

# Distribution of the Error in Estimated Numbers of Fixed Points of the Discrete Logarithm

Joshua Holden

Department of Mathematics

Rose-Hulman Institute of Technology, Terre Haute, IN, 47803-3999, USA

holden@rose-hulman.edu

## Abstract

Brizolis asked the question: does every prime  $p$  have a pair  $(g, h)$  such that  $h$  is a fixed point for the discrete logarithm with base  $g$ ? This author and Pieter Moree, building on work of Zhang, Cobeli, and Zaharescu, gave heuristics for estimating the number of such pairs and proved bounds on the error in the estimates. These bounds are not descriptive of the true situation, however, and this paper is a first attempt to collect and analyze some data on the distribution of the actual error in the estimates.

## 1 Introduction

Paragraph F9 of [6] includes the following problem, attributed to Brizolis: given a prime  $p > 3$ , is there always a pair  $(g, h)$  such that  $g$  is a primitive root of  $p$ ,  $1 \leq h \leq p - 1$ , and

$$g^h \equiv h \pmod{p} ? \tag{1}$$

In other words, is there always a primitive root  $g$  such that the discrete logarithm  $\log_g$  has a fixed point? This question has now been settled affirmatively by Campbell and Pomerance in [2]. The answer relies on an estimate for the number  $N(p)$  of pairs  $(g, h)$  which satisfy the equation, have  $g$  is primitive root, and also have  $h$  a primitive root which thus must be relatively prime to  $p - 1$ . This result seems to have been discovered and proved by Zhang in [10] and later, independently, by Cobeli and Zaharescu in [3].

In [7] and [8], Pieter Moree and this author applied the same methods to estimate the number of solutions to (1) given no conditions on  $g$  and  $h$ . Unfortunately, the error term involved in this estimate was completely unsatisfactory. It was also shown in [8] that for a positive proportion of primes a better error estimate can be obtained, and it was conjectured that one could do even better. The object of this note is to collect and analyze some data on the distribution of the actual error in these estimates.

The idea of repeatedly applying the function  $x \mapsto g^x \pmod{p}$  is used in the famous cryptographically secure pseudorandom bit generator of Blum and Micali. ([1]; see also [9] and [5], among others, for further developments.) If one could predict that a pseudorandom generator was going to fall

into a fixed point or cycle of small length, this would obviously be detrimental to cryptographic security. We hope that the investigation of the cycle structure of the discrete logarithm will therefore eventually be of some use to those interested in the field of cryptography.

Using the same notation as in the previously cited papers, we will refer to an integer which is a primitive root modulo  $p$  as PR and an integer which is relatively prime to  $p - 1$  as RP. An integer which is both will be referred to as RPPR and one which has no restrictions will be referred to as ANY.

All integers will be taken to be between 1 and  $p - 1$ , inclusive, unless stated otherwise. If  $N(p)$  is, as above, the number of solutions to (1) such that  $g$  is a primitive root and  $h$  is a primitive root which is relatively prime to  $p - 1$ , then we will say  $N(p) = F_{g_{\text{PR}}, h_{\text{RPPR}}}(p)$ , and similarly for other conditions. We will use  $d(n)$  for the number of divisors of  $n$  and  $\sigma(n)$  for the sum of the divisors of  $n$ . All other notations should be fairly standard.

## 2 Heuristics, Conjectures, and Previous Results

The fundamental observation at the heart of the estimation of  $F_{g_{\text{PR}}, h_{\text{RPPR}}}(p)$  is that if  $h$  is a primitive root modulo  $p$  which is also relatively prime to  $p - 1$ , then there is a unique primitive root  $g$  satisfying (1), namely  $g = h^{\bar{h}}$  reduced modulo  $p$ , where  $\bar{h}$  denotes the inverse of  $h$  modulo  $p - 1$  throughout this note. Thus to estimate  $N(p)$ , we only need to count the number of such  $h$ ;  $g$  no longer has to be considered. We observe that there are  $\phi(p - 1)$  possibilities for  $h$  which are relatively prime to  $p - 1$ , and we would expect each of them to be a primitive root with probability  $\phi(p - 1)/(p - 1)$ . This heuristic uses the assumption that the condition of being a primitive root is in some sense “independent” of the condition of being relatively prime.

We will actually need the following slightly more general heuristic:

**Heuristic 1 (Heuristic 2.6 of [7])** *The order of  $x$  modulo  $p$  is independent of the greatest common divisor of  $x$  and  $p - 1$ , in the sense that for all  $p$ , and all divisors  $e$  and  $f$  of  $p - 1$ ,*

$$\begin{aligned} \frac{1}{p-1} \#\left\{x \in \{1, \dots, p-1\} : \gcd(x, p-1) = e, \quad \text{ord}_p(x) = \frac{p-1}{f}\right\} \\ \approx \frac{1}{p-1} \#\{x \in \{1, \dots, p-1\} : \gcd(x, p-1) = e\} \\ \times \frac{1}{p-1} \#\left\{x \in \{1, \dots, p-1\} : \text{ord}_p(x) = \frac{p-1}{f}\right\}. \end{aligned}$$

The following lemma makes this heuristic rigorous; it was stated and proved in [7] using the ideas in [3].

**Lemma 1 (Lemma 2.7 of [7])** *Let  $e$  and  $f$  be divisors of  $p - 1$ , and  $N$  a multiple of  $p - 1$ . Let  $\mathcal{P} = \{1, \dots, N\}$  and*

$$\mathcal{P}' = \left\{x \in \mathcal{P} : \gcd(x, p-1) = e, \quad \text{ord}_p(x) = \frac{p-1}{f}\right\}.$$

Then

$$\left| \#\mathcal{P}' - \frac{N}{(p-1)^2} \phi\left(\frac{p-1}{f}\right) \phi\left(\frac{p-1}{e}\right) \right| \leq d\left(\frac{p-1}{f}\right) d\left(\frac{p-1}{e}\right) \sqrt{p}(1 + \ln p) \\ \leq d(p-1)^2 \sqrt{p}(1 + \ln p).$$

Using this lemma with  $e = f = 1$  it is straightforward to prove Cobeli and Zaharescu's version of Zhang's result.

**Theorem 1 (Theorem 1 of [3])**

$$\left| F_{g_{\text{PR}}, h_{\text{RPPR}}}(p) - \frac{\phi(p-1)^2}{p-1} \right| \leq d(p-1)^2 \sqrt{p}(1 + \ln p).$$

For the situation with no conditions on  $g$  and  $h$ , we see that (1) can be solved exactly when  $\gcd(h, p-1) = e$  and  $h$  is a  $e$ -th power modulo  $p$ , and in fact there are exactly  $e$  such solutions. Thus

$$F_{g_{\text{ANY}}, h_{\text{ANY}}}(p) = \sum_{e|p-1} e T(e, p). \quad (2)$$

where

$$T(e, p) = \#\left\{ h \in \mathcal{P}(1, 1, p-1)^{(e)} : \gcd(h, p-1) = e \right\}.$$

Applying the lemma with  $e = f = 1$  gives

**Proposition 1 (Proposition 4.2 of [7])** *Let  $e \mid p-1$ . Then*

$$(a) \left| T(e, p) - \frac{1}{e} \phi\left(\frac{p-1}{e}\right) \right| \leq d\left(\frac{p-1}{e}\right) \sqrt{p}(1 + \ln p).$$

$$(b) T(1, p) = \phi(p-1).$$

$$(c) T(p-1, p) = T\left(\frac{p-1}{2}, p\right) = 0.$$

$$(d) 0 \leq T(e, p) \leq \phi\left(\frac{p-1}{e}\right).$$

(e)

$$|F_{g_{\text{ANY}}, h_{\text{ANY}}}(p) - (p-3)| \\ \leq d(p-1) \left( \sigma(p-1) - \frac{3}{2}(p-1) \right) \sqrt{p}(1 + \ln p).$$

Unfortunately, the “error” term in Part (e) will be larger than the main term for infinitely many  $p$ . Using the deep result of Fouvry (see, e.g., [4]) that  $\gg x/\ln x$  primes  $p \leq x$  are such that  $p-1$  has a prime factor larger than  $p^{0.6687}$ , it was proved that:

**Theorem 2 (Theorem 4.8 of [7])** *There are  $\gg x/\ln x$  primes  $p \leq x$  such that*

$$F_{g \text{ ANY}, h \text{ ANY}}(p) = (p - 1) + O(p^{5/6}).$$

*More specifically, there are  $\gg x/\ln x$  primes  $p \leq x$  such that*

$$|F_{g \text{ ANY}, h \text{ ANY}}(p) - (p - 1)| \leq p^{0.8313} d(p - 1)^2 (2 + \ln p).$$

It was also noted in [7] that if Fouvry's assertion holds true with 0.6687 replaced by some larger  $\theta$  (up to  $\theta = 3/4$ ), then in Theorem 2 the exponents  $5/6$  and  $0.8313$  can be replaced by  $3/2 - \theta + \delta$  and  $3/2 - \theta$  for any  $\delta > 0$ .

On the other hand, we also expect that for many primes the error term cannot be set too small. According to Heuristic 1, we can model  $T(e, p)$  using a set of independent random variables  $X_1, \dots, X_{p-1}$  such that

$$X_h = \begin{cases} \gcd(h, p - 1) & \text{with probability } \frac{1}{\gcd(h, p - 1)}; \\ 0 & \text{otherwise.} \end{cases}$$

Then the heuristic suggests that  $F_{g \text{ ANY}, h \text{ ANY}}(p)$  is approximately equal to the expected value of  $X_1 + \dots + X_{p-1}$ , which is clearly  $p - 1$ . On the other hand, the variance  $\sigma^2$  is the expected value of

$$\left( \sum_{h=1}^{p-1} X_h - (p - 1) \right)^2.$$

Note that the expected value of  $X_h X_j$  is  $\gcd(h, p - 1)$  if  $h = j$  and 1 otherwise. Using this, an easy computation shows that

$$\sigma^2 = \sum_{h=1}^{p-1} \gcd(h, p - 1) - (p - 1) = \sum_{d|p-1} d \phi\left(\frac{p-1}{d}\right) - (p - 1).$$

In particular, the standard deviation  $\sigma$  is less than  $p^{1/2+\epsilon}$  for every  $\epsilon > 0$  (for sufficiently large  $p$ ). Thus we have the following:

**Conjecture 1 (Conjecture 3.6 of [7])** *There are  $o(x/\ln x)$  primes  $p \leq x$  for which*

$$|N_{(1), g \text{ ANY}, h \text{ ANY}}(p) - (p - 1)| > p^{1/2+\epsilon}$$

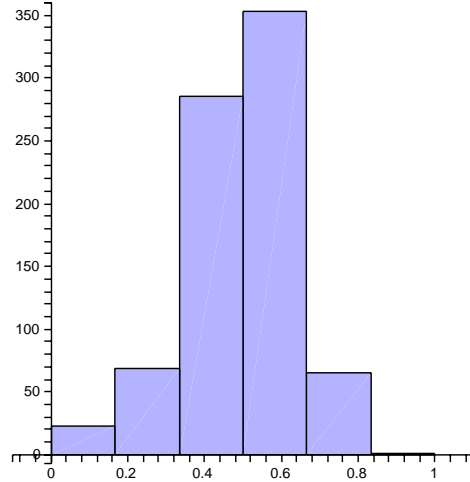
*for every  $\epsilon > 0$ .*

### 3 Data and Analysis

Since a factor of the form  $p^\alpha$  dominates all of the proven and conjectured bounds on the error given above, we decided to collect data on the values of  $\delta = N_{(1), g \text{ ANY}, h \text{ ANY}}(p) - (p - 3)$  for the first 1800 primes (3 through 15413). The data was then tallied based on the value of  $\log_p |\delta|$ . Table 1 and Figure 1 give the data for  $\delta \geq 0$ , while Table 2 and Figure 2 give the data for  $\delta < 0$ . The case  $\delta = 0$

Table 1: Values of  $\delta \geq 0$  for  $3 \leq p \leq 15413$ 

$\log_p  \delta $	0-1/6	1/6-1/3	1/3-1/2	1/2-2/3	2/3-5/6	5/6-1	total
# of $p$	23	69	285	353	65	1	796

Figure 1: Plot of values of  $\delta \geq 0$  for  $3 \leq p \leq 15413$ Table 2: Values of  $\delta < 0$  for  $3 \leq p \leq 15413$ 

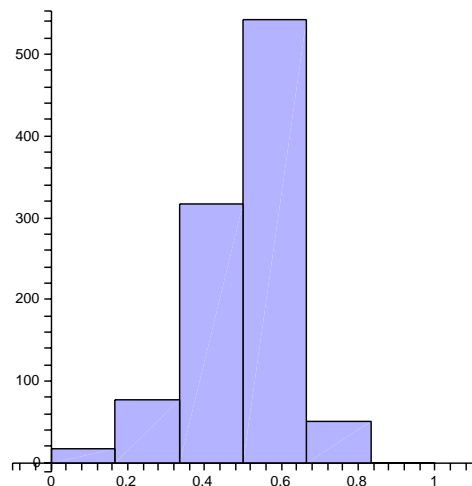
$\log_p  \delta $	0-1/6	1/6-1/3	1/3-1/2	1/2-2/3	2/3-5/6	5/6-1	total
# of $p$	17	78	316	542	51	0	1004

did not actually occur in this sample. Likewise, there were no cases where  $|\delta| > p$ , although this is certainly not ruled out for  $\delta > 0$ .

It is not clear whether the greater number of negative values of  $\delta$  is significant, or a coincidence of this particular data set. The mean for Table 1 is 0.4943 and the mean for Table 2 is 0.5050. This may reflect the same apparent bias towards negative values of  $\delta$ .

Table 3 and Figure 3 give the values of  $|\delta|$  for all computed values of  $\delta$ . The mean for this table is 0.5003, which suggests that the expected value of  $\log_p |\delta|$  may in fact be  $1/2$ , i.e., that the values of  $\delta$  may cluster around  $\sqrt{p}$ . It is not immediately clear how to derive this from the heuristics. The sample standard deviation can be calculated to be 0.1374, but the data does not appear to be precisely normally distributed. This is confirmed by a chi-squared test for goodness of fit, which returns the extremely small  $p$ -value of  $7.8039 \cdot 10^{-34}$ .<sup>1</sup> A sample skewness of  $-0.6785$  and a sample

<sup>1</sup>The  $p$ -value here can be interpreted as the chance that a random sample taken from the predicted distribution

Figure 2: Plot of values of  $\delta < 0$  for  $3 \leq p \leq 15413$ 

kurtosis of 3.6516 can also be computed. This reflects an asymmetric longer left tail (toward smaller values of  $\log_p |\delta|$ ) and a somewhat sharper peak than a normal distribution.

Table 3: All values of  $|\delta|$  for  $3 \leq p \leq 15413$ 

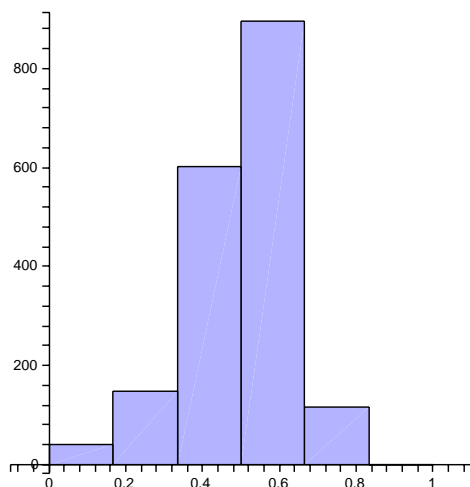
$\log_p  \delta $	0-1/6	1/6-1/3	1/3-1/2	1/2-2/3	2/3-5/6	5/6-1	total
# of $p$	40	147	601	895	116	1	1800

A log-normal distribution was also investigated by taking the exponential function of the mid-points of each of the class intervals. This resulted in a mean of 1.6643, a sample standard deviation of 0.2196, a sample skewness of  $-0.2366$  and a sample kurtosis of 3.2065. Thus the shape of the distribution looks more like a normal distribution; however the chi-squared goodness of fit test still gives an extremely small  $p$ -value of  $2.2243 \cdot 10^{-10}$ . Thus this still does not seem to be the correct distribution. More investigation is clearly necessary, both theoretical and statistical.

The data sets from the tables were collected on a Beowulf cluster, using 16 nodes, each consisting of 2 Pentium III processors running at 1 Ghz. The programming was done in C, using MPI, OpenMP, and OpenSSL libraries. The collection took approximately 60 hours for the 1800 primes between 3 and 15413 (inclusive).

---

would deviate from the distribution as a whole at least as much as the observed data did. Thus this set of data is an extremely bad match for the prediction. We are using statistical language in this note even though the data sets do not come from random variables, and are in fact deterministic. Thus, all of the statistical results in this note should be taken with a very large grain of salt.

Figure 3: Plot of all values of  $|\delta|$  for  $3 \leq p \leq 15413$ 

## 4 Conclusion and Future Work

This note is clearly a preliminary effort. The fact that we were unable to interpret the data as any sort of normal distribution is unsatisfying, if not perhaps surprising. We hope in the future to provide at least a conjectural explanation of this data. A better theoretical understanding of the error terms in the theorems we have cited would of course be helpful in this.

The project of extending our analysis to three-cycles and more generally  $k$ -cycles for small values of  $k$ , mentioned in previous papers, still remains to be done. Along similar lines, Igor Shparlinski has suggested attempting to analyze the average length of a cycle. Daniel Cloutier, a student at the Rose-Hulman Institute of Technology, has recently begun to collect data which we hope will shed light on both of these problems.

## Acknowledgments

The author would like to thank Pieter Moree for providing the heuristics for estimating the standard deviation  $\sigma$  of  $F_{g, \text{ANY}, h, \text{ANY}}(p)$  and for several other results cited in this note. He would also like to thank Diane Evans of the Rose-Hulman Institute of Technology Mathematics Department for statistical advice. Finally, he would like to thank the editor for several helpful suggestions.

## References

- [1] Manuel Blum and Silvio Micali. How to generate cryptographically strong sequences of pseudorandom bits. *SIAM J. Comput.*, 13(4):850–864, 1984.
- [2] Mariana Campbell. On fixed points for discrete logarithms. Master’s thesis, University of California at Berkeley, Spring 2003.

- [3] Cristian Cobeli and Alexandru Zaharescu. An exponential congruence with solutions in primitive roots. *Rev. Roumaine Math. Pures Appl.*, 44(1):15–22, 1999.
- [4] Étienne Fouvry. Théorème de Brun-Titchmarsh: Application au théorème de Fermat. *Invent. Math.*, 79(2):383–407, 1985.
- [5] Rosario Gennaro. An improved pseudo-random generator based on discrete log. In Mihir Bellare, editor, *Advances in Cryptology — CRYPTO 2000*, pages 469–481. Springer, 2000.
- [6] Richard K. Guy. *Unsolved Problems in Number Theory*. Springer-Verlag, 1981.
- [7] Joshua Holden and Pieter Moree. New conjectures and results for small cycles of the discrete logarithm. In Alf van der Poorten and Andreas Stein, editors, *High Primes and Misdemeanours: lectures in honour of the 60th birthday of Hugh Cowie Williams*, number 41 in Fields Institute Communications, pages 245–254. American Mathematical Society, 2004. <http://xxx.lanl.gov/abs/math.NT/0305305>.
- [8] Joshua Holden and Pieter Moree. Some heuristics and results for small cycles of the discrete logarithm. *Mathematics of Computation*, 2005. To appear. <http://xxx.lanl.gov/abs/math.NT/0401013>.
- [9] Sarvar Patel and Ganapathy S. Sundaram. An efficient discrete log pseudo-random generator. In Hugo Krawczyk, editor, *Advances in Cryptology — CRYPTO '98*, pages 304–317. Springer, 1998.
- [10] Wen Peng Zhang. On a problem of Brizolis. *Pure Appl. Math.*, 11(suppl.):1–3, 1995.