Ball State Undergraduate Mathematics Exchange
`https://digitalresearch.bsu.edu/mathexchange`
Vol. 17, No. 1 (Fall 2023)
Pages 2 – 14

An investigation into the law of small numbers using R

*Yasir Zubayr Barlas, Dudley Stark**

Yasir Zubayr Barlas pursued his undergraduate studies in the field of mathematics at Queen Mary, University of London. His academic interests encompass various areas of mathematics, with a specific focus on probability and statistics. The research presented in this paper was carried out during his undergraduate years.

Dudley Stark received his Ph.D. from University of Southern California in 1994 and is a Reader (Associate Professor) in Mathematics and Probability at Queen Mary, University of London. His research interests lie in the fields of probability and combinatorics. He enjoys teaching a variety of modules in financial mathematics, statistics, and pure mathematics.

Abstract

The Law of Small Numbers states that the Binomial distribution converges to the Poisson distribution. Using the programming language R, we investigate the total variation distance between Binomial$(n, c/n)$ and Poisson(c) when we fix c and n individually. We also look at the asymptotics for nd_{TV} for a fixed c, where nd_{TV} is the total variation distance d_{TV} multiplied by increasing values of n. Several properties of d_{TV} are looked at in this paper.

1 Introduction

'The Law of Small Numbers' is a book written by Ladislaus von Bortkiewicz [1]. Quine and Seneta [2] state that there is much misconception about the book and its contents. Assume we have a short series of N independent observations with a Poisson(λ_i) for $i \in \{1,\ldots,N\}$. Bortkiewicz found that these observations act as if they are from a sample of size N with a Poisson distribution, even with unequal λ_i's. It is known that in certain circumstances, the Binomial distribution converges to the Poisson distribution.

***Corresponding author:** d.s.stark@qmul.ac.uk

In our study, we will be using Binomial$(n, c/n)$ and Poisson(c) for our investigation of the total variation distance between the two distributions.

Definition 1. Total variation distance measures the closeness between two distributions. The distance is defined by

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Y)) = \frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Y=j)|$$

where X and Y are discrete random variables and $\mathscr{L}(X)$ and $\mathscr{L}(Y)$ denotes their distributions. The state space of these discrete random variables is $\{0, 1, 2, \dots\}$. It is an important statistical distance measure, which in layman's terms measures the difference between two probability distributions. It is part of a wider field that too measures the difference between two probability distributions, called 'f-divergence' [3].

We wished to find higher order expansions of the total variation distance, but this was not possible using the programming language R. Instead, we look at the first order asymptotics for nd_{TV}, where c is fixed and n is increasing. This paper reports on several plots of the total variation distance for the law of small numbers.

For interested readers, we review a number of properties of the total variation distance which include calculating d_{TV} as a finite sum and the metric axioms. We also provide the R code of our plots, if a reader would like to use the code in their own research.

In our research, we have looked at several scenarios for our calculation of the total variation distance. We manipulated n and c to observe d_{TV}, and later nd_{TV}. One scenario involves finding a convergence towards a certain value of the total variation distance multiplied by n. nd_{TV} is this scenario, where c is fixed and $n \to \infty$. The convergence can be seen when n is small and in some cases when n is large, however this depends on the value of c.

2 Properties of the Total Variation Distance

The following properties seen in this section are important when it comes to calculating the total variation distance. They support the validity of our results and are part of the research project conducted.

2.1 d_{TV} as a finite sum

The total variation distance can be found in a shorter way, by looking at only the positive parts of the original formula. The new method of calculating the total variation distance will only look at a finite number of terms, hence why it is shorter.

Definition 2. A real function $f(x)$ can be split into two parts,

$$f^+(x) = \begin{cases} f(x), & \text{if } f(x) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$f^-(x) = \begin{cases} -f(x), & \text{if } f(x) < 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $f(x) = f^+(x) - f^-(x)$.

Theorem 1. The total variation distance (d_{TV}) can be found by looking only at the positive parts

$$d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \sum_{j=0}^{n} (P(X=j) - P(Y=j))^+.$$

Proof: We can separate functions into two different parts, one being the positive part and the other being the negative part. The positive part looks only at the positive values of a function and vice versa for the negative part.

By **Definition 2.**,

$$d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \frac{1}{2} \sum_{j=0}^{\infty} |P(X=j) - P(Y=j)|$$

$$|P(X=j) - P(Y=j)| = ((P(X=j) - P(Y=j))^+ + (P(X=j) - P(Y=j))^- \tag{3}$$

$$(P(X=j) - P(Y=j)) = ((P(X=j) - P(Y=j))^+ - (P(X=j) - P(Y=j))^- \tag{4}$$

The distributions are discrete and we assume that we have a fixed n number of trials.

Looking at (3),

$$d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \frac{1}{2} \sum_{j=0}^{n} (P(X=j) - P(Y=j))^+ + \frac{1}{2} \sum_{j=0}^{n} (P(X=j) - P(Y=j))^-$$

$$2 \cdot d_{TV}(\mathscr{L}(X), \mathscr{L}(Y)) = \sum_{j=0}^{n} (P(X=j) - P(Y=j))^+ + \sum_{j=0}^{n} (P(X=j) - P(Y=j))^-$$

Now looking at (4),

$$\sum_{j=0}^{n}(P(X=j)-P(Y=j))=\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{+}-\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{-}$$

Left hand side,

$$\sum_{j=0}^{n}P(X=j)-\sum_{j=0}^{n}P(Y=j)=1-1=0.$$

Implying that,

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{+}=\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{-}.$$

From this, we can now calculate the distance as follows

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=\sum_{j=0}^{n}(P(X=j)-P(Y=j))^{+}.$$

2.2 Metric Axioms

Theorem 2. The total variation distance (d_{TV}) is a metric on the space of distributions.
Proof: We review the metric axioms in order to prove this theorem.

Definition 3. A metric satisfies three axioms,

(i) Non-negativity, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))\geq 0$

(ii) The identity of indiscernibles, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=0$ if and only if $\mathscr{L}(X)=\mathscr{L}(Y)$

(iii) Symmetry, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=d_{TV}(\mathscr{L}(Y),\mathscr{L}(X))$

(iv) The triangle inequality, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Z))\leq d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))+d_{TV}(\mathscr{L}(Y),\mathscr{L}(Z))$

Axiom 1: Straightforward from the definition of total variation distance, notice that the absolute value ensures that we take the positive result.

Axiom 2: For the identity of indiscernibles, it can be easy to see that $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))=0$ if and only if $\mathscr{L}(X)=\mathscr{L}(Y)$. We know that $|P(X=j)-P(Y=j)|\geq 0$ for all j. We need the left hand side to be equal to zero, so we make $P(X=j)=P(Y=j)$. The two distributions are then identical, $\mathscr{L}(X)=\mathscr{L}(Y)$, so $|P(X=j)-P(X=j)|=0$.

Axiom 3: For the axiom of symmetry, we see that $|P(X=j)-P(Y=j)|\geq 0$. If we were to swap $P(X=j)$ and $P(Y=j)$ around for $|P(Y=j)-P(X=j)|\geq 0$, the absolute value allows for the same result.

Axiom 4: For the triangle inequality, we need to show that $d_{TV}(\mathscr{L}(X),\mathscr{L}(Z))\leq d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))+d_{TV}(\mathscr{L}(Y),\mathscr{L}(Z))$. We can first display

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Z))=\frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Z=j)|.$$

We can then see the following

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Z)) \leq \frac{1}{2}\sum_{j=0}^{\infty}\big\{|P(X=j)-P(Y=j)|+|P(Y=j)-P(Z=j)|\big\}.$$

The right hand side can be separated into

$$\frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Y=j)|+\frac{1}{2}\sum_{j=0}^{\infty}|P(Y=j)-P(Z=j)|.$$

So then

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Z)) \leq d_{TV}(\mathscr{L}(X),\mathscr{L}(Y))+d_{TV}(\mathscr{L}(Y),\mathscr{L}(Z)).$$

2.3 d_{TV} as a finite sum

Theorem 3. The total variation distance (d_{TV}) has property $0 \leq d_{TV} \leq 1$.

Proof: We have already seen that $d_{TV} \geq 0$. By **Definition 1.** the total variation distance (d_{TV}) is defined by

$$d_{TV}(\mathscr{L}(X),\mathscr{L}(Y)) = \frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Y=j)|$$

The absolute value being taken ensures that there are no negative values, so we know that $d_{TV} \geq 0$. To find out why d_{TV} does not exceed 1, we have

$$\frac{1}{2}\sum_{j=0}^{\infty}|P(X=j)-P(Y=j)| \leq \frac{1}{2}\sum_{j=0}^{\infty}(P(X=j)+P(Y=j))$$

Looking at the right hand side,

$$\frac{1}{2}\sum_{j=0}^{\infty}P((X=j)+P(Y=j)) = \frac{1}{2}\sum_{j=0}^{\infty}P(X=j)+\frac{1}{2}\sum_{j=0}^{\infty}P(Y=j)$$

$$\frac{1}{2}\sum_{j=0}^{\infty}P(X=j)+\frac{1}{2}\sum_{j=0}^{\infty}P(Y=j) = \frac{1}{2}+\frac{1}{2} = 1$$

It is seen that $0 \leq d_{TV} \leq 1$.

A value of 0 for the total variation distance tells us that both random variables are identical. This is seen from our first metric axiom, where $d_{TV}(\mathscr{L}(X),\mathscr{L}(Y)) = 0$ if and only if $\mathscr{L}(X) = \mathscr{L}(Y)$.

A value of 1 can be seen as the opposite of having the total variation distance equal to 0. For a value of 1, both random variables have disjoint sample spaces.

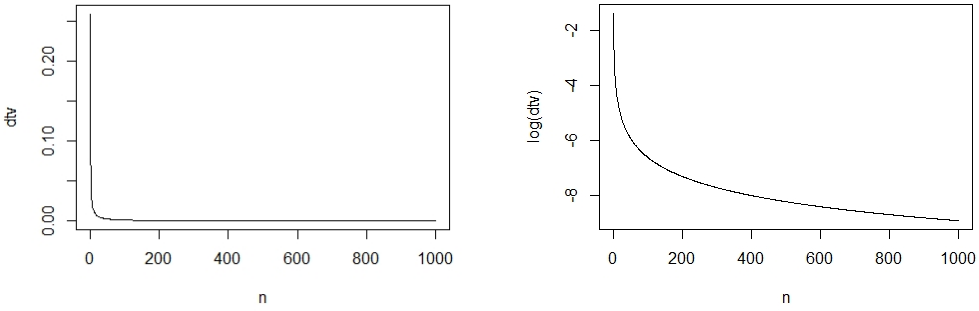
3 Main Results

We use the programming language R [4] to produce our results. The function 'TotalVarDist' from the package 'distrEx' [5] is used to calculate the total variation distance without having to write longer pieces of code. Our random variables X and Y remain the same throughout, with $X \sim \text{Binomial}(n, c/n)$ and $Y \sim \text{Poisson}(c)$.

3.1 d_{TV} for multiple values of n

We keep c as a fixed value here and we make n increase seemingly to infinity. Kennedy and Quine [6] have found an exact expression for $n \geq 1$ and $0 < np \leq 2 - \sqrt{2}$. They call this $f_1(p)$ in their paper, where $p = \frac{c}{n}$ in our calculations

$$f_1\left(\frac{c}{n}\right) = c\left(1 - \frac{c}{n}\right)^{n-1} - ce^{-c}.$$



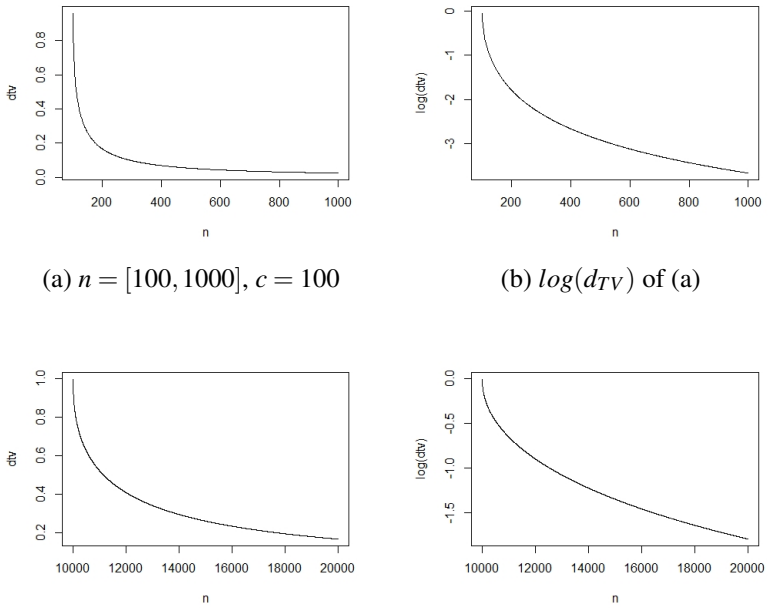
(a) d_{TV} (b) $log(d_{TV})$

Figure 1: Plot for $c = 2 - \sqrt{2}$ for increasing values of n

R Code for Figure 1:

```
#dtv for multiple n
library(distrEx)
library(distr)
n = 1
results1 = list()
while (n <= 1000){
  c = 2 - sqrt(2)
  x = TotalVarDist(Binom(size = n, prob = c/n), Pois(c))
  results1 = c(results1, x)
  #once n = 1000 in the program, produce a plot of the results
  if (n == 1000){
    plot(unlist(results1), type="l", xlab = "n", ylab = "dtv")}
n = n + 1}
```

We see here that as n increases, we move rapidly to a total variation distance of zero. Changing the value of c (and n where required) has little effect to the form of the plot shown in Figure 1. Additional plots are shown in Figure 2 for different c.



(a) $n = [100, 1000]$, $c = 100$ (b) $log(d_{TV})$ of (a)

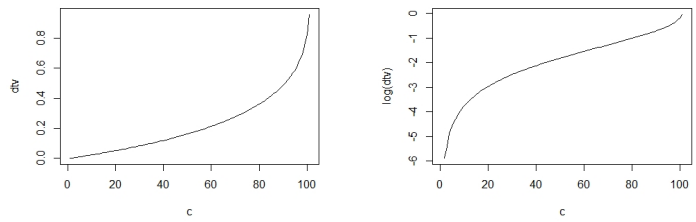
(c) $n = [10000, 20000]$, $c = 10000$ (d) $log(d_{TV})$ of (c)

Figure 2: Plots of different values of c

Comparing d_{TV} to $log(d_{TV})$, we see a similar behaviour in the plots. This behaviour is consistent at higher values of n and c, as seen by the plots. Therefore, as we increase n for a fixed c, the total variation distance decreases in value close to zero for larger values of n.

3.2 d_{TV} for multiple values of c

We keep n as the fixed value here and we increase c to a certain value of n. We cannot make c a greater value than n because that would make the probability c/n greater than one. Figure 3 shows a plot for $0 \leq c \leq 100$ and $n = 100$.



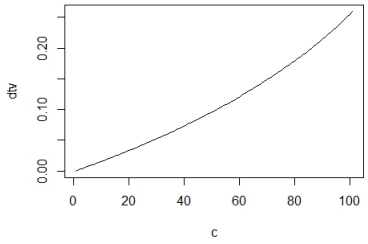
(a) d_{TV} (b) $log(d_{TV})$

Figure 3: Plot for $n = 100$ for increasing values of c

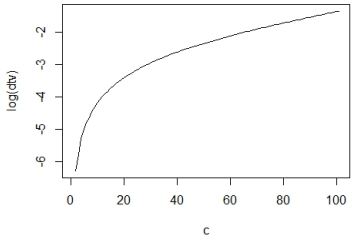
R Code for Figure 3:

```
[language=R]
#dtv for multiple c
library(distrEx)
library(distr)
c = 0
results2 = list()
while (c <= 100){
  n = 100
  x = TotalVarDist(Binom(size = n, prob = c/n), Pois(c))
  results2 = c(results2, x)
  #once c = 100 in the program, produce a plot of the results
  if (c == 100){
    plot(unlist(results2), type="l", xlab = "c", ylab = "dtv")}
c = c + 1}
```

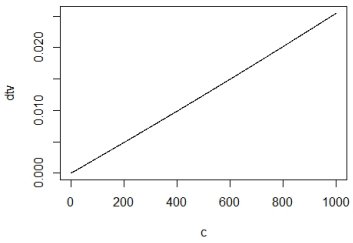
The shape of this plot in Figure 3 will not be consistent for all values of n (and c where required). When n is equal to the largest value of c, we see the behaviour of an exponential function. As n increases with the same boundaries for c, the plot eventually becomes linear. This is visualised in Figure 4. However, it remains that as c increases for a fixed n that the total variation distance increases.



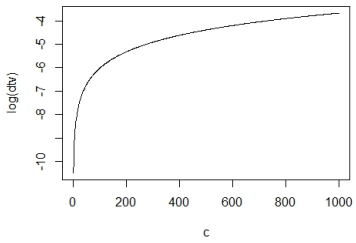
(a) $n = 150, c = [0, 100]$



(b) $log(d_{TV})$ of (a)



(c) $n = 10000, c = [0, 1000]$



(d) $log(d_{TV})$ of (c)

Figure 4: Plots of different values of n

3.3 nd_{TV} for multiple values of n

This builds up on Section 3.1, where c is fixed and n is increasing. nd_{TV} has similar behaviour to what we had done in Section 3.1, as n increases we have a smaller total variation distance. Figure 5 shows a plot $7 \leq n \leq 1000$ and $c = 7$.

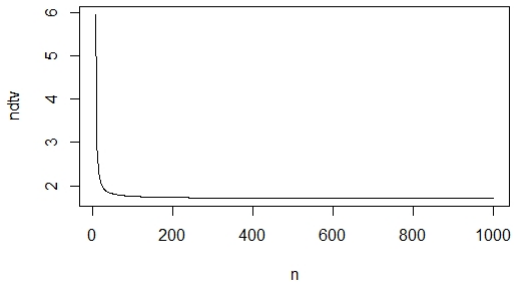


Figure 5: Plot for nd_{TV} for multiple values of n

R Code for Figure 5:

```
[language=R]
#ndtv for multiple values of n
library(distrEx)
library(distr)
n = 7
xaxis = seq(7,1000)
results3 = list()
while (n <= 1000){
  c = 7
  x = TotalVarDist(Binom(size = n, prob = c/n), Pois(c))
  nx = n*x
  results3 = c(results3, nx)
  if (n == 1000){
    plot(xaxis, unlist(results3), type="l", xlab = "n",
    ylab = "ndtv")} n = n + 1}
```

On the plot, it seems that there is a convergence towards a certain value. A result from a research paper by Prokhorov [7] tells us that

$$\lim_{n\to\infty} nd_{TV} = \frac{c}{\sqrt{2e\pi}}.$$

We investigated this and found that such a convergence exists for this particular plot at approximately 1.703 to 3 decimal places. It appears that convergence is a common feature in plots of this kind, with its visibility increasing at sufficiently large enough n. Our plot agrees with Prokhorov's result and shows that nd_{TV} is quite rapid.

3.4 nd_{TV} for multiple values of c

This builds up on Section 3.2, where we have a fixed n and c is increasing. As with Section 3.2, as we increase c we have a larger total variation distance. Figure 6 shows a plot for $0 \leq c \leq 1000$ and $n = 1000$.

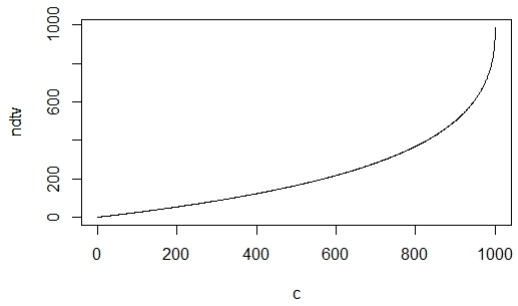


Figure 6: Plot for nd_{TV} for multiple values of c

R Code for Figure 6:

```
[language=R]
#ndtv for multiple values of c
library(distrEx)
library(distr)
c = 0
results4 = list()
while (c <= 1000){
  n = 1000
  x = TotalVarDist(Binom(size = n, prob = c/n), Pois(c))
  nx = n*x
  results4 = c(results4, nx)
  if (c == 1000){
    plot(unlist(results4), type="l", xlab = "c",
    ylab = "ndtv")} c = c + 1}
```

If we were to keep n constant, there would be no clear convergence here. The final value of nd_{TV} is 987.382 at $n = 1000$ and $c = 1000$, having $c > 1000$ would cause the probability c/n in the Binomial to be greater than one. We cannot continue calculating the total variation distance here and so we conclude that there is no convergence for a fixed n. For large enough n, we will find convergence.

3.5 The second term in the asymptotic expansion of d_{TV}

Definition 4. We have functions $f(i)$ and $g(i)$. We introduce the "little-o" notation for a function $o(g(i))$, where informally we can say "little o of g of i". $f(i) = o(g(i))$ means that

$$\forall k > 0 \, \exists m > 0 \, \forall i \geq m : 0 \leq f(i) < kg(i). \tag{5}$$

The value of m does not depend on i, but it may depend on k [8].

$f(i) = o(g(i))$ also means that

$$\lim_{i\to\infty} \frac{f(i)}{g(i)} = 0 \quad [9]. \tag{6}$$

We now define an asymptotic sequence using the “little-o" notation. A finite or infinite sequence of functions $\phi_i(z)$, $i = 1,2,\ldots$ and $z \in \mathbb{C}$ is defined to be an asymptotic sequence as $z \to z_0$ if,

$$\phi_{i+1}(z) = o(\phi_i(z)) \tag{7}$$

and also that $\lim_{z\to z_0} \frac{\phi_{i+1}(z)}{\phi_i(z)} = 0$.

From our definition of an asymptotic sequence, we say that $\sum_{i=1} a_i\phi_i(z)$, where the a_i are constants, is an asymptotic expansion or an asymptotic approximation for a function $f(z)$ if for every N

$$f(z) = \sum_{i=1}^{N} a_i\phi_i(z) + o(\phi_N(z)) \quad [10]. \tag{8}$$

The following displays a recursive method for finding the first two terms of the asymptotic expansion for d_{TV}:

$$\phi_1(n) = \frac{1}{n}$$
$$\phi_2(n) = \frac{1}{n^2}$$

$$a_1 = \lim_{n\to\infty} nd_{TV} = \frac{c}{\sqrt{2e\pi}}$$

$$a_2 = \lim_{n\to\infty} \left(n^2 d_{TV} - a_1 n\right) \tag{9}$$

$$d_{TV} \approx \frac{a_1}{n} + \frac{a_2}{n^2}$$

With $c = 7$, the convergence in (9) certainly occurs at $n = 100000$. The plot seen in Section 3.3 clearly indicates a convergence towards a certain value and it was found for $7 \leq n \leq 1000$. Further calculations using R prove that the error of convergence is very small, so we can use $n = 100000$. Therefore, we choose this as our n. We now substitute our values into the equations above.

$$\phi_1(100000) = \frac{1}{100000} = 0.00001$$
$$\phi_2(100000) = \frac{1}{(100000)^2} = 0.0000000001$$
$$a_1 = \frac{7}{\sqrt{2e\pi}} = 1.6937950716$$
$$a_2 \doteq d_{TV}(100000)^2 - 100000a_1 = 945.68$$

Thus, our prediction is that a_2 is about 946 for $c = 7$.

4 Conclusion

Our program allowed us to see how the total variation distance behaves for our parameters and distributions. We proved the property that the total variation distance can be written as a finite sum. Two other properties were also looked at, being that the distance is on the metric space and that $0 \leq d_{TV} \leq 1$. We have managed to find the asymptotics for a scenario with nd_{TV}.

In the future, we hope to address the following:

- The convergence is very fast for nd_{TV} where c is fixed. A program with more precision would be useful when finding say a_2.
- We would like to extend the paper by Prokhorov [7] further by finding higher order expansions.

5 Acknowledgements

Yasir Barlas is appreciative to Queen Mary, University of London, the 'Faculty of Science and Engineering' and the 'School of Mathematical Sciences' for being selected on the internship programme for this project. The 'BAME Undergraduate Research Internship' is sponsored by IBM and allows for undergraduate students to be involved in a research project. We are grateful to the referee for the recommended revisions during the publication process.

Bibliography

[1] Bortkiewicz, L. von. (1898) Das Gesetz der kleinen Zahlen [The law of small numbers]. Available at: `https://archive.org/details/dasgesetzderklei00bortrich/`.

[2] Quine M. P. and Seneta E. (1987) Bortkiewicz's data and the law of small numbers. *International Statistical Review / Revue Internationale de Statistique*, **55**(2):173–181, 1987.

[3] Sason, I. (2018) On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, **20**(5):383.

[4] R Core Team. (2021) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.*

[5] Ruckdeschel P., Kohl M., Stabla T. and Camphausen F. (2006) S4 classes for distributions. *R News*, **6**(2):10-13.

[6] Kennedy, J. E., and Quine, M. P. (1989) The total variation distance between the binomial and Poisson distributions. *The Annals of Probability*, **17**(1).

[7] Prokhorov, Yu. V. (1953) Asymptotic behavior of the binomial distribution. *Uspekhi Mat. Nauk*, **8**(3(55)):135-142.

[8] Black, P. E. (2019) little-o notation. *Dictionary of Algorithms and Data Structures* [online]. Available at: `https://www.nist.gov/dads/HTML/littleOnotation.html`.

[9] de Bruijn, N. G. (1981) Asymptotic Methods in Analysis. *Bibliotheca mathematica, Dover Publications.*

[10] Murray, J. D. (1984) Asymptotic Analysis: Asymptotic expansions. *Applied Mathematical Sciences*, **48**:1-18.