MATEMATIKA, 2018, Volume 34, Number 2, 187–204

Range-Box Plotting Relating to Discrete Distribution

[1]Mohd Bakri Adam*, [2]Babangida Ibrahim Babura and [3]Kathiresan Gopal

[1,3]Institute for Mathematical Research, Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

[3]Department of Mathematics, Federal University Dutse

*Corresponding author: bakri@upm.edu.my

Article history
Received: 22 September 2016
Received in revised form: 14 June 2017
Accepted: 31 October 2017
Published on line: 1 December 2018

Abstract The box plot has been used for a very long time since 70s in checking the existence of outliers and the asymmetrical shape of data. The existing box plot is constructed using five values of statistics calculated from either the discrete or continous data. Many improvement of box plots have deviated from the elegant and simplier approach of exploratory data analysis by incorporating many other statistic values resulting the turning back of the noble philosophy behind the creation of box plot. The modification using range value with the minimum and maximum values are being incorporated to suit the need of selected discrete distribution when outliers is not an important criteria anymore. The new modification of box plot is not based on the asymmetrical shape of distribution but more on the spreading and partitioning data into range measure. The new propose name for the box plot with only three values of statistics is called range-box plot.

Keywords Exploratory data analysis; box plot; range-box plot; discrete distribution

Mathematics Subject Classification

1 Introduction

The reputation of box plot is phenomenon. Although it has been introduced by Tukey [1] or slightly earlier, many of it users have forgotten to cite the great contributor of box plot. Many researchers using it without referencing it to Tukey [1] as it's become a cliche. In this research, some characters of discrete distributions i.e. binomial will be explored using the box plot. By looking at the box plot alone we can distinguish and roughly recognise what is the representation distribution for any discrete data set. In this paper the modification of the box plot using range of the data is used to construct new type of box plot.

Tukey [1] works have not be limited ony to the usage of box plot for the continuous data type. Shitan and Vazifedan [2] use the discrete data as their examples and exercises to construct classical box plot. Velleman and Hoaglin [3] also do not set the constraints for the type of data. Hubert and Vandervieren [4], McGill *et al.* [5] and Nuzzo [6] discuss about the variability of

the usage and changes for the classical box plots but still stick to the five values of statistic in their new recommendations. McGill et al. [5] and Nuzzo [6] ignore the skewness characteristic of any distribution when constructing any box plot. Hubert and Vandervieren [4] incorporate the skewness aspect but not an impressive move as the approaches is becoming more complexes in contructing the boxplot. Babura *et al.* [7] also do some modification to the construction of box plot but focusing on the extreme data applications. Ferreira *et al.* [8] use the box plots in their research to present some of the results using existing box plot.

2 Development of Box Plot

Since 1977 there are a few developments to improve the construction of the box plot. Previously, Tukey [1] introduces box plot with only five value of statistics and easier to use, that's the forgotten spirit and soul relating closely to the philosophy of exploratory data analysis. Box plot when was introduced by Tukey [1] is a simplier graphical statistical tool with a non-parametric approaches with no intention of usage any inference intention.

Schwertman *et al.* [9] introduce a more simple general box plot method for identifying outliers by estabishing the probabilistics basis for fences but no attempt to draw the box plot graphically but focusing more on identified the outliers and extreme values. Schwertman & de Silva [10] once again further their research by incorporating identification of multiple outliers with simulation comparison of sequential fences to the box plot rule and generalized extreme standardized deviation test but still not shown it graphically in their findings.

Carter *et al.* [11] compare the two methods in indentifying the existence of outliers also without any sketch of the box plots. Although improvement have been done but the structure of the works becoming complicated as many other value of statistics have been used. Still the classical box plot has a great and remarkable contribution to the data analysis works in other to obtain important information. Many users of box plot for many data set, there are possibility of the existance of outliers but that assumption is not really true especially when come to any simulation study or from data which have the characteristics of heavy tail distributions. Outliers serve as the indicator for allocating suitable distributions. Robustness characteristics is not statistically a main issue anymore.

Hubert and Vandervieren [4] introduce an adjusted box plot for skewed distribution by incorporating a robust measure of skewness with the use of medcouple. The adjusted box plot is becoming more complicated when more complex mathematics are sued which is not a favor movement in exploratory data analysis philosophy which is *simple and straight forward to construct* when the box plot is being introduced earlier by Tukey [1]. Similar complex and difficult approaches have been done by many researches, [9–13] which deviated from the earlier concept of exploratory data analysis.

3 Classical Box Plot

Let $x_1, x_2, \ldots, x_n$ or $\{\boldsymbol{x}\}$ be observations from any discrete distributed data. The box plot can be constructed from $x_1, x_2, \ldots, x_n$ using the five statistic values i.e. the median, x_m, first quartile, Q_1, third quartile, Q_3, the minimum value, x_{min}, and the maximum, x_{max}, values. Later, other values such as inter-range quartile, IQR, inner fence, f_i and outer fence, f_o can

be calculated. On the other hand, outliers are the observations which appear to be higher than upper f_i or lower f_i i.e. the outliers are to be different from the bulk of the data. A moderate outlier is an observation which lies between f_i but within f_o. An extreme outlier is an observation which has bigger value than the upper f_o or smaller value than lower f_o. See [2, 7].

In order to get x_m, we need to sort the observations to be $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where the x_m is the value of the middle of the sort data. The value of x_m is depend on either the number of observations, n, is odd, $x_m = x_{\left(\frac{n+1}{2}\right)}$, or when n is even, $x_m = \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right]$, but when the data becoming larger then percentile approaches is used to get the x_m value. Later the first quartile Q_1 and the third quartile Q_3 of the data should be obtained. All the materials here only require a basic knowledge of statistics. The inter-quartile range is IQR where $IQR = Q_3 - Q_1$. Then the lower inner fence is defined as,

$$f_i^L = Q_1 - \mathbf{1.5} \times IQR$$

while the upper inner fence is given as

$$f_i^U = Q_1 + \mathbf{1.5} \times IQR.$$

The lower outer fence is defined as,

$$f_o^L = Q_1 - \mathbf{3.0} \times IQR$$

while the upper outer fence is given as

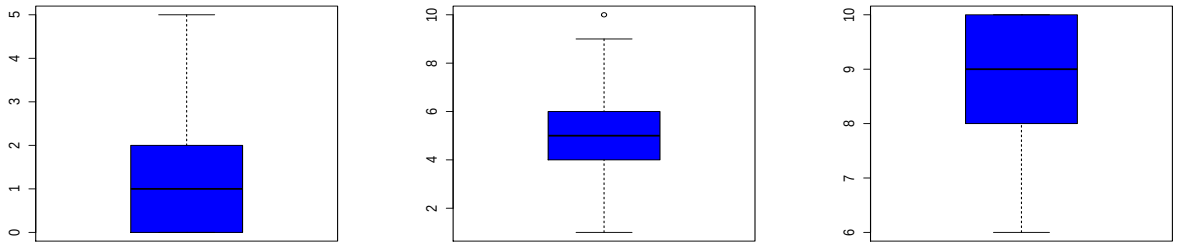
$$f_o^L = Q_1 + \mathbf{3.0} \times IQR.$$

Although, Tukey [1] introduce the box plot to the world and the plot is very popular among researcher but the justification of selecting the formula for inner and outer fences are not clearly defined i.e. how the selection of the multiplicative values of 1.5 and 3.0 in the formula are still not clearly being justified.

3.1 Constructing the Box Plot

The box plot consists of a box frame with the edge boundaries are between the Q_1 and Q_3, and 50 percent of data populated inside this frame. The x_m is drawn as a line inside the box frame. From the boundaries of the frame box, a whiskers line (if exist) is drawn stretching from either the wall of Q_1 to minimum value of x_i before or equal to the f_i^L and the other whisker line is drawn from the wall of Q_3 up to the largest value of x_i but lower or equal to f_i^U. Observation(s) outside the f_i is (are) considered as an outlier(s) or extreme outlier(s). Here the only focus is on the skeletal box plot i.e. the basic structure of box plot. Outliers usually have been marked with a circle or other symbols.

If all the observations i.e. 25% of the lower data is equal to the Q_1 or all the observations i.e. 25% of the upper data is equal to Q_3 then no whiskers line is drawn. If all values in Q_1, Q_2, Q_3 are ties then no box frame is drawn. Carefully use of the box plot should be imposed if the previous scenario happened. Refer to Figure 1.



(a) (b) (c)

Figure 1: (a). Typical Box Plot for $\mathbf{B}(n_b = 10, p = 0.1)$; (b). Typical Box Plot for $\mathbf{B}(n_b = 10, p = 0.5)$; (c). Typical Box Plot for $\mathbf{B}(n_b = 10, p = 0.9)$ with Number of Simulation Data is 1000.

4 Introducing Range-Box Plot

In constructing range-box plot, the usage of only three value of statistics are range, minimum and maximum values compared to five value of statistics for the classical box plot and this range-box plot can be used when the outliers are not really an important criterium in the analysis. As in many cases, many data set did not come from an ideal normal distribution. By incorporating range, simpler approaches is introduced in constructing the new contruction of an alternative box plot. Range is the difference between the biggest value and the smallest value in the data set. In this paper, we will introduce two type of range-box plots i.e. Range-box plot Type I and Range-box plot Type II.

4.1 Range-Box Plot Type I

Similar notation has been incorporated for median, x_m. Instead of using the IQR value which is based on the Q_1 and Q_3 values, in this study as the search for outliers is not the main priority, we are using the range value to get the new IQR_{range} using the following formulation. Our range is defined as $R = x_{(n)} - x_{(1)}$. Then we have the first range quartile $Q_1^{R_1} = x_{min} + 0.25 \times R$ and the third range quartile $Q_3^{R_1} = x_{max} - 0.25 \times R$ of the data. Next, we can calculate and incorporate the total number of observations fall in each partition. By ignoring the procedure of identifying outliers and extreme values, the step of calculating lower inner fence, upper inner fence, lower outer fence and upper outer fence i.e. f_i^L, f_i^U, f_o^L and f_o^U are eliminated. The modification of the boxplot will only ended with the whisker lines for both side of the frame box if available. The frame of box will be drawn where the median, x_m, is located. The data is partitioned into three parts from x_{min} to $Q_1^{R_1}$, $Q_1^{R_1}$ to $Q_3^{R_1}$ and $Q_3^{R_1}$ to x_{max} with some interpretation with care should be imposed when either $x_{\min} = x_m = Q_1^{R_1}$ or $x_{\max} = x_m = Q_3^{R_1}$. We name this new box plot as **range-box plot Type I**. Process of constructing the range-box plot is much simpler and faster compared to the classical box plot. The program in constructing the new range-box plot is written in R programming.

4.2 Range-Box Plot Type II

The construction for the box plot Type II, the first range quartile $Q_1^{R_2} = x_{min} + \frac{1}{3} \times R$ and the third range quartile $Q_3^{R_2} = x_{max} - \frac{1}{3} \times R$ of the data have been calculated. Similar to the Type I, the frame of box will be drawn where the median, x_m, is located. The data are partitioned into three parts as in Type I by replacing the R_1 notation with R_2.

Tukey [1] uses a few examples in introducing classical box plot. Here we will show the limitation of the existing classical box plot. One of the common limitation is that the classical box plot has divided the data set into four portions. The existing classical box plot not suitable for discrete data and also for long tail type of data.

5 Discrete Distributions

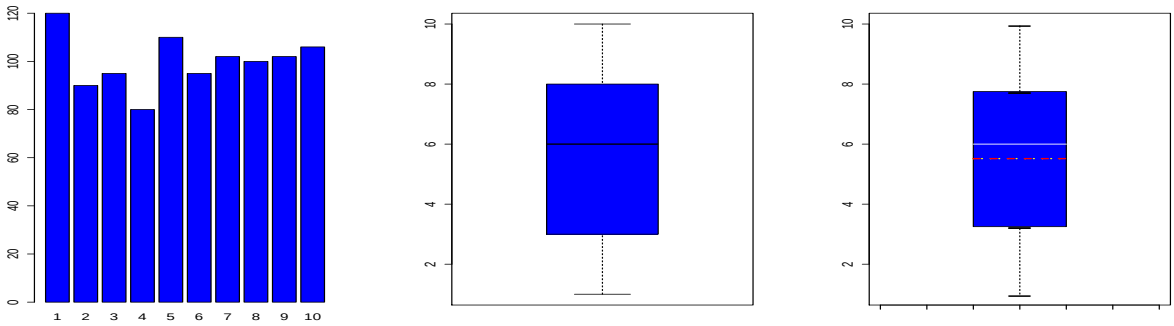
In this paper we are presenting all the box plots for discrete uniform, bernoulli, poisson, negative binomial, geometric, hyper-geometric, binomial and binomial's approximation i.e. poisson and normal distributions. Later, we are only focusing on Binomial Distribution as it's can be related to the other distributions by changing some character of the parameters, when the sample data are becoming bigger (asymptotic distribution) or from some transformations of data. The relationship between these distributions can be referred in [14] and [15].

Table 1: Selected Discrete Distributions

Distribution	Notation	Probability Mass Function, pmf	
Uniform	$\mathbf{U}(a, b)$	$\frac{1}{n}$,	$x = x_1, x_2, \ldots, x_n$
Bernoulli	$\mathrm{B}(1, p)$	$p^x(1-p)^{1-x}$,	$x = 0, 1$
Poisson	$\mathrm{P}(\mu)$	$\frac{\mu^x \exp(-\mu)}{x!}$,	$x = 0, 1, 2, \ldots.$
Geometric	$\mathrm{G}(p)$	$p(1-p)^{x-1}$,	$x = 1, 2, \ldots.$
Hypergeometric	$\mathrm{HG}(n, N, K)$	$\frac{C_x^K C_{n-x}^{N-K}}{C_n^N}$,	$x = \max\{n - N + K, 0\}, \ldots, \min\{K, n\}.$
Binomial	$\mathrm{B}(n, p)$	$\binom{n_b}{x} p^x (1-p)^{n_b - x}$,	$x = 1, 2, \ldots, n_b.$
Negative Binomial	$\mathrm{BN}(p)$	$C_{k-1}^{x-1} p^k (1-p)^{x-k}$,	$x = k, k+1, \ldots.$

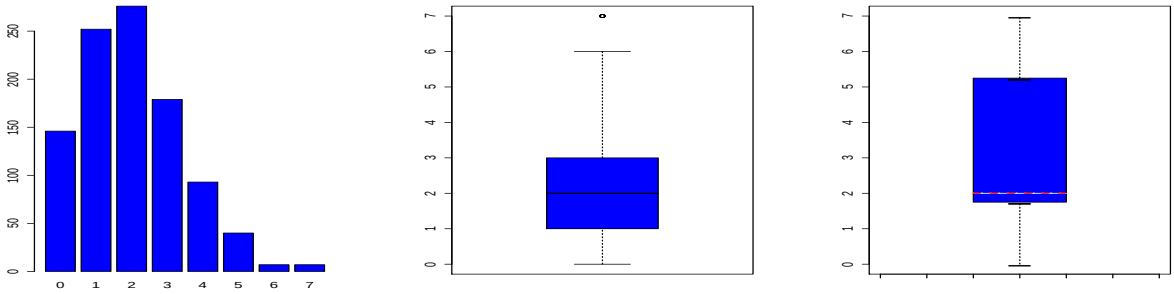
Notes: The notations used are from the standard way of Wiley series in probability and statistics

The discrete uniform distribution is a distribution that give a finite number of outcomes equally likely to happen. The ties always exist and the outliers are possible to be occured. The famous example for discrete uniform distribution is from throwing a fair die. The probability of a given scores of values 1, 2, 3, 4, 5 and 6 is $\frac{1}{6}$. The mean and median are the same with no mode value. The skewnes of the distribution is zero i.e. a symmetric probability distribution.



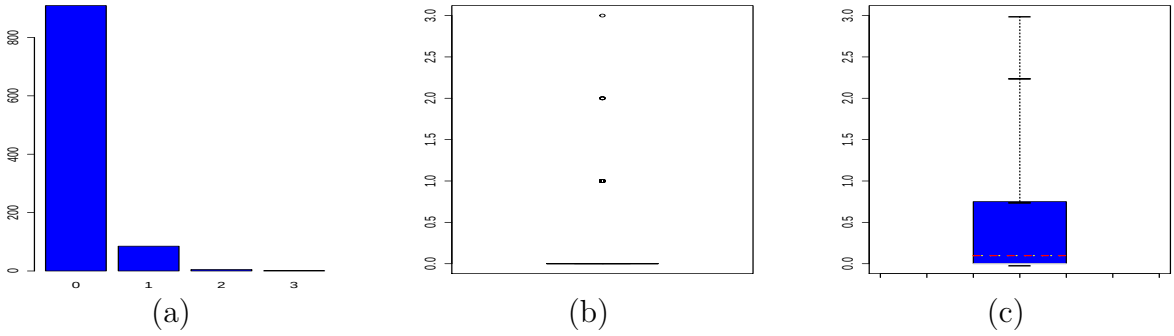
(a) (b) (c)

Figure 2: (a). Bar Plot for $\mathbf{U}(a = 1, b = 10)$; (b). Box Plot for $\mathbf{U}(a = 1, b = 10)$; (c). Range-box Plot Type I for $\mathbf{U}(a = 1, b = 10)$ with Number of Simulation Data is n =1000.



(a) (b) (c)

Figure 3: (a). Bar Plot for $\mathbf{P}(\mu = 2)$; (b). Box Plot for $\mathbf{P}(\mu = 2)$; (c). Range-box Plot Type I for $\mathbf{Poi}(\mu = 2)$ with Number of Simulation Data is n =1000.



(a) (b) (c)

Figure 4: (a). Bar Plot for $\mathbf{G}(p = 0.9)$; (b). Box Plot for $\mathbf{G}(p = 0.9)$; (c). Range-box Plot Type I for $\mathbf{G}(p = 0.9)$ with Number of Simulation Data is n =1000.

In the later research we will only focusing on binomial and its approximation distributions. The binomial distribution can be linked with other distributions such as Poisson, Bernoulli, Hypergeometric and Normal distributions. The link between the binomial distribution with other distributions is in stages indirect-relationship leading to Discrete Uniform and Geometric distributions. While Geometric and Negative Binomial is interchangable depend on its param-

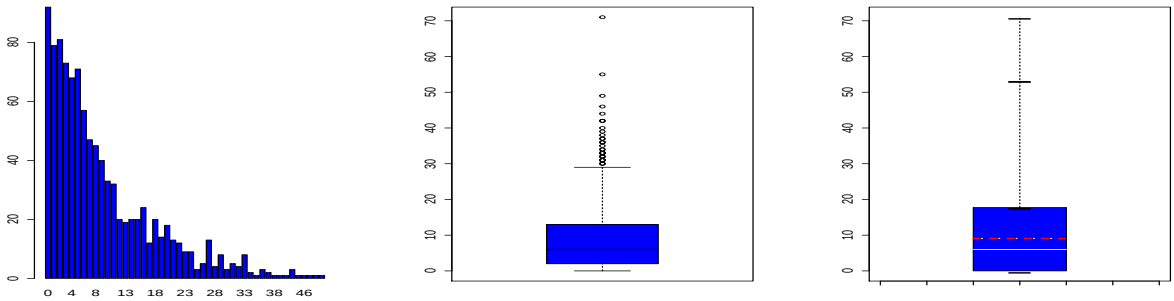


Figure 5: (a). Bar Plot for $\mathbf{G}(p = 0.1)$; (b). Box Plot for $\mathbf{G}(p = 0.1)$; (c). Range-box Plot Type I for $\mathbf{G}(p = 0.1)$ with Number of Simulation Data is n =1000.

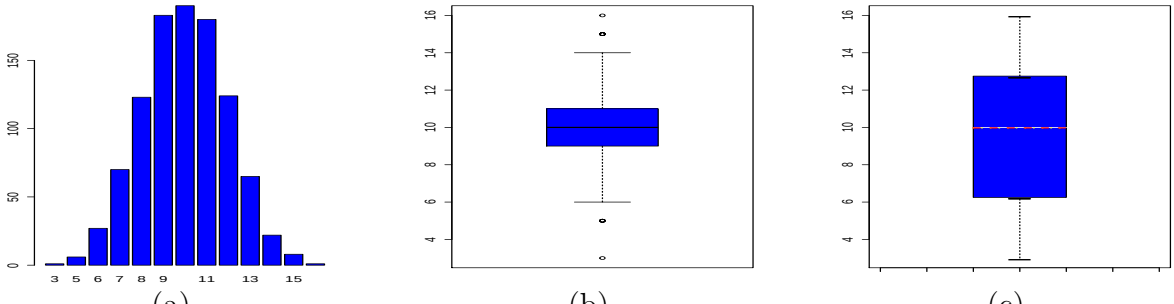


Figure 6: (a). Bar Plot for $\mathbf{HG}(50, 50, 20)$; (b). Box Plot for $\mathbf{HG}(50, 50, 20)$; (c). Range-box Plot Type I for $\mathbf{HG}(50, 50, 20)$ with Number of Simulation Data is n =1000.

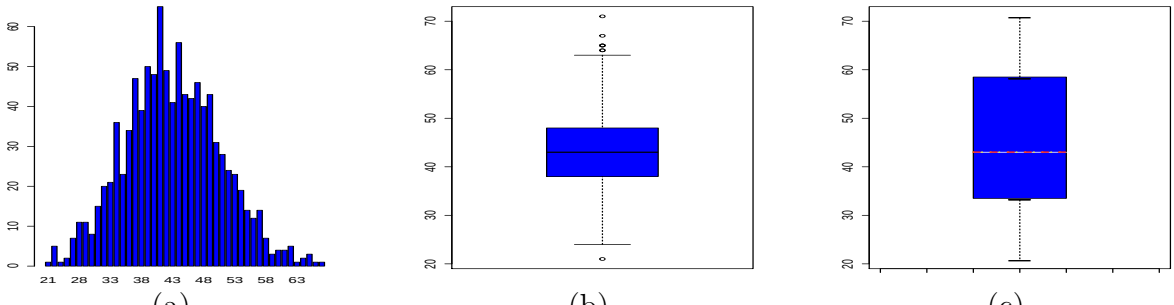
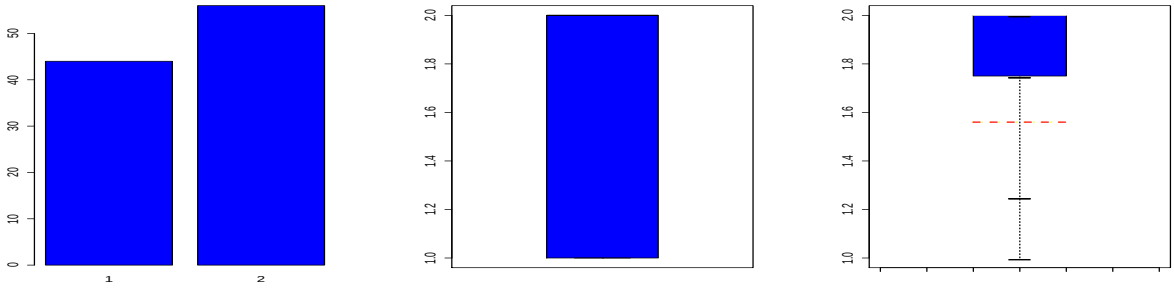


Figure 7: (a). Bar Plot for $\mathbf{NB}(n_{nb} = 100, p = 0.7)$; (b). Box Plot for $\mathbf{NB}(n_{nb} = 100, p = 0.7)$; (c). Range-box Type I Plot for $\mathbf{NB}(n_{nb} = 100, p = 0.9)$ with Number of Simulation Data is n =1000.

eter values. The box plots and range-box plots for all these discrete distributions are shown in Figures 2-8. For uniform distribution, both plots in Figure 2 are confusing and give wrong interpretation. Figure 3 shows the existing of a point of outliers which misleading as this is a simulated data from Poisson distribution. The existing box plot cannot give any information for Geometry distribution compared to range-box plot, see Figure 4. While in Figure 5 the



(a) (b) (c)

Figure 8: (a). Bar Plot for $\mathbf{B}(n = 1, p = 0.4)$; (b). Box Plot for $\mathbf{B}(n = 1, p = 0.4)$; (c). Range-box Plot Type I for $\mathbf{B}(n = 1, p = 0.4)$ with Number of Simulation Data is n =1000.

box plot gives more outliers which is important to characterize the geometry distribution with small p value. Similar discussion for Figures 6, 7 and 8.

The binomial distribution also can be related to approximation distribution either discrete or continuous distribution i.e. the approximation binomial to either poisson or normal distribution depending on the value of p. The characteristic of binomial distribution is as follows:

(i) Two possible outcomes with constant probability of success, $0 < p < 1$, for each trial.

(ii) Each trial is independent.

(iii) It's happen in n_b times. The n_b is used here to differentiate with the size of sample n.

As we appointed out that for the binomial distribution condition, the point of outlier is not really critical and important in profiling the shape of the distribution, as the profile of the distribution solely depend on the value of p parameter. The tail of the distribution is not one of the important criteria or issue for binomial distribution. The usage of box plot now is not to reveal the robustness characteristics but more into given extra information about the data. From Equation (1) we clearly can check that the p^x and $(1-p)^{n_b-x}$ are going to be zero very fast as $n_b \to \infty$ at the same time $\binom{n_b}{x}$ is becoming very large. When $n_b \to \infty$, the approximation distributions are recommended.

Figure 1 shows three simulation of binomial data from three different p values. We observe that the p values give three slightly different box plots. If $0 \leftarrow p$ then the posible outliers will be on the right, where $p \to 1$ then the possible outliers will be on the left, while the p value around 0.5 will give less outliers with symmetrical shape of the box plot, but the outliers in the discrete binomial distribution not give any important character to this distribution as binomial distribution is not a heavy tail distribution type.

Table 2 shows the allocation of data when we increase the p value from 0.1 to 0.9, the character of skewness change i.e. skew to right to skew to the left. We also can find similar result if we increase the n_b from 10 to 100 or ∞.

In theoretical statistics usually when either $0 \leftarrow p$ or $p \to 1$ and $n_b \to \infty$, the discrete binomial distribution will be approximated using poisson distribution $\mathbf{P}(\mu = n_b \times p)$ where $\mu = \sigma^2$. Wherelse if $p \approx 0.5$ the usual practice is to approximate the binomial with Normal distribution where $\mu = n_b p$ and $\sigma^2 = n_b p(1-p)$.

Table 2: The Range of Percentage Allocation of Discrete Values $(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$ Simulate from the Binomial Distribution $X \sim \mathbf{B}(n_b = 10, p)$ for n =1000 (sample size).

$p \backslash n_b$	0	1	2	3	4	5
0.1	25-45	26-51	13-30	1-12	0-4	0-2
0.3	0-9	6-19	13-40	14-38	11-33	3-19
0.5	0-2	0-8	1-15	4-25	12-28	11-34
0.7	-	-	0-1	0-3	0-12	2-28
0.9	-	-	-	-	-	0-1
$p \backslash n_b$	6	7	8	9	10	
0.1	0-1	-	-	-	-	
0.3	0-9	0-3	0-1	0-1	-	
0.5	9-29	2-20	0-12	0-5	0-4	
0.7	9-30	16-38	9-34	1-21	0-9	
0.9	0-5	2-12	10-29	26-51	23-49	

5.1 Ties Phenomena

Another challenge in box plotting for the binomial discrete is the occurance of repeated integer values generated especially when $p \to 1$ or $0 \leftarrow p$. For example if we simulate the data from $\mathbf{B}(n_b = 10, p)$ with n =1000 observations, what we get is in Table 2, many occurance of ties. These repeated values creating the existance of too many ties. As in Table 2, the ties occur significantly around the median (mode and mean) value. We can see that, when we simulate 1000 data from $X \sim \mathbf{B}(n_b = 10, p = 0.1)$ approximately 26 to 51% the values are 1, the value of 0 will be range from 25-45% from the total number of observations. This is an obvious characters of binomial discrete data. The tie values will follow the selection of the p value. We can see clearly that for $X_p \sim \mathbf{B}(n_b = 10, p)$ the $X_p \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, but when $p = 0.1$ then $X_{0.1} \in \{0, 1, 2, 3, 4, 5, 6\}$ while when $p = 0.9$ then $X_{0.9} \in \{5, 6, 7, 8, 9, 10\}$. The $X_{0.1}$ and $X_{0.9}$ are subsets to X_p i.e. $X_{0.1} \subset X_p$ and $X_{0.9} \subset X_p$. It is why the ties phenomena appeared in the discrete binomial data need to be address when constructing the box plot. Saying there is an outliers in the discrete binomial data is really an awkward thing to do.

The existance of ties, will resultant to the disappearance of the whisker line for both side when either $0 \leftarrow p$ or $p \to 1$. See Figures 1 (a) and 1(c). The existing of outliers in Figure 1 (b) also misleading. Another observation is 75% of data fall inside the frame box as in Figures 1 (a) and 1(c) but it is not clearly shown as the ties happen on $x_{(1)}, x_{(2)}$ and $x_{(3)}$ for $0 \leftarrow p$ and $x_{(n-2)}, x_{(n-1)}$ and $x_{(n)}$ for $p \to 1$.

5.2 Skewness of Data

The skewness of the binomial distribution is depend on the value of p and the existance of outliers depend on the value of p. Figure 9 shows that from classical box plot, for both cases $p = 0.1$ and $p = 0.9$, its tend to give the indicator of the existence of outliers, but for the binomial discrete distribution, the outliers are not really an important character. It will lead to the misconception about the role of box plot. In Figure 9, we simulate three set of data from

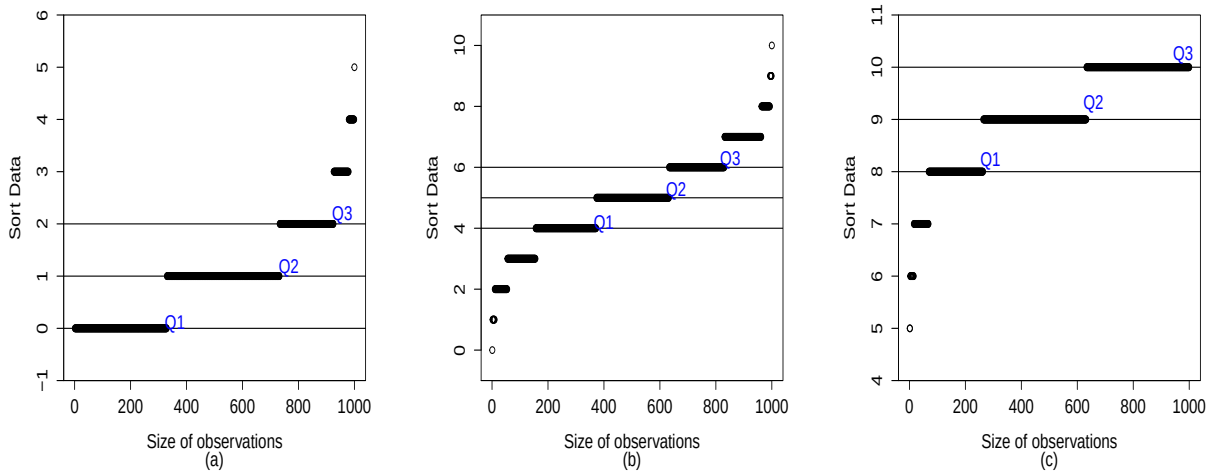


Figure 9: (a). Scatter Plot for Simulation of n =1000 Data from $X \sim \mathbf{B}(n_b = 10, p = 0.1)$ (b). Scatter Plot for Simulation of n =1000 Data From $X \sim \mathbf{B}(n_b = 10, p = 0.5)$ (c). Scatter Plot for Simulation of n =1000 Data from $X \sim \mathbf{B}(n_b = 10, p = 0.9)$

$\mathbf{B}(n_b = 10, p)$ with 1000 observations. It is very obvious that we can see the pattern of skewness. For the case of $\mathbf{B}(n_b = 10, p)$, we observe that $Q_1 = x_{(1)} = x_{\min} = 0$, $Q_2 = x_{(2)} = 1$ and $Q_3 = x_{(3)} = 2$ for $0 \leftarrow p$ where else $Q_1 = x_{(9)} = 8$, $Q_2 = x_{(10)} = 9$ and $Q_3 = x_{(11)} = x_{\max} = 10$ for $p \rightarrow = 1$.

Figure 10 shows that the highest peak or mode for the distribution will be less than 50%. Table 2 also shows that for $p = 0.1$ the mode is 1, $p = 0.3$ the mode is 3, $p = 0.7$ the mode is 7 and for $p = 0.9$ the mode is 9. The theoretical mode for binomial $\mathbf{B}(n_b, p)$ distribution is as follows

$$\text{mode} = \begin{cases} \lfloor (n_b + 1)p \rfloor & \text{if } (n_b + 1)p \text{ is 0 or noninteger,} \\ (n_b + 1)p \text{ and } (n_b + 1)p - 1 & \text{if } (n_b + 1)p \in \{1, \ldots, n_b\}, \\ n_b & \text{if } (n_b + 1)p = n_b + 1. \end{cases}$$

where $\lfloor . \rfloor$ is the floor function. For $\mathbf{B}(n_b = 10, p)$, the mode and median for the simulated data give the same value with slightly different value of mean. When $n_b \rightarrow \infty$, some adjustment need to be carried out, but the mode become clearer if $n \rightarrow \infty$.

5.3 Properties of Range-Box Plot for Discrete Binomial Data

The properties of range-box plot are mostly equivalent in purpose with the features of classical box plot i.e. which contained the box and the whisker lines except for detecting the outliers. The only main different is the range-box plot will ignore the existance of outliers as the outlier is not important to characterise, the binomial distribution. Classical box plot actually divided the data into four partition, from x_{min} to Q_1, Q_1 to Q_2, Q_2 to Q_3 and Q_3 to x_{max}. The whisker exist in either first partition or the last partition but ended only to the lowest(highest) in the inner fence. Later the span of Q_1 up Q_3 is joined together to have the IQR span. Outliers should be in either the first partition or last partition if the value is lower than lower inner fence or outer than outer inner fence.

Meanwhile for the range-box plot, we considered using the range which also been partitioned by four but later the second and third partition are joint together to analog with the classical box plot to represent the inter-range quartile, between Q_1 and Q_3. The box is draw around the median value. The classical box plot divide the data into n_1, $n_2 + n_3$ and n_4 number of observations, where the total number of sample is $n = n_1+n_2+n_3+n_4$ where $n_1 = n_2 = n_3 = n_4$. For the range-box plot the numbers of observation is depend on where the data fall. The range-box plot can show clearly the skewness of the data compared to the classical data. In Figure 11 for the range-box plot, the box will be draw around the location of the median, later we also find graphically that the mean and mode values also fall in the drawn box. No identification of outliers is carried out as it is not an important criterium to be looked upon. The inter-range quartile is half range distance centred at the sample median and form the entire height of the box in the range-box plot. This is a replacement of quartile box in the standard classical box plot.

Let $X \sim \mathbf{B}(n_b, p)$. As $n \to \infty$ the asymptotic distribution of X will be partitioned into three distinct portions by the display of the range-box plot. This partition will vary in-terms of the density of sub-population when p takes values from 0 to 0.5 or from 0.5 to 1. To investigate the sub population density in each partition by the range-box plot we vary the binomial parameter

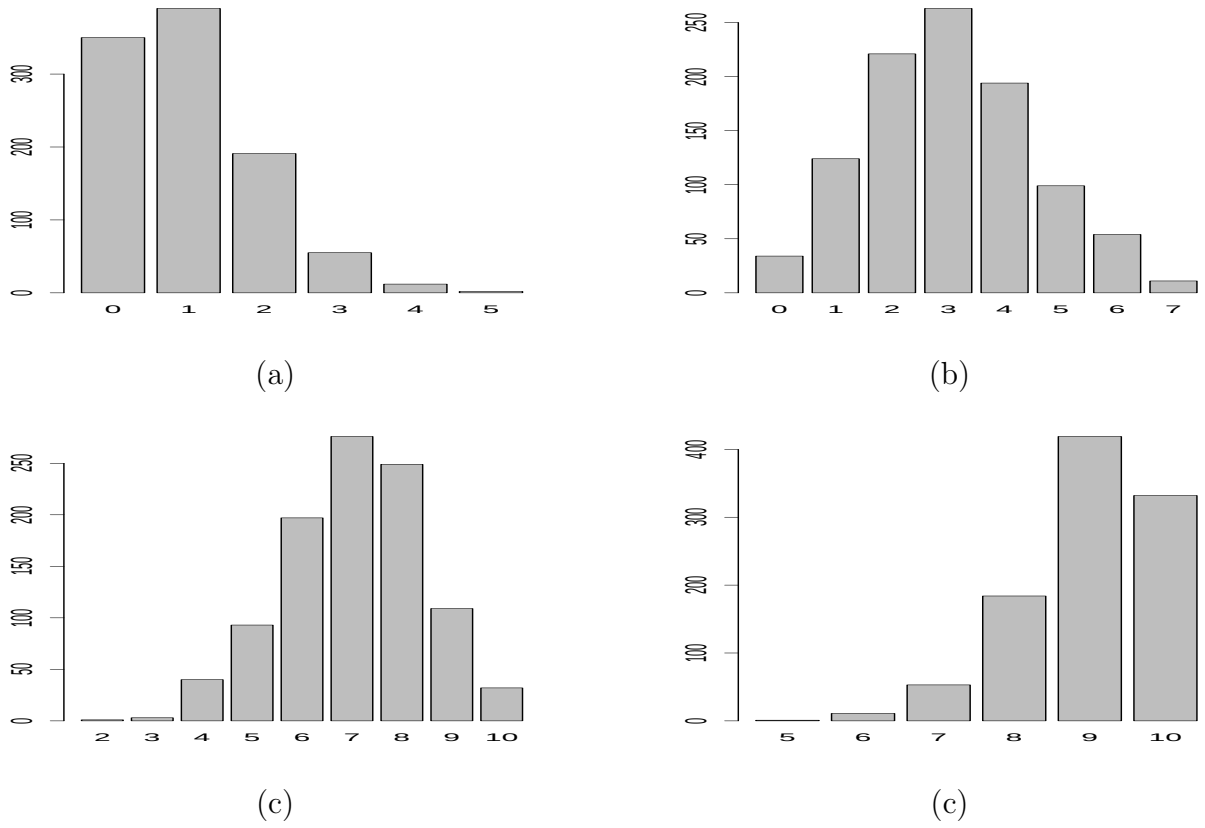
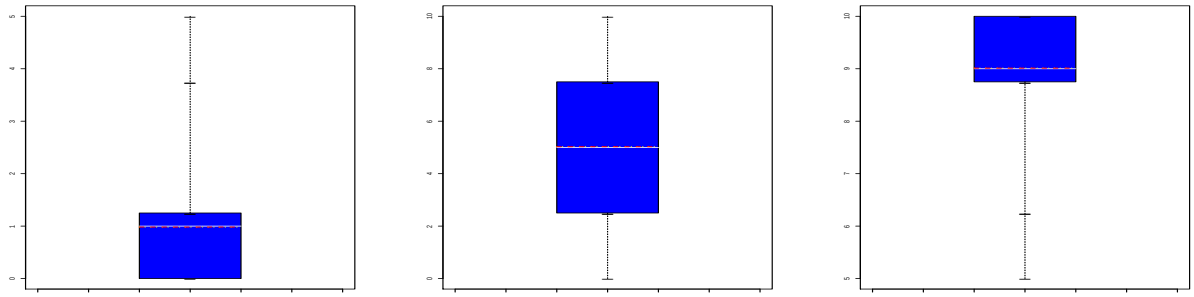


Figure 10: (a). Typical Bar Plot for $\mathbf{B}(n_b = 10, p = 0.1)$; (b). Typical Bar Plot for $\mathbf{B}(n_b = 10, p = 0.3)$; (c). Typical Bar Plot for $\mathbf{B}(n_b = 10, p = 0.7)$; (d) Typical Bar Plot for $\mathbf{B}(n_b = 10, p = 0.9)$ with Number of Simulation Data is 1000.



(a) (b) (c)

Figure 11: (a). Range-box Plot for $\mathbf{B}(n_b = 10, p = 0.1)$; (b). Range-box Plot for $\mathbf{B}(n_b = 10, p = 0.5)$; (c). Range-box Plot for $\mathbf{B}(n_b = 10, p = 0.9)$ with Number of Simulation Data is n =1000.

p from 0 to 0.5 and record the average density in each partition for a 1000 replicated samples for each fixed n_b and p values.

Table 3 shows that the standard classical box plot will not look at the partitional of data, but really on the existance of outliers. The number of observation are divided into three groups with the ratio of 25:50:25. In the range-box plot, we can observe than the partition of data is more sensitive to the value of p. When $0.1^+ \leftarrow p$ about 70-75% of the data will be in the first partition, about 24-28% will be in the middle partition i.e. $p \rightarrow 0.5^-$ or $0.5^+ \leftarrow p$, and 1-2% of data will be at the third partition when $p \rightarrow 0.9^-$. This trend changes when $p \rightarrow 0.1$ and $p \rightarrow 0.9$. The character of skewness can be observed clearly in the range-box plot compared to the classical box plot.

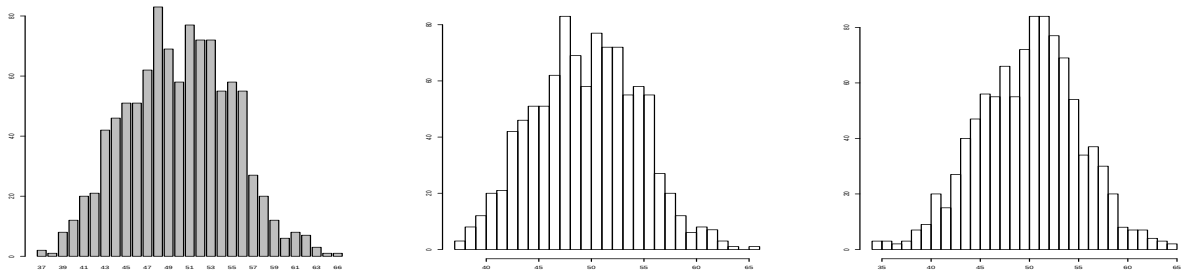
Another advantages of range-box plot is that all possible mode, mean and median of binomial discrete distribution can be captured inside the rectangular box. Figure 11 shows very clearly that the mean, mode and median values are located in each box frame for all p values.

5.4 Effect of Box Plot's Modification to Normal Approximation and Poisson Approximation for Binomial Distribution

It is more convenient for binomial distribution $\mathbf{B}(n_b, p)$ to be approximated with Normal distribution when $n_b \rightarrow \infty$ and p is approaching the value of 0.5. Additional requirements are $n_b p > 5$ and $n_b(1-p) > 5$ where $\frac{5}{n_b} < p < 1 - \frac{5}{n_b}$ [16]. Although some continuous correction to be imposed when changing the structure form of the data from discrete to continuous, the ap-

Table 3: The Portion of Observations Allocation on Each Quartile (IQR or IQR_{range}) for $n_b = 10$

	n_1	$n_2 + n_3$	n_4
Classical box plot	25%	50%	25%
Range-box plot $[p = 0.1]$	70-75%	24-28%	1-2%
Range-box plot $[p = 0.5]$	5-10%	77-88%	5-10%
Range-box plot $[p = 0.9]$	1-2%	24-28%	70-75%



(a) (b) (c)

Figure 12: (a). Bar Plot Plot for $\mathbf{B}(n_b = 100, p = 0.5)$; (b). Histogram for $\mathbf{B}(n_b = 10, p = 0.5)$; (c). Histogram for $\mathbf{N}(\mu = 50, \sigma^2 = 25)$ with Number of Simulation Data is 1000.

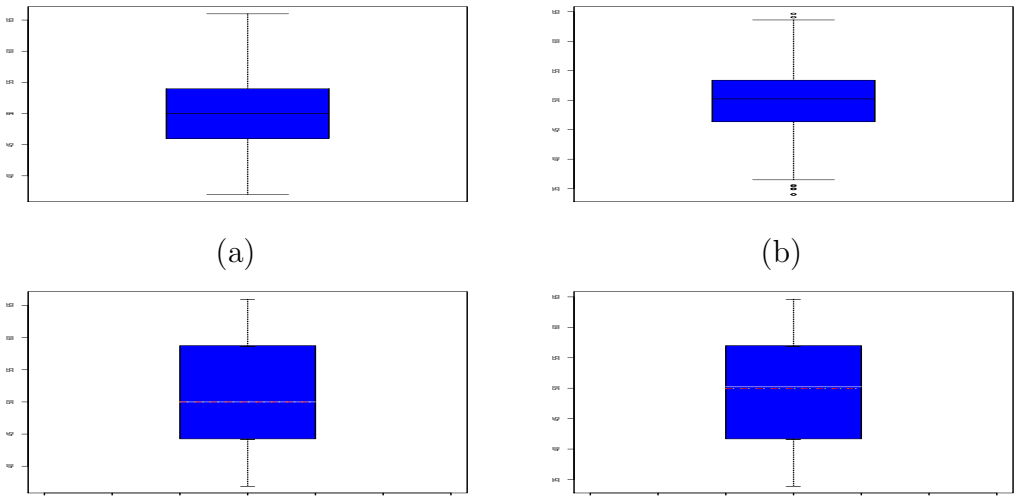
proximation seem to be working very well when the size of observations is increasing. Figure 12 shows the simulation data from the binomial and approximate normal using 1000 observations. Hypothesis testing for

$$H_0 : \mu_{\mathbf{B}(n_b,p)} = \mu_{\mathbf{N}(n_b p, n_b pq)}$$

versus

$$H_1 : \mu_{\mathbf{B}(n_b,p} \neq \mu_{\mathbf{N}(n_b p, n_b pq)}$$

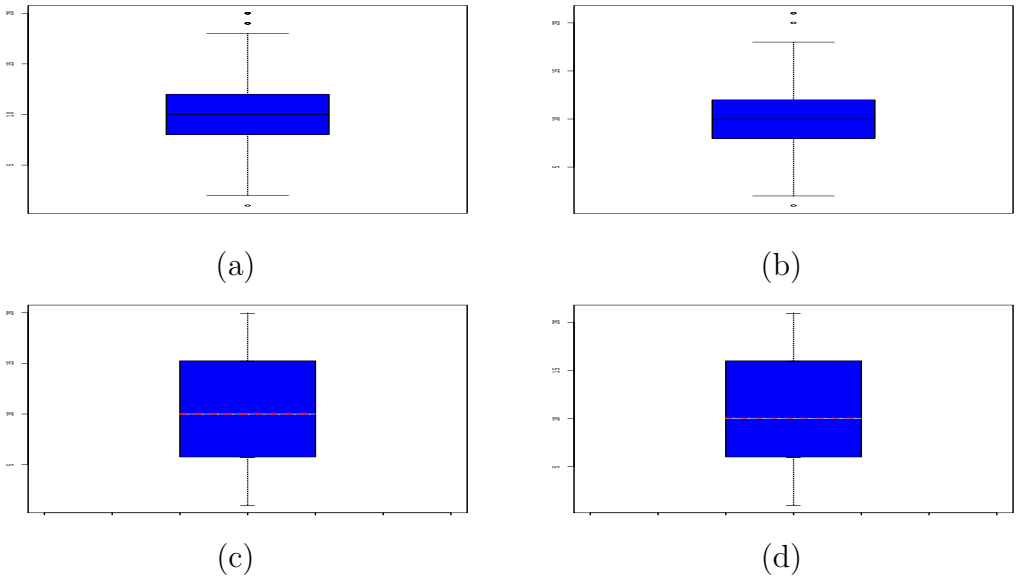
give a decision that cannot reject the H_0. These two distributions are the same.



(a) (b) (c) (d)

Figure 13: (a). Box Plot for $\mathbf{B}(n_b = 100, p = 0.5)$; (b). Box Plot for $\mathbf{N}(\mu = 50, \sigma^2 = 25)$; (c) Range-box Plot for $\mathbf{B}(n_b = 100, p = 0.5)$; (d). Range-box Plot for $\mathbf{N}(\mu = 50, \sigma^2 = 25)$ with Number of Simulation Data is 1000.

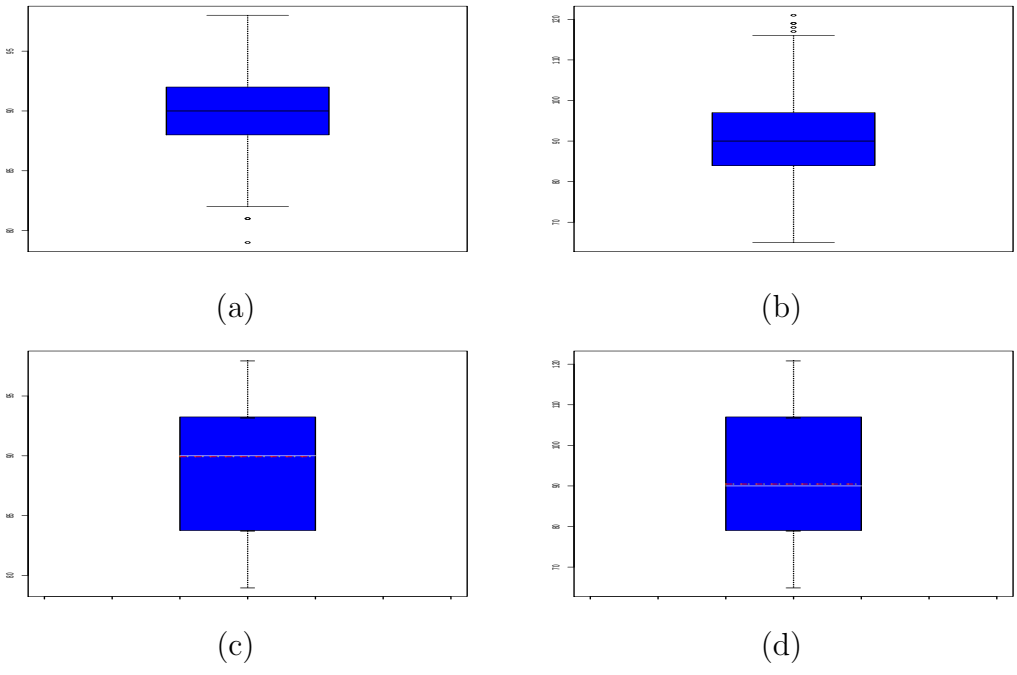
We also presenting the data in Figure 12 using classical box plot and range-box plot as in Figure 13. The classical box plot for the simulation data from normal distribution tend to give some outliers which is not the case for range-box plot. The variability for both either binomial and Normal approximation also quiet similar.



(a) (b)

(c) (d)

Figure 14: (a). Box Plot for $\mathbf{B}(n_b = 100, p = 0.1)$; (b). Box Plot for $\mathbf{P}(\lambda = 10)$; (c). Range-box Plot for $\mathbf{B}(n_b = 100, p = 0.1)$; (d). Range-box Plot for $\mathbf{P}(\lambda = 10)$ with Number of Simulation Data is 1000.



(a) (b)

(c) (d)

Figure 15: (a). Box Plot for $\mathbf{B}(n_b = 100, p = 0.9)$; (b). Box Plot for $\mathbf{P}(\lambda = 90)$; (c) Range-box Plot for $\mathbf{B}(n_b = 100, p = 0.9)$; (d). Range-box Plot for $\mathbf{P}(\lambda = 90)$ with Number of Simulation Data is 1000.

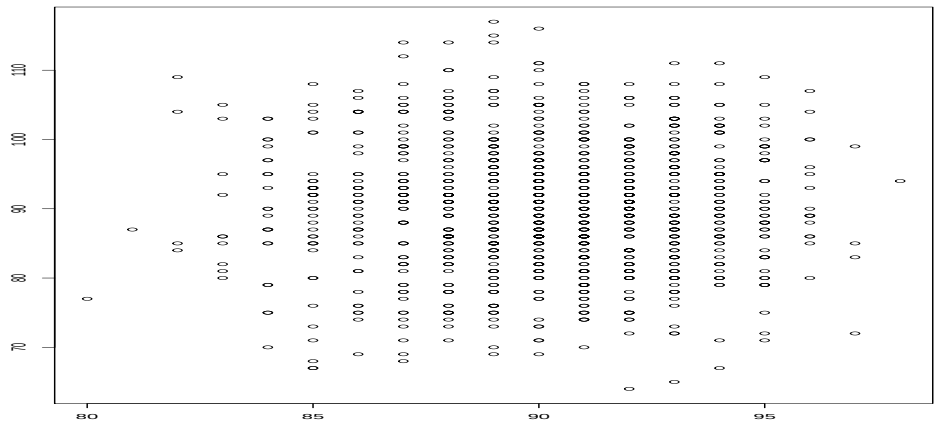
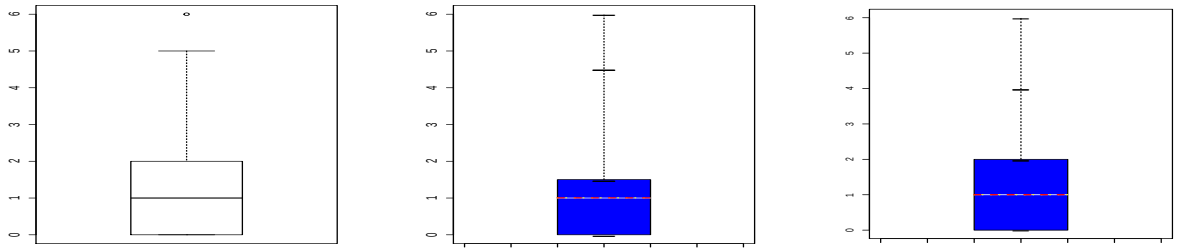
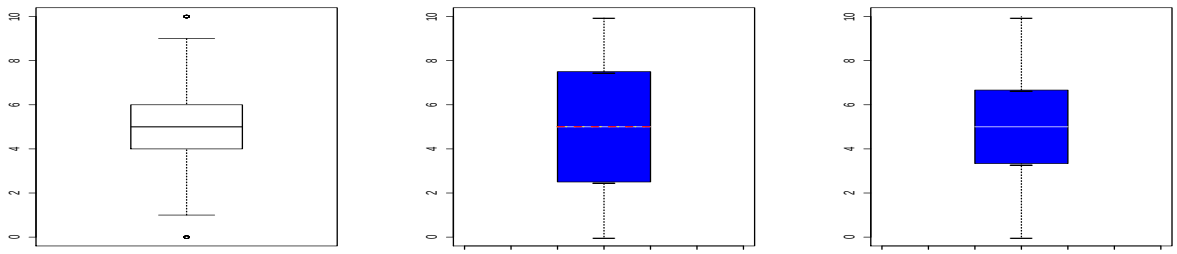


Figure 16: The Simulated Data from $\mathbf{B}(n_b = 100, p = 0.9)$ Versus Simulated Data from $\mathbf{P}(\lambda = 90)$



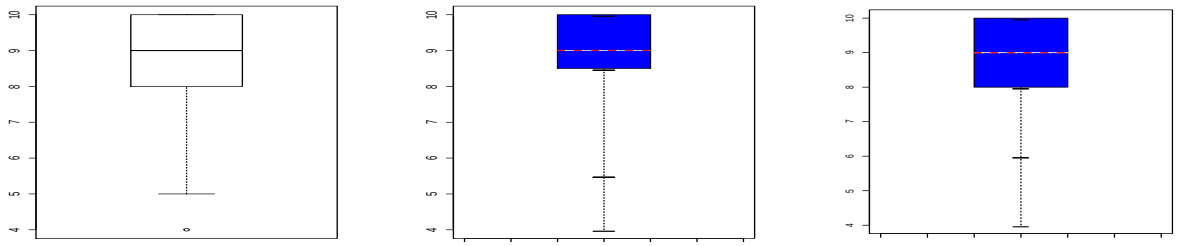
(a) (b) (c)

Figure 17: (a). Classical Plot for $\mathbf{B}(n_b = 10, p = 0.1)$; (b). Range-Box Plot Type I for $\mathbf{B}(n_b = 10, p = 0.1)$; (c). Range-box Plot Type II for $\mathbf{B}(n_b = 10, p = 0.1)$ with Number of Simulation Data is 1000.



(a) (b) (c)

Figure 18: (a). Classical Plot for $\mathbf{B}(n_b = 10, p = 0.5)$; (b). Range-Box Plot Type I for $\mathbf{B}(n_b = 10, p = 0.5)$; (c). Range-box Plot Type II for $\mathbf{B}(n_b = 10, p = 0.5)$ with Number of Simulation Data is 1000.



(a) (b) (c)

Figure 19: (a). Classical Plot for $\mathbf{B}(n_b = 10, p = 0.9)$; (b). Range-Box Plot Type I for $\mathbf{B}(n_b = 10, p = 0.9)$; (c). Range-box Plot Type II for $\mathbf{B}(n_b = 10, p = 0.9)$ with Number of Simulation Data is 1000.

When either $\frac{5}{n_b} \leftarrow p$ or $p \rightarrow 1 - \frac{5}{n_b}$ and $n \rightarrow \infty$, binomial distribution can be axpproximated using another discrete distribution i.e. Poisson distribution. For the first cases which is $\frac{5}{n_b} \leftarrow p$, we found that the $\mathbf{B}(n_b = 100, 0 \leftarrow p) \approx \mathbf{P}(\lambda = n_b \times p = 10)$. The variability is also quite similar. See Figure 14. For the later cases where $p \rightarrow 1 - \frac{5}{n_b}$, we find that the variability for the approximation distribution is greater. See Figure 15 and Figure 16. We also observed that the data is populated around the middle partition of $Q_1^R - Q_3^R$. However for the hypothesis testing using Wilcoxon test (λ is the mean for Poisson distribution) is

$$H_0 : \text{Median for } \mathbf{B}(n_b, p) = \text{Median for } \mathbf{P}(n_b p)$$
$$\text{versus}$$
$$H_1 : \text{Median for } \mathbf{B}(n_b, p) \neq \text{Median for } \mathbf{P}(n_b p)$$

also give a decision that cannot reject the H_0. These two distributions have the same median.

The median (represent by the straight line in the frame) and mean (represent by the dashed line) are correctedly captured in the range-box plot type I. Although the classical box plot managed to carried out similar task, the existence of outliers will mislead the reader. See Figure 15 (b).

As we can observed that all the range-box plots type I frame boxes tend to be smaller in size. The range-box plot type II is being introduced by modifiying the value of where to divide range appropriately. We suggest this type II range-box plot is simplier than type I and the frame box size at the same time comparable with the existing classical box plot. The characters of the type I modification is still maintained. The amount of the most dense observations still in the frame box and still manage to show the skewness of the data but excluding the detection for the outliers as it is a misleading criteria for a non symmetrical distributions.

6 Conclusion

The new modification of range-box plot type I and type II have some better properties compared to classical box plot to serve the nature of discrete binomial data and other related discrete distributions. No outliers are being considered in the plot as it is creating a misleading as the data come from simulation of real probability distribution. The grouping of data according to

the classification group of quartile also not really well justified. The allocation of partitioning the sample size is now playing an important rule in studying the characteristics of the discrete data. The range-box plot type I and type II gives more stable structure for the selected discrete distributions compared to the classical box plot. The existing of outliers in classical box plot contribute to the misleading interpretation of the discrete distributions have been removed permanently. The classical box plot was meant for asymmetrical continuous data not the discrete data like binomial, geometry and other distributions.

Range-box type I and type II plots can be used more informatively in discovering the characters of discrete binomial data. Any uniformly distributed data, the usage of classical box plot should be avoided. It will not reveal any information about the uniform data.

When $\frac{5}{n_b} \leftarrow p$ and $n_b \rightarrow \infty$ the approximation using Poisson distribution is equivalent, similar to when $p \approx 0.5$ and $n_b \rightarrow \infty$, the approximation using Normal distribution tend to give the same result. When $p \rightarrow 1 - \frac{5}{n_b}$ and $n_b \rightarrow \infty$, the approximation using Poisson distribution should be carried out cautiously as the simulated data using the approximated data give bigger variability. Range-box plot manages to capture the median and mean of the data for the approximated distributions.

7 Acknowledgement

This project is partially funded from Grant FRGS 02-1-15-1741FR.

References

[1] Tukey, J. W. *Exploratory Data Analysis. Stochastic Ordering among Order Statistics and Sample Spacings. Technical Report, Indian Statistical Institute.* Reading, Massachusetts: Addison-Wesley. 1977.

[2] Shitan, M. and Vazifedan, T. *Exploratory Data Analysis: for Almost Anyone*. Serdang: UPM Press. 2011.

[3] Velleman, P. F. and Hoaglin, D. C. *Applications, Basics and Computing of Exploratory Data Analysis*. Boston, Massachusetts: Duxbury Press. 1981.

[4] M. Hubert, and E. Vandervieren, An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, Vol. 52, 5186-5201 (2008).

[5] McGill, R., Tukey, J. W. and Larsen, W. A. Variations of box plots. *The American Statistician*. 1978. 32(1): 12-16.

[6] Nuzzo, R. L. The box plots alternative for visualizing quantitative data. *Statistically Speaking*. 2016. PM R 8. 268-272.

[7] Babura, B. I., Adam, M. B., Fitrianto, A. and Abdul Samad, A. R. Modified boxplot for extreme data. *The 3rd ISM International Statistical Conference 2016, AIP Conference Proceeding*. 1842. 030034(2017); doi: 10.1063/1.4982872.

[8] Ferreira, J. E. V., Pinheiro, M. T. S., Dos Santos, W. R. S. and R. Da Silva Maia, Graphical representation of chemical periodicity of main elements through boxplot. 2016 *Educacion Quimica*. 27: 209-216.

[9] N. C. Schwertman, M. A. Owens, and R. Adnan, A simple more general boxplot method for identifying outliers. *Computational Statistics & Data Analysis.* 2004. 47: 165-174.

[10] Schwertman, N. C. and De Silva, R. Identifying outliers with sequential fences. *Computational Statistics & Data Analysis.* 2007. 51: 3800-3810.

[11] Carter, N. J., Schwertman, N. C. and Kiser, N. C. A Comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry. *Statistical Methodology.* 2009. 6: 604-621.

[12] Bruffaerts, C., Verardi, V. and Vermandele. C. A generalized boxplot for skewed and heavy-tailed distributions. *Statistics and Probability Letters.* 2014. 95: 110-117.

[13] Marmolejo-Ramos, F. and Tian, T. S. The shifting boxplot, a boxplot based on essential summary statistics around the mean. *International Journal of Psychological Research.* 2010. 3(1): 37-45.

[14] Leemis, L. M. and Mc Question, J. T. Univariate distribution relationships. *The American Statistician* 2008. 62(1): 45-55.

[15] Song, W. T. Relationships among some univariate distributions. *IIE Transactions.* 2005. 37: 651-656.

[16] W. Tang, H. He, and X. M. Tu, *Applied Categorical and Count Data Analysis.* New York: CRC Press. 2012.