Identifying the Probability Distributions of Highest Exponents of Factors in the Positive Integers

Richard Han
gtID: 903757484
Georgia Tech
rhan70@gatech.edu

Abstract

In this paper, I explore the probability distribution of the highest exponent of 2 that occurs in a set of positive integers. I apply the chi-squared goodness-of-fit test to determine whether or not the distribution is binomial, poisson, or geometric. It is determined experimentally and analytically that the distribution is geometric.

Keywords Probability Distributions · Number Theory · Chi-squared Goodness-of-fit Test · Maximum Likelihood Estimation

1 Background/Problem Statement

The positive integers begin as $1, 2, 3, 4, 5, 6, 7, 8, ...$ Each number has a highest power of 2 that evenly divides the number. For instance, the odd numbers all have factors of 2^0, where 0 is the highest power of 2 that evenly divides them. The even number 2 has 1 as the highest power of 2 that divides it, the even number 4 has 2 as the highest power of 2 that evenly divides it, and the even number 8 has 3 as the highest power of 2 that evenly divides it. In this paper, I would like to explore the probability of any given positive integer having a certain highest power of 2 that evenly divides it. Since every other positive integer is odd, we would expect the probability of any given positive integer having highest power 0 of 2 to be about $1/2$. What about the probability of any given positive integer having highest power 1 of 2? Or the probability of having highest power 2 of 2, etc? This is the question I will explore in this paper.

To answer this question, I will look at the first 1000 positive integers and apply a chi-squared goodness-of-fit test to the set of highest powers of 2 that occur in each number. I will test whether the set of highest powers of 2 that occur in the set is distributed according to a binomial distribution, a Poisson distribution, or a geometric distribution. We will see that the distribution is geometric.

To provide further evidence supporting the claim that the distribution is geometric, I will generate $10,000$ random integers between 1 and $100,000$ and apply the chi-squared goodness-of-fit test to the set of highest powers of 2 that occur to test whether the distribution is binomial, Poisson, or geometric.

2 Main Findings

Let X be a random variable that denotes the largest power of 2 that evenly divides a positive integer. Our goal is to determine what the probability mass function is for this random variable. In the following subsections, we will test whether the distribution is binomial, Poisson or geometric, respectively. In each subsection, we will start off by finding the maximum likelihood estimator for any relevant parameter, then apply a chi-squared goodness-of-fit test.

2.1 Binomial Distribution

Suppose $H_0 : X_1, ..., X_m \overset{\text{iid}}{\sim} Bin(n, p)$.

Form the likelihood function

$$
\begin{aligned}
L(p) = \prod_{i=1}^{m} f(x_i) &= \prod_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} \\
&= \left(p^{\sum_{i=1}^{m} x_i}\right)(1-p)^{mn-\sum_{i=1}^{m} x_i} \prod_{i=1}^{m}\binom{n}{x_i}
\end{aligned}
$$

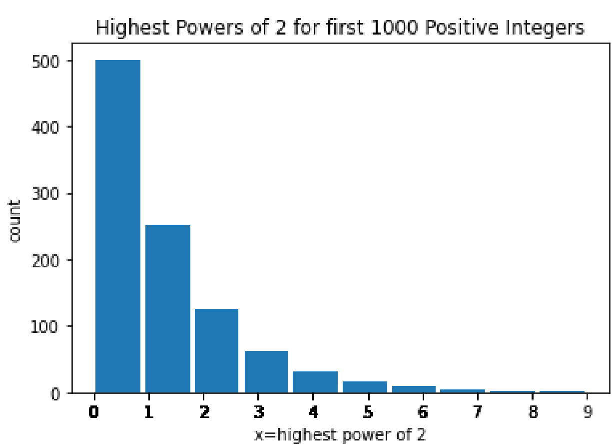
We want the p that maximizes $L(p)$. This is equivalent to finding p that maximizes $\ln L(p)$.

$$
\begin{aligned}
\ln L(p) &= \ln(p^{\sum_{i=1}^{m} x_i}) + \ln(1-p)^{mn-\sum_{i=1}^{m} x_i} + \ln \prod_{i=1}^{m}\binom{n}{x_i} \\
&= \left(\sum_{i=1}^{m} x_i\right)\ln p + (mn - \sum_{i=1}^{m} x_i)\ln(1-p) + \ln \prod_{i=1}^{m}\binom{n}{x_i}
\end{aligned}
$$

$$
\begin{aligned}
&\implies \frac{d\ln L(p)}{dp} = \frac{\sum_{i=1}^{m} x_i}{p} - \frac{mn - \sum_{i=1}^{m} x_i}{1-p} \equiv 0 \\
&\implies \frac{\sum_{i=1}^{m} x_i}{p} = \frac{mn - \sum_{i=1}^{m} x_i}{1-p} \\
&\implies \left(\sum_{i=1}^{m} x_i\right)(1-p) = (mn - \sum_{i=1}^{m} x_i)p \\
&\implies \sum_{i=1}^{m} x_i = mnp \\
&\implies \hat{p} = \frac{\sum_{i=1}^{m} x_i}{mn}
\end{aligned}
$$

Now, I want to look at the first 1000 positive integers and the set of highest powers of 2 that occur as factors of them. I want to apply a chi-squared goodness-of-fit test to test whether the distribution for this set is $Bin(n, p)$. In this context, n is the largest power of 2 that occurs as a factor for the first 1000 positive integers; n happens to be 9.

Here is a histogram of the counts for the highest powers of 2 that occur as factors for the first 1000 positive integers:



Note that about half of the numbers have highest power 0 because about half of them are odd numbers; subsequently, the counts are approximately half of the immediately preceding count. The sample mean $\overline{X} = 0.994$, and the sample variance $S^2 \approx 1.92$.

In order to calculate the chi-squared statistic, we need E_x and O_x for each $x = 0, 1, .., n$. The O_x's are found by counting how many 0's occur, how many 1's occur, etc., as highest powers of 2 in the factors for the first 1000 positive integers. If O is the vector of all O_x's, then $O = \begin{bmatrix} 500 \\ 250 \\ 125 \\ 63 \\ 31 \\ 16 \\ 8 \\ 4 \\ 2 \\ 1 \end{bmatrix}$.

$E_x = m\hat{P}(X = x) = m\binom{n}{x}\hat{p}^x(1-\hat{p})^{n-x} \quad x = 0, 1, ..., n.$

The chi-squared statistic is given by $\chi_0^2 = \sum_{x=0}^{n} \frac{(O_x - E_x)^2}{E_x}$. Let $k = n + 1$, $s = 1$ and $\alpha = 0.05$.

If $\chi_0^2 > \chi^2_{\alpha,k-1-s}$, reject H_0. Otherwise, we accept H_0.

In the case of the positive integers from 1 to 1000, let $m = 1000$ and $n = 9$.

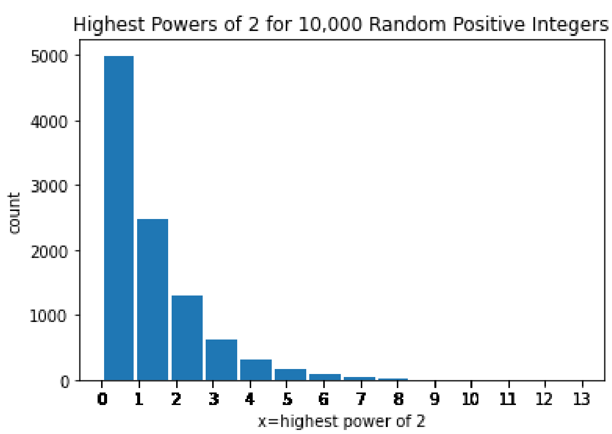
We find that $\hat{p} = \frac{994}{9000}$. Plugging this into the formula for E_x gives:

$E_x = 1000\binom{9}{x}(\frac{994}{9000})^x(1 - \frac{994}{9000})^{9-x} \quad x = 0, 1, ..., 9.$

We now compute the chi-squared statistic to be $\chi_0^2 \approx 435267 > \chi^2_{0.05,8} \approx 15.51$. So, we reject H_0. The X_i's are not $Bin(9, \hat{p})$.

To provide further evidence that the X_i's are not binomial distributed, I generated $10,000$ random integers between 1 and $100,000$. In this case, $n = 13$ and $m = 10,000$.

Here is a histogram of the counts for the highest powers of 2 that occur as factors for the 10,0000 random integers:



Again, we can see that the distribution looks geometric. The sample mean $\overline{X} = 1.0083$, and the sample variance $S^2 \approx 2.0488$.

We have $\hat{p} = \frac{10083}{130000} \approx 0.07756$. Let $k = n + 1$, $s = 1$, and $\alpha = 0.05$.

If $\chi_0^2 > \chi^2_{\alpha,k-1-s}$, reject H_0.

$\chi_0^2 \approx 109633647913 > 21.03 \approx \chi^2_{0.05,12}$. So, we reject H_0, and the X_i's are not $Bin(n, \hat{p})$.

2.2 Poisson Distribution

Suppose $H_0 : X_1, ..., X_m \overset{\text{iid}}{\sim} Pois(\lambda)$.

Form the likelihood function

$$L(\lambda) = \prod_{i=1}^{m} f(x_i) = \prod_{i=1}^{m} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$
$$= \frac{e^{-mx}\lambda^{\sum_{i=1}^{m} x_i}}{\prod_{i=1}^{m} x_i!}$$

We want the λ that maximizes $L(\lambda)$. This is equivalent to finding λ that maximizes $\ln L(\lambda)$.

$$\ln L(\lambda) = -m\lambda + (\sum_{i=1}^{m} x_i)\ln\lambda - \ln(\prod_{i=1}^{m} x_i!)$$

$$\implies \frac{d\ln L(\lambda)}{d\lambda} = -m + \frac{\sum_{i=1}^{m} x_i}{\lambda} \equiv 0$$
$$\implies m = \frac{\sum_{i=1}^{m} x_i}{\lambda}$$
$$\implies \hat{\lambda} = \frac{\sum_{i=1}^{m} x_i}{m}$$

As before, I want to look at the first 1000 positive integers and the set of highest powers of 2 that occur as factors of them. I want to apply a chi-squared goodness-of-fit test to test whether the distribution for this set is $Pois(\lambda)$. In this context, n is the largest power of 2 that occurs as a factor for the first 1000 positive integers; n happens to be 9.

In order to calculate the chi-squared statistic, we need E_x and O_x for each $x = 0, 1, .., n$. As before, $O = \begin{bmatrix} 500 \\ 250 \\ 125 \\ 63 \\ 31 \\ 16 \\ 8 \\ 4 \\ 2 \\ 1 \end{bmatrix}$.

$E_x = m\hat{P}(X = x) = m\frac{e^{-\hat{\lambda}}\hat{\lambda}^x}{x!} \quad x = 0, 1, ..., n.$

The chi-squared statistic is given by $\chi_0^2 = \sum_{x=0}^{n} \frac{(O_x - E_x)^2}{E_x}$. Let $k = n + 1, s = 1$ and $\alpha = 0.05$.

If $\chi_0^2 > \chi^2_{\alpha,k-1-s}$, reject H_0. Otherwise, we accept H_0.

In the case of the positive integers from 1 to 1000, let $m = 1000$ and $n = 9$.

We find that $\hat{\lambda} = \frac{994}{1000}$. Plugging this into the formula for E_x gives:

$E_x = 1000e^{-0.994}\frac{(0.994)^x}{x!} \quad x = 0, 1, ..., 9.$

We now compute the chi-squared statistic to be $\chi_0^2 \approx 1994.47 > \chi^2_{0.05,8} \approx 15.51$. So, we reject H_0. The X_i's are not $Pois(\hat{\lambda})$.

To provide further evidence that the X_i's are not Poisson distributed, we consider the $10,000$ random numbers generated earlier. In this case, $n = 13$ and $m = 10,000$.

We have $\hat{\lambda} = \frac{10083}{10000} \approx 1.0083$. Let $k = n + 1, s = 1$, and $\alpha = 0.05$.

If $\chi_0^2 > \chi^2_{\alpha,k-1-s}$, reject H_0.

$\chi_0^2 \approx 7141100 > 21.03 \approx \chi^2_{0.05,12}$. So, we reject H_0, and the X_i's are not $Pois(\hat{\lambda})$.

2.3 Geometric Distribution

For this subsection, we say $X \sim Geom(r)$ if the pmf $f(x) = (1-r)^x r \quad x = 0, 1, ...$

Some facts for this geometric distribution are: $E[X] = \frac{1-r}{r}, Var(X) = \frac{1-r}{r} + \left(\frac{1-r}{r}\right)^2$, mgf $M_X(t) = \frac{r}{1-(1-r)e^t}$.

Suppose $H_0 : X_1, ..., X_m \overset{\text{iid}}{\sim} Geom(r)$.
Form the likelihood function

$$L(r) = \prod_{i=1}^{m} f(x_i) = \prod_{i=1}^{m} (1-r)^{x_i} r$$
$$= r^m (1-r)^{\sum_{i=1}^{m} x_i}$$

We want the r that maximizes $L(r)$. This is equivalent to finding r that maximizes $\ln L(r)$.

$$\ln L(r) = m \ln r + \left(\sum_{i=1}^{m} x_i\right) \ln(1-r)$$

$$\implies \frac{d \ln L(r)}{dr} = \frac{m}{r} - \frac{\sum_{i=1}^{m} x_i}{1-r} \equiv 0$$
$$\implies \frac{m}{r} = \frac{\sum_{i=1}^{m} x_i}{1-r}$$
$$\implies \hat{r} = \frac{m}{m + \sum_{i=1}^{m} x_i}$$

As before, I want to look at the first 1000 positive integers and the set of highest powers of 2 that occur as factors of them. I want to apply a chi-squared goodness-of-fit test to test whether the distribution for this set is $Geom(r)$. In this context, n is the largest power of 2 that occurs as a factor for the first 1000 positive integers; n happens to be 9.

In order to calculate the chi-squared statistic, we need E_x and O_x for each $x = 0, 1, .., n$. As before, $O = \begin{bmatrix} 500 \\ 250 \\ 125 \\ 63 \\ 31 \\ 16 \\ 8 \\ 4 \\ 2 \\ 1 \end{bmatrix}$.

$E_x = m\hat{P}(X = x) = m(1-\hat{r})^x \hat{r} \quad x = 0, 1, ..., n.$

The chi-squared statistic is given by $\chi_0^2 = \sum_{x=0}^{n} \frac{(O_x - E_x)^2}{E_x}$. Let $k = n + 1, s = 1$ and $\alpha = 0.05$.

If $\chi_0^2 > \chi^2_{\alpha, k-1-s}$, reject H_0. Otherwise, we accept H_0.

In the case of the positive integers from 1 to 1000, let $m = 1000$ and $n = 9$.

We find that $\hat{r} = \frac{1000}{1994} \approx \frac{1}{2}$. Plugging this into the formula for E_x gives:

$E_x = 1000\left(1 - \frac{1000}{1994}\right)^x \left(\frac{1000}{1994}\right) \quad x = 0, 1, ..., 9.$

We now compute the chi-squared statistic to be $\chi_0^2 \approx 0.06387 >^? \chi^2_{0.05,8} \approx 15.51$. No, the inequality does not hold. So, we accept the null hypothesis that the X_i's are $Geom(\hat{r})$.

To provide further evidence that the X_i's are Geometric distributed, we consider the $10,000$ random numbers generated earlier. In this case, $n = 13$ and $m = 10,000$.

We have $\hat{r} \approx 0.4979 \approx \frac{1}{2}$. Let $k = n + 1, s = 1$, and $\alpha = 0.05$.

If $\chi_0^2 > \chi^2_{\alpha,k-1-s}$, reject H_0.

$\chi_0^2 \approx 18.08 >^? 21.03 \approx \chi^2_{0.05,12}$. No, the inequality does not hold. So, we accept the null hypothesis that the X_i's are $Geom(\hat{r})$.

2.4 True Probability Mass Function for X

In the above analysis, we found evidence to believe that X has the probability distribution $Geom(r)$ for some r. In the analysis, we found the MLE $\hat{r}$ for r, and it turned out to be about $1/2$. Actually, we know what the true r should be; it should be $1 - \frac{1}{p}$ where p is the factor we're considering. In the above analysis, $p = 2$ so that $1 - \frac{1}{p} = \frac{1}{2}$ and $X \sim Geom(1/2)$. Earlier, we saw that $\hat{r} \approx 1/2$, which makes sense.

Claim: Let $X =$ the largest power of p that divides a positive integer. Then, $X \sim Geom(1 - \frac{1}{p})$.

Arguing informally, the powers p^x occur once every p^x numbers; so the probability of a number having p^x as a factor is $\frac{1}{p^x}$. However, the powers p^{x+1} occur once every p^{x+1} numbers; so the probability of a number having p^{x+1} as a factor is $\frac{1}{p^{x+1}}$. Hence, in order to get the probability of a number having p^x as the highest power of p as factor (that is, $P(X) = x$), we need to subtract $\frac{1}{p^{x+1}}$ from $\frac{1}{p^x}$. Therefore, the pmf $f(x) = \frac{1}{p^x} - \frac{1}{p^{x+1}} \quad x = 0, 1, ...$

This can be rewritten as $f(x) = \left(\frac{1}{p}\right)^x (1 - \frac{1}{p}) = (1-r)^x r$, where $r = 1 - \frac{1}{p}$.

So, $X \sim Geom(r)$, where $r = 1 - \frac{1}{p}$.

In the experiments we performed, we also saw that the sample mean was approximately 1 and that the sample variance was approximately 2; this makes sense since, when $r = \frac{1}{2}$, $E[X] = \frac{1-r}{r} = 1$ and $Var(X) = \frac{1-r}{r} + \left(\frac{1-r}{r}\right)^2 = 2$.

3 Conclusion

In this paper, I have explored the probability distribution of the highest powers of 2 that are factors of the first 1000 positive integers. Chi-squared goodness-of-fit tests were performed to test whether or not the distribution is binomial, Poisson, or geometric. The evidence supports the claim that the distribution is geometric, and $10,000$ random numbers between 1 and $100,000$ were tested to further support this claim. Finally, an explicit formula for the probability mass function for the general case of a factor p, not just the factor 2, was discussed.