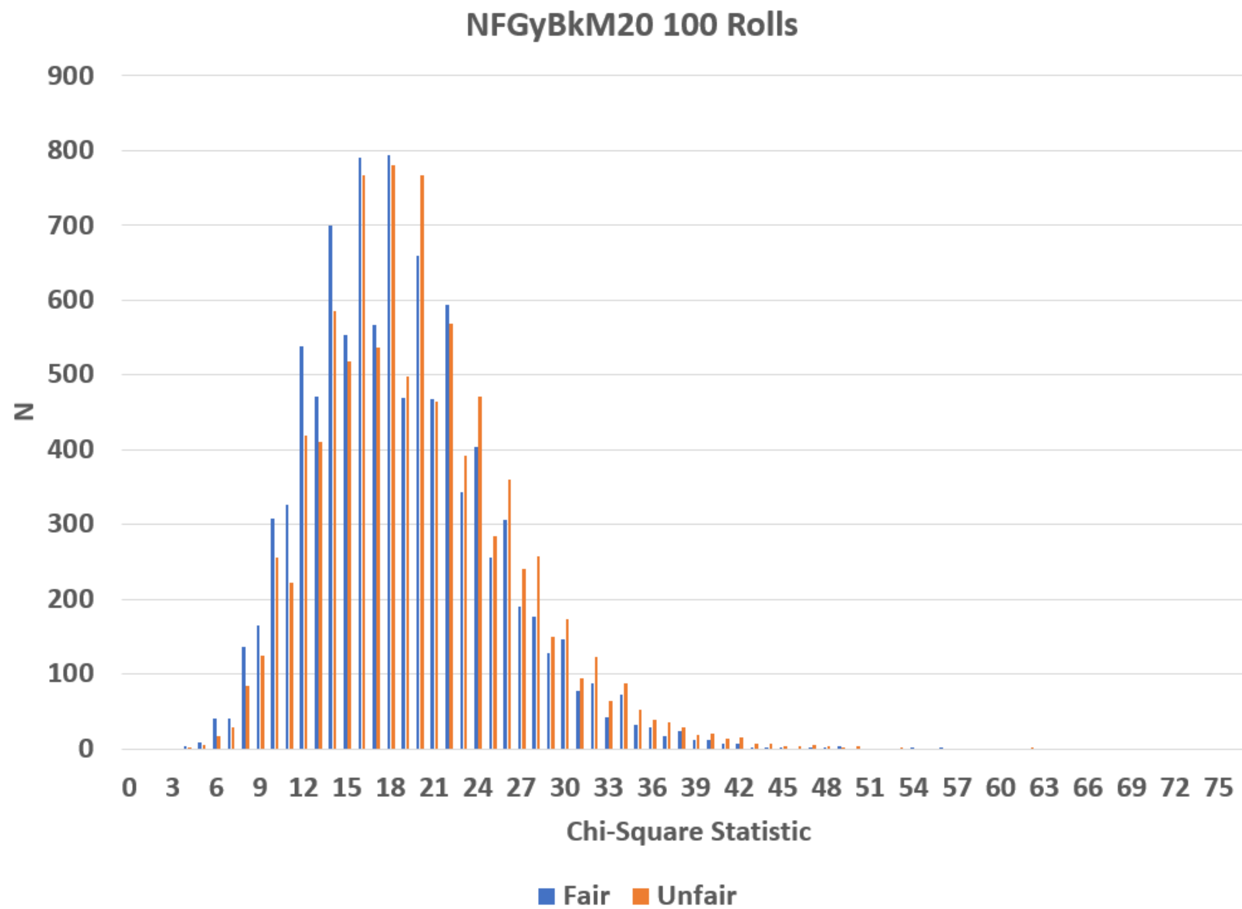

Figure 11. Chi-square distribution from 10,000 simulations of 100 rolls of a D20

Figure 13 provides the same curves for 1000 rolls. Now the power is 51% so you have a coin toss chance of correctly determining that the die is unfair after 1000 rolls. At 2000 rolls, you have an 87% chance. At 3000 rolls, you have a 98% chance of correctly identifying the die as unfair. Figure 14 shows that chi-square distribution. Notice that as the number of rolls increases, the mean of the chi-square distribution moves away from that of the fair die and the spread of the distribution also increases. For a die that is slightly unfair, 3000 rolls are needed to achieve a 98 percent level of confidence in your determination that the die is unfair.

One hundred rolls of dice are clearly not sufficient to determine the fairness or unfairness of dice. This can be seen in Figure 15 which is the running χ^2 chart for our most unfair die. The curve only crosses the critical line and stays above it after 442 rolls.

Figure 16 illustrates how the probability distribution of χ^2 changes as a function of the number of rolls. Each plot is based on 10,000 simulations of the number of rolls indicated in the legend. Figure 16a is for the most unfair die with a chi-square value of 371.72 after 3000 rolls. Figure 16b is for a die with a chi-square of 35.56. The mean of χ^2 increases as does the standard deviation. Figures 17 and 18 show a plot of the change in the mean and standard deviation as a function of the number of rolls. The mean changes linearly in accordance with Equation 3.

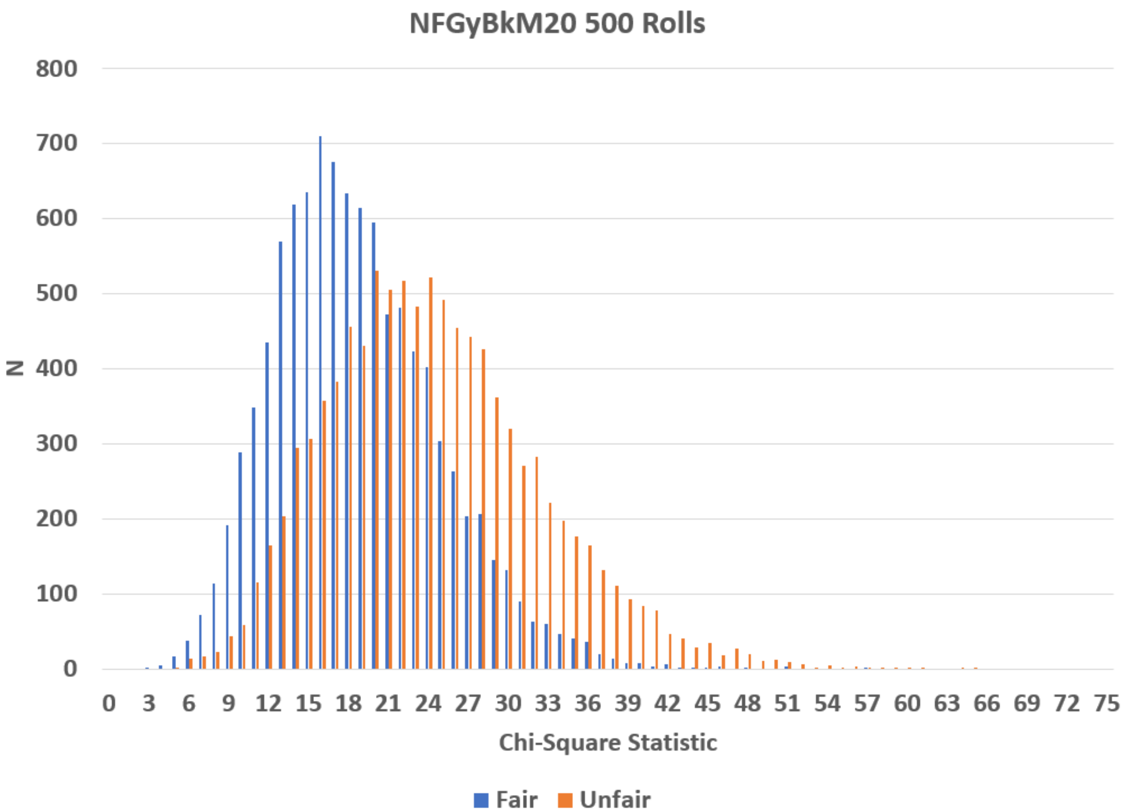

Figure 12. Chi-square distribution from 10,000 simulations of 500 rolls of a D20

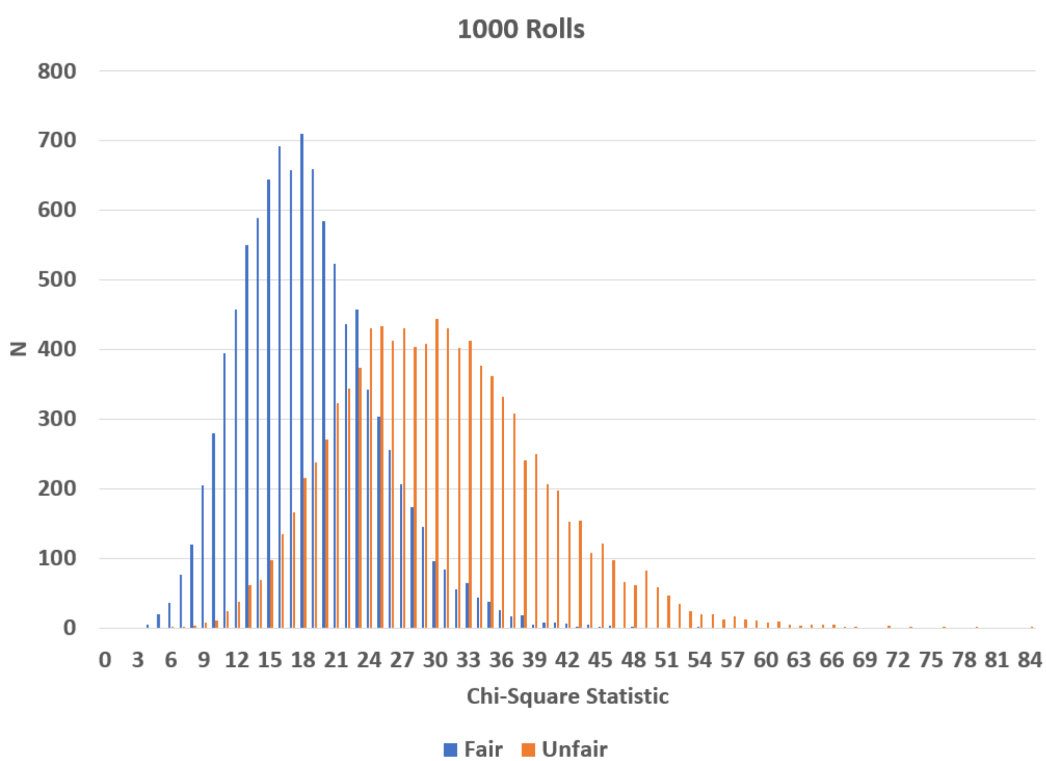

Figure 13. Chi-square distribution from 10,000 simulations of 1000 rolls of a D20

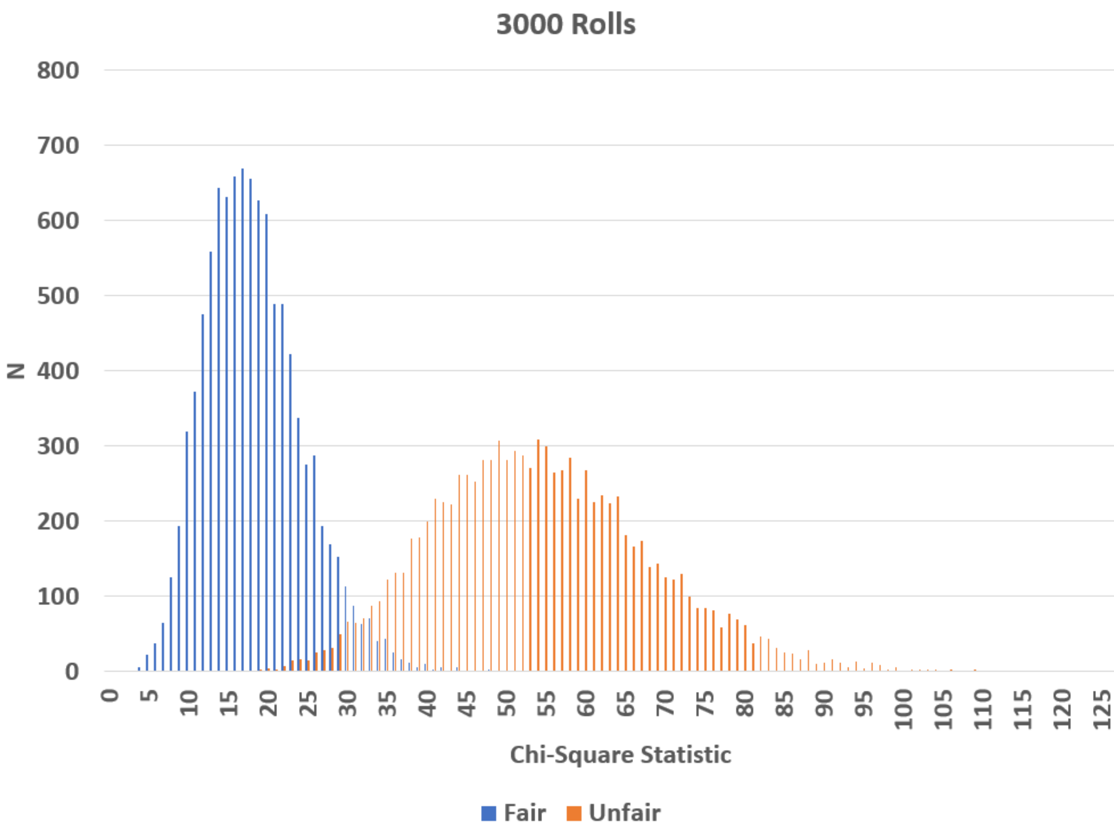

Figure 14. Chi-square distribution from 10,000 simulations of 3000 rolls of a D20

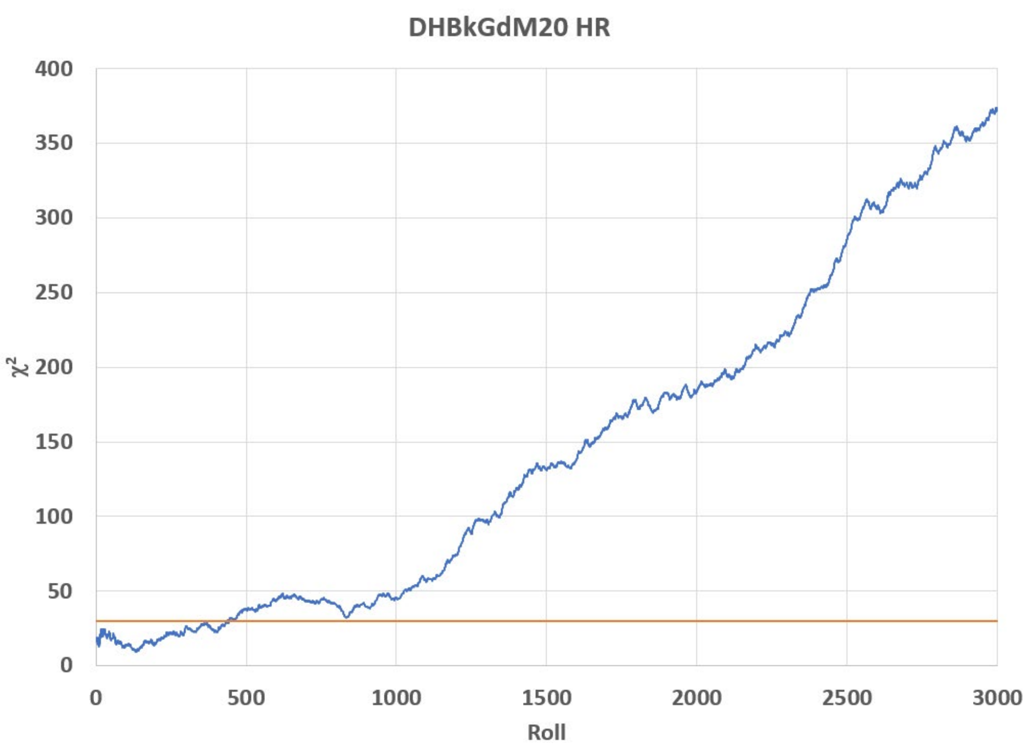

Figure 15. Running χ^2 chart for the most unfair die tested.

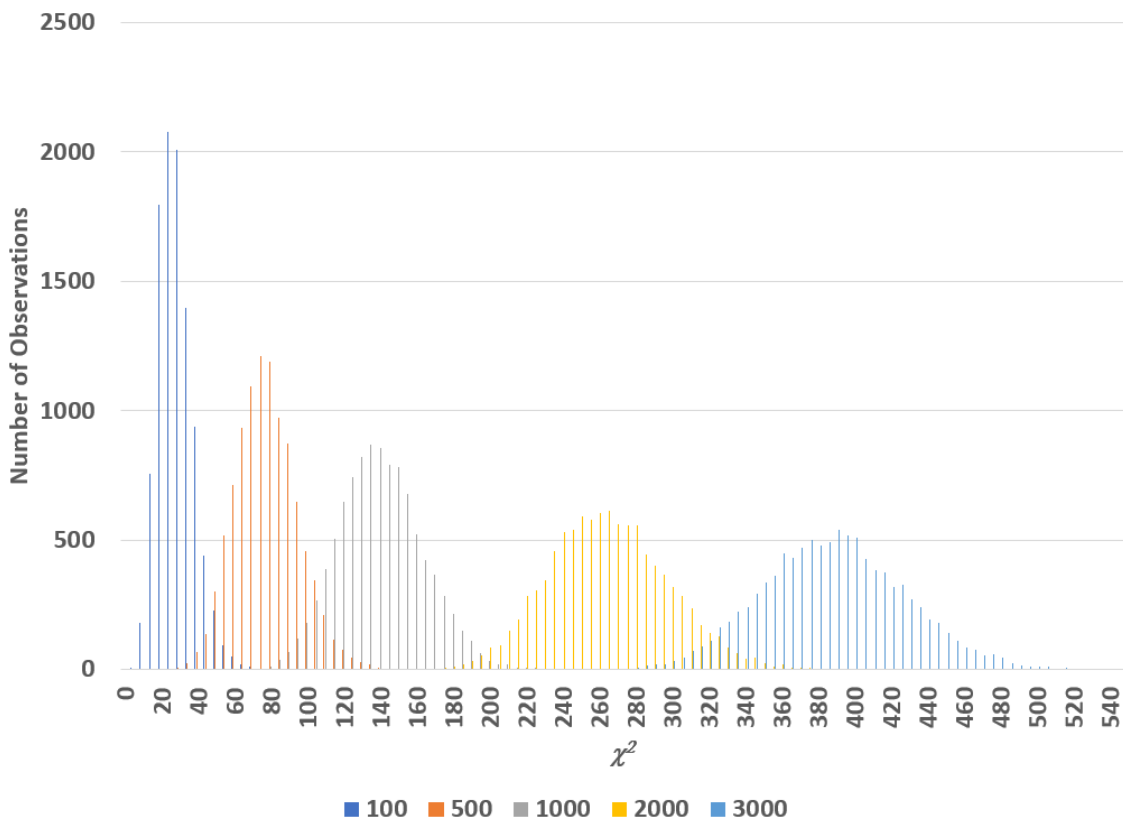

Figure 16a. Variation of χ^2 probability distributions with the number of rolls for DHBkGdM20 HR

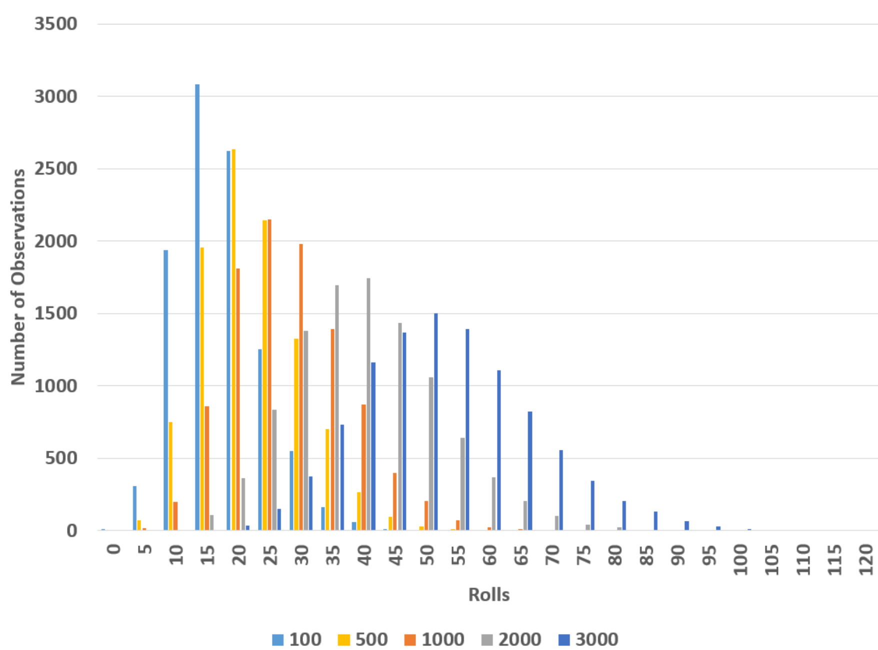

Figure 16b. Variation of χ^2 probability distributions with the number of rolls for NFGyBkM20 HR

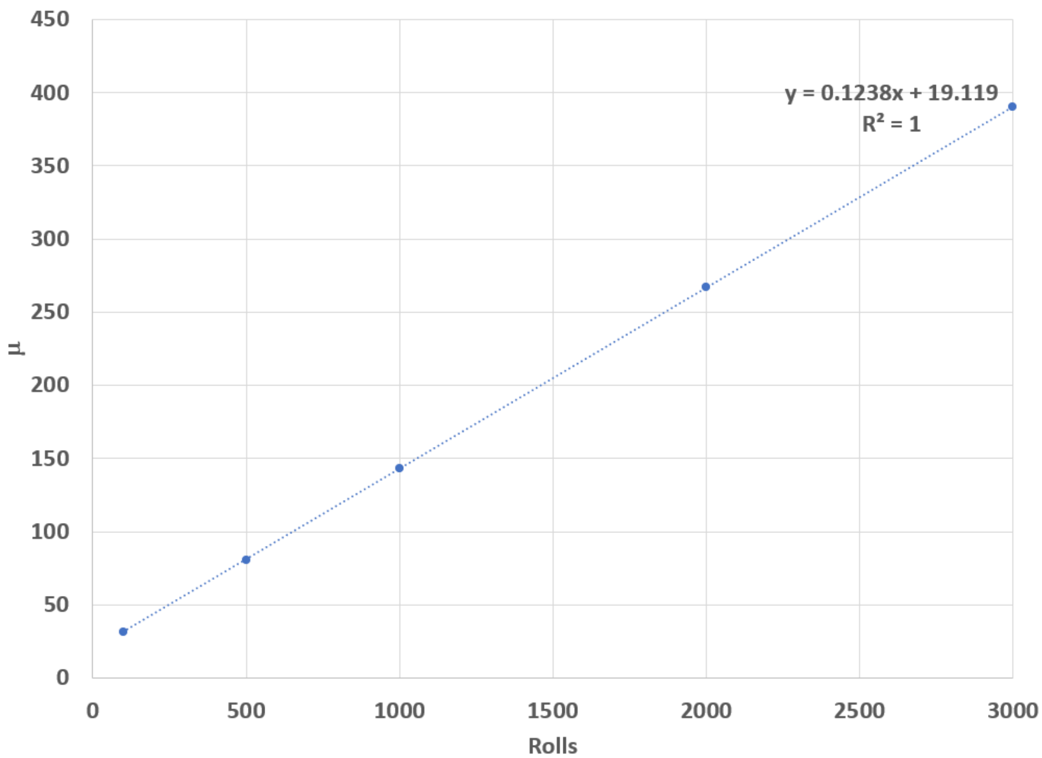

Figure 17. Variation of the mean with the number of rolls for DHBkGdM20 HR

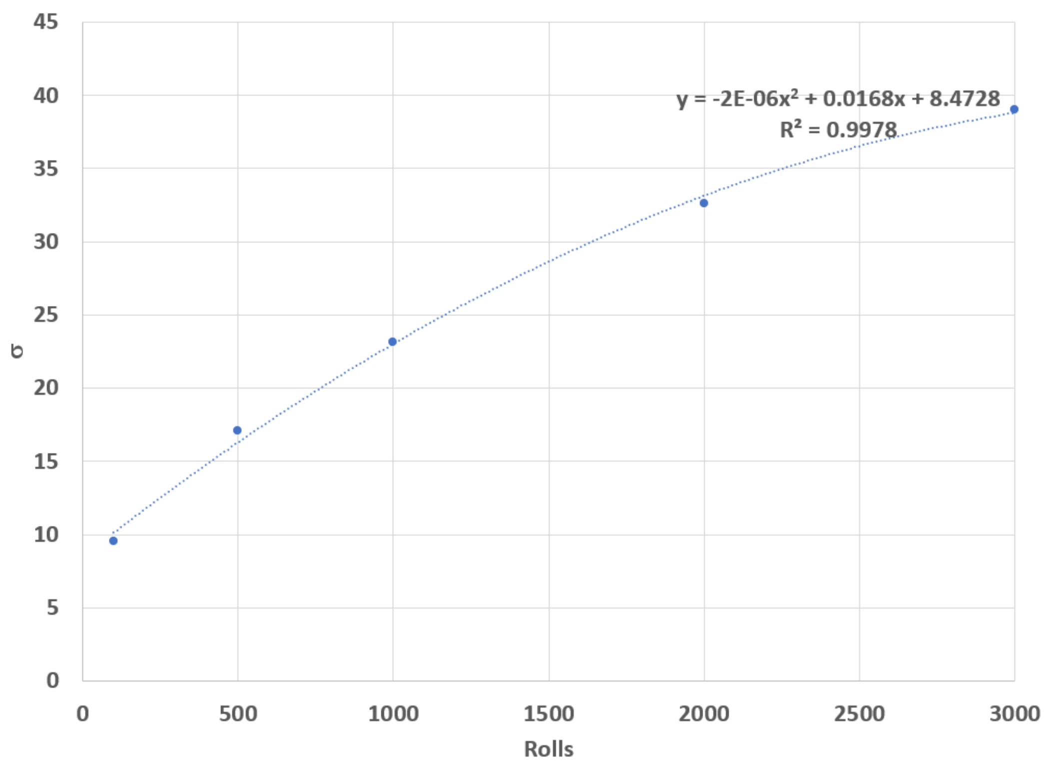

Figure 18. Variation of the standard deviation with the number of rolls for DHBkGdM20 HR

Summary

The running chi-square test was the best of the three methods we tried for testing dice. It almost always had better power than the modified Kolmogorov-Smirnov and the double binomial tests.

Because the distribution of the sums of multiple dice rolls approaches the normal distribution, dice mechanics involving sums generally do better than dice mechanics involving one unfair die. The sum of six sets of 3D20s tested fair after 3000 rolls in every case tested, including one set with the three most unfair dice tested.

Four sets of Fate dice were rolled 3000 times and the rolls of each die were recorded. Of the 16 dice, only one tested unfair. However, the roll sums of it and the other 3 Fate dice tested fair. However, individually all four dice of one set tested with very low values of chi-square after 3000 rolls, but their sums tested unfair. Simulations of Fate dice rolls show significant variation of the chi-square statistic even after 3000 rolls. While the standard deviation of the chi-square distribution is $(2\nu)^{1/2}$ where ν is the number of degrees of freedom, the chi-square distribution of unfair dice is much broader after 3000 rolls.

For nominally unfair dice, 100 rolls are not sufficient to test a D20 die for unfairness using the chi-square test. However, 3000 rolls are sufficient to obtain a high level of confidence in the unfairness determination. The probability distribution of χ^2 changes with the number of rolls of an unfair die, though it does not change for a fair die. The mean of the distribution increases linearly with a slope given by Equation 3. The standard deviation also increases with the number of rolls but with a gradually decreasing slope.

References

Campbell, Warren and Wimsatt, Hunter, "Dice Testing with the Running Chi-Square Distribution" (2022). SEAS Faculty Publications. Paper 7. https://digitalcommons.wku.edu/seas_faculty_pubs/7

Campbell, Warren and Dolan, William P., "Dice Questions Answered" (2020). SEAS Faculty Publications. Paper 2. https://digitalcommons.wku.edu/seas_faculty_pubs/2

Campbell, C. Warren and Dolan, William P., "Dice Mythbusters" (2019). Student Research Conference Select Presentations. Paper 48. https://digitalcommons.wku.edu/sel_pres/48