Some Thoughts on Mean - Median

July 12, 2017

Abstract

The mean median difference test is an attractive gerrymandering standard for many reasons. It meets constitutional requirements and is arguably the simplest metric to calculate and conceptualize. While it does not work for states that are not close to even partisanship, the worst gerrymandering occurs in close states. These states are also of high strategic value for those seeking to take control of a statehouse in advance of redistricting.

Here we consider two issues related to the test, namely whether the standardized or unstandardized mean median difference should be used, and whether the sampling variance of mean median difference should be estimated from historical data or using the asymptotic variance from an assumed null distribution.

1 Introduction

The mean median difference test [1, 2, 3], fits with existing rulings since it measures partisan bias without regard to proportionality and is simple enough to be calculated by a judge without an expert witness. While it is only effective in measuring bias for states which are close to even in terms of partisanship, [2], close states account for a major portion of the gerrymandering that occurs in the US. The mean median difference statistic can be standardized or unstandardized. We denote these two statistics as C_s and C_u respectively, and they are defined as follows

$$C_s = \frac{\bar{X} - \theta}{S} \tag{1}$$

$$C_u = \bar{X} - \theta \tag{2}$$

Where $\bar{X}$ is the mean district vote percentage, θ is the median district vote percentage, and S is the standard deviation of the district vote percentages. The standardized statistic C_s is therefore simply C_u divided by the standard deviation.

In section 2, we consider how the variance for estimating p-values of these statistics should be calculated. Wang has proposed the asymptotic variance for a unit normal distribution for C_s, which was derived by [4]. Similar calculations for C_u can be found in [4] and [5]. We consider two separate but related aspects of this. First, if we accept the assumption of independent and identically distributed districts that is required for deriving an asymptotic variance, then what is the most appropriate null distribution to use? Second, we consider whether an asymptotic variance should be used with a null distribution at all. Districts vote percentages are neither independent nor identically distributed, but a normal distribution is still a reasonable assumption for the sampling distribution of C_s and C_u. As an alternative to an asymptotic variance, we estimate the variance using an empirical Bayesian simulation.

In section 3, we will show that C_u is preferable since C_s allows bad actors to manipulate the standard by increasing the standard deviation of the district vote percentages. Not only does this allow gerrymanderers to skirt a potential standard based on C_s, but it provides an incentive for them to create even more polarized districts, which is the result of increasing the standard deviation of district voter percentages. In contrast, C_u can not be manipulated in the same fashion.

2 The Sampling Variance

For a continuous symmetric distribution with density function f, mean μ, variance σ^2, and median ν, the asymptotic sampling variances σ_s^2 and σ_u^2 for $\sqrt{n}C_s$ and $\sqrt{n}C_u$ respectively are [4, 5].

$$\sigma_s^2 = 1 + \frac{1}{4\sigma^2 f^2(\nu)} - \frac{\tau}{\sigma f(\nu)} \tag{3}$$

and

$$\sigma_u^2 = \sigma^2 + \frac{1}{4f^2(\nu)} - \frac{\tau}{f(\nu)} \tag{4}$$

where

$$\tau = \mu - 2\int_{-\infty}^{\nu} x f(x)\, dx \tag{5}$$

These variances can be computed readily for common symmetric distributions, but can additionally be estimated from data by using a kernel density estimate of the data and numerical quadrature to approximate the above integrals. The asymptotic variance for the unit normal distribution is used quite commonly and proposed by Wang for use as a gerrymandering standard, but other distributions can be considered as well to try to better reflect voting data in actual elections. In addition we consider a uniform distribution on the unit interval, a uniform distribution on the interval .15 - .85, and a KDE estimate of all contested elections in US congress from 1972-2016, symmetrized about 0.5. The variances are tabulated in table 1 and KDE estimate for the voting data is in figure 1.

Table 1: Asymptotic variances for C_u and C_s for a few distributions

Distribution	σ_s^2	σ_u^2
Unit Normal	0.5708	0.5708
Uniform (0,1)	1.0	0.0833
Uniform (0.15,0.85)	1.0	0.0408
Symmetrized Voting Data	1.2022	0.0395

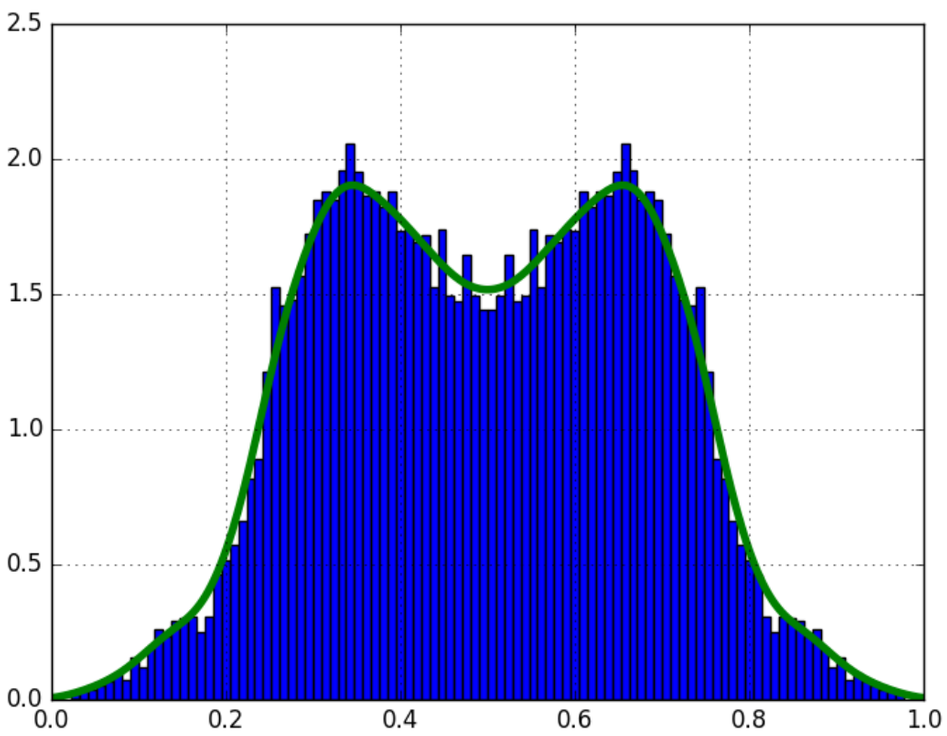

Figure 1: Histogram and KDE estimate of symmetrized voting data

An alternative procedure for estimating these variances is to simulate elections and compute the variance of C_u and C_s from the result. This is done by estimating a beta distribution for each district using the method of moments and a hyperprior for the variance, and simulating 10,000 elections. This removes the assumption of identically distributed districts and also removes the discrepancy between the asymptotic and actual sampling variance at small sample sizes. The sampling variances are then estimated by computing the variance that best fits the simulated variances. While the mean median difference is only effective for measuring gerrymandering for close states, here we are only interested in the variance of the statistic from election to election so close states and partisan states are both included in the fit. Plots are shown in figures 2 and 3, and the results of fitting are 0.2345 for C_s and 0.0057 for C_u. Both of these are substantially smaller than the asymptotic results from table 1, although for C_s the standard normal variance seems like a reasonable high end estimate. Additionally for C_s the smallest possible asymptotic variance for a unimodal distribution is 0.25. These results suggest that the sampling variance of the mean median difference, regardless if it is standardized or not, should be estimated from historical data rather than an asymptotic variance from some null distribution. The method used here combines historical data with an empirical model, but that is not the only possibility. More sophisticated models could be used, or no model could be used at all, and the sampling variance could be estimated directly from the data.

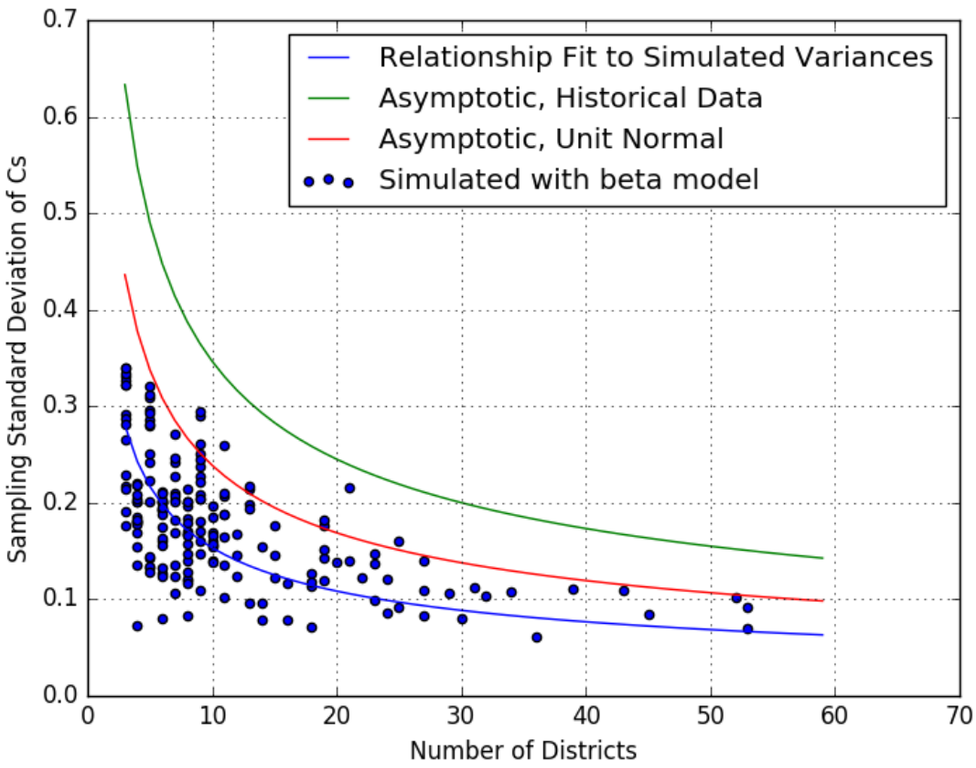

Figure 2: Asymptotic and fit variances for the standardized mean median difference

3 Gaming the Mean Median Difference

The adoption of any standard to reign in gerrymandering will lead to attempts to skirt around it. One approach to gaming C_s is to simply increase the standard deviation of the district vote percentages, which decreases the p-value of the test. This can be done while keeping C_s essentially unchanged. To illustrate this, we use a multi objective optimization technique, where we minimize the district standard deviation while maximizing the p-value for C_s using the asymptotic variance from the unit normal. We use 18 districts and require that one party get at least 14 safe wins, where a safe win is defined as having at least 55% of the vote. The mean of the district votes must lie within 0.5% of 50.0%, and the vote percentages in any district must be within 15% and 75%. At some point, decreasing the

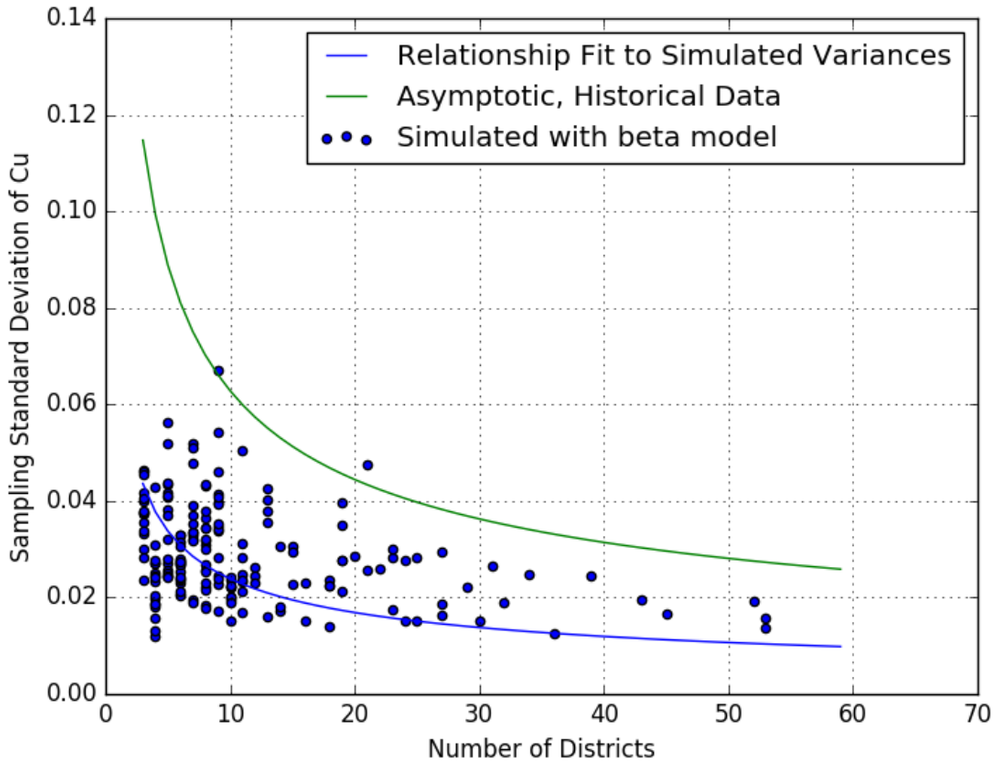

Figure 3: Asymptotic and fit variances for the unstandardized mean median difference

district standard deviations must come at the expense of also decreasing the p-value of the test. This results in a family of non dominated or Pareto optimal solutions representing the trade off between these two objective. The reason for this procedure is to show how C_s can be gamed solely by changing the standard deviation as opposed to some combination of changing C_u as well as the standard deviation, since we will look at attempting to game C_u in a moment.

Figure 4 shows the resulting Pareto front, and under the assumptions of this simple analysis, it is possible to safely win 14 of 18 seats (77.8%) safely with 50% of the vote while manipulating C_s to obtain a p-value above the usual threshold of 5%. Further, in each of the Pareto optimal solutions the mean median difference is essentially the same. This is shown in figure 5, where each pareto optimal solution for the vote percentages is plotted. Is a different sampling variance was adopted on the basis of the results of section 2, then this would improve the situation somewhat since the resulting standard would be stricter, but this change alone would not resolve the vulnerability of C_s based standards.

It is also evident that if the gerrymanderers were forced to reduce C_u then this would require them to either give up some of their safe wins or lower their safe win threshold from 55% to something lower. This is what we want, since it results in less bias and more competition in the state's elections. To illustrate this we again use optimization, but with a slightly different approach from before. Instead of multiobjective optimization, we do a series of single objective optimizations with different numbers of safe wins and safe win thresholds while trying to maximize the p value of the C_u test using the sampling variance from the previous section. For 18 seats, only 9 safe wins could be achieved while passing the test for C_u with a safe win threshold of 55%. In order to achieve more than 9 safe wins while passing the test, the safe win threshold would need to be reduced, making those wins less secure. This is shown in figure 6

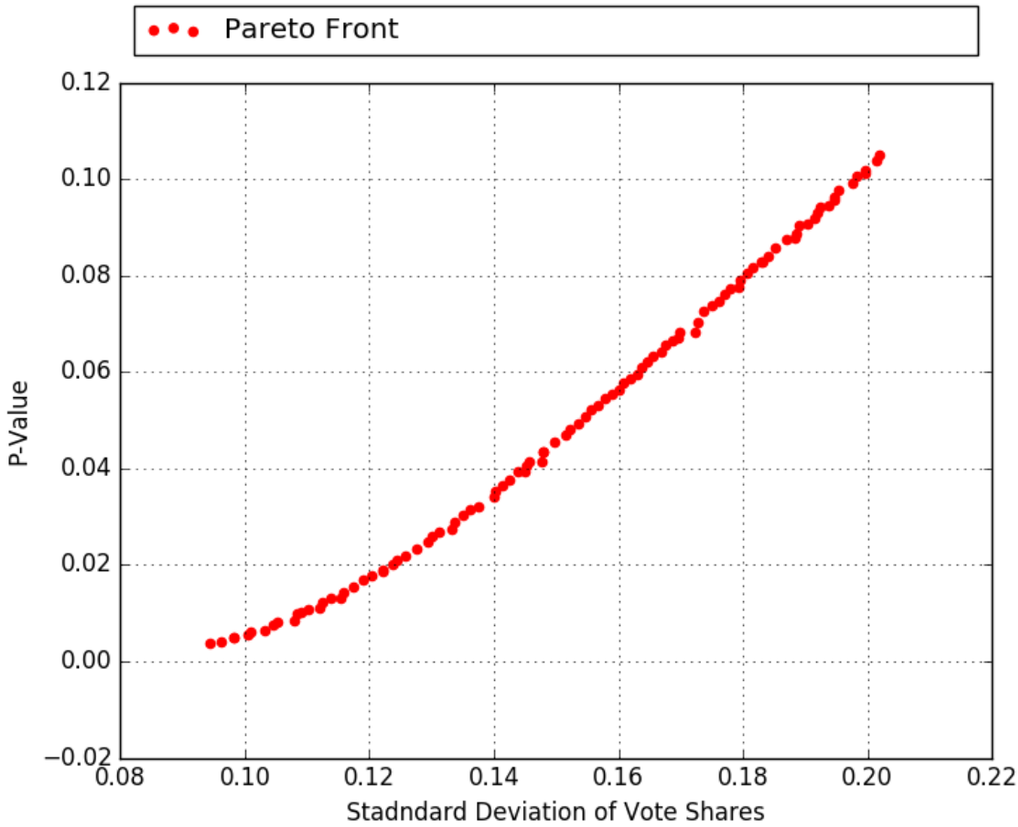

Figure 4: Pareto front for gaming the standardized mean median difference

Acknowledgment

References

[1] Samuel Wang. Three Tests for Practical Evaluation of Partisan Gerrymandering. *Stanford Law Review*, 68(6):1263.

[2] Samuel S.-H. Wang. Three Practical Tests for Gerrymandering: Application to Maryland and Wisconsin. *Election Law Journal: Rules, Politics, and Policy*, 15(4):367–384, October 2016.

[3] Michael D. McDonald and Robin E. Best. Unfair Partisan Gerrymanders in Politics and Law: A Diagnostic Applied to Six Cases. *Election Law Journal: Rules, Politics, and Policy*, 14(4):312–330, November 2015.

[4] Paul Cabilio and Joe Masaro. A simple test of symmetry about an unknown median. *Canadian Journal of Statistics*, 24(3):349–361, September 1996.

[5] Antonietta Mira. Distribution-free test for symmetry based on Bonferroni's measure. *Journal of Applied Statistics*, 26(8):959–972, December 1999.

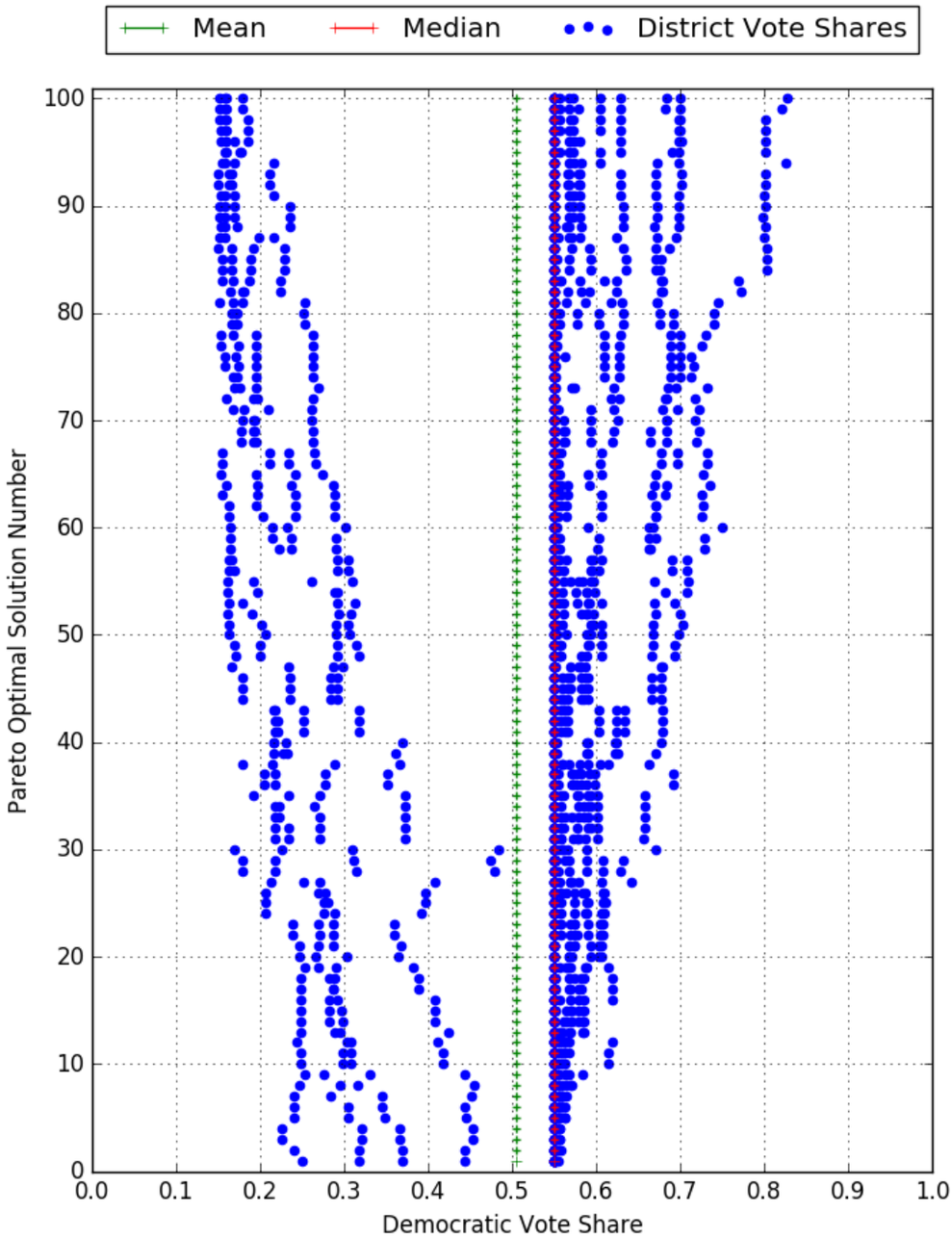

Figure 5: Pareto optimal solutions shown in each row of the plot. As the rows move up from the bottom the p-value of the test increases along with the standard deviation of the vote percentages

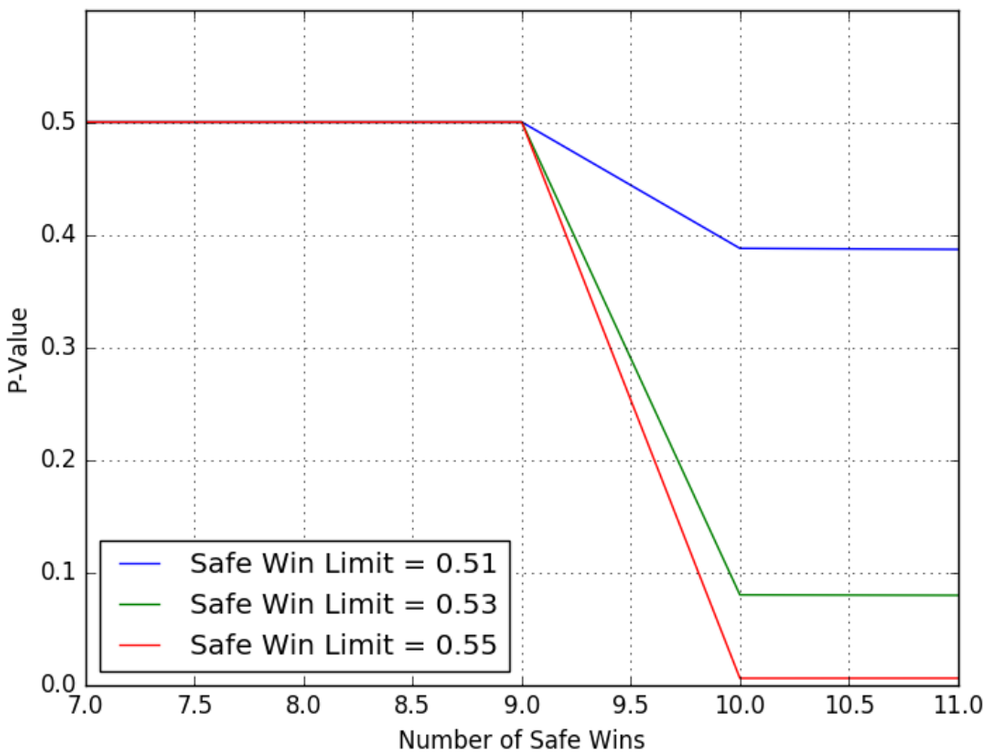

Figure 6: Number of achievable save wins for different safe win thresholds