

Regression To The Mean for a New Bivariate Binomial  
Distribution



By

Zaka Ullah

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2024



*In the Name of Allah The Most Merciful and The Most Beneficent*

**Regression To The Mean for a New Bivariate Binomial  
Distribution**



**By**

**Zaka Ullah**

*A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF MASTER OF PHILOSOPHY IN  
STATISTICS*

**Supervised By**

**Dr. Manzoor Khan**

**Department of Statistics**

**Faculty of Natural Sciences**

**Quaid-i-Azam University, Islamabad**

**2024**

# Declaration

I “Zaka Ullah” hereby solemnly declare that this thesis titled, “Regression To The Mean for a New Bivariate Binomial Distribution”.

- This work was done wholly in candidature for a degree of M.Phil Statistics at this University.
- Where I got help from the published work of others, this is always clearly stated.
- Where I have quoted from the work of others, the source is always mentioned. Except for such quotations, this thesis is entirely my research work.
- Where the thesis is based on work done by myself jointly with my supervisor, I have made clear exactly what was done by others and what I have suggested

Dated:\_\_\_\_\_

Signature:\_\_\_\_\_

# CERTIFICATE

Regression to the mean for a new Bivariate Binomial  
distribution

By

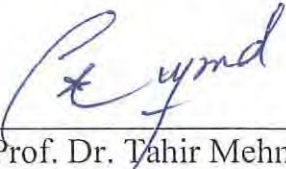
Zaka ullah


(Reg. No. 02222211016)

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF M.PHIL. IN  
STATISTICS

*We accept this thesis as conforming to the required standards*

1.   
Dr. Manzoor Khan  
(Supervisor)

2.   
Prof. Dr. Tahir Mehmood  
(External Examiner)

3.   
Prof. Dr. Ijaz Hussain  
(Chairman) 2.6/04/24

DEPARTMENT OF STATISTICS  
QUAID-I-AZAM UNIVERSITY  
ISLAMABAD, PAKISTAN  
2023

# Dedication

*I am feeling great honor and pleasure to dedicate this research work to*

**My beloved Parents and Family**

*Whose endless affection, prayers, and wishes have been a great source of comfort  
for me during my whole education period and my life*

# Acknowledgments

All praises to Almighty Allah (SWT), the light of Heavens and Earths, The One Who put good thoughts in one's mind, turn them into determinations, and then make the way towards their fulfillment by showering all His Blessings throughout the journey. Best of praises and Peace be upon all the Sacred Messengers and especially for the Last of them Hazrat Muhammad (SAW) are the minarets of knowledge for all the mankind. It is a matter of great honor for me to express gratitude to my Supervisor, Dr. Manzoor Khan, for his supervision and constant support. His invaluable help of constructive comments and suggestions throughout the thesis works have contributed to the success of this research. I would like to express my appreciation to all those teachers, especially Dr. Akbar Ali, who have enlightened my knowledge throughout my academic career and guidance throughout research work. I also thank my friends and colleagues, especially Muhammad Yousaf, Muhammad Umair, Aimel Zafar, and Imam Hussain for their help and words of encouragement

# Abstract

When measurements are made twice on the same subject / person, RTM is noticed when relatively high or low observations are likely to be followed by less extreme observations that are closer to the true mean. When subjects are selected based on some cut-off points, the observed mean difference of the pre-post variable is called the total effect. The total effect is equal to the sum of the RTM effect and treatment effect and should be accounted for RTM.

Bi-variate binomial-binomial distribution models the data when the two groups, say  $i$  and  $j$ , have two possible outcomes in a fixed number of trials in which the number of success follows a binomial distribution. It is the result of the convolution of two independent binomial marginals. The study considers the Bi-variate binomial-binomial distribution. Formula for the total, treatment and RTM effect are derived. Using R's optimize function, the log-likelihood function was maximized to provide the maximum likelihood estimators. The results of the simulation study showed that the maximum likelihood (ML) estimators of RTM are unbiased and consistent.

The proposed method is applied to the real data of the number of patients dealt with in one of two boxes of critical care Emergency service of San Agustin Hospital, in Linares (Spain). The parameters of real data are estimated through MLE and substitute in the derived RTM formula. This shows that the Total Effect = Treatment Effect + RTM Effect.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Identification of RTM through graphs . . . . .	3
1.2	Consequences of RTM . . . . .	5
1.3	Addressing RTM at the design stage . . . . .	5
1.3.1	Randomized control trials . . . . .	5
1.3.2	Selection based on multiple measurements . . . . .	6
1.4	Research Motivation . . . . .	6
1.5	Problem Statement . . . . .	6
1.6	Organization of the thesis . . . . .	7
<b>2</b>	<b>Literature Review</b>	<b>8</b>
2.1	RTM Effect Under Bivariate Normal Distribution . . . . .	8
2.2	Bivariate binomial distribution . . . . .	11
2.3	Bivariate Beta-Binomial distribution . . . . .	13
<b>3</b>	<b>Derivation of Regression to the mean</b>	<b>15</b>
3.1	Bi-variate Binomial-Binomial Distribution . . . . .	15
3.2	RTM, total, and treatment effects . . . . .	16
3.2.1	Case 1: Subjects in the right extreme . . . . .	17
3.2.2	Case 2: Subjects in the left extreme . . . . .	19
3.2.3	Variance of total effect $Tk(i_0; i, j)$ . . . . .	21
3.3	Derivation of RTM and treatment effects . . . . .	24
3.4	The effect of cut-off point, $i_0$ , on RTM . . . . .	26
3.5	Effect of parameters on RTM . . . . .	27
3.5.1	RTM as a function of $p$ . . . . .	27
3.5.2	RTM as a function of $\theta$ . . . . .	28
3.5.3	RTM as a function of $b$ . . . . .	29
<b>4</b>	<b>Estimation and Simulation Study</b>	<b>31</b>
4.1	Maximum Likelihood Estimation of RTM . . . . .	31
4.2	Data generation and simulation study . . . . .	31

---

4.2.1	Estimation of RTM and intervention effect under different sample sizes . . . . .	33
4.2.2	Empirical properties of RTM . . . . .	34
4.3	Data Example . . . . .	36
4.4	Discussion . . . . .	37
<b>5</b>	<b>Conclusion</b>	<b>39</b>
	<b>References</b>	<b>42</b>

# List of Tables

4.1	Total, Treatment, and RTM Effects . . . . .	36
4.2	Observed frequencies of the number of patients in two boxes of the Critical Care and Emergency Service in the San Agustin Hospital, in Linares (Spain) . . . . .	37

# List of Figures

1.1	RTM effect in diastolic blood pressure in baseline and follow-up measurement with true mean and variation . . . . .	2
1.2	RTM effect in diastolic blood pressure in baseline and follow-up measurement with true mean and variation . . . . .	4
3.1	Graph of the derived formula of RTM for less than or greater than cut-off points, when the distribution under study is bivariate binomial-binomial with parameters $n_1=15, n_2=13, P=0.5, \theta =0.5, m =14, b=0.1, n=10000$ . . . . .	27
3.2	RTM is a function of parameter $p$ (Right cut off) . . . . .	28
3.3	RTM is a function of parameter $\theta$ (Right cut off) . . . . .	29
3.4	RTM is a function of parameter $b$ (Right cut off) . . . . .	30
4.1	The distribution of RTM effect via Normal Q-Q plot for $n_1=20, n_2=20, P=0.6, \theta =0.3, m =30, b=0.3, n=10000$ . . . . .	35
4.2	Sampling distribution and estimates of RTM for different sample sizes $n_1=20, n_2=20, P=0.6, \theta =0.3, m =30, b=0.3, n=10000$ . . . . .	36

# Chapter 1

## Introduction

When measurements are made twice on the same subject or person, RTM is noticed when relatively high or low observations are likely to be followed by less extreme observations that are closer to the true mean. The concept of RTM was first introduced by Sir Francis Galton ([Galton, 1886](#)) in the 19th century, who studied the inheritance of height in humans. He noticed that the heights of children tended to regress towards the average height of their parents, rather than being equal to or more extreme than them as shown in [Fig 1.1](#). He called this phenomenon regression towards mediocrity in hereditary stature.

As in [Fig 1.1](#) the standardized height  $z = (x - \mu)/\sigma$ , where  $\mu$  and  $\sigma$  are the population mean and standard deviation, respectively. While the heights of children are very near the genuine mean, those of parents are on the extremes of left and right. The amount of the RTM effect is shown by the arrow.

Galton's idea of RTM was later formalized by Karl Pearson, who developed the mathematical theory of correlation and regression analysis. Pearson showed that the correlation coefficient between two variables is equal to the slope of the regression line that best fits the data. He also proved that the regression line always passes through the mean of both variables, which explains why extreme values tend to move toward the mean.

RTM is a common source of error and bias in many fields of research, especially when the selection of subjects or units is based on an initial measurement that is subject to random error. For example, if a researcher selects the most intelligent students based on a test score and then administers another test to them, the average score of the second test is likely to be lower than the first one, even if there is no real change in the student's abilities. This is because the first test score may have been inflated by random factors, such as guessing, luck, or mood, and the second test score may have been closer to the true mean of the student's intelligence.

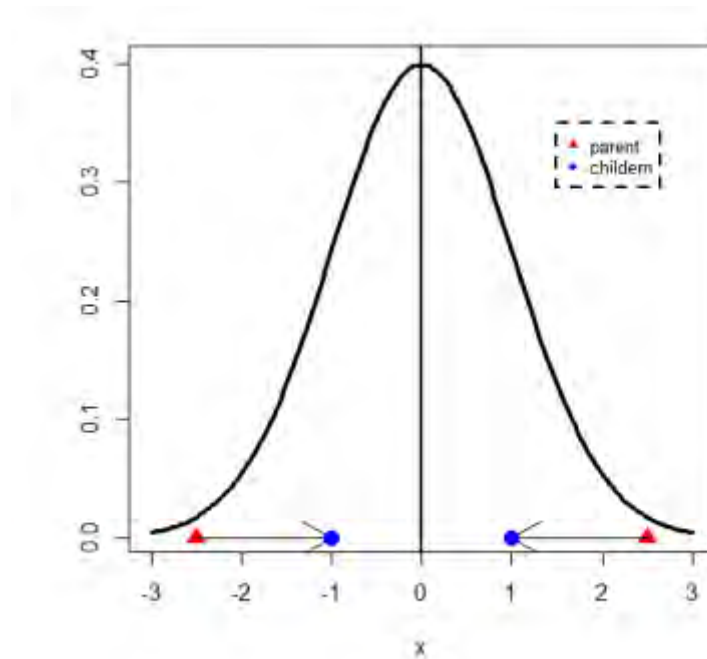


Figure 1.1: RTM effect in diastolic blood pressure in baseline and follow-up measurement with true mean and variation

(James, 1973) examined the effects of RTM in control clinical trials, where observations were made both before and after the application of treatment but without the presence of a control group. In his analysis, he found that there were two types of regression effects: one resulting from biological variation over time and the other from measurement variance. The author emphasized the significance of the control group and made a compelling case for randomized clinical trials. It is essential to distinguish the effects of the RTM from the treatment effects in order to avoid making inaccurate findings.

According to (Barnett et al., 2005), the RTM effect results from measurements of observations that contain random error; the higher the random error term, the bigger the RTM impact. RTM is a common difficulty in data analysis due to the rarity of data without random error. When the chosen measurements are at the extreme of the distribution, RTM can occur in groups as well as on an individual level (Johnson and George, 1991).

Studies, where repeated measurements or observations are gathered, have been influenced by RTM throughout a wide range of research fields. For instance, (Yu and Chen, 2015) in social psychology offered data in support of the effectiveness of social conformity and unrealistic optimism effects, however, the effects were no longer perceptible after adjusting for RTM.

(Prior et al., 2005) adjusted RTM in the medical area to reduce cardiovascular risk factors through health programs by implementing minor interventions without

a control group. The authors concluded that the intervention had a significant impact and decreased the risk of cardiovascular disease by reducing smoking, high blood pressure, and hypercholesterolemia in high-risk individuals. The patients who were at low risk, on the other hand, and their decrease in systolic blood pressure was brought about by RTM.

In the field of economics, abnormally high rates of economic growth are rarely long-lasting and are frequently followed by abrupt declines. As a result, it can be inaccurate to predict economic growth without taking RTM into account (Pritchett and Summers, 2014). After grabbing into consideration the effects of RTM, Pritchett and Summers provide caution over the potential of a reduction in the present growth rates of the Asian giants, China and India..

## 1.1 Identification of RTM through graphs

The RTM effect can be visualized by a simple scatter plot of the difference between follow-up and baseline measurement against baseline measurement. Figure 1.2 shows the scatter plot of high diastolic blood pressure patient data in which the effect size (Follow-up minus Baseline measurement) is plotted against baseline measurement. The solid line shows no change in baseline and follow-up, while the regression line is estimated for both groups. The difference between the two regression lines is the estimated treatment effect. The plot shows the possibility of the RTM effect such as the patients whose blood pressure is normal but rises on follow-up measurement then this change is above the solid line. Similarly, in patients whose diastolic BP is very high and drops on the next measurement, the change is lower than the zero line. The placebo group highlights that the difference between the two measurements is scattered around zero and this fluctuation is due to RTM.

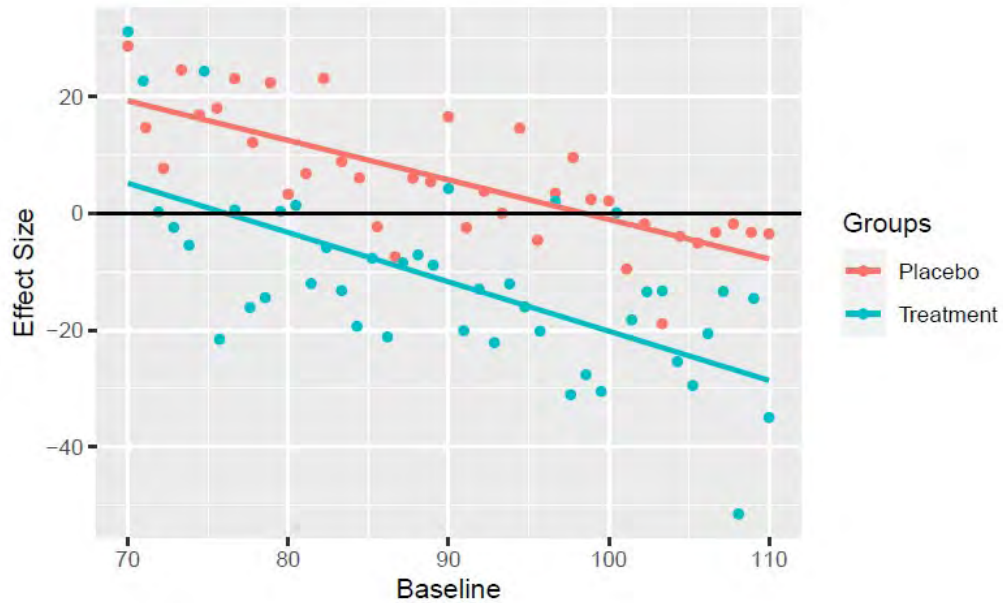


Figure 1.2: RTM effect in diastolic blood pressure in baseline and follow-up measurement with true mean and variation

## Existing methods for regression to the mean

When choosing treatment subjects based on a truncation point, say  $x_0$ , there are several ways to measure and estimate RTM and intervention effects. With  $X_i = X_0 + E_i$ , where  $X_0$  and  $E_i$  are independent of one another, and  $X_0$  is normally distributed as  $N(\mu, \sigma_0^2)$ , these methods assume an additive model of the true component  $X_0$  and a random error component  $E_i$ . The error terms  $E_i$  are identically distributed as  $N(\mu, \sigma_e^2)$ , for  $i = 1, \dots, n$ . Consequently,  $X_i$  has the same distribution as  $N(\mu, \sigma^2)$ , where  $\sigma^2$  is equal to  $\sigma_0^2 + \sigma_e^2$ .

To derive regression to the mean equations, many current approaches (James 1973; Gardner and Heady 1973; Davis 1976; Johnson and George 1991) assume a bivariate normal distribution. Regression to the mean equations has been derived by these authors using model (1) or some adaptations.

James (1973) estimated RTM in uncontrolled clinical studies where the bivariate normal variables are stationary and strictly positively correlated ( $\rho > 0$ ). Davis (1976) developed an approach that was useful in reducing the RTM effect. The approach consists of taking multiple subjects before applying the treatment to them. Senn et al. (1985) extended the derivation of James (1973) and derived the maximum likelihood estimators for the parameters to different types of sampling techniques related to the bivariate normal distribution.

In many real-life situations, the variable under study could be the number of



successes or counts of an event. In the recent past, [Khan and Olivier \(2018\)](#) and [Khan and Olivier \(2019\)](#) developed methods for dealing with the problem posed by the RTM effect using the bivariate Poisson and binomial distributions.

## 1.2 Consequences of RTM

RTM can produce inaccurate results regarding the impact of an intervention or treatment, particularly if subjects are chosen using baseline criteria. The following are some potential effects of RTM if it is not Addressed:

- Overestimating the effectiveness of a treatment that is given to subjects with low initial values, such as patients with severe symptoms or poor test scores [Morton and Torgerson \(2003\)](#).
- Underestimating the effectiveness of a treatment that is given to subjects with high initial values, such as athletes or students with exceptional performance.
- Attributing a causal relationship between two variables that are unrelated, such as the "Sports Illustrated jinx" that claims that athletes featured on the cover of the magazine will perform poorly in their next game.
- Misinterpreting natural variation as real change, such as the "hot hand fallacy" that assumes that a player who has a streak of successful shots will continue to do so.

## 1.3 Addressing RTM at the design stage

Understanding the RTM effect in intervention studies can be helped by the study design ([Yudkin and Stratton \(1996\)](#),[Linden \(2013\)](#)). The preceding subsections include descriptions of several well-known study designs and their possible impacts on RTM.

### 1.3.1 Randomized control trials

To reduce selection bias and level the impact of RTM among groups, individuals can be randomly assigned to treatment groups (such as placebo and therapy). The RTM impact can be estimated using the mean change in the placebo group. The difference between the mean change in the treatment group and the mean change in the placebo group can therefore be used to determine how the treatment effect accounts for RTM. Randomization, however, isn't always feasible because of moral or practical issues.

### 1.3.2 Selection based on multiple measurements

RTM is proportional to measurement variability, according to the RTM formula. Reducing measurement variability can be achieved by choosing two or more baseline measurements. Assuming the RTM effect occurred between the first and subsequent measurements, the study selection criterion can then be based on the average of multiple measurements (Gardner and Heady, 1973), (Davis, 1976). When an intervention is implemented, reducing variability can be utilized to provide a more accurate estimate of each subject's true component. Taking many measurements is a decision that is based on the cost of gathering and is not always feasible when resources are constrained.

## 1.4 Research Motivation

The current methodologies for regression to the mean (RTM) predominantly rely on continuous or fixed-probability discrete distributions, restricting their adequacy when success or failure probabilities fluctuate across trials—a common scenario in medical studies influenced by factors like age, treatment, and conditions. To accommodate this variability, the bivariate binomial-binomial distribution emerges as a promising model. This discrete distribution, capturing success and failure counts across independent trial groups with varying probabilities drawn from distinct binomial distributions, addresses overdispersion, correlations, and fluctuating success probabilities. Despite its potential, this distribution remains underexplored compared to its counterpart, the bivariate beta-binomial, which has garnered substantial attention across fields like ecology, genetics, and epidemiology. This research aims to fill this gap by investigating RTM within the bivariate binomial-binomial distribution and devising methods to analyze and correct this phenomenon within this framework, paving the way for improved understanding and application of RTM in this nuanced context.

## 1.5 Problem Statement

This research aims to address this gap by investigating the complexities of regression to the mean in a bivariate BB distribution setting. The study will focus on elucidating how varying success probabilities and the number of trials across independent groups influence the convergence of extreme observations toward their respective means. By offering a comprehensive understanding of these dynamics, this research endeavors to refine statistical inference techniques and predictive models, catering to diverse fields reliant on categorical data, including but not

limited to healthcare and social sciences

## 1.6 Organization of the thesis

This thesis is composed of five chapters, which cover the following topics: Chapter 1 provides a brief history, applications, detection, impacts, and prevention of RTM. Chapter 2 reviews the existing methods for RTM analysis in the literature. Chapter 3 presents the technical details, derivation of RTM, its variance, co-variances, estimated the treatment, total, and RTM effects, and examined the effect of parameters on RTM. chapter 4 provide calculation of maximum likelihood estimations (MLEs) under bi-variate binomial-binomial distribution, results of simulation studies and also applies the proposed method to real data from the medical field. Chapter 5 summarizes the main findings and contributions of this research.

# Chapter 2

## Literature Review

When measuring the same subject at two or more time periods, RTM has significant consequences for evaluating treatment methods. Many methods have been proposed in past research to estimate the effect of RTM in various cases. Different authors have suggested their approaches to reducing the effect of RTM and estimate treatment effects. In this chapter, we give a brief overview of all these existing methods and techniques.

### 2.1 RTM Effect Under Bivariate Normal Distribution

(James, 1973) worked early on the RTM effect for bivariate normal distribution in clinical studies. He argued that the observed variable is composed of true and random error components. Let  $X_i$  be the effect size on  $i^{th}$  measurement of the same subject and  $x_0$  is the true measurement,

$$X_i = X_0 + e_i$$

where  $e_i$  is random error and  $i = 1, 2, \dots$  (James, 1973) found that the treatment effect was a sum up of two effects i.e., the biological effect and measurement error, and urged to separate RTM from the true treatment effect to avoid unreliable results. He argued that the control group should be used to avoid the serious effect of RTM. James derived the estimator of RTM by assuming that the pre and post-variables  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  with  $cov = \sigma_0^2$ . Assuming the null treatment effect, the RTM effect is equal to the condition difference of the pre and post-treatment mean. The derived RTM effect formula is.

$$R(x_0) = \frac{\sigma(1 - \rho)\phi(z_0)}{1 - \Phi(z_0)} = \frac{\sigma_e^2}{\sqrt{\sigma_0^2 + \sigma_e^2}} \cdot \frac{\phi(z_0)}{1 - \Phi(z_0)} \quad (2.1)$$

James analyzed that if the treatment was effective, then the model relating the pre-post variables could be written as

$$X_2 - \mu = \gamma\rho(X_1 - \mu) + e_1 \quad (2.2)$$

For  $\gamma < 1$ , the treatment was considered effective. The observed difference under bivariate normal distribution is

$$E(Z_1 - Z_2 | Z_1 > z_0) = \frac{(1 - \gamma\rho)\phi(z_0)}{1 - \Phi(z_0)} \quad (2.3)$$

where  $Z_i$  is standardized variables for  $i = 1, 2$ . James established the above formula in Eq. 2.3 for the overall proportionate decrease attributable to RTM and treatment as well as to RTM alone. This will partition the total effect into true and RTM effects. However, it fails when the pre-post measurements variables are independent.

Similar to (James, 1973), (Gardner and Heady, 1973) worked on the derivation of the RTM effect, but along with bivariate measures, the authors also investigated the effect of multiple measurements on RTM. The authors assumed the normal distribution of pre-post variables with  $\rho = \sigma_o / \sqrt{\sigma_0^2 + \sigma_1^2}$ .

The subjects selected based on the right cut-off point, i.e.  $X_i > x_o$  follow the univariate truncated normal distribution with mean.

$$E(X_i | X_i > x_0) = \mu + \sigma \frac{\phi(z_0)}{1 - \Phi(z_0)}. \quad (2.4)$$

Similarly, the mean of  $X_0$  given that the observation is greater than the cutoff point is.

$$E(X_0 | X_i > x_0) = \mu + \frac{\sigma_0^2}{\sigma} \cdot \frac{\phi(z_0)}{1 - \Phi(z_0)}. \quad (2.5)$$

Since  $\sigma > \sigma_0^2/\sigma$  unless  $\sigma_e^2 = 0$ , it is therefore clear from the above equations 2.4 and 2.5 that the observed mean of observation is always greater than their true mean due to the presence of the RTM effect.

(Gardner and Heady, 1973) derived the RTM formula for multiple measurement  $n$  on the same subject and is given by

$$R(x_0) = E(\bar{X} - X_0 | \bar{X} > x_0) = \frac{\sigma_e^2/n}{\sqrt{\sigma_0^2 + \sigma_e^2/n}} \cdot \frac{\phi(z_{0n})}{1 - \Phi(z_{0n})}, \quad (2.6)$$

where  $\bar{X} = \sum x_i/n$  is the sample mean of  $n$  multiple measurements. James' derivation of RTM is a special case of equation 2.6 for  $n = 1$ . However, the RTM effect approaches zero when  $n$  becomes sufficiently large. (Davis, 1976) worked

on the study design to reduce the RTM effect. Let the mean of all multiple measurements and a follow-up observation be  $\bar{X}$  and  $X^*$ , respectively such that  $\bar{X} \sim N(\mu, \sigma_o^2 + \sigma_e^2/n)$  and  $X^* \sim N(\mu, \sigma_o^2 + \sigma_e^2)$ . The RTM effect is derived as shown below:

$$R(x_0, n) = E(\bar{X} - X^* | \bar{X} > x_0) = \frac{\sigma_e^2/n}{\sqrt{\sigma_o^2 + \sigma_e^2/n}} \cdot \frac{\phi(z_{0n})}{1 - \Phi(z_{0n})} \quad (2.7)$$

which is the same as derived by (Gardner and Heady, 1973). Using the first observation  $X_1$  as a classification baseline measurement, i.e., choosing a subject based on the event  $X_1 > x_0$ , and the second observation  $X_2$  on the same subject as the baseline from which the treatment effect may be assessed could be useful to mitigate the RTM effect (Davis, 1976). Let  $X_3$  be the post-treatment measurement, then the author derived the RTM formula by taking the conditional expectation of truncated distribution;

$$R(x_0, \rho_{12}, \rho_{13}) = E(X_2 - X_3 | X_1 > x_0) = (\rho_{12} - \rho_{13}) \cdot \sigma \frac{\phi(z_0)}{1 - \Phi(z_0)}. \quad (2.8)$$

Where the correlation coefficient between  $(X_1, X_i)$  are represented by  $\rho_{1j}$  for  $j = 2, 3$ . The RTM effect becomes zero when the two correlation coefficients are equal with baseline measurement, thereby not requiring multiple measurements for reducing the RTM effect. So far, the observed values were assumed to have consisted of two components, measurement error and the second is biological variables such as emotional and other influences during the recording of observation. (Johnson and George, 1991) extend the previous model and included the subject effect,  $S_i \sim N(0, \sigma_s^2)$ . Hence the model becomes;

$$Y_{ij} = X_0 + S_i + E_{ij} \quad (2.9)$$

where  $i = 1, 2, \dots, m$ ,  $j = 1, 2, \dots, n$  and  $Y_{ij}$  represents the  $j^{th}$  replicate measurement at the  $i^{th}$  study time. The correlation between  $S_i$  and  $S_k$  is positive and independent of random error  $E_i$  and baseline measurement  $X_0$ .

Under Equation.2.9 the RTM formula derived by (Johnson and George, 1991) is

$$R_T(y_0) = \frac{(1 - \rho_s) \sigma_s^2 + \sigma_e^2/n}{m\sigma_{\bar{y}}} \cdot \frac{\phi(z_1)}{1 - \Phi(z_1)} \quad (2.10)$$

The above equation 2.10 represents the total RTM effect due to measurement error and subject effect. One can decrease the RTM measurement error by either increasing the number of repeated measurements or by increasing the number of

replication  $n$ . On the other hand, a greater number of time-dependent, repeated measurements  $m$  minimize the impact of subject variability on the regression.

The detailed work on RTM under normal distribution was recently done by (Khan and Olivier, 2023). The authors partitioned the total effect into true treatment and RTM effects, derived the MLE, and checked their properties such as unbiasedness, consistency, and normality. The RTM effect was depicted for both positive and negative correlations. They derived the RTM effect in a pre-post measurement case in which the pre-variable is composed of true and random parts i.e.,  $X_1 = X_0 + \epsilon_1$  and post-variable  $X_2 = a + bX_0 + \epsilon_2$ . Where  $a + bX_0$  is the true part and  $\epsilon_2$  is the random component. The RTM effect is estimated as.

$$T(x_0, \theta) = (\mu_1 - \mu_2) + (\sigma_1 - \rho\sigma_2) \frac{\phi(z)}{1 - \Phi(z)} \quad (2.11)$$

The first part on the right-hand side,  $(\mu_1 - \mu_2)$ , of the above equation is the average treatment effect, and the second term is the RTM effect. The authors also derived the variance of RTM as shown below.

$$\text{var}(X_1 - X_2 \mid X_1 > x_0) = \sum_{i=1}^2 \text{var}(X_i \mid X_1 > x_0) - 2 \times \text{cov}(X_1, X_2 \mid X_1 > x_0). \quad (2.12)$$

The maximum likelihood estimators of the total, RTM, and true effect were derived.

$$\hat{T}_r(x_0, \mathbf{x}) = \hat{\mu}_1 - \hat{\mu}_2 + \frac{\phi(\hat{z}_0)}{1 - \Phi(\hat{z}_0)} \cdot (\hat{\sigma}_1 - \hat{\rho}\hat{\sigma}_2), \quad (2.13)$$

$$\hat{R}_r(x_0; \mathbf{x}) = (\hat{\sigma}_1 - \hat{\rho}\hat{\sigma}_2) \cdot \frac{\phi(\hat{z}_0)}{1 - \Phi(\hat{z}_0)}, \text{ and} \quad (2.14)$$

$$\hat{\delta}(\mathbf{x}) = \hat{\mu}_1 - \hat{\mu}_2. \quad (2.15)$$

The distribution of RTM and true treatment  $\hat{\delta}(x)$  were shown to be asymptotically normal and unbiasedness and consistency of the estimators were established. The simulation study shows that the RTM and intervention estimates are close to the true value in all cases while James (1973) method gives poor estimates.

## 2.2 Bivariate binomial distribution

The quantification of the RTM effect before and after treatment involving binomial experiments is derived by (Khan and Olivier, 2019). The authors proposed expressions for the RTM when the variable of interest is the number of successes in

the fixed number of trials. The successive number of successes can be expressed as

$$X_1 = Y_0^{(1)} + Y_1, \quad X_2 = Y_0^{(2)} + Y_2 \quad (2.16)$$

where  $X_i$  is the total number of successes before and after study design where the number of trials is fixed, say  $n$ ,  $Y_0^{(i)}$  are an actual number of successes and  $Y_i$  are a random number of successes for  $i = 1, 2$ . (Khan and Olivier, 2019) used bivariate binomial distribution which was first discussed by (Aitken and Gonin, 1936) and given by

$$f_{X_1, X_2}(x_1, x_2, n) = \sum_{\alpha=0}^{\min(x_1, x_2)} f(\cdot),$$

where

$$f(\cdot) = \binom{n}{\alpha, x_1 - \alpha, x_2 - \alpha, n + \alpha - x_1 - x_2} \theta_0^\alpha \theta_1^{x_1 - \alpha} \theta_2^{x_2 - \alpha} (1 - \theta_0 - \theta_1 - \theta_2)^{n + \alpha - x_1 - x_2}$$

with  $\text{cov}(X_1, X_2) = n(\theta_0 - (\theta_0 + \theta_1)(\theta_0 + \theta_2))$ . The expected truncated difference of pre and post-variables is equal to the RTM effect. For a null intervention left and right RTM are obtained by putting  $\theta_1 = \theta_2$  as

$$\begin{aligned} R_l(x_0; \boldsymbol{\theta}) &= E(X_1 - X_2 | X_1 \leq x_0), \\ &= n\theta_1 \cdot \frac{P_{n-1}(X_1 = x_0)}{F_n(x_0 | \theta_0 + \theta_1)}, \end{aligned}$$

and

$$\begin{aligned} R_r(x_0; \boldsymbol{\theta}) &= E(X_1 - X_2 | X_1 > x_0), \\ &= n\theta_1 \cdot \frac{P_{n-1}(X_1 = x_0)}{1 - F_n(x_0 | \theta_0 + \theta_1)}. \end{aligned}$$

(Khan and Olivier, 2019) derived formulae for the total effect which is the sum of the intervention effect and RTM effect. If the selection is based on all measurements greater than the baseline point  $x_0$ , then total effect is given by

$$T_r(x_0, \boldsymbol{\theta}) = n \cdot \frac{\theta_1 (1 - F_{n-1}(x_0 | \theta_0 + \theta_1)) - \theta_2 [1 - F_n(x_0 | \theta_0 + \theta_1) - (\theta_0 + \theta_1) P_{n-1}(X_1 = x_0)]}{1 - F_n(x_0 | \theta_0 + \theta_1)}.$$

On the other hand, if the selection is based on all measurements less than or equal to the baseline point  $x_0$ , then the total effect is given by



$$T_l(x_0, \boldsymbol{\theta}) = n \cdot \frac{\theta_2 [F_n(x_0|\theta_0 + \theta_1) + (\theta_0 + \theta_1)P_{n-1}(X_1 = x_0)] - \theta_1 F_{n-1}(x_0 - 1|\theta_0 + \theta_1)}{F_n(x_0|\theta_0 + \theta_1)}$$

where  $P_{n-1}(X_1 = x_0) = \binom{n-1}{x_0}(\theta_0 + \theta_1)^{x_0}(1 - \theta_0 - \theta_1)^{n-1-x_0}$ . (Khan and Olivier, 2019) derived the maximum likelihood estimates (MLE) of the parameters when  $\alpha$  is known as

$$\hat{\theta}_0 = \sum_{j=1}^k \alpha_j/k, \quad \hat{\theta}_1 = \sum_{j=1}^k (x_{1j} - \alpha_j)/k, \quad \hat{\theta}_2 = \sum_{j=1}^k (x_{2j} - \alpha_j)/k.$$

When  $\alpha$  is unknown, the parameters of the bivariate binomial distribution were derived using numerical methods.

## 2.3 Bivariate Beta-Binomial distribution

The chance of success ( $p$ ) in the binomial distribution is taken to be constant from trial to trial. This presumption might not be true in many circumstances. In these cases, an alternate distribution called the beta-binomial distribution is applied. A binomial distribution is referred to as a beta-binomial distribution if the success probability is randomly selected from the beta distribution rather than being fixed at each  $n$  trail. The probability of success is fixed in a binomial distribution and not fixed in a beta-binomial distribution, which is the key distinction between the two.

The probability mass function (PMF) of bivariate beta-binomial distribution is given by

$$P(X_1, X_2) = \frac{\binom{n_1}{x_1} \binom{n_2}{x_2} B(x_1 + x_2 + \alpha, n_1 + n_2 + \beta - x_1 - x_2)}{B(\alpha, \beta)} \quad (2.17)$$

where  $x_1 = 0, 1, 2, \dots, n_1$ ,  $x_2 = 0, 1, 2, \dots, n_2$ ,  $n_1, n_2 = 1, 2, \dots$ , and  $\alpha, \beta > 0$ .  $B(\cdot, \cdot)$  is the beta function and  $\alpha$  and  $\beta$  are the parameters. The marginal PMF of  $X_1$  and  $X_2$  are the univariate beta-binomial with parameter  $n_i$ ,  $\alpha$  and  $\beta$ , respectively, for  $i = 1, 2$ . The respective mean and variance of the univariate beta-binomial distribution are

$$E(X_i) = \frac{n\alpha}{\alpha + \beta}$$

and

$$\text{Var}(X_i) = \frac{n\alpha\beta(n + \alpha + \beta)}{(\alpha + \beta + 1)(\alpha + \beta)^2} \quad \text{for } i = 1, 2$$

The correlation coefficient of bivariate beta-binomial distribution is

$$\rho_{x_1x_2} = \left[ \left( 1 + \frac{\alpha + \beta}{n_1} \right) \left( 1 + \frac{\alpha + \beta}{n_2} \right) \right]^{-\frac{1}{2}} \quad (2.18)$$

For  $n_1 = n_2 = n$ , the correlation coefficient  $\rho$  of the symmetrical bivariate beta-binomial distribution is

$$\rho_{x_1x_2} = \left\{ \frac{1 + \alpha + \beta}{N} \right\}^{-1} = \frac{N}{N + \alpha + \beta} \quad (2.19)$$

$$\begin{aligned} Rr(x_0; \alpha, \beta) &= \mathbb{E}(X_1 - X_2 | X_1 > x_0) \\ &= \frac{\alpha}{(\alpha + \beta)} \left[ \frac{n_1(1 - F_{n_1-1}(x_0 - 1 | \alpha', \beta)) - n_2(1 - F_{n_2}(y_0 | \alpha', \beta))}{1 - F_{n_1}(x_0 | \alpha, \beta)} \right] \end{aligned}$$

and

$$\begin{aligned} Rr(x_0; \alpha, \beta) &= \mathbb{E}(X_1 - X_2 | Y \leq y_0) \\ &= \frac{\alpha}{(\alpha + \beta)} \left[ \frac{n_2(1 - F_{n_2}(x_0 | \alpha', \beta)) - n_1(1 - F_{n_1-1}(x_0 - 1 | \alpha', \beta))}{1 - F_{n_1}(x_0 | \alpha, \beta)} \right] \end{aligned}$$

The treatment or intervention effect in pre-post studies is the expected difference between the pre-post observations.

$$\delta_i(\alpha, \beta) = E(X_1 - X_2) = \frac{n_1\alpha}{\alpha + \beta} - \frac{n_2\alpha}{\alpha + \beta} \quad (2.20)$$

The treatment effect will be zero when  $n_1$  and  $n_2$  are equal, and RTM is accountable for the total effect. The bivariate beta-binomial distribution is restricted in this manner when the parameters in  $X_1$  and  $X_2$  are identical.

The total effect for the right cut-off point under the bivariate beta-binomial distribution is.

$$Tr(x_0; \alpha, \beta) = \frac{\alpha}{(\alpha + \beta)} \left[ \frac{n_1(1 - F_{n_1-1}(y_0 - 1 | \alpha', \beta)) - n_2(1 - F_{n_2}(x_0 | \alpha', \beta))}{1 - F_{n_1}(x_0 | \alpha, \beta)} \right]$$

The total effect for the left cut-off point is defined as

$$Tl(x_0; \alpha, \beta) = \frac{\alpha}{(\alpha + \beta)} \left[ \frac{n_2(F_{n_1}(x_0 | \alpha', \beta)) - n_1(F_{n_1-1}(y_0 - 1 | \alpha', \beta))}{F_{n_1}(y_0 | \alpha, \beta)} \right]$$

# Chapter 3

## Derivation of Regression to the mean

Bi variate binomial-binomial distribution models the data when the two groups, say  $i$  and  $j$ , have two possible outcomes in a fixed number of trials in which the number of success follows a binomial distribution. It is the result of the convolution of two independent binomial marginals. Previous work on RTM quantification was done by [Khan and Olivier \(2019\)](#) when the samples were drawn from binomial distribution, and Khan and Aimal worked on bi-variate beta-binomial when the probability of success is drawn randomly from beta distribution, but it has two defects One is that the distribution has identical marginals, therefore we cannot estimate the true treatment effect. The other is that when the probability of success is not constant throughout the trial and follow a beta distribution, then the existing method under the bi-variate binomial distribution of RTM is not appropriate. In this chapter, we derive the RTM effect under the bi-variate BB distribution, its variance, covariance, and maximum likelihood when the observation is selected based on a cutoff point either on the left or right side. The bi-variate binomial-binomial distribution is also derived as a mixture model which is useful for computer sampling from the distribution and facilitates computation of the joint probabilities. The bi-variate binomial-binomial distribution is shown to be positive quadrant dependent. The chi-square goodness of fit for the bi-variate binomial-binomial distribution is better than the bi-variate Generalized Poisson distribution ([Famoye and Consul, 1995](#)) and bi-variate Conway Maxwell Poisson distributions ([Ong et al., 2021](#)).

### 3.1 Bi-variate Binomial-Binomial Distribution

The probability mass function (PMF) of bivariate binomial-binomial distribution is given by

$$p(i, j) = \sum_{k=0}^m \binom{n_1 + k}{i} p^i q^{n_1 + k - i} \binom{n_2 + k}{j} \theta^j (1 - \theta)^{n_2 + k - j} \binom{m}{k} b^k (1 - b)^{m - k} \quad (3.1)$$

where  $i = 1, 2, \dots, n_1 + k$ ,  $j = 1, 2, \dots, n_2 + k$ ,  $n_1, n_2$  and  $m$  are positive integers. The marginal PMF of  $i$  and  $j$  are the univariate binomial distribution with parameters  $n_1, P, n_2, \theta, m$  and  $b$  respectively. Then the respective mean and variance of the univariate binomial-binomial distribution are

$$E(i) = n_1 p + m b p,$$

$$E(j) = n_2 \theta + m b \theta,$$

$$\text{Var}(I) = n_1 p - n_1 p^2 + m b p - m b^2 p^2,$$

$$\text{Var}(j) = n_2 \theta - n_1 \theta^2 + m b \theta - m b^2 \theta^2,$$

The covariance between  $i$  and  $j$  is follow

$$\text{Cov}(I, J) = p \theta m b - p \theta m b^2$$

## 3.2 RTM, total, and treatment effects

In intervention studies like epidemiology, medicine, subjects/patients with measurements below or above some specified truncation or cut-off point, say  $i_0$ , are selected for intervention or treatment. Let  $i$  and  $j$  be some characteristics of interest before and after treatment on the same subject. The joint distribution of the before and after treatment measurements at truncated point  $i_0$  is given by

$$f_r(i, j) = \frac{f(i, j)}{f(i > i_0)} \quad \text{where } i_0 < i < \infty, -\infty < j < \infty$$

where the subscript  $r$  represents the right truncated in  $f_r(i, j)$ . The total effect  $T(i_0, \theta)$  is defined as the conditional expectation of the difference between before and after treatment variables and is given by

$$\begin{aligned}
T(i_0, \theta) &= E(i - j | i > i_0) \\
&= \sum_{m=i_0}^{\infty} \sum_{n=1}^{\infty} (i_m - j_n) f(i_m, j_n | i > i_0)
\end{aligned} \tag{3.2}$$

where  $\theta$  is the parameter vector. Depending on how well the treatment effect works, the total effect may be partially or totally driven by the RTM effect. The treatment effect is zero when both  $i_m$  and  $j_n$  have the same distribution, i.e.,  $E(i_m - j_n) = 0$ . The RTM effect is defined as the conditional expectation of the difference between  $i_m$  and  $i_n$ , and it is given by

$$R(i_0, \theta) = E(i - j | i > i_0) \tag{3.3}$$

The difference between the unconditional means of  $i$  and  $j$  is defined as the average treatment effect.

$$\delta(\lambda) = E(i - j)$$

Thus, the total effect  $T(i_0, \theta)$  can be written as

$$T(i_0, \theta) = R(i_0, \theta) + \delta(\lambda), \tag{3.4}$$

where  $\lambda = (n_1, n_2, p, \theta, m, b)$  is a function of the means of  $i$  and  $j$ . The treatment or intervention may be applied to the extreme cases above and below some specified cut-off point, denoted as  $i_0$ . The cut-off point may be either on the right or left side of the distribution. In this chapter, both the right and the left extreme or cut-off points are considered, and their derivations are given in the following subsection.

### 3.2.1 Case 1: Subjects in the right extreme

Let the application of an intervention be decided on the basis that the initial value  $i$  is greater than some cut-off point, say  $i_0$ , then the truncated bivariate binomial-binomial distribution is given by

$$P_t(i, j) = \frac{1}{1 - P(i \leq i_0)} \sum_{k=0}^m \binom{m}{k} b^k (1 - b)^{m-k} \binom{n_1+k}{i} p^i q^{n_1+k-i} \binom{n_2+k}{j} \theta^j (1 - \theta)^{n_2+k-j}$$

The expectation of  $i$  conditioned on the event,  $i > i_0$ , is

$$E(i | i > i_0) = \frac{1}{1 - P(i \leq i_0)} \sum_{i=i_0+1}^{n1+k} \sum_{j=0}^{n2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} i \binom{n1+k}{i} p^i q^{n1+k-i} \\ \times \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j}$$

After some algebraic manipulation, the expectation reduces to

$$E(i | i > i_0) = \frac{1}{1 - P(i \leq i_0)} \left[ n_1 p \sum_{i=i_0}^{n1+k-1} \sum_{j=0}^{n2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \right. \\ \cdot \binom{n1+K-1}{i} p^i q^{n1+k-i-1} \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j} + mpb \\ \left. \sum_{i=i_0}^{n1+k-1} \sum_{j=0}^{n2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \right. \\ \left. \cdot \binom{n1+K-1}{i} p^i q^{n1+k-i-1} \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j} \right], \quad (3.5)$$

$$\text{where } 1 - P(i \leq i_0) = \sum_{i=i_0+1}^{n1+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+K}{i} p^i q^{n1+k-i}$$

and is the cumulative distribution function (CDF) of the uni-variate binomial-binomial distribution. Now considering the conditional expectation of  $(j | i > i_0)$ , we have

$$E(j | i > i_0) = \frac{1}{1 - P(i \leq i_0)} \left[ \sum_{i=i_0+1}^{n1+k} \sum_{j=0}^{n2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+K}{i} p^i q^{n1+k-i} \right. \\ \left. j \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j} \right]$$

After some algebraic manipulation, the expectation reduces to

$$\begin{aligned}
E(j | i > i_0) = & \frac{1}{1 - P(i \leq i_0)} \left[ n_2 \theta \sum_{i=i_0+1}^{n_1+k} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \right. \\
& \cdot \binom{n_1+K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} + m\theta b \\
& \sum_{i=i_0+1}^{n_1+k} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\
& \left. \cdot \binom{n_1+K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} \right], \quad (3.6)
\end{aligned}$$

To get the total effect for the right cut-off point under the bi-variate binomial-distribution subtract equation 3.4 from 3.5, we get

$$\begin{aligned}
T_r(i_o, i, j) = & \frac{1}{1 - P(i \leq i_0)} \left[ n_1 p \sum_{i=i_0}^{n_1+k-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \right. \\
& \cdot \binom{n_1+K-1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} + mpb \\
& \sum_{i=i_0}^{n_1+k-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\
& \cdot \binom{n_1+K-1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} - n_2 \theta \\
& \sum_{i=i_0+1}^{n_1+k} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \\
& \cdot \binom{n_1+K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} - m\theta b \\
& \sum_{i=i_0+1}^{n_1+k} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\
& \left. \cdot \binom{n_1+K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} \right]. \quad (3.7)
\end{aligned}$$

### 3.2.2 Case 2: Subjects in the left extreme

When the subjects selected for treatment are in the left tail/end of a distribution, i.e.,  $i$  is less than or equal to the cut-off point, say  $i_0$ , then, the truncated probability distribution function of  $i$  and  $j$  is

$$P_{i \leq i_0}(i, j) = \frac{1}{P(i \leq i_0)} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n_1+k}{i} p^i q^{n_1+k-i} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j}$$

The expectation of  $i$  conditioned on the event  $i \leq i_0$  is

$$E(i | i \leq i_0) = \frac{1}{P(i \leq i_0)} \sum_{i=0}^{i_0} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} i \binom{n_1+k}{i} p^i q^{n_1+k-i} \\ \times \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j},$$

The conditional expectation upon simplification becomes

$$E(i | i \leq i_0) = \frac{1}{P(i \leq i_0)} \left[ n_1 p \sum_{i=0}^{i_0-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \right. \\ \cdot \binom{n_1+K-1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} + mpb \\ \left. \sum_{i=0}^{i_0-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \right. \\ \left. \cdot \binom{n_1+K-1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} \right], \quad (3.8)$$

Now considering the conditional expectation of  $j | i \leq i_0$ , we have

$$E(j | i \leq i_0) = \frac{1}{P(i \leq i_0)} \left[ \sum_{i=0}^{i_0} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n_1+K}{i} p^i q^{n_1+k-i} \right. \\ \left. j \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} \right],$$

Solving the  $E(j | I \leq i_0)$  we get,

$$E(j | i \leq i_0) = \frac{1}{P(i \leq i_0)} \left[ n_2 \theta \sum_{i=0}^{i_0} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \right. \\ \cdot \binom{n_1+K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} + m \theta b \sum_{i=0}^{i_0} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\ \left. \cdot \binom{n_1+K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} \right], \quad (3.9)$$



The total effect for the left cut-off point is defined as a

$$T_l(i_0; i, j) = E(j | i \leq i_0) - E(i | i \leq i_0), \quad (3.10)$$

where the subscript  $l$  is for the left cut-off point. An expression for  $T_l(i_0; i, j)$  can be obtained by subtracting equation (3.8) from equation (3.7), as

$$\begin{aligned} T_l(i_0; i, j) = & \frac{1}{P(i \leq i_0)} \left[ n_1 p \sum_{i=0}^{i_0-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \right. \\ & \cdot \binom{n_1 + K - 1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} + mpb \\ & \sum_{i=0}^{i_0-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\ & \cdot \binom{n_1 + K - 1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} - n_2 \theta \\ & \sum_{i=0}^{i_0} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \\ & \cdot \binom{n_1 + K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} - m \theta b \\ & \sum_{i=0}^{i_0} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\ & \left. \cdot \binom{n_1 + K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} \right]. \end{aligned}$$

### 3.2.3 Variance of total effect $Tk(i_0; i, j)$

For statistical inferences, the derivation of variance is important. The respective variances of the total effect can be obtained by combining variances of  $i$  and  $j$  and covariance of  $(i, j)$  as

$$\text{var}(i - j | i > i_0) = \text{var}(i | i > i_0) + \text{var}(j | i > i_0) - 2\text{cov}(i, j | i > i_0), \quad (3.11)$$

and

$$\text{var}(j - i | i \leq i_0) = \text{var}(i | i \leq i_0) + \text{var}(j | i \leq i_0) - 2\text{cov}(i, j | i \leq i_0), \quad (3.12)$$

where

$$\text{var}(i|i > i_0) = \mathbb{E}(i(i-1)|i > i_0) + \mathbb{E}(i|i > i_0) - (\mathbb{E}(i|i > i_0))^2,$$

$$\text{var}(j|i > i_0) = \mathbb{E}(j(j-1)|i > i_0) + \mathbb{E}(j|i > i_0) - (\mathbb{E}(j|i > i_0))^2,$$

$$\text{Cov}(i, j|i > i_0) = \mathbb{E}(i, j|i > i_0) - \mathbb{E}(i|i > i_0) \cdot \mathbb{E}(j|i > i_0).$$

After completing derivation equation 3.11 becomes

$$\begin{aligned} \text{var}(i-j|i > i_0) = & \frac{1}{1-p(i \leq i_0)} \\ & (n_1^2 p^2 C + 2n_1 mb P^2 D - n_1 p^2 C + p^2 m^2 b^2 E - mb^2 p^2 E + n_1 p A + pmb B) \\ & - \left[ \frac{1}{1-P(i \leq i_0)} (n_1 p A + pmb B) \right]^2 \\ & + \frac{1}{1-P(i \leq i_0)} \\ & (n_2^2 \theta^2 H + 2n_2 mb \theta^2 I - n_2 \theta^2 H + \theta^2 m^2 b^2 J - mb^2 \theta^2 J + n_2 \theta F + \theta mb G) \\ & - \left[ \frac{1}{1-P(i \leq i_0)} (n_2 \theta F + \theta mb G) \right]^2 \\ & - 2 \left[ \frac{1}{1-P(i \leq i_0)} \right. \\ & (n_1 n_2 p \theta K + n_1 pmb \theta L + n_2 pmb \theta L + m^2 b^2 p \theta M - mb^2 p \theta M + p \theta mb L) \\ & \left. - \left[ \frac{1}{p(i \leq i_0)} (n_1 p A + pmb B) (n_2 \theta F + mb \theta G) \right] \right] \end{aligned}$$

where

$$A = \sum_{i=i_0}^{n_1+k-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n_1+k-1}{i} p^i q^{n_1+k-1-i} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j}$$

$$B = \sum_{i=i_0}^{n_1+k-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n_1+k-1}{i} p^i q^{n_1+k-1-i} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j}$$

$$C = \sum_{i=i_0-1}^{n_1+k-2} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n_1+k-2}{i} p^i q^{n_1+k-2-i} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j}$$

$$D = \sum_{i=i_0-1}^{n_1+k-2} \sum_{j=0}^{n_2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n_1+k-2}{i} p^i q^{n_1+k-2-i} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j}$$

$$E = \sum_{i=i_0-1}^{n_1+k-2} \sum_{j=0}^{n_2+k} \sum_{k=0}^{m-2} \binom{m-2}{k} b^k (1-b)^{m-k-2} \binom{n_1+k-2}{i} p^i q^{n_1+k-2-i} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j}$$

$$\begin{aligned}
F &= \sum_{i=i_o+1}^{n1+k} \sum_{j=0}^{n2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1} \\
G &= \sum_{i=i_o+1}^{n1+k} \sum_{j=0}^{n2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1} \\
H &= \sum_{i=i_o+1}^{n1+k} \sum_{j=0}^{n2+k-2} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-2}{j} \theta^j (1-\theta)^{n2+k-j-2} \\
I &= \sum_{i=i_o+1}^{n1+k} \sum_{j=0}^{n2+k-2} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-2}{j} \theta^j (1-\theta)^{n2+k-j-2} \\
J &= \sum_{i=i_o+1}^{n1+k} \sum_{j=0}^{n2+k-2} \sum_{k=0}^{m-2} \binom{m-2}{k} b^k (1-b)^{m-k-2} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-2}{j} \theta^j (1-\theta)^{n2+k-j-2} \\
K &= \sum_{i=i_o}^{n1+k-1} \sum_{j=0}^{n2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+k-1}{i} p^i q^{n1+k-i-1} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1} \\
l &= \sum_{i=i_o}^{n1+k-1} \sum_{j=0}^{n2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n1+k-1}{i} p^i q^{n1+k-i-1} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1} \\
M &= \sum_{i=i_o}^{n1+k-1} \sum_{j=0}^{n2+k-1} \sum_{k=0}^{m-2} \binom{m-2}{k} b^k (1-b)^{m-k-2} \binom{n1+k-1}{i} p^i q^{n1+k-i-1} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1}
\end{aligned}$$

Similarly, an expression of the variance for left cut-off point is given by

$$\begin{aligned}
\text{var}(j-i|i \leq i_o) &= \frac{1}{p(i \leq i_o)} \\
&\quad (n_1^2 p^2 C * + 2n_1 m b P^2 D * - n_1 p^2 C * + p^2 m^2 b^2 E * - m b^2 p^2 E * + n_1 p A * + p m b B *) \\
&\quad - \left[ \frac{1}{P(i \leq i_o)} (n_1 p A * + p m b B *) \right]^2 \\
&\quad + \frac{1}{P(i \leq i_o)} \\
&\quad (n_2^2 \theta^2 H * + 2n_2 m b \theta^2 I * - n_2 \theta^2 H * + \theta^2 m^2 b^2 J * - m b^2 \theta^2 J * + n_2 \theta F * + \theta m b G *) \\
&\quad - \left[ \frac{1}{P(i \leq i_o)} (n_2 \theta F * + \theta m b G *) \right]^2 \\
&\quad - 2 \left[ \frac{1}{P(i \leq i_o)} \right. \\
&\quad (n_1 n_2 p \theta K * + n_1 p m b \theta L * + n_2 p m b \theta L * + m^2 b^2 p \theta M * - m b^2 p \theta M * + p \theta m b L *) \\
&\quad \left. - \left[ \frac{1}{p(i \leq i_o)} (n_1 p A * + p m b B *) (n_2 \theta F * + m b \theta G *) \right] \right]
\end{aligned}$$

where

$$\begin{aligned}
A^* &= \sum_{i=0}^{i_0-1} \sum_{j=0}^{n2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+k-1}{i} p^i q^{n1+k-1-i} \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j} \\
B^* &= \sum_{i=0}^{i_0-1} \sum_{j=0}^{n2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n1+k-1}{i} p^i q^{n1+k-1-i} \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j} \\
C^* &= \sum_{i=0}^{i_0-2} \sum_{j=0}^{n2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+k-2}{i} p^i q^{n1+k-2-i} \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j} \\
D^* &= \sum_{i=0}^{i_0-2} \sum_{j=0}^{n2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n1+k-2}{i} p^i q^{n1+k-2-i} \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j} \\
E^* &= \sum_{i=0}^{i_0-2} \sum_{j=0}^{n2+k} \sum_{k=0}^{m-2} \binom{m-2}{k} b^k (1-b)^{m-k-2} \binom{n1+k-2}{i} p^i q^{n1+k-2-i} \binom{n2+k}{j} \theta^j (1-\theta)^{n2+k-j} \\
F^* &= \sum_{i=0}^{i_0} \sum_{j=0}^{n2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1} \\
G^* &= \sum_{i=0}^{i_0} \sum_{j=0}^{n2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1} \\
H^* &= \sum_{i=0}^{i_0} \sum_{j=0}^{n2+k-2} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-2}{j} \theta^j (1-\theta)^{n2+k-j-2} \\
I^* &= \sum_{i=0}^{i_0} \sum_{j=0}^{n2+k-2} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-2}{j} \theta^j (1-\theta)^{n2+k-j-2} \\
J^* &= \sum_{i=0}^{i_0} \sum_{j=0}^{n2+k-2} \sum_{k=0}^{m-2} \binom{m-2}{k} b^k (1-b)^{m-k-2} \binom{n1+k}{i} p^i q^{n1+k-i} \binom{n2+k-2}{j} \theta^j (1-\theta)^{n2+k-j-2} \\
K^* &= \sum_{i=0}^{i_0-1} \sum_{j=0}^{n2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \binom{n1+k-1}{i} p^i q^{n1+k-i-1} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1} \\
L^* &= \sum_{i=0}^{i_0-1} \sum_{j=0}^{n2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \binom{n1+k-1}{i} p^i q^{n1+k-i-1} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1} \\
M^* &= \sum_{i=0}^{i_0-1} \sum_{j=0}^{n2+k-1} \sum_{k=0}^{m-2} \binom{m-2}{k} b^k (1-b)^{m-k-2} \binom{n1+k-1}{i} p^i q^{n1+k-i-1} \binom{n2+k-1}{j} \theta^j (1-\theta)^{n2+k-j-1}
\end{aligned}$$

### 3.3 Derivation of RTM and treatment effects

The treatment or intervention effect in pre-post studies is the expected difference between the pre-post observations. Let  $Ri(i_0; i, j)$  and  $\delta_i(i, j)$  be the RTM and intervention effects, respectively for  $i = r, l$ . The average treatment effect can be written as

$$\delta_i(i, j) = E(i - j) = n_1p + mbp - n_2\theta - mb\theta$$

So, expressions of RTM for right and left extreme/cut-off are, respectively,

$$RTM = Totaleffect - Treatmenteffect \quad (3.13)$$

$$\begin{aligned}
R_r(i_0, i, j) &= E(i - j | i > i_0) \\
&= \frac{1}{1 - P(i \leq i_0)} \left[ n_1p \sum_{i=i_0}^{n_1+k-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \right. \\
&\quad \cdot \binom{n_1+K-1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} + mpb \\
&\quad \sum_{i=i_0}^{n_1+k-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\
&\quad \cdot \binom{n_1+K-1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} - n_2\theta \\
&\quad \sum_{i=i_0+1}^{n_1+k} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \\
&\quad \cdot \binom{n_1+K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} - m\theta b \\
&\quad \sum_{i=i_0+1}^{n_1+k} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\
&\quad \cdot \binom{n_1+K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} \left. \right] \\
&\quad - (n_1p + mbp - n_2\theta - mb\theta).
\end{aligned} \quad (3.14)$$

and

$$\begin{aligned}
R_l(i_0, i, j) &= E(i - j \mid i \leq i_0) \\
&= \frac{1}{P(i \leq i_0)} \left[ n_1 p \sum_{i=0}^{i_0-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \right. \\
&\quad \cdot \binom{n_1 + K - 1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} + m p b \\
&\quad \sum_{i=0}^{i_0-1} \sum_{j=0}^{n_2+k} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\
&\quad \cdot \binom{n_1 + K - 1}{i} p^i q^{n_1+k-i-1} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} - n_2 \theta \\
&\quad \sum_{i=0}^{i_0} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^m \binom{m}{k} b^k (1-b)^{m-k} \\
&\quad \cdot \binom{n_1 + K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} - m \theta b \\
&\quad \sum_{i=0}^{i_0} \sum_{j=0}^{n_2+k-1} \sum_{k=0}^{m-1} \binom{m-1}{k} b^k (1-b)^{m-k-1} \\
&\quad \left. \cdot \binom{n_1 + K}{i} p^i q^{n_1+k-i} \binom{n_2+k-1}{j} \theta^j (1-\theta)^{n_2+k-j-1} \right] - (n_1 p + m b p - n_2 \theta - m b \theta).
\end{aligned}$$

### 3.4 The effect of cut-off point, $i_0$ , on RTM

The severity of RTM is dependent on the cut-off point and behaves differently for different distributions. Using the formula  $R_r(i_0; i, j)$  and  $R_l(i_0; i, j)$ , a graph for various cut-off points is depicted in Figure 3.1. For demonstrative purposes, we fixed the values of parameters ( $n_1 = 15$ ,  $n_2 = 13$ ,  $P = 0.5$ ,  $\theta = 0.5$ ,  $m = 14$ ,  $b = 0.1$ ,  $n = 10000$ ). As the graph shows, increasing the value of  $i_0$ , RTM for the right cut-off point increases steeper than a linear relation. Increasing the left cut-off causes RTM to decrease and approach zero for larger values of  $i_0$ .

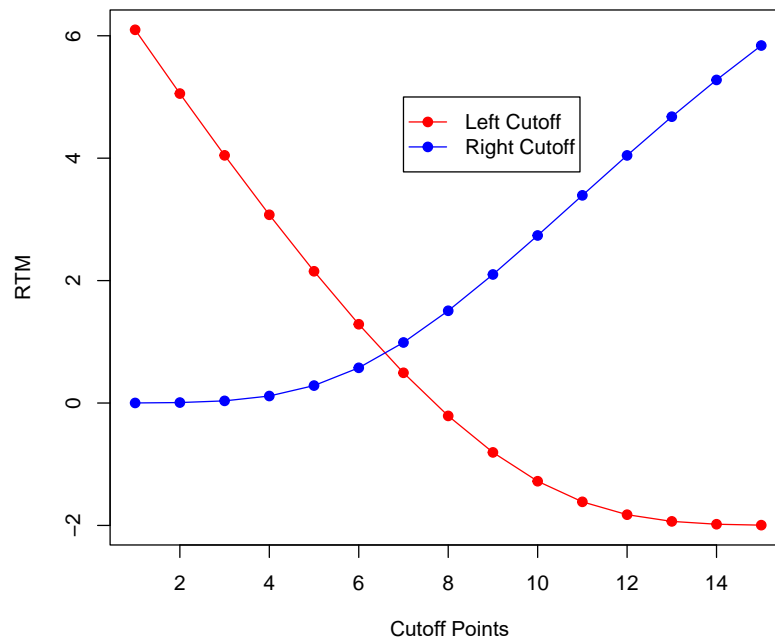


Figure 3.1: Graph of the derived formula of RTM for less than or greater than cut-off points, when the distribution under study is bivariate binomial-binomial with parameters  $n_1=15$ ,  $n_2=13$ ,  $P=0.5$ ,  $\theta =0.5$ ,  $m =14$ ,  $b=0.1$ ,  $n=10000$

## 3.5 Effect of parameters on RTM

Usually, the RTM effect decreases with increasing correlation between the successive variables. For the binomial-binomial distribution, the correlation is a function of  $i$ ,  $j$ , and  $b$ , and decreases with increasing values of the parameters. To see this effect, we separately plot RTM as a function of each parameter.

### 3.5.1 RTM as a function of $p$

RTM is a function of parameter  $p_1$  shown in Figure 3.2. For right side, we can fix cut off point as well as other parameters too ( $n_1=20$ ,  $n_2=20$ ,  $\theta =0.3$ ,  $m =30$ ,  $b=0.3$ ,  $n=10000$ ,  $i_o=15$ ). As seen in the graph when the value of  $p_1$  increases, the RTM value decreases.

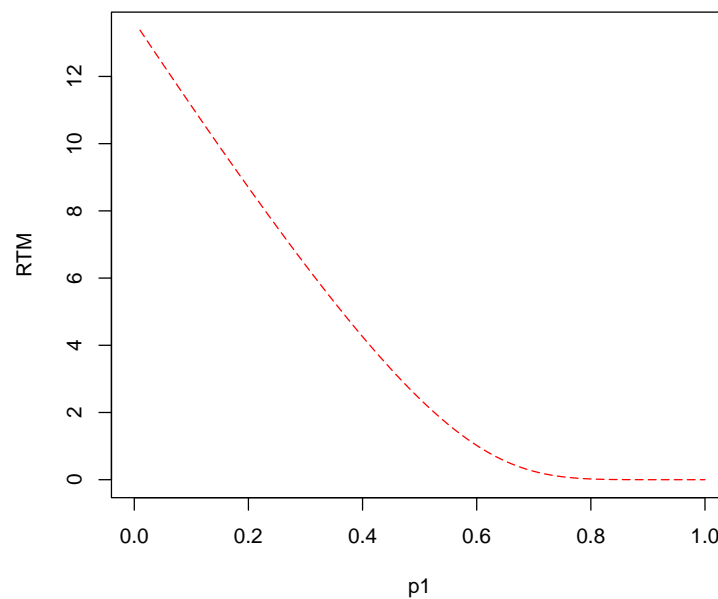


Figure 3.2: RTM is a function of parameter  $p$  (Right cut off)

### 3.5.2 RTM as a function of $\theta$

RTM is a function of parameter  $\theta$  is shown in Figure 3.3. For the right side, we can fix cut off point as well as other parameters too ( $n_1=20, n_2=20, p=0.6, m=30, b=0.3, n=10000, i_o=15$ ). As seen in the graph as the value of  $\theta$  increases, the RTM value increases.



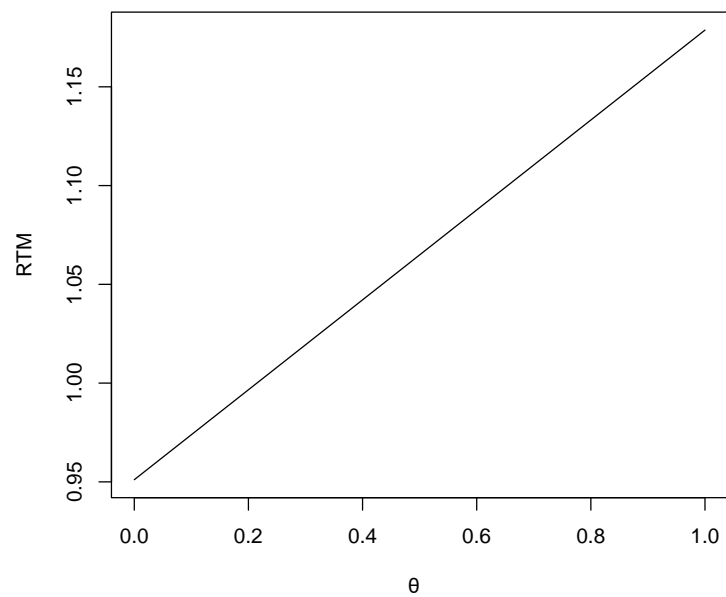


Figure 3.3: RTM is a function of parameter  $\theta$  (Right cut off)

### 3.5.3 RTM as a function of $b$

RTM is a function of parameter  $b$  shown in Figure 3.4. For the right side, we can fix cut off point as well as other parameters too ( $n_1=20, n_2=20, p=0.6, \theta =0.3, m=30, n=10000, i_o=15$ ) as seen in the graph when the value of  $b$  increases, the RTM value decreases.

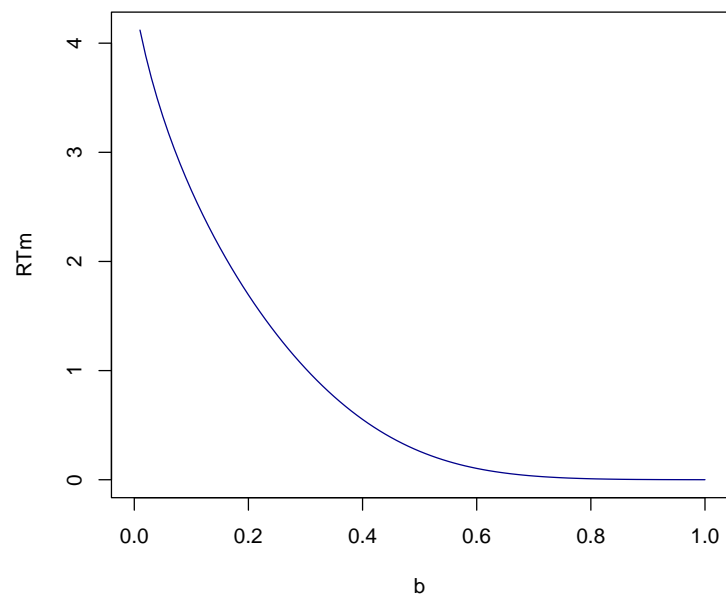


Figure 3.4: RTM is a function of parameter  $b$  (Right cut off)

# Chapter 4

## Estimation and Simulation Study

### 4.1 Maximum Likelihood Estimation of RTM

Let  $(i_1, j_1), (i_2, j_2), \dots, (i_n, j_n)$  be the pairs of pre and post measurements of size  $n$  from the bivariate binomial-binomial distribution. For simplicity, let us consider  $P(I = i, J = j | i > i_0)$  by  $P_T(i, j)$  for brevity. The respective likelihood and log-likelihood functions can be written as,

$$L(i, j; x) = \prod_{i=1}^n P_T(i_i, j_i)$$

and

$$l(i, j; x) = \sum_{i=1}^n \log(P_T(i_i, j_i))$$

where  $P_T(i, j)$  is the truncated bi-variate binomial-binomial probability mass function. The log-likelihood function can be maximized by the ‘optim’ function in R since there exists no explicit solution, and a numerical method could be used.

### 4.2 Data generation and simulation study

Data generation is mandatory for conducting a simulation study. For bivariate binomial-binomial distribution, first generate a variate  $r$  from  $B(m, b)$ . Next, we generate number of pre successes from  $B(n_1 + r, p)$  and  $Y$  from  $B(n_2 + r, \theta)$  respectively.

The density function from which the random numbers are generated is shown below

$$p(i, j) = \sum_{k=0}^m \binom{n_1+k}{i} p^i q^{n_1+k-i} \binom{n_2+k}{j} \theta^j (1-\theta)^{n_2+k-j} \binom{m}{k} b^k (1-b)^{m-k}$$

Random data can be generated from the above distribution using the builtin functions in R, as

$$r = \text{rbinom}(n, m, b)$$

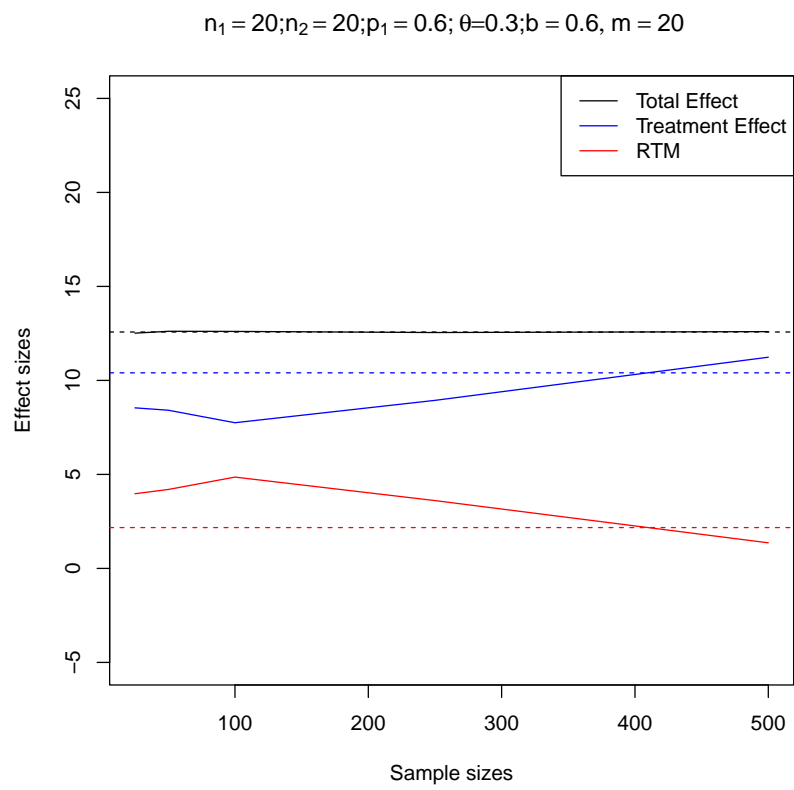
$$X = \text{rbinom}(n, r + n_1, p_1)$$

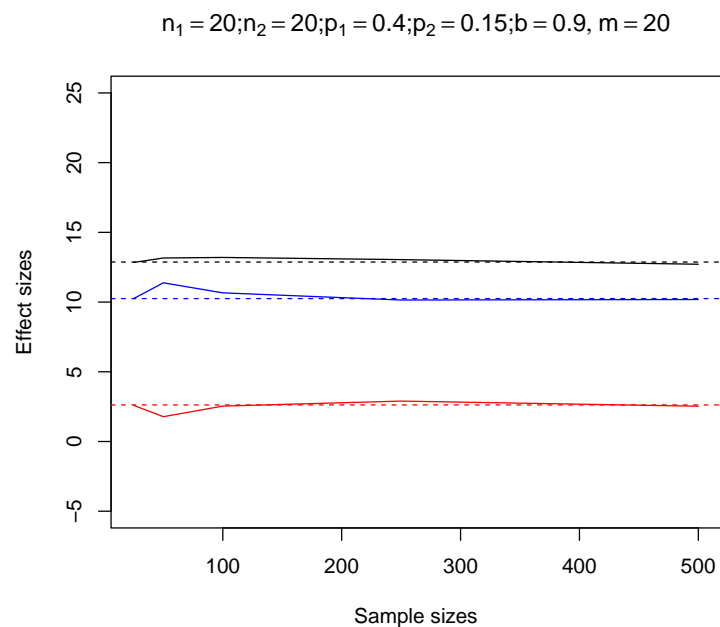
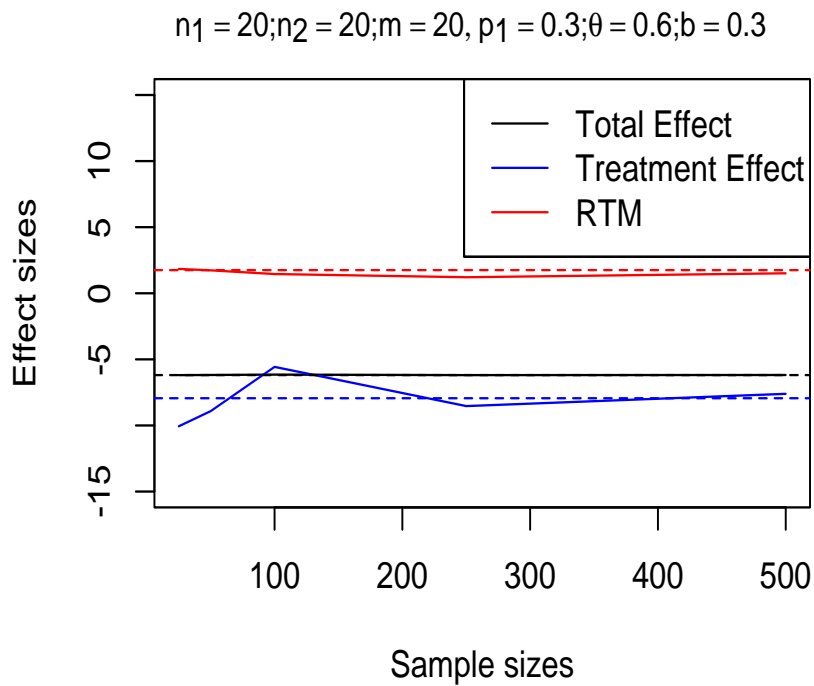
$$Y = \text{rbinom}(n, r + n_2, \theta)$$

The following steps are followed to generate bivariate samples.

- First a random sample of size  $m$  was generated from binomial distribution with parameter  $m = 30$  and probability of success  $b = 0.3$
- Then two random samples of size  $n$  were generated from the binomial distribution using the parameters  $n_1 = 20, p = 0.6, n_2 = 20$  and  $\theta = 0.3$
- Then the pre and post-variables were generated from the univariate binomial distribution with parameters  $r + n_1, p_1$  and  $r + n_2, \theta$  respectively.
- $i$  and  $j$  were considered pre and post-observations of an intervention study.
- After generating a random number the observation which is below or above the specific cutoff point the selected sample was considered as truncated sample.
- The above steps are replicated 1000 times and the corresponding total effect, treatment effect, and RTM effect are estimated using the maximum likelihood estimation

### 4.2.1 Estimation of RTM and intervention effect under different sample sizes





### 4.2.2 Empirical properties of RTM

To check the normality of the sampling distribution of  $\hat{R}_k(i_0; i, j)$ , normal Q-Q plots are constructed from 1000 repeated samples and are given in Figure 4.1. The graph indicates that the sampling distribution of  $\hat{R}_k(i_0; i, j)$  for the right cut-off

point  $i_0 = 15$  and for sample sizes 50, 150, and 250 are approximately normal. The sampling distribution of RTM was also determined to be normal at various cut-off points and parameters but is not given here for brevity.

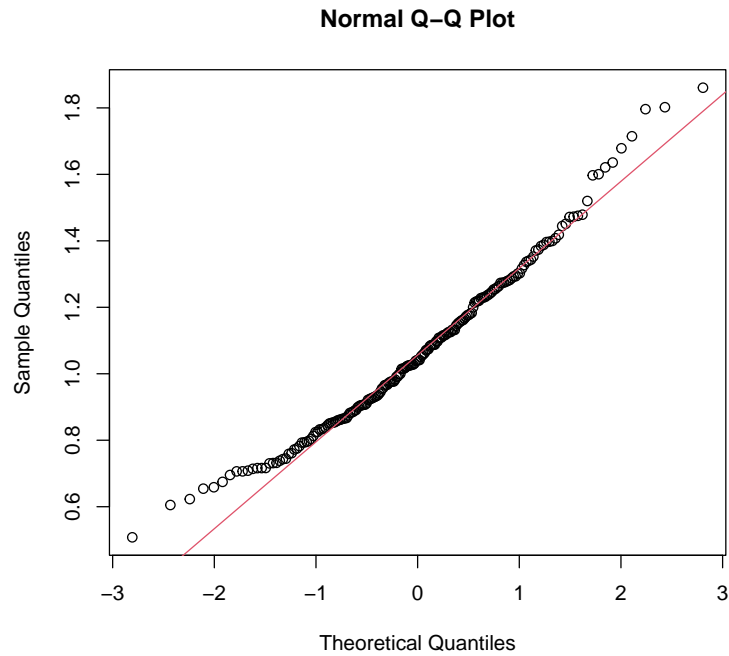


Figure 4.1: The distribution of RTM effect via Normal Q-Q plot for  $n_1=20$ ,  $n_2=20$ ,  $P=0.6$ ,  $\theta =0.3$ ,  $m =30$ ,  $b=0.3$ ,  $n=10000$

To check the empirical properties such as unbiasedness and consistency of RTM, the estimated RTM for the different sample sizes are depicted in Figure 4.2. As seen in the graph, the mean estimated RTM (the blue dots) is very close to the true RTM (the red dotted line) revealing the asymptotic unbiasedness of RTM. Furthermore, it is shown that if the sample size increases the variance of the estimate of RTM is reduced and shrinks toward the mean of the RTM estimate which suggests that the estimate is consistent.

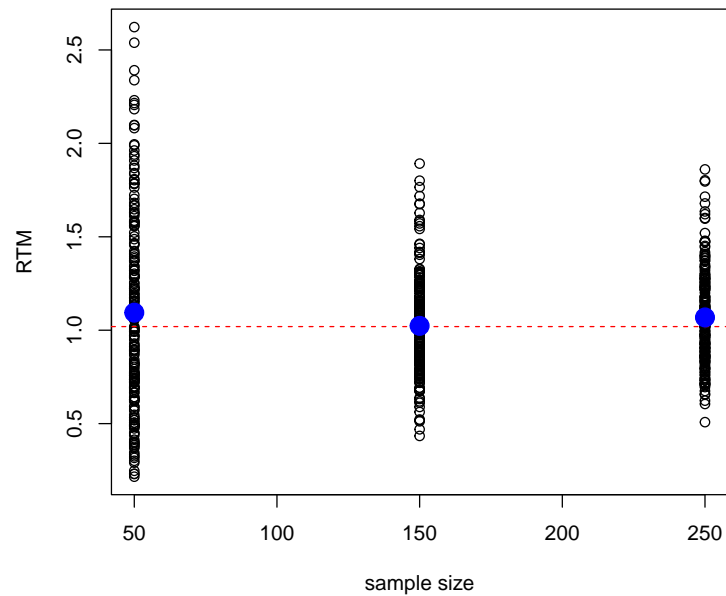


Figure 4.2: Sampling distribution and estimates of RTM for different sample sizes  $n_1=20, n_2=20, P=0.6, \theta =0.3, m =30, b=0.3, n=10000$

### 4.3 Data Example

The proposed method is applied to the real data of the number of patients dealt with in one of two boxes of critical care Emergency service at San Agustin Hospital, in Linares (Spain) (Rodriguez et al., 2023) The total number of patients admitted in each box is estimated by the proposed Method. The cutoff point is chosen at 2 which means that the total number of Patients in box 1 is at least two. The Estimated total, Treatment, and RTM Effect are shown in the following Table 4.1.

	Total	Treatment	RTM
Proposed	0.762	0.084	0.678

Table 4.1: Total, Treatment, and RTM Effects

The total effect is estimated is 0.762 which is the additive component of RTM and Treatment effect. RTM was major part of the total effect and could have exaggerated the treatment effect.



		Box 2						
Box 1		0	1	2	3	4	5	6
0		0	0	0	1	0	0	0
1		0	1	4	5	3	1	0
2		0	3	10	13	11	3	0
3		1	4	12	11	11	5	1
4		1	3	10	10	7	3	1
5		0	1	4	4	4	1	0
6		0	0	1	1	1	0	0

Table 4.2: Observed frequencies of the number of patients in two boxes of the Critical Care and Emergency Service in the San Agustin Hospital, in Linares (Spain)

The parameters of the data are estimated through MLE which shows that  $p = 0.150$ ,  $\theta = 0.496$ , and  $b = 0.01$ .

## 4.4 Discussion

Regression to the mean, or RTM, is a significant problem in data analysis that could produce incorrect conclusion in a study; for this reason, it's necessary to consider RTM to prevent incorrect conclusions. In literature, RTM expressions are available for bi-variate Normal, binomial, and beta binomial distribution. However the expression of RTM for bi-variate binomial binomial distribution are missing in literature . Therefore, quantifying the RTM effect for bi-variate binomial binomial distribution and developing a method are the main objectives discussed in this chapter. Expression for the RTM effect was derived for both left and right cut-off/extreme points. RTM happens when subjects in a pre-post-study design are selected at the extreme points of distribution. The severity of the RTM effect increases when the extreme/cut-off points are farther in the tail of a distribution.

As the cutoff point ( $i_0$ ) increases the RTM effect increases exponentially for the right cut-off and vice versa.

When cut off point ( $i_0$ ) and all other Parameters ( $n_1, n_2, \theta, m, b, n$ ) are fixed the RTM effect for right cut off decreases as we increase the value of  $p$ . Similarly when the value of  $\theta$  increases, the RTM effect for the right cutoff also increases. parameter b causes the decrease in the RTM effect for the right cutoff when its value increases.

Using R's optimize function, the log-likelihood function was maximized to provide the maximum likelihood estimators. The results of the simulation study

showed that the RTM's maximum likelihood (ML) estimators are unbiased and consistent.

We perform a simulation study to demonstrate the behavior of different effects. The process is replicated 100 times on different sample sizes. As we increase the sample sizes the estimated Effects such as total, treatment, and RTM effect approaches to true values.

Finally, we apply the proposed method to the real data of a number of patients dealt with in one of two boxes of critical care emergency service at San Agustin Hospital, in Linares (Spain). The parameters of real data are estimated through MLE and substituted in the derived RTM formula. This shows that the Total Effect = Treatment Effect + RTM Effect. The total effect was mostly driven by the RTM effect. Overlooking the RTM effect could result in declaring an ineffective treatment as an effective one.

# Chapter 5

## Conclusion

In literature, the RTM effect is estimated when the probability of success is drawn from beta distribution also their marginal density are identical due to which the treatment effect cannot be estimated therefore in this work we derive RTM its properties when the marginal density is the convolution of two independent marginal binomial distribution.

In conclusion, the discussion emphasizes the significance of Regression to the Mean (RTM) as a crucial concern in data analysis, particularly in the context of studies where pre-post designs involve selecting subjects from extreme points of distribution. The chapter addresses a gap in the literature by deriving expressions for RTM effects in the case of bivariate binomial distributions, both for left and right cut-off points.

The severity of the RTM effect is highlighted, noting that it increases when extreme or cut-off points are farther in the tail of a distribution. The relationship between the cut-off point ( $i_0$ ) and the exponential increase in the RTM effect for right cut-off points is established. Additionally, the discussion explores the impact of various parameters, such as  $p$ ,  $\theta$ , and  $b$ , on the RTM effect for the right cut-off, demonstrating how changes in these parameters influence the RTM phenomenon.

The application of R's optimize function to maximize the log-likelihood function is discussed, showing that the Maximum Likelihood (ML) estimators for RTM are unbiased and consistent, as demonstrated in a simulation study replicated across different sample sizes.

The simulation study further reveals that as sample sizes increase, estimated effects such as Total, Treatment, and RTM approach their true values. Finally, the proposed method is applied to real data from the critical care Emergency service of San Agustin Hospital, demonstrating that the Total Effect is equal to the sum of the Treatment Effect and the RTM Effect. Overall, the chapter provides a comprehensive exploration of RTM in the context of bivariate binomial distributions, offering valuable insights and a practical methodology for addressing

this issue in data analysis.

# References

- Aitken, A. and Gonin, H. (1936). Xi.—on fourfold sampling with and without replacement. *Proceedings of the Royal Society of Edinburgh*, 55:114–125.
- Barnett, A. G., Van Der Pols, J. C., and Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International journal of epidemiology*, 34(1):215–220.
- Davis, C. (1976). The effect of regression to the mean in epidemiologic and clinical studies. *American journal of epidemiology*, 104(5):493–498.
- Famoye, F. and Consul, P. (1995). Bivariate generalized poisson distribution with some applications. *Metrika*, 42:127–138.
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.
- Gardner, M. and Heady, J. (1973). Some effects of within-person variability in epidemiological studies. *Journal of Chronic Diseases*, 26(12):781–795.
- James, K. E. (1973). Regression toward the mean in uncontrolled clinical studies. *Biometrics*, pages 121–130.
- Johnson, W. D. and George, V. T. (1991). Effect of regression to the mean in the presence of within-subject variability. *Statistics in Medicine*, 10(8):1295–1302.
- Khan, M. and Olivier, J. (2018). Quantifying the regression to the mean effect in poisson processes. *Statistics in Medicine*, 37(26):3832–3848.
- Khan, M. and Olivier, J. (2019). Regression to the mean for the bivariate binomial distribution. *Statistics in medicine*, 38(13):2391–2412.
- Khan, M. and Olivier, J. (2023). Regression to the mean: Estimation and adjustment under the bivariate normal distribution. *Communications in Statistics-Theory and Methods*, 52(19):6972–6990.

- Linden, A. (2013). Assessing regression to the mean effects in health care initiatives. *BMC medical research methodology*, 13:1–7.
- Morton, V. and Torgerson, D. J. (2003). Effect of regression to the mean on decision making in health care. *Bmj*, 326(7398):1083–1084.
- Ong, S. H., Gupta, R. C., Ma, T., and Sim, S. Z. (2021). Bivariate conway–maxwell poisson distributions with given marginals and correlation. *Journal of Statistical Theory and Practice*, 15:1–19.
- Prior, J. O., van Melle, G., Crisinel, A., Burnand, B., Cornuz, J., and Darioli, R. (2005). Evaluation of a multicomponent worksite health promotion program for cardiovascular risk factors—correcting for the regression towards the mean effect. *Preventive Medicine*, 40(3):259–267.
- Pritchett, L. and Summers, L. H. (2014). Asiaphoria meets regression to the mean. Technical report, National Bureau of Economic Research.
- Rodriguez, J., Conde, A., Sáez, A. J., and Olmo, M. J. (2023). Some aspects of bivariate gaussian discrete distributions. *No Journal*.
- Senn, S. J., Brown, R. A., and James, K. (1985). Estimating treatment effects in clinical trials subject to regression to the mean. *Biometrics*, 41(2):555–560.
- Yu, R. and Chen, L. (2015). The need to control for regression to the mean in social psychology studies. *Frontiers in Psychology*, 5:1574.
- Yudkin, P. and Stratton, I. (1996). How to deal with regression to the mean in intervention studies. *The Lancet*, 347(8996):241–244.