Spring 2019

Essays on Mixture Models

Trevor R. Camper

ESSAYS ON MIXTURE MODELS

by

TREVOR CAMPER

(Under the Direction of Stephen Carden)

ABSTRACT

When considering statistical scenarios where one can sample from populations that are not of interest for the purposes of a study, bivariate mixture models can be used to study the effect that this missampling can have on parameter estimation. In this thesis, we will examine the behavior that bivariate mixture models have on two statistical constructs: Cronbach's alpha [10], and Spearman's rho [39]. Chapter 1 will introduce notions of mixture models and the definition of bias under mixture models which will serve as the central concept of this thesis. Chapter 2 will investigate a particular psychometric issue known as insufficient effort responding (IER), which we model as a mixture model, while Chapter 3 will deal with mixture models in a more general setting. Chapter's 2 and 3 will demonstrate that the sign of the bias and the bias under bivariate mixture models for Cronbach's alpha and Spearman's rho, respectively, are polynomial functions in the mixing proportions of the underlying distributions. This will be followed in each chapter by simulation results and observations.

INDEX WORDS: Cronbach's alpha, Spearman's rho, Reliability, Mixture models, Rank-based Correlation, Insufficient Effort Responding

2009 Mathematics Subject Classification: 62H20, 62D05, 62J12

ESSAYS ON MIXTURE MODELS

by

TREVOR CAMPER

B.B.A., Georgia Southern University, 2016

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial

Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

STATESBORO, GEORGIA

ESSAYS ON MIXTURE MODELS

by

TREVOR CAMPER

Major Professor:	Stephen Carden
Committee:	Divine Wanduku
	Nicolas Holtzman
	Arpita Chatterjee

Electronic Version Approved:
May 2019

DEDICATION

This thesis is dedicated to my niece, Quinn. You're going to grow up to be a very bright and talented physicist/medical doctor/ballerina one day. Let this thesis serve as a reminder that as long as you work hard, anything is possible.

ACKNOWLEDGMENTS

I wish to acknowledge everyone who has contributed to my academic development thus far. These people include: my parents, Ken and Rebecca Camper, for the continual support even with my sporadic changes in academic fields; my ever-supporting academic advisors, Stephen Carden and Tharanga Wickramarachchi, for all of the advice, letters of recommendation, flexed office hours, and for letting a kid with no mathematical experience whatsoever take Probability in Fall 2015; NOS^{TM}Energy Drink, for providing me with unbounded energy during the creative spells needed to complete this thesis; and last but not least, the members of office 2016, for without them, graduate school would have been pretty boring.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

In many statistical settings, there is the possibility that when attempting to sample from a population of interest that observations from a secondary population can occur randomly. For example, consider surveying individuals in City A, with the intent of estimating the mean income of the area. However, City A happens to be a tourist destination for City B. In the absence of screening questions, when surveying individuals on the streets of City A, it is possible that some denizens of City B could be asked to perform the survey. If it happens that City A and City B tend to have different mean incomes, then the resulting presence of survey responses from denizens of City B may result in a biased estimate of the mean income of City A. This example demonstrates a particular phenomena of when one random quantity, the estimated mean income of City A, is dependent upon two random quantities, namely the incomes of City A *and* City B. This phenomena leads naturally into our first definition.

Definition 1. [[38]] A random variable X is said to have a *mixture distribution* if the distribution of X depends on a quantity that also has a distribution.

In the example presented above, one can think of the incomes coming from City A as coming from a target, or valid, population, while incomes coming from City B are from an undesired, or contaminating population. Going forward, we shall associate random variables (or vectors) V and C with these valid and contaminating classes, respectively. Similar situations as described in the introductory paragraph tend to occur in many sampling scenarios. Whereas in the example, we see our contaminating population coming from one source, this is not always the case, as this contamination can ultimately arise from many possible sources. However, in constructing a probability model to use in this situation, one can simply treat the sources of contamination as coming from a random variable C that has a mixture distribution. This formulation then results in there only being two components

of interest: those associated with the valid distribution, and those associated with a class of contamination. The next definition provides a more mathematical construction of the concepts described thus far.

Definition 2. Let M, V, and C be random variables, and let $W \sim Bernoulli(p)$. We say that M is a bivariate mixture of V and C if the distribution of M is a mixture distribution of V and C. Formally, we say that M is a bivariate mixture of V and C with mixing proportion p if

$$M = (1 - W)V + WC.$$

Now that we have described a proper probability model to describe the mis-sampling issue, we can now begin to discuss how bivariate mixture models affect the population forms of particular statistics. Namely, practitioners would be interested in knowing if the population form of a statistic under the mixture model is larger or smaller than that under the target population, and what conditions on the contaminating distribution result in a larger or smaller value of the statistic. The notion that the statistic can be larger or smaller under a mixture model than for the target population leads us into our next definition.

Definition 3. Let θ be the population form of some statistic $\hat{\theta}$. The *bias under bivariate mixture* is defined as a function of the mixing proportion p as follows:

$$\text{Bias}(p) = \theta_M - \theta_V$$

where θ_M is the statistic under the bivariate mixture model M described in Definition 2, and θ_V is the statistics under the target population.

Definition 3 will serve as the main area of focus for the remainder of this thesis. In the second Chapter of this thesis, we will explore the effect of bivariate mixture models on Cronbach's alpha [10], a measure of internal consistency often used when constructing surveys. In particular, we will investigate a particular form of mis-sampling, commonly

known as insufficient effort responding, via a bivariate mixture model where the contaminating distribution is represented as a distribution of responses for which little effort is supplied by the survey respondent. This will allow us to provide a characterization of the sign of the bias under this particular mis-sampling via the underlying means, variances, and covariances of the valid and contaminating distributions. Chapter 2 will be concluded by including commentary and observations dealing with particular types of insufficient effort responding.

Following Chapter 2, Chapter 3 will discuss Spearman's rho [39], a rank-based measure of correlation. First, however, the population form of Spearman's rho involves a particular probabilistic notion known as a copula, which we define at the start of Chapter 3. We then demonstrate that the bias in Spearman's rho defined in Definition 3 is a cubic equation in the mixing proportion p. In addition to this, we briefly discuss the dynamics of a cubic equation, with particular care for the cubic's root behavior. We follow this by demonstrating via simulation that all possible mathematical situations for a cubic equation can exist for the bias. Chapter 4 will conclude this thesis and discuss future routes for new research. This concludes Chapter 1.

CHAPTER 2

CRONBACH'S ALPHA AND IER

When administering self-report surveys, there is often a class of respondents that fail to provide accurate and thoughtful responses [1]. This type of responding, referred to as *careless responding* or *insufficient effort responding* (IER) [2], can contaminate otherwise accurate data and bias statistical summaries. Intuition suggests that IER would weaken measures of the association of variables, and often it does [3]. However, recent research [4, 5] has observed and discussed the non-intuitive phenomena of IER *inflating* measures of association.

The negative consequences of inflated measures of association due to IER are many. The effect of IER on the linear correlation coefficient has been well documented [6, 7, 4], and conditions causing the magnitude of correlation to inflate have been characterized. Other studies have suggested that patterned careless responses in surveys with positively and negatively keyed items can lead to misleading conclusions about the dimensionality of constructs [8, 9].

Cronbach's alpha [10] is also subject to possible inflation under IER [11, 5]. Whether justified or not [12], Cronbach's alpha is often the only reported measure of reliability. McNeish [13] found that of 118 studies published in a 21-month period in American Psychological Association flagship journals, 109 used Cronbach's alpha as the sole assessment of reliability. Therefore the possibility of inflation due to careless responses is particularly insidious, as it is possible that researchers will overestimate an instrument's reliability. Ultimately, this can have downstream effects on the evaluation of the study, and can make studies seem more sound than they actually are.

2.0.1 Work Related to Cronbach's Alpha Under IER or Mixture Models

Previous investigations into the effect of IER on Cronbach's alpha fall into three categories: those that remove suspected IER from real data and see how the value of Cronbach's alpha changes [2, 14], those that simulate a combination of valid and careless responses and examine alpha and other statistics [5], and those that proceed by mathematical derivations. Previous works in the last category include Attali's [15] investigation of reliability in the context of speeded multiple-choice questions, and Fong, Ho, & Lam's [11] study which considered IER as consisting of either random or straight-lining responses, derived a formula for the bias in Cronbach's alpha, and plotted the bias for various proportions of each kind of IER.

The present chapter primarily consists of mathematical derivations with a small simulation component. The main result is a characterization of the behavior of Cronbach's alpha under a mixture of two distributions, representing valid responses and IER. An important earlier work in this area is that of Waller [16], in which he derived an expression for the value of Cronbach's alpha under a mixture model and illustrated through several examples how the mixture can create either a negative or positive bias. Our analysis will extend his work in two directions. First, we will relate the sign of the bias in Cronbach's alpha to a quadratic function of the mixing proportion. The roots and leading coefficient of this function yield five mathematical possibilities. Simulation is used to show that all five possibilities are potential realities by identifying sampling scenarios that correspond to each. Second, we relate the result to IER. Six distinct observations will be made, which include not only general confirmations of previous observations from simulation studies, but also the existence of a case which we have not seen mentioned in the literature.

The intent of this chapter is to demonstrate via mathematical proof all of the ways that Cronbachs alpha can be biased, including some possibilities that are non-intuitive. As

an educational aid, we include a link to a simulation app to help visualize the impact of varying proportions of IER.

The remaining sections will derive the mathematics, apply the main result in the context of IER, and conclude.

2.0.2 Mathematics for General Result

As the result builds on Waller's [16], we adopt his notation wherever possible. Kuder and Richardson [17] defined a measure of internal reliability for binary choices (commonly known as KR-20), which was generalized by Hoyt [18] and Guttman [19] and popularized by Cronbach [10] to the form bearing his name. Let $\mathbf{V} = (V_1, V_2, \ldots, V_k)$ represent responses from a multivariate probability distribution to k items on an instrument. The notation $\mathbf{V}$ is used to represent the *valid* distribution. Cronbach's alpha is defined as

$$\alpha_V = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \text{var}(V_i)}{\text{var}\left(\sum_{i=1}^{k} V_i\right)}\right). \tag{2.1}$$

An alternate formulation in terms of average variances and covariances will be convenient. Define $\overline{\sigma_{iV}^2}$ to be the average variance of components of $\mathbf{V}$, and $\overline{\sigma_{ijV}}$ to be the average covariance between distinct components of $\mathbf{V}$. Specifically,

$$\overline{\sigma_{iV}^2} = \frac{\sum_{i=1}^{k} \text{var}(V_i)}{k}$$

and

$$\overline{\sigma_{ijV}} = \frac{\sum\sum_{i\neq j} \text{cov}(V_i, V_j)}{k(k-1)}.$$

Then Cronbach's alpha may be expressed in the form [16, 20]

$$\alpha = \frac{k\overline{\sigma_{ijV}}}{(k-1)\overline{\sigma_{ijV}} + \overline{\sigma_{iV}^2}}. \tag{2.2}$$

Now consider an instrument with k items given to a population with two distinct subgroups. The first subgroup has a response distribution denoted by $\mathbf{V} = (V_1, V_2, \ldots, V_k)$, and the second subgroup has a response distribution denoted by $\mathbf{C} = (C_1, C_2, \ldots, C_k)$. The notation $\mathbf{C}$ is used to represent the *contaminating* IER.

Let W be a Bernoulli random variable with parameter p, where p is a value between zero and one representing the probability of observing a response from the contaminating class. That is, W is a random variable which takes value one with probability p and zero with probability $1 - p$. The responses actually recorded on the instrument are described by the multivariate distribution $\mathbf{M}$, defined by

$$\mathbf{M} = (1 - W)\mathbf{V} + W\mathbf{C}. \tag{2.3}$$

The notation $\mathbf{M}$ is used to emphasize that it is a *mixture* of the valid and contaminating responses. Because W is either zero or one, each individual gives responses from one of the two response distributions. With probability p an individual will give contaminating responses, and with probability $1 - p$ an individual will give valid responses.

By adopting this model, it is assumed that a respondent will either respond attentively to all items, or respond in an invalid manner to all items. We acknowledge that this assumption does not perfectly model real-life data; responses may be partially invalid [21] and are more likely to be invalid at the end of a survey [22]. However, we believe (and there is precedent in the literature [16]) this assumption represents a reasonable trade-off between the realism of the assumptions and the complexity of the model. Furthermore, the usual data cleaning methods used by a practitioner to remove suspected IER operate at the respondent level rather than the item level.

The goal is to find when $\alpha_M > \alpha_V$; that is, when contamination inflates Cronbach's alpha. First, notation and two results that will aid in the comparison are introduced.

Let μ_{iV} and μ_{iC} denote the respective means of responses to item i from the valid

and contaminating distributions. The differences in these means are called "item validities" in the taxometrics literature and are denoted by $\Delta_i = \mu_{iV} - \mu_{iC}$ [16]. As with the variances and covariances, only averages are needed. Specifically, the average product of item validities for distinct items, and the average of squared item validities:

$$\overline{\Delta_i \Delta_j} = \frac{\sum \sum_{i \neq j} \Delta_i \Delta_j}{k(k-1)} = \frac{\sum \sum_{i \neq j} (\mu_{iV} - \mu_{iC})(\mu_{jV} - \mu_{jC})}{k(k-1)},$$
$$\overline{\Delta_i^2} = \frac{\sum_{i=1}^{k} \Delta_i^2}{k} = \frac{\sum_{i=1}^{k} (\mu_{iV} - \mu_{iC})^2}{k}.$$

The first of the two needed results is known as the *general covariance mixture theorem* [23, 24]. Here, it will be expressed in terms of averages.

Lemma 2.0.1. *Let **M** be defined as a mixture of **V** and **C** as in Equation* (2.3)*, where p is the probability of observing a response from **C**. Assume the random quantities **V**, **C**, and W are independent. Then*

1. $\overline{\sigma_{ijM}} = (1-p)\overline{\sigma_{ijV}} + p\overline{\sigma_{ijC}} + p(1-p)\overline{\Delta_i \Delta_j}$
2. $\overline{\sigma_{iM}^2} = (1-p)\overline{\sigma_{iV}^2} + p\overline{\sigma_{iC}^2} + p(1-p)\overline{\Delta_i^2}$

A proof is in Appendix A of Meehl [23].

The second result is an inequality between average variances and covariances of items within a distribution, and will be used to investigate special cases during the discussion.

Lemma 2.0.2. *The average of covariances between distinct components of a multivariate distribution **V** is less than or equal to the average variance. Symbolically,*

$$\overline{\sigma_{ijV}} \leq \overline{\sigma_{iV}^2}.$$

The proof is in Appendix A.1. The main result can now be stated and proved.

Theorem 2.1. *Let* $\boldsymbol{V}$ *and* $\boldsymbol{C}$ *be multivariate distributions with* k *components representing potential responses to an instrument. Let* W *be a Bernoulli random variable with parameter* p *between zero and one. Define* $\boldsymbol{M} = (1-W)\boldsymbol{V} + W\boldsymbol{C}$ *as a mixture of* $\boldsymbol{V}$ *and* $\boldsymbol{C}$*. Assume that* $\boldsymbol{V}$*,* $\boldsymbol{C}$*, and* W *are independent. The behavior of Cronbach's alpha under the mixture can be broken down into five categories.*

1. *Cronbach's alpha does not change for any mixing proportion.* $\alpha_M = \alpha_V$ *for all* p*.*

2. *Cronbach's alpha inflates for any mixing proportion.* $\alpha_M > \alpha_V$ *for all* p*.*

3. *Cronbach's alpha deflates for any mixing proportion.* $\alpha_M < \alpha_V$ *for all* p*.*

4. *Cronbach's alpha inflates for small mixing proportions, but deflates for large mixing proportions. There is a value* p_0 *in the interval* $(0, 1)$ *such that* $\alpha_M > \alpha_V$ *for* $p < p_0$*, but* $\alpha_M < \alpha_V$ *for* $p > p_0$*.*

5. *Cronbach's alpha deflates for small mixing proportions, but inflates for large mixing proportions. There is a value* p_0 *in the interval* $(0, 1)$ *such that* $\alpha_M < \alpha_V$ *for* $p < p_0$*, but* $\alpha_M > \alpha_V$ *for* $p > p_0$*.*

Furthermore, there exist distributions that will yield each of the above cases, including when the item scale is continuous, discrete, or binary.

Proof. The general strategy is to derive that the sign of the bias in Cronbach's alpha has the same sign as a quadratic function of the mixing proportion p, and then invoke elementary properties of quadratic functions. Begin by finding conditions under which Cronbach's alpha inflates, or when $\alpha_M - \alpha_V > 0$. Apply Equation (2.2), the alternate form of Cronbach's alpha.

$$\frac{k\overline{\sigma_{ijM}}}{(k-1)\overline{\sigma_{ijM}} + \overline{\sigma_{iM}^2}} - \frac{k\overline{\sigma_{ijV}}}{(k-1)\overline{\sigma_{ijV}} + \overline{\sigma_{iV}^2}} > 0.$$

Combine into a single fraction with a common denominator.

$$k\frac{\overline{\sigma_{ijM}}\left((k-1)\overline{\sigma_{ijV}}+\overline{\sigma_{iV}^2}\right)-\overline{\sigma_{ijV}}\left((k-1)\overline{\sigma_{ijM}}+\overline{\sigma_{iM}^2}\right)}{\left((k-1)\overline{\sigma_{ijM}}+\overline{\sigma_{iM}^2}\right)\left((k-1)\overline{\sigma_{ijV}}+\overline{\sigma_{iV}^2}\right)}>0.$$

Expand the numerator. The term $(k-1)\left(\overline{\sigma_{ijM}}\right)\left(\overline{\sigma_{ijV}}\right)$ will cancel.

$$\frac{k}{\left((k-1)\overline{\sigma_{ijM}}+\overline{\sigma_{iM}^2}\right)\left((k-1)\overline{\sigma_{ijV}}+\overline{\sigma_{iV}^2}\right)}\left(\overline{\sigma_{iV}^2}\,\overline{\sigma_{ijM}}-\overline{\sigma_{ijV}}\,\overline{\sigma_{iM}^2}\right)>0. \tag{2.4}$$

The denominator of (2.4) is a scaled product of variances and must be positive, so it will not affect the sign of the bias. Thus to determine if Cronbach's alpha has inflated, it suffices to find when

$$\overline{\sigma_{iV}^2}\,\overline{\sigma_{ijM}}-\overline{\sigma_{ijV}}\,\overline{\sigma_{iM}^2}>0.$$

Apply the general covariance mixture theorem to the average variances and covariances of **M**.

$$\overline{\sigma_{iV}^2}\left((1-p)\overline{\sigma_{ijV}}+p\overline{\sigma_{ijC}}+p(1-p)\overline{\Delta_i\Delta_j}\right)-\overline{\sigma_{ijV}}\left((1-p)\overline{\sigma_{iV}^2}+p\overline{\sigma_{iC}^2}+p(1-p)\overline{\Delta_i^2}\right)>0.$$

Expand and group terms that include p^2 and p. The term $(1-p)\overline{\sigma_{iV}^2}\,\overline{\sigma_{ijV}}$ will cancel. The result is

$$\left(\overline{\sigma_{ijV}}\,\overline{\Delta_i^2}-\overline{\sigma_{iV}^2}\,\overline{\Delta_i\Delta_j}\right)p^2+\left(\overline{\sigma_{iV}^2}\,\overline{\sigma_{ijC}}-\overline{\sigma_{ijV}}\,\overline{\sigma_{iC}^2}-\overline{\sigma_{ijV}}\,\overline{\Delta_i^2}+\overline{\sigma_{iV}^2}\,\overline{\Delta_i\Delta_j}\right)p>0.$$

Define a and b as

$$a=\overline{\sigma_{ijV}}\,\overline{\Delta_i^2}-\overline{\sigma_{iV}^2}\,\overline{\Delta_i\Delta_j}, \tag{2.5}$$

$$b=\overline{\sigma_{iV}^2}\,\overline{\sigma_{ijC}}-\overline{\sigma_{ijV}}\,\overline{\sigma_{iC}^2}-\overline{\sigma_{ijV}}\,\overline{\Delta_i^2}+\overline{\sigma_{iV}^2}\,\overline{\Delta_i\Delta_j}=\overline{\sigma_{iV}^2}\,\overline{\sigma_{ijC}}-\overline{\sigma_{ijV}}\,\overline{\sigma_{iC}^2}-a. \tag{2.6}$$

It is now easy to see that Cronbach's alpha inflates when $f(p)=ap^2+bp>0$. $f(p)$ is a quadratic with no constant term. This is a simple family of functions, though the

coefficients are not simple. Momentarily ignore how a and b are defined (we will return to this soon), and consider the possible behaviors of functions of the form $f(p) = ap^2 + bp$. In particular, we are interested in the sign for values of p in the interval $(0, 1)$, as this will determine when Cronbach's alpha inflates or deflates. This behavior can be characterized in terms of the concavity and roots of $f(p)$. The roots are at 0 and $-b/a$ (as long as $a \neq 0$). Consider each case in turn.

1. Case one is a trivial possibility when $a = b = 0$. $f(p) = 0$ for all p in $(0, 1)$.

2. Case two is $f(p) > 0$ for all p in $(0, 1)$. This has two subcases: if $a \geq 0$ and $b \geq 0$ (but not both $a = b = 0$), or if $0 < -a \leq b$.

3. Case three is $f(p) < 0$ for all p in $(0, 1)$. This has two subcases: if $a \leq 0$ and $b \leq 0$ (but not both $a = b = 0$), or if $b \leq -a < 0$.

4. Case four occurs when $0 < b < -a$. The non-zero root p_0 lies in the interval $(0, 1)$, meaning $f(p)$ changes sign at p_0. Because $a < 0$ the function is concave down, so the bias changes from positive to negative.

5. Case five occurs when $-a \leq b < 0$. The non-zero root p_0 is in the interval $(0, 1)$, except now $a > 0$ and the function is concave up. The bias changes from negative to positive.

Examples of each non-trivial case, including the subcases for two and four, are included in Figure 3.2.

Because the sign of the bias is derived to be the same as the sign of a quadratic function with a root at zero, these scenarios are exhaustive. For example, a scenario in which Cronbach's first deflates, then inflates, and deflates again as p increases would require three crossings of the horizontal axis. This is not possible for a quadratic and is logically excluded.

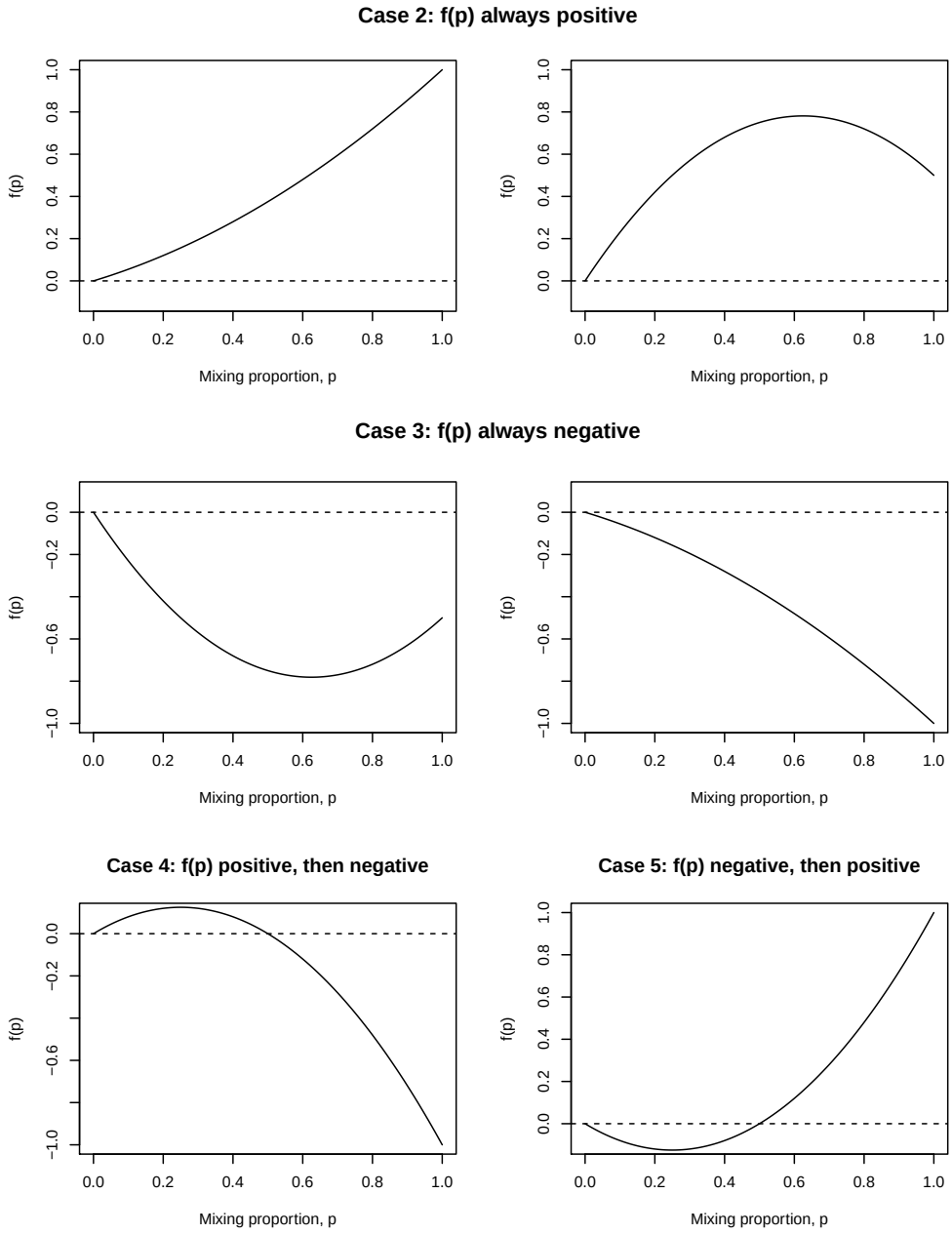

Figure 2.1: Graphs showing the general behavior of cases two through five for $f(p) = ap^2 + bp$, where p is the mixing proportion representing the probability of seeing a contaminating response. The sign of this function (positive or negative) on each region is equivalent to the sign of the bias of Cronbach's alpha.

To complete the proof, remember that the values a and b are not arbitrary, but defined in Equations (2.5) and (2.6) in terms of summaries of two multivariate probability distributions which represent responses to an instrument. Item validities, representing differences in means, are bounded by the item scale. Variances and covariances are also limited by the range of the scale and inequalities such as Cauchy-Schwarz [25]. Furthermore, if the scale has a small number of discrete options or is binary, then means, variances, and covariances are not independent parameters. A natural question is: are all five cases actually possible? The answer is yes, and the following two paragraphs describe how to obtain each.

Consider an instrument with 20 items on a five-point scale from one to five. No questions use negative keying. Discrete data is produced by first generating multivariate normal observations with a given mean vector and covariance matrix, and then rounding. The covariance matrix is constructed by using the average variance for the diagonal entries and the average covariance for the off-diagonal entries. The multivariate normal observations are rounded to the nearest integer in the scale. Forcing discrete responses by rounding will, of course, change the means, variances, and covariances, but in the particular cases considered, the change is not enough to alter the characterization of Cronbach's alpha. Two data sets are produced, representing valid responses and mixed responses. Cronbach's alpha is calculated for each, and the bias is recovered as the difference. This simulation is repeated for values of p, the mixing proportion, in increments of .025 between zero and one. Table 2.1 describes the means, variances, and covariances of the multivariate normal values which were rounded to obtain $\mathbf{V}$ and $\mathbf{C}$ in order to reproduce each case. Figure 2.2 shows, for each case, the bias of the simulation as a solid black line, the exact bias before discretization as a dotted blue line, and the function $f(p)$ as a dot-dash red line. To aid in seeing when the sign changes, there is a dashed line for the horizontal axis. The simulations used 10,000 respondents at each value of p.

Reproducing all cases when all items are binary options is trickier, but still possible.

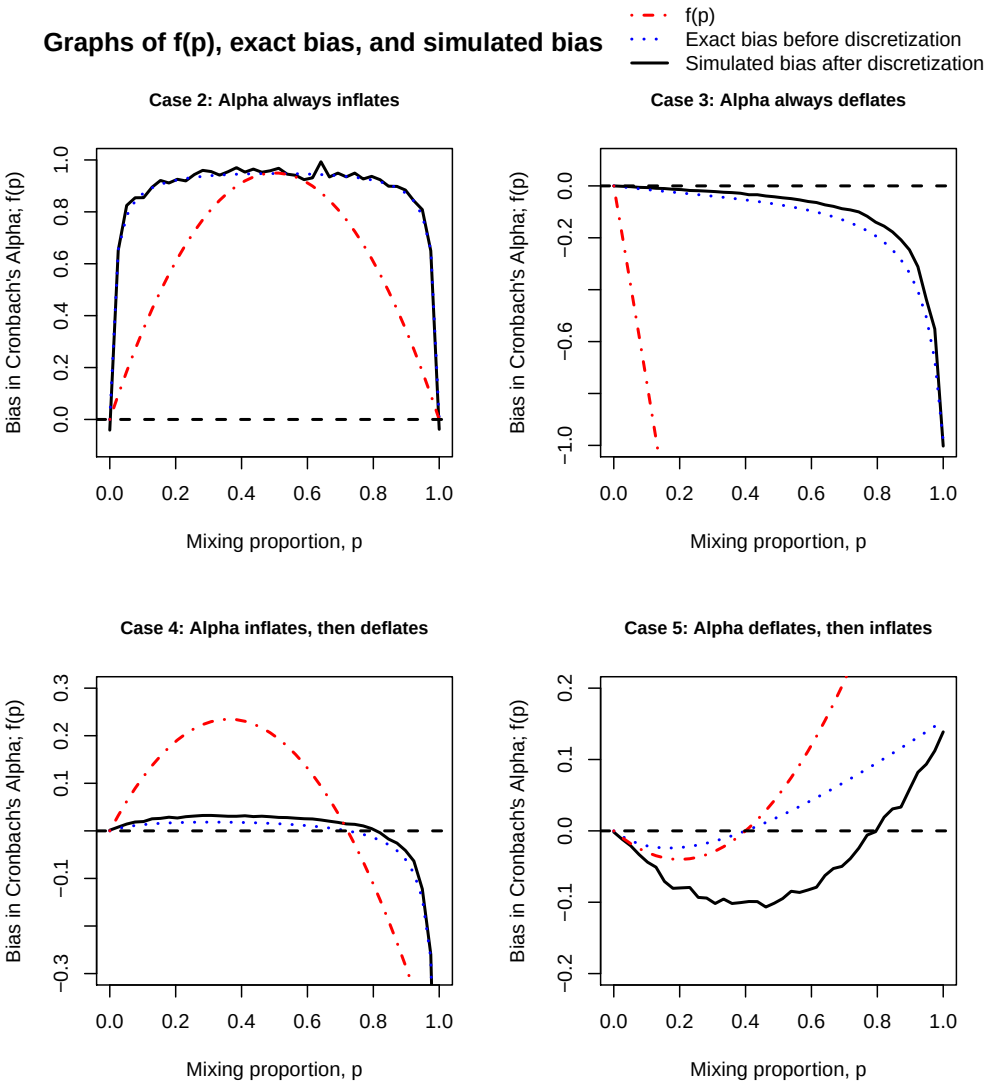

Figure 2.2: Graphs of $f(p)$, exact bias before discretization, and simulated bias after discretization for each case. The distributions described in Table 2.1 were used to produce these plots. 10,000 respondents were simulated for each value of p.

Table 2.1: Summaries of **V** and **C** (before discretization) for producing each case in the context of a scale with five options.

	Means of **V**	$\overline{\sigma^2_{iV}}$	$\overline{\sigma_{ijV}}$	Means of **C**	$\overline{\sigma^2_{iC}}$	$\overline{\sigma_{ijC}}$
Case 1	$\mu_{iV} = 3$ for all items	2	1	$\mu_{iC} = 3$ for all items	2	1
Case 2	$\mu_{iV} = 2$ for all items	1	0	$\mu_{iC} = 4$ for all items	1	0
Case 3	$\mu_{iV} = 2$ for all items	2	1.5	$\mu_{iC} = 4$ for all items	6	0
Case 4	$\mu_{iV} = 2$ for all items	1	.5	$\mu_{iC} = 4$ for all items	1	0
Case 5	$\mu_{iV} = 3$ for all items	1	.2	$\mu_{iV} = 1$ for odd items; $\mu_{iV} = 5$ for even items	1	.8

Table 2.2: Summaries of **V** and **C** (before discretization) for producing each case in the context of a scale with binary options.

	Means of **V**	$\overline{\sigma^2_{iV}}$	$\overline{\sigma_{ijV}}$	Means of **C**	$\overline{\sigma^2_{iC}}$	$\overline{\sigma_{ijC}}$
Case 1	$\mu_{iV} = .5$ for all items	1	.5	$\mu_{iC} = .5$ for all items	1	.5
Case 2	$\mu_{iV} = .4$ for all items	1	0	$\mu_{iC} = .6$ for all items	1	.8
Case 3	$\mu_{iV} = .4$ for all items	.5	.4	$\mu_{iC} = .6$ for all items	1	0
Case 4	$\mu_{iV} = .2$ for all items	.3	.1	$\mu_{iC} = .8$ for all items	.3	0
Case 5	$\mu_{iV} = .5$ for all items	1	.3	$\mu_{iC} = 0$ for odd items; $\mu_{iC} = 1$ for even items	.5	.2

Data was simulated in the same manner as before, except now responses are rounded to the nearest of zero or one. Table 2.2 describes the summaries of the multivariate normal values which were rounded to obtain **V** and **C** in order to reproduce each case. A larger number of simulations was necessary to reduce the sampling variability and clearly see the bias in Cronbach's alpha. We found 20,000 to be sufficient for all except case five, which used 100,000 simulations. The graphs for the binary case are not significantly different from the five-option case, and are omitted. This completes the proof. □

We illustrate one of the non-intuitive possibilities, case 2, through an example with

simulated data. Table 3.4 contains data from ten respondents for an instrument with five items. The data were generated such that there was a $p = .5$ chance of a contaminating response. In this sample, the result was six valid responses and four contaminating responses. Each class has a variance of one for all items and a covariance of zero (due to independence) between any pair of items. Because any pair of items has independent responses, the exact value for Cronbach's alpha is zero, which is estimated from this sample to be .13 for the valid class and .063 for the contaminating class. However, because the valid class has a mean of two and the contaminating class has a mean of four, responses *appear* to be consistent when the two classes are combined into a single dataset. The resulting estimate of Cronbach's alpha for the entire sample is .87, a value generally seen as desirable, yet we see it is only due to contamination in the sample.

Now let us relate this example to Theorem 2.1. The contaminating classes have summaries $\mu_{iV} = 2$, $\mu_{iC} = 4$, $\overline{\sigma_{iV}^2} = 1$, $\overline{\sigma_{iC}^2} = 1$, $\overline{\sigma_{ijV}} = 0$, and $\overline{\sigma_{ijC}} = 0$. Applying Equations (2.5) and (2.6), we see:

$$a = \overline{\sigma_{ijV}}\,\overline{\Delta_i^2} - \overline{\sigma_{iV}^2}\,\overline{\Delta_i\Delta_j} = 0 \cdot 4 - 1 \cdot 4 = -4,$$
$$b = \overline{\sigma_{iV}^2}\,\overline{\sigma_{ijC}} - \overline{\sigma_{ijV}}\,\overline{\sigma_{iC}^2} - a = 1 \cdot 0 - 0 \cdot 1 - (-4) = 4.$$

As $0 < -a = b$, this is case 2 of Theorem 2.1, so Cronbach's alpha will inflate for any proportion p of contamination.

For this example we estimated Cronbach's alpha using Equation (2.1) with sample estimates of variance, but most software solutions have built-in functions with helpful features. For example, the `alpha()` function in the `psych` package [26] in R [27] produces a confidence interval for alpha and an analysis of how alpha will change if items are removed from this instrument.

At this point, no claim is made as to how likely each case is, only that all are possible. In the discussion, we will demonstrate that each of these cases could potentially be arrived at through specific types of IER.

Table 2.3: An example of data simulated from case 2. Each class is inconsistent, with estimates of Cronbach's alpha being .13 for the valid class and .06 for the contaminating class, yet the combined data set estimates alpha as .87.

Respondent	Response class	Q1	Q2	Q3	Q4	Q5
1	Contaminating	4	5	4	3	4
2	Valid	3	1	3	1	1
3	Valid	1	1	1	1	2
4	Valid	1	2	4	2	1
5	Contaminating	4	4	3	2	4
6	Valid	1	1	1	1	2
7	Valid	2	3	1	2	2
8	Contaminating	3	5	3	5	5
9	Contaminating	3	5	3	4	3
10	Valid	2	2	4	2	1

Because of the removal of the positive term in Equation (2.4), $f(p)$ does not give the magnitude of the bias, but is a function with the same sign as the bias. The cancelled term includes covariances of $\mathbf{M}$ which implicitly depend on p, so the magnitude of the bias is a ratio of polynomials in p and is more difficult to analyze. Figure 2.2 makes it clear that $f(p)$ and the magnitude of the exact bias share the same roots and sign but potentially very different magnitudes. This is why the cases in Tables 2.1 and 2.2 and Figure 2.2 do not differentiate between $f(p)$ being concave or convex; that property is not necessarily shared with the magnitude of the bias.

The simulation code that produced Figure 2.2 is available through a Shiny R [28] web app found at https://alphaier.shinyapps.io/cronbachs_alpha_under_ier/. The app allows users to investigate how Cronbach's alpha will behave under a mixture model consisting of discretized multivariate normal distributions with any mean vector, average variance, and average covariance (as long as the resulting covariance matrix is valid). We see two potential uses for this tool. The first is educational, as it allows the user to visualize the effects of contamination on Cronbach's alpha and test the effect when the valid and contaminating distribution are altered. Second, the forthcoming discussion relates Theorem 2.1 to the two main types of IER, but IER could potentially manifest through myriad possible response distributions. For any other hypothesized pattern of IER, if the means, variances, and covariances can be specified, this tool can immediately determine the effect of that contamination on Cronbach's alpha.

The reader is reminded that the preceding is true for any mixture of two distributions, whether the interpretation of "valid" and "contaminating" holds or not.

2.1 Discussion and Special Cases

This section refers repeatedly to the quadratic coefficients a and b, which were defined as

$$a = \overline{\sigma_{ijV}}\,\overline{\Delta_i^2} - \overline{\sigma_{iV}^2}\,\overline{\Delta_i\Delta_j},$$
$$b = \overline{\sigma_{iV}^2}\,\overline{\sigma_{ijC}} - \overline{\sigma_{ijV}}\,\overline{\sigma_{iC}^2} - \overline{\sigma_{ijV}}\,\overline{\Delta_i^2} + \overline{\sigma_{iV}^2}\,\overline{\Delta_i\Delta_j} = \overline{\sigma_{iV}^2}\,\overline{\sigma_{ijC}} - \overline{\sigma_{ijV}}\,\overline{\sigma_{iC}^2} - a.$$

With effort, these complicated coefficients yield much information about the behavior of Cronbach's alpha under mixture models.

Now we relate the behavior of Cronbach's alpha to IER. Studying IER in any generality is difficult because IER can take so many forms. IER is defined more by what the responses are not, rather than by what they are. For this reason, past investigations [11, 5] have typically considered two extreme forms through which IER may manifest:

- *Random responding*: Item responses are uniformly and independently chosen from those available.

- *Straight-lining*: The respondent chooses the same option for all items, either in an attempt to complete the instrument as quickly as possible or operating on the belief that all questions are sufficiently similar to the first. Different respondents may choose different options, but each respondent repeats their choice without deviation.

A strength of the current approach is that any form of IER can be investigated as long as the means, average variance, and average covariance can be determined. Some of the following observations will refer specifically to random responding or straight-lining, and the fifth observation will deal with a potential form of IER we have not seen studied in the literature.

Observation 1: If all components of **V** have a common mean, and all components of **C** have a common mean, then the quadratic coefficient a cannot be positive. Cases one

through four are still possible, but case five is not.

Suppose that all means are equal within a response distribution, that is, $\mu_{iV} = \mu_{jV}$ and $\mu_{iC} = \mu_{jC}$ for all items i and j. This implies $\overline{\Delta_i^2} = \overline{\Delta_i \Delta_j}$, and so Equation (2.5) can be simplified as

$$a = \left(\overline{\sigma_{ijV}} - \overline{\sigma_{iV}^2}\right) \overline{\Delta_i^2}. \tag{2.7}$$

$\overline{\Delta_i^2}$ is an average of squares and is clearly positive, while Lemma 2.0.2 implies the term in parentheses is negative. Thus a is non-positive, precluding the possibility of case five.

Actually, the assumption of common means within a distribution is stronger than necessary, as it is sufficient for all item validities to be equal. However, the case of common means within a distribution is an important special case for many of the following observations, so that is the form in which the observation is stated.

Observation 2: The forces pressuring Cronbach's alpha to inflate are:

1. Increasing the differences between means of **V** and **C** when item validities have the same sign,

2. Increasing the ratio of average covariance to variance for the contaminating distribution,

3. Decreasing the ratio of average covariance to variance for the valid distribution.

Likewise, forces in the opposite direction will pressure Cronbach's alpha to deflate.

The first part of this observation is easiest to see when response distributions have common means, so consider $a \neq 0$ as expressed in Equation (2.7). As the difference in means grows, a becomes more negative and b becomes more positive. This moves in the direction of cases two and four, so alpha tends to inflate. This is pertinent to situations in which the content of the survey will lead the mean of attentive responses to be close to

an extreme bound. Consider a survey attempting to detect a rare trait like psychopathy. The mean of attentive responses is expected to be low, but if careless responses are chosen randomly and have a mean close to the midpoint of the scale, it is possible alpha will inflate due to IER. This suggests that measures of extreme psychopathology may report inflated values of Cronbach's alpha.

The second part of this observation is intuitively plausible, as it corresponds to contaminating with a highly internally consistent distribution, such as straight-lining. As the average covariance of **C** increases, b becomes more positive, moving away from case three towards case two, possibly moving through cases four and five.

The third part of this observation corresponds to a valid distribution with low internal consistency, leaving ample opportunity for inflation. As the average covariance of **V** decreases, a becomes more negative, which by itself would pressure alpha towards deflation, but b includes $-a$ in the sum, so b is becoming more positive. Also the second term in b includes a negative average covariance of **V**. Thus b is increasing faster than a is decreasing, moving in the direction of case two and possibly case four.

Observation 3: If means are equal across items and distributions and contamination consists purely of random responses, then Cronbach's alpha must deflate (except for the unusual case that $\alpha_V \leq 0$). However, if the distributions have different means, either inflation or deflation is possible.

The key characteristic of random responding is the independence between responses, so $\overline{\sigma_{ijC}} = 0$. All means being equal implies all item validities are zero, thus $a = 0$. Combining these observations, $b = -\overline{\sigma_{ijV}}\overline{\sigma^2_{iC}}$. The assumption that $\alpha_V > 0$ implies $\overline{\sigma_{ijV}} > 0$, so b is negative. This is case three, in which Cronbach's alpha deflates. The example of case three in Table 2.1 illustrates this exact situation.

However, if the means of **V** and **C** are not equal, then deflation is not guaranteed. Consider case two in Table 2.1 and Figure 2.2, in which the contaminating distribution has

independent responses, yet alpha always inflates due to the difference in means. Case four also uses a contaminating distribution with independent observations, but whether alpha inflates or deflates depends on the exact mixing proportion p. Case two is noteworthy because both of the distributions contributing to the mixture have average covariances of zero (thus $\alpha_V = \alpha_C = 0$), yet the mixture has a positive value of Cronbach's alpha. This scenario is discussed by Waller [16] as admittedly contrived and non-realistic, but useful as an example of the non-intuitive nature of reliability measures under mixtures.

The scenario of contamination with random responses is included in the simulation study of DeSimone et al. [5]. Prior to the study, the authors state the expectation that random responding will reduce alpha (p. 312). Figure 2 of the same article confirms that for their particular situation, random responses did indeed result in a strict decrease in Cronbach's alpha. However, the results of the present article show that this is not the only possibility, and that random responses can increase Cronbach's alpha if the means of the valid and contaminating distributions are sufficiently different.

Observation 4: Assume the valid distribution has a common mean and no questions use reverse keying. If contamination consists purely of straight-lining, then alpha is guaranteed to inflate.

The key characteristic of straight-lining respondents is that covariance equals the variance, which can be seen from applying $C_i = C_j$ to the definition of covariance. Furthermore, straight-lining forces a common mean. Thus the item validities are all identical, and from Observation 1 a is negative. Combining with the fact $\overline{\sigma_{ijC}} = \overline{\sigma_{iC}^2}$ and Lemma 2.0.2, b can be simplified as

$$b = \overline{\sigma_{iC}^2}\left(\overline{\sigma_{iV}^2} - \overline{\sigma_{ijV}}\right) - a \geq -a > 0.$$

This is case two, so Cronbach's alpha can only inflate. This confirms generally the expectation by DeSimone et al. [5] that pure straight-lining will inflate alpha.

Observation 5: If contamination is of a form that alternates between extremes, then case five is a possibility. Cronbach's alpha deflates for small p, but inflates for larger p.

Consider a mischievous responder who deliberately alternates between the first and last option in a scale for the entirety of the instrument. This form of IER can produce case five. We are not aware of (nor would we expect) any studies of this contrived style of response (though anecdotally, one of the authors observed a classmate exhibit this behavior on a standardized test in secondary school). This is case five in Table 2.1 and Figure 2.2. This could also be produced by straight-lining respondents when the survey alternates between regular and reverse keying.

Observation 6: To investigate the effects of multiple types of IER occurring simultaneously, mixture models and the general covariance mixture theorem can be applied iteratively.

In reality, IER rarely consists exclusively of purely random or straight-lining responses. It is more likely that non-valid responses from $\mathbf{C}$ are themselves a mixture of random, straight-lining, and perhaps other kinds of IER. Therefore the general covariance mixture theorem (Lemma 2.0.1 in the present chapter) can be applied repeatedly to obtain the parameters of $\mathbf{C}$, at which point Theorem 2.1 can be applied to determine whether Cronbach's alpha will deflate or inflate.

Consider the following illustrative example. An instrument has five questions, with each having five options. Of the respondents, 75% will answer in a valid manner, 20% will answer randomly, and 5% will straight-line. The valid responses follow a discrete uniform with $\alpha_V = .4$, corresponding to a common mean $\mu_{iV} = 3$, an average variance of $\overline{\sigma_{iV}^2} = 2$ and an average covariance of $\overline{\sigma_{ijV}} = .235$. Careless responses come in two forms. The 20% of respondents who respond randomly (denote corresponding quantities with the subscript R) have a common mean $\mu_{iR} = 3$, an average variance of $\overline{\sigma_{iR}^2} = 2$, and an average covariance of $\overline{\sigma_{ijR}} = 0$. The 5% of respondents who straight-line (denote corresponding

quantities with the subscript S) chose the first item uniformly, so these responses have a common mean $\mu_S = 3$, an average variance of $\overline{\sigma_{iS}^2} = 2$, and an average covariance of $\overline{\sigma_{ijS}} = 2$. All means are identical, so item validities are zero. Within the contaminating class, 80% comes from random responses and 20% are straight-line responses, so an application of Lemma 2.0.1 yields $\overline{\sigma_{iC}^2} = .8 \cdot 2 + .2 \cdot 2 = 2$, and $\overline{\sigma_{ijC}} = .8 \cdot 0 + .2 \cdot 2 = .04$. Next, applying Equations (2.5) and (2.6) yields $a = 0$ and $b = 2 \cdot 2 - 0 \cdot 2 = .04 > 0$, so this is case two, where Cronbach's alpha will always inflate. It is interesting that for this example, only the mixing proportion *within* the contaminating class is important. Because the valid distribution has relatively low internal consistency, the few straight-lining responders have a larger effect than the more numerous random responders. Changing the mixture proportion between the valid and contaminating class may change the magnitude of the bias, but will not alter the fact that Cronbach's alpha will inflate due to contamination.

In the previous example, the random responding followed a uniform distribution (a sort of pure randomness). There are infinite other potential response distributions, but fortunately, a full specification of the exact distribution is not necessary. The behavior of Cronbach's alpha depends on the valid and contaminating distributions only through the variances, covariances, and differences in means. So in a partial sense, the realm of possibilities is reduced. Any multivariate distribution of responses with identical means, variances, and covariances will bias Cronbach's alpha in the same manner.

How do real-life manifestations of careless responses tend to affect Cronbach's alpha? For an answer, we defer to studies with real data where alpha is calculated with and without suspected IER. Huang et al. [2] compared alpha for 30 facets and found that generally alpha decreases as a result of careless responses, with a notable exception: one section with eight positively keyed and only two negatively keyed items manifested in an increase in alpha, which the authors attribute to straight-lining respondents, which is consistent with the present analysis. Wertheimer [14] conducted a similar analysis for multiple data sets

and classified respondents as conscientious, random, or patterned. In summary, removing random respondents tended to increase alpha, removing patterned respondents tended to decrease alpha, and removing both tended to increase alpha, but to a lesser degree. This agrees with the results of the present chapter, lending evidence that the mathematical assumptions are not too unrealistic.

2.2 Conclusion

This chapter has presented a mathematical analysis of the behavior of Cronbach's alpha when responses are contaminated with a secondary distribution, with a discussion emphasizing the implications for contamination from careless responses.

One limitation is the assumption that a respondent will answer all questions in either a valid or non-valid manner. In reality, some individuals will become weary part-way through a survey and begin to answer carelessly, or give careless answers to items that demand a greater deal of thought. On one hand, this simplification makes a mathematical analysis feasible, reveals the possibility of behavior not mentioned in previous IER literature (case 5), and offers mathematical certainty whenever the assumption is met. On the other hand, because the model excludes the possibility of partial IER by an individual, the conclusions reached in this investigation will not apply perfectly to real-life manifestations of IER. This is a weakness likely to affect any conceptual research into IER, as any model general enough to encompass all possible IER patterns is probably too broad to reach any specific conclusions. More complex models may allow the relaxation of this assumption, such as specifying probabilities of careless responses based on cognitive load and distance into the instrument, or modeling the number of items a respondent will answer validly before answering carelessly from that point on. Such models may not allow for a tractable analysis, and might be better suited for simulation studies. Another possibility is applying person-fit statistics [29, 30] and item-response theory [31] to investigations of reliability

under IER.

This chapter does not address the issue of how a researcher should deal with IER. The reader is referred to one of the many articles detailing methods for detecting and removing IER [32, 2, 33, 34, 35].

Though not the only measure of reliability, Cronbach's alpha is the most common. This chapter should not be viewed as an indictment of a deficiency unique to Cronbach's alpha, but alpha is the first natural choice for investigating the effect of IER on internal consistency. Investigation into the effect of IER on other measures of reliability, including beta [36] or ω_h [37], is a possible avenue for future research.

Hopefully, the analysis in this chapter will increase understanding of how Cronbach's alpha will behave under IER, and convince practitioners that IER is a threat to high quality research. In particular, except for the special cases of straight-lining when no questions use reverse keying and random responding with the same mean, IER can cause Cronbach's alpha to behave in non-intuitive and unpredictable ways. Because Cronbach's alpha can inflate due to IER, practitioners should be aware that a high value of alpha does not imply respondents were sufficiently attentive; it may be due to straight-lining or random responding with a different mean. In the other direction, random responding with a mean similar to the valid responses can decrease Cronbach's alpha, underestimating the reliability of an instrument. We suggest a best practice is that measures for prevention, detection and removal of IER take place before analysis [32], especially calculations of reliability.

CHAPTER 3

BIAS IN SPEARMAN'S RHO UNDER BIVARIATE MIXTURE

3.1 INTRODUCTION

When considering data that presents some form of relationship that cannot be described as a primarily linear, there are many different correlation measures that can be used. One particular class of correlation measures, used when the data has some form of monotonic relationship, is the class of rank-based correlation methods. In particular, a famous rank-based measure of monotonic correlation is that of Spearman's Rho [39]. A population form of Spearman's Rho was not discovered until later by Kruskal and Lehmann [40, 41]. The population form of Spearman's Rho involves the notion of a copula [44], which we define below.

Definition 4. A *two-dimensional copula* is a function $C : \mathbb{I}^2 \to \mathbb{I}$ satisfying the following properties:

1. For every $u, v \in \mathbb{I}$,

   $C(u, 0) = 0 = C(0, v)$ and

   $C(u, 1) = u$, $C(1, v) = v$

2. For every $u_1, u_2, v_1, v_2 \in \mathbb{I}$, satisfying $u_1 \leq u_2$ and $v_1 \leq v_2$

   $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$

Originally, copulas were discovered in the study of probabilistic metric spaces [42], and their relation to random variables can be summarized in the following theorem.

Theorem 3.1 (Sklar's Theorem [42])**.** *Let H be a joint distribution function with marginals F and G. Then there exists a copula C such that for all $x, y \in \overline{\mathbb{R}}$,*

$$H(x, y) = C(F(x), G(y)).$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on $RanF \times RanG$. Conversely, if C is a copula and F and G are distribution functions, then the function H defined prior is a joint distribution function with marginals F and G.

Essentially, Sklar's theorem tells us that a copula serves as "linking" function between the marginal distribution functions of two random variables. This "linking" function, in a sense, captures the dependence structure between two random variables. Thus, since a copula captures dependence, we can discuss measures of correlation in terms of copulas. The population form of Spearman's Rho as a function of a copula is as follows:

$$\rho_s = 12 \iint_{\mathbb{I}^2} C(u_1, u_2) du_1 du_2 - 3 \tag{3.1}$$

Of particular interest to practitioners is the effect that sampling from an undesired, or contaminating, population can have on statistical summaries. Usually, this sort of sampling from an undesired population can be represented by a bivariate mixture model, as defined in Definition 3, where the mixing proportion represents the probability of an observation coming from the contaminating distribution. The approach of using mixture models as a representation of mis-sampling has been used in many settings with regards to statistical summaries. In particular, the resulting difference between the "contaminated" versions of statistical summaries, which we shall consider as the bias resulting from the mixture model (as defined in Definition 3), has been of particular interest to researchers in the area of quantitative psychology (see [16]). In addition to this, characterizations of the resulting behavior of the bias for different values of the mixing proportion has drawn the attention of researchers. Results along these lines have been demonstrated for Cronbach's Alpha, a measure of internal consistency (see [16, 43]).

In this chapter, we aim to provide some form of a characterization of the bias of Spearman's Rho under a bivariate mixture model. In Section 3.2, we demonstrate that the bias in Spearman's Rho resulting from a bivariate mixture model can be represented as a

cubic function of the mixing proportion. We then derive some mathematical cases for a cubic function that would be of particular interest to practitioners in Section 3.3. Finally, in Section 3.4, due to the complexity of the coefficients of the cubic function, we perform simulations for multivariate normal random variables to demonstrate that there exist mean vectors and covariance matrices such that all mathematical cases for a cubic function as described in Section 3.2 are possible for the forms that our coefficients take.

3.2 Mathematics of Spearman's Rho Result

We will now begin demonstrating that the bias of Spearman's Rho under a bivariate mixture model can be represented as a cubic function. First, we begin with some algebraic and probabilistic lemmas that will aid in this demonstration. For the following lemmas and the main theorem, we shall treat $\mathbf{V} = (V_1, V_2)$ as representing the desired, or valid, distribution, and $\mathbf{C} = (C_1, C_2)$ as representing the undesired, or contaminating distribution.

Lemma 3.2.1. *Let $\boldsymbol{C} = (C_1, C_2)$, $\boldsymbol{V} = (V_1, V_2)$ be continuous random vectors with support on $\mathbb{S} \subset \mathbb{R}^2$. For $\boldsymbol{M} = W\boldsymbol{C} + (1 - W)\boldsymbol{V}$, with $W \sim Bernoulli(p)$ and independent of $\boldsymbol{C}$ and $\boldsymbol{V}$,*

$$C_M(u_1, u_2) = pF_C(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)) + (1 - p)F_V(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)).$$

Proof. By Sklar's Theorem [42], we have by continuity

$$\begin{aligned} F_M(u_1, u_2) =& C_M(F_{M_1}(u_1), F_{M_2}(u_2)) \\ \implies& F_M(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)) = C_M(F_{M_1}(F_{M_1}^{-1}(u_1)), F_{M_2}(F_{M_2}^{-1}(u_2))) = C_M(u_1, u_2). \end{aligned}$$

Thus, using this expression for the copula of $\mathbf{M}$, we have that

$$\begin{aligned} C_M(u_1, u_2) =& F_M(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)) \\ =& \mathbb{P}(M_1 \leq F_{M_1}^{-1}(u_1), M_2 \leq F_{M_2}^{-1}(u_2)) \end{aligned}$$

By applying the Law of Total Probability, we arrive at

$$\begin{aligned}
&= p\mathbb{P}(C_1 \leq F_{M_1}^{-1}(u_1), C_2 \leq F_{M_2}^{-1}(u_2)) + (1-p)\mathbb{P}(V_1 \leq F_{M_1}^{-1}(u_1), V_2 \leq F_{M_2}^{-1}(u_2)) \\
&= pF_C(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)) + (1-p)F_V(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)).
\end{aligned}$$

□

The above demonstrates that we can represent the copula of a bivariate mixture model as a convex combination in the mixing proportion of the joint distribution functions of V and C. The next lemma demonstrates that a similar result holds for marginal distribution and density functions.

Lemma 3.2.2. *Let C, V, and M be described as in Lemma 3.2.1. Then, for $i \in \{1, 2\}$,*

$$F_{M_i}(u_i) = pF_{C_i}(u_i) + (1-p)F_{V_i}(u_i)$$

and

$$f_{M_i}(u_i) = pf_{C_i}(u_i) + (1-p)f_{V_i}(u_i). \tag{3.2}$$

Proof. By the Law of Total Probability,

$$\begin{aligned}
F_{M_i}(u_i) &= \mathbb{P}(M_i \leq u_i) \\
&= p\mathbb{P}(C_i \leq u_i) + (1-p)\mathbb{P}(V_i \leq u_i) \\
&= pF_{C_i}(u_i) + (1-p)F_{V_i}(u_i)
\end{aligned}$$

Thus, the first equality is satisfied. Applying a derivative, we arrive at:

$$\begin{aligned}
&\implies \frac{d}{du_i}F_{M_i}(u_i) = \frac{d}{du_i}pF_{C_i}(u_i) + \frac{d}{du_i}(1-p)F_{V_i}(u_i) \\
&\implies f_{M_i}(u_i) = pf_{C_i}(u_i) + (1-p)f_{V_i}(u_i).
\end{aligned}$$

□

The final lemma is simply for algebraic purposes. It will allow us to simplify some of the expressions that we will encounter in the main theorem.

Lemma 3.2.3. *Let C, V, and M be described as in Lemma 3.2.1. Then,*

$$f_{M_1}(u_1)f_{M_2}(u_2) = \Delta_2\Delta_1 p^2 + (f_{V_2}(u_2)\Delta_1 + f_{V_1}(u_1)\Delta_2)p + f_{V_1}(u_1)f_{V_2}(u_2)$$

where

$$\begin{aligned}\Delta_2 =& f_{C_2}(u_2) - f_{V_2}(u_2)\\ \Delta_1 =& f_{C_1}(u_1) - f_{V_1}(u_1)\end{aligned}$$

Proof. Applying Lemma 3.2.2,

$$\begin{aligned}f_{M_1}(u_1)f_{M_2}(u_2) =& (pf_{C_1}(u_1) + (1-p)f_{C_2}(u_2))(pf_{C_2}(u_2) + (1-p)f_{V_2}(u_2))\\ =& p^2 f_{C_1}(u_1)f_{C_2}(u_2) + p(1-p)f_{C_1}(u_1)f_{V_2}(u_2) + p(1-p)f_{C_2}(u_2)f_{V_1}(u_1)\\ &+ (1-p)^2 f_{V_1}(u_1)f_{V_2}(u_2)\\ =& p^2(f_{C_1}(u_1)f_{C_2} - f_{C_1}(u_1)f_{V_2}(u_2) - f_{C_2}(u_2)f_{V_1}(u_1) + f_{V_1}(u_1)f_{V_2}(u_2))\\ &+ p(f_{C_1}(u_1)f_{V_2}(u_2) - 2f_{V_1}(u_1)f_{V_2}(u_2) + f_{C_2}(u_2)f_{V_1}(u_1))\\ &+ f_{V_1}(u_1)f_{V_2}(u_2)\\ =& p^2(f_{C_2}(u_2) - f_{V_2}(u_2))(f_{C_1}(u_1) - f_{V_1}(u_1))\\ &+ p(f_{V_2}(u_2)(f_{C_1}(u_1) - f_{V_1}(u_1)) + f_{V_1}(u_1)(f_{C_2}(u_2) - f_{V_2}(u_2))\\ &+ f_{V_1}(u_1)f_{V_2}(u_2)\\ =& \Delta_2\Delta_1 p^2 + (f_{V_2}(u_2)\Delta_1 + f_{V_1}(u_1)\Delta_2)p + f_{V_1}(u_1)f_{V_2}(u_2).\end{aligned}$$

□

Now that we have proved the above lemmas, we can now begin to prove our main result.

Theorem 3.2. *Let C, V, and M be defined as in Lemma 3.2.1. Then the bias in Spearman's Rho as defined in Definition 3 is a cubic function of the mixing proportion of the form:*

$$\textit{Bias}(p) = ap^3 + bp^2 + cp$$

where

$$\begin{aligned}
a =& 12 \iint_{\mathbb{S}} (F_C(v_1, v_2) - F_V(v_1, v_2))\Delta_2\Delta_1 dv_1 dv_2, \\
b =& 12 \iint_{\mathbb{S}} (F_C(v_1, v_2) - F_V(v_1, v_2))(f_{V_2}(v_2)\Delta_1 + f_{V_1}(v_1)\Delta_2) dv_1 dv_2 \\
&+ 12 \iint_{\mathbb{S}} F_V(v_1, v_2)\Delta_2\Delta_1 dv_1 dv_2, \\
c =& 12 \iint_{\mathbb{S}} (F_C(v_1, v_2) - F_V(v_1, v_2)) f_{V_1}(v_1) f_{V_2} v_2) dv_1 dv_2 \\
&+ 12 \iint_{\mathbb{S}} F_V(v_1, v_2)(f_{V_2}(v_2)\Delta_1 + f_{V_1}(v_1)\Delta_2) dv_1 dv_2.
\end{aligned}$$

Proof. Applying Lemma 3.2.1, and by Equation 3.1, we have the following expression for bias:

$$\begin{aligned}
\rho_M - \rho_V =& 12 \iint_{\mathbb{I}^2} C_M(u_1, u_2) du_1 du_2 - 3 - (12 \iint_{\mathbb{I}^2} C_V(u_1, u_2) du_1 du_2 - 3) \\
=& 12 \iint_{\mathbb{I}^2} C_M(u_1, u_2) du_1 du_2 - 12 \iint_{\mathbb{I}^2} C_V(u_1, u_2) du_1 du_2 \\
=& 12p \iint_{\mathbb{I}^2} F_C(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)) du_1 du_2 - 12p \iint_{\mathbb{I}^2} F_V(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)) du_1 du_2 \\
&+ 12 \iint_{\mathbb{I}^2} F_V(F_{M_1}^{-1}(u_1), F_{M_2}^{-1}(u_2)) du_1 du_2 - 12 \iint_{\mathbb{I}^2} C_V(u_1, u_2) du_1 du_2.
\end{aligned}$$

Noting that the marginal distribution functions are continuous and strictly increasing and thus the inverse distribution functions exist, we apply the following variable substitution to the first three integrals of the bias:

$$\begin{aligned}
v_1 = F_{M_1}^{-1}(u_1), v_2 = F_{M_2}^{-1}(u_2) \\
\implies u_1 = F_{M_1}(v_1), u_2 = F_{M_2}(v_2).
\end{aligned}$$

Note that these are univariate substitutions. Thus, for the Jacobian, off-diagonal terms are zero. Thus, we have:

$$|\mathbb{J}| = \left| \begin{bmatrix} \frac{\partial u_1}{\partial v_1} & 0 \\ 0 & \frac{\partial u_2}{\partial v_2} \end{bmatrix} \right| = f_{M_1}(v_1) f_{M_2}(v_2)$$

Thus, our bias becomes:

$$\begin{aligned} =& 12p \iint_{\mathbb{S}} F_C(v_1, v_2) f_{M_1}(v_1) f_{M_2}(v_2) dv_1 dv_2 \\ & - 12p \iint_{\mathbb{S}} F_V(v_1, v_2) f_{M_1}(v_1) f_{M_2}(v_2) dv_1 dv_2 \\ & + 12 \iint_{\mathbb{S}} F_V(v_1, v_2) f_{M_1}(v_1) f_{M_2}(v_2) dv_1 dv_2 - 12 \iint_{\mathbb{I}^2} C_V(u_1, u_2) du_1 du_2. \end{aligned}$$

Applying Lemma 3.2.3, we have:

$$\begin{aligned} =& 12p \iint_{\mathbb{S}} F_C(v_1, v_2) \Big(\Delta_2 \Delta_1 p^2 + (f_{V_2}(v_2)\Delta_1 + f_{V_1}(v_1)\Delta_2)p + f_{V_1}(v_1) f_{V_2}(v_2) \Big) dv_1 dv_2 \\ & - 12p \iint_{\mathbb{S}} F_V(v_1, v_2) \Big(\Delta_2 \Delta_1 p^2 + (f_{V_2}(v_2)\Delta_1 + f_{V_1}(v_1)\Delta_2)p + f_{V_1}(v_1) f_{V_2}(v_2) \Big) dv_1 dv_2 \\ & + 12 \iint_{\mathbb{S}} F_V(v_1, v_2) \Big(\Delta_2 \Delta_1 p^2 + (f_{V_2}(v_2)\Delta_1 + f_{V_1}(v_1)\Delta_2)p + f_{V_1}(v_1) f_{V_2}(v_2) \Big) dv_1 dv_2 \\ & - 12 \iint_{\mathbb{I}^2} C_V(u_1, u_2) du_1 du_2. \end{aligned}$$

We now group terms in order of their respective powers of the mixing proportion p.

$$
\begin{aligned}
=&\left(12\iint_{\mathbb{S}} F_C(v_1,v_2)\Delta_2\Delta_1 dv_1dv_2 - 12\iint_{\mathbb{S}} F_V(v_1,v_2)\Delta_2\Delta_1 dv_1dv_2\right)p^3\\
&+\Bigg(12\iint_{\mathbb{S}} F_C(v_1,v_2)(f_{V_2}(v_2)\Delta_1 + f_{V_1}(v_1)\Delta_2)dv_1dv_2\\
&\qquad - 12\iint_{\mathbb{S}} F_V(v_1,v_2)(f_{V_2}(v_2)\Delta_1 + f_{V_1}(v_1)\Delta_2)dv_1dv_2\\
&+12\iint_{\mathbb{S}} F_V(v_1,v_2)\Delta_2\Delta_1 dv_1dv_2\Bigg)p^2\\
&+\Bigg(12\iint_{\mathbb{S}} F_C(v_1,v_2)f_{V_1}(v_1)f_{V_2}(v_2)dv_1dv_2\\
&-12\iint_{\mathbb{S}} F_V(v_1,v_2)f_{V_1}(v_1)f_{V_2}v_2)dv_1dv_2\\
&+12\iint_{\mathbb{S}} F_V(v_1,v_2)(f_{V_2}(v_2)\Delta_1 + f_{V_1}(v_1)\Delta_2)dv_1dv_2\Bigg)p\\
&+12\iint_{\mathbb{S}} F_V(v_1,v_2)f_{V_1}(v_1)f_{V_2}(v_2)dv_1dv_2\\
&-12\iint_{\mathbb{I}^2} C_V(u_1,u_2)du_1du_2.
\end{aligned}
$$

Note that the above expression is almost the desired cubic. Thus, it is sufficient to show that

$$12\iint_{\mathbb{S}} F_V(v_1,v_2)f_{V_1}(v_1)f_{V_2}(v_2)dv_1dv_2 = 12\iint_{\mathbb{I}^2} C_V(u_1,u_2)du_1du_2$$

in order to show that our Bias(p) takes the desired form. With this in mind, consider the first integral. First, apply Sklar's Theorem [42] on the distribution function. This yields:

$$
\begin{aligned}
&12\iint_{\mathbb{S}} F_V(v_1,v_2)f_{V_1}(v_1)f_{V_2}(v_2)dv_1dv_2\\
=&12\iint_{\mathbb{S}} C_V(F_{V_1}(v_1),F_{V_2}(v_2))f_{V_1}(v_1)f_{V_2}(v_2)dv_1dv_2. \qquad (3.3)
\end{aligned}
$$

We now perform a bivariate variable substitution of the form:

$$
\begin{aligned}
w_1 = F_{V_1}(v_1), w_2 = F_{V_2}(v_2)\\
\implies v_1 = F_{V_1}^{-1}(w_1), v_2 = F_{V_2}^{-1}(w_2)
\end{aligned}
$$

Note that these are univariate substitutions. Thus, for the Jacobian, off-diagonal terms are zero and the remaining diagonal derivatives are derivatives of inverse functions. Since these functions are non-zero on their common support $\mathbb{S}$, we have by the Inverse Function Theorem [45]:

$$|\mathbb{J}| = \left| \begin{bmatrix} \frac{\partial v_1}{\partial w_1} & 0 \\ 0 & \frac{\partial v_2}{\partial w_2} \end{bmatrix} \right| = \frac{1}{f_{V_1}(F_{V_1}^{-1}(w_1))f_{V_2}(F_{V_2}^{-1}(w_2))}.$$

Thus, (3.3) now becomes:

$$\begin{aligned} &12 \iint_{\mathbb{S}} C_V(F_{V_1}(v_1), F_{V_2}(v_2)) f_{V_1}(v_1) f_{V_2}(v_2) dv_1 dv_2 \\ =&12 \iint_{\mathbb{I}^2} C_V(w_1, w_2) \frac{f_{V_1}(F_{V_1}^{-1}(w_1)) f_{V_2}(F_{V_2}^{-1}(w_2))}{f_{V_1}(F_{V_1}^{-1}(w_1)) f_{V_2}(F_{V_2}^{-1}(w_2))} dw_1 dw_2 \\ =&12 \iint_{\mathbb{I}^2} C_V(w_1, w_2) dw_1 dw_2. \end{aligned}$$

Thus, we have proved our claim. This completes the proof. □

3.3 Mathematics of a Cubic Equation and Practical Scenarios

Now that we have demonstrated that the bias in Spearman's Rho under a bivariate mixture model can be described by a cubic equation of the mixing proportion, of particular interest is the resulting behavior of the bias. In particular, it would be useful to know how varying levels of the mixing proportion changes the level of the bias, for what values of the mixing proportion the bias is positive or negative, and perhaps if changing the mixing proportion can result in our bias going from positive to negative or vice versa. In addition to this, it would be of use to practitioners to know if this resulting behavior could be characterized by the underlying coefficients of our cubic equation. Thus, it is the aim of this section to consider possible cases for our bias that would be interesting for practitioners, while also providing sets of inequalities on the coefficients of the cubic equation that would result in these cases. We will now begin by enumerating possible scenarios of the bias, following this by the mathematical possibilities that would result in these cases.

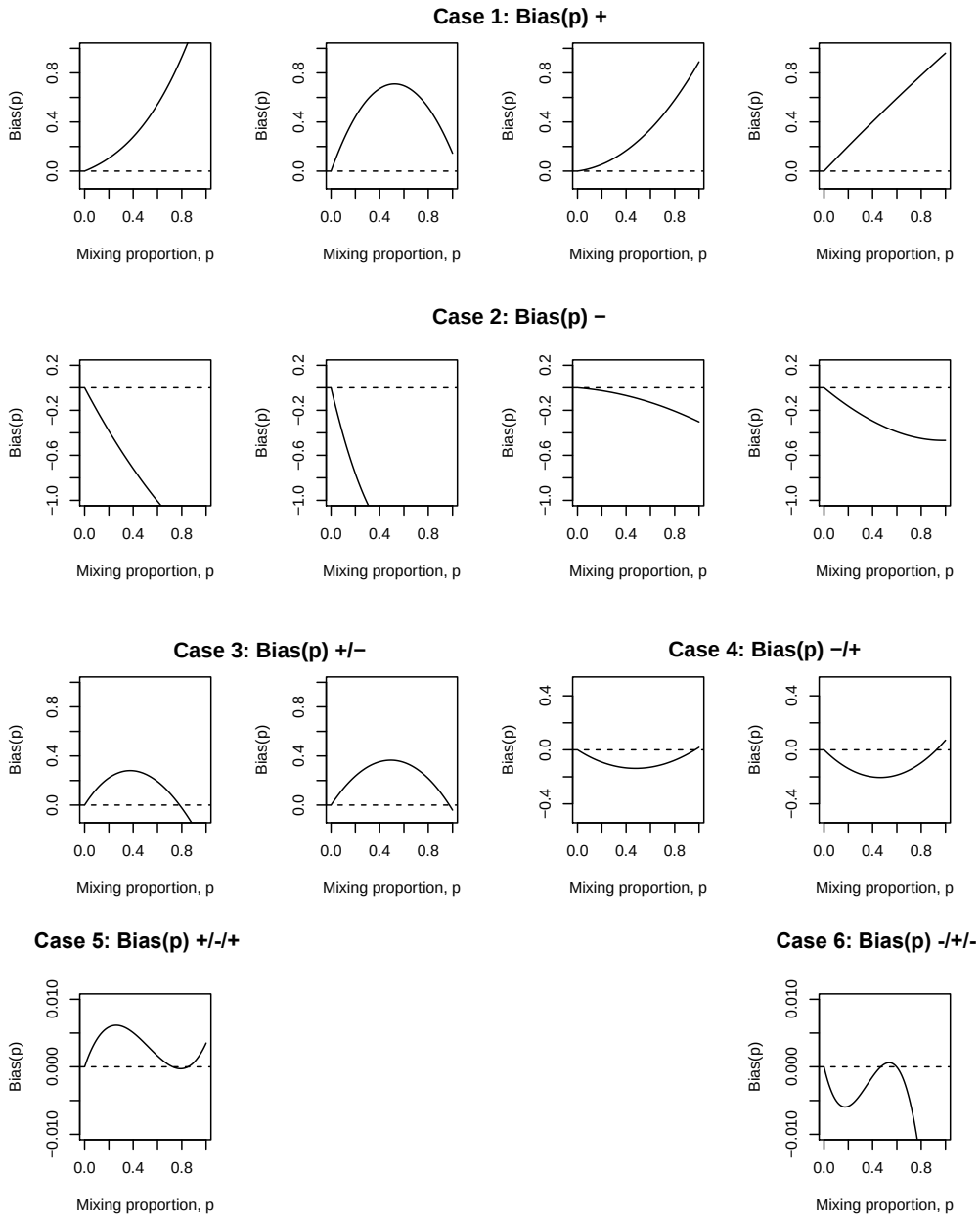

Figure 3.1: Graphs showing the general behavior of cases one through six for $Bias(p) = ap^3 + bp^2 + cp$, where p is the mixing proportion representing the probability of seeing a contaminating response.

3.3.1 Cases 1 & 2: Inflate Always and Deflate Always

Of first interest would be when any change in the mixing proportion does not result in a change in the sign of the bias. Mathematically, this occurs when either

$$\text{Bias}(p) > 0 \qquad \forall p \in (0,1)$$

or when

$$\text{Bias}(p) < 0 \qquad \forall p \in (0,1).$$

Going forward, we shall denote the first of the cases as "Inflate Always" and the latter as "Deflate Always", as they imply that the mixed Spearman's Rho is larger or smaller than the value under the target population V, respectively. Since our bias is a cubic equation, the above two conditions are equivalent to (respectively):

$$ap^3 + bp^2 + cp > 0 \qquad \forall p \in (0,1)$$

and

$$ap^3 + bp^2 + cp < 0 \qquad \forall p \in (0,1).$$

Thus, we can now begin to discuss the mathematical possibilities of a cubic equation that would result in this behavior. First, note that our bias is a cubic function without a constant term, and thus must have a root at zero. Hence, for inflate always and deflate always, there are four mathematical possibilities that can occur for each. These include: complex non-zero roots; one non-zero root greater than one and one non-zero root less than zero; two non-zero roots greater than one; and two non-zero roots that are less than zero. The mathematical possibilities that would result in these cases for inflate always are as follows:

1(a). Complex Roots: $a > 0, |b| < 2\sqrt{ac}, c > 0$

1(b). Two Roots Greater than 1: $a > 0,\ |b| > 2\sqrt{ac},\ b \leq -2a,\ c \geq a,\ a + b + c \geq 0$

1(c). Two Roots Less than 0: $a > 0,\ b > 0,\ c > 0,\ |b| > 2\sqrt{ac}$

1(d). One Positive and One Negative Root: $a < 0,\ c > 0,\ a + b + c \geq 0$

As for deflate always, the mathematic possibilities are as follows:

2(a). Complex Roots: $a < 0, |b| < 2\sqrt{ac}, c < 0$

2(b). Two Roots Greater than 1: $a < 0,\ |b| > 2\sqrt{ac},\ b \geq -2a,\ c \leq a,\ a + b + c \leq 0$

2(c). Two Roots Less than 0: $a < 0,\ b < 0,\ c < 0,\ |b| > 2\sqrt{ac}$

2(d). One Positive and One Negative Root: $a > 0,\ c < 0,\ a + b + c \leq 0$

The derivations of these sets of inequalities with respect to inflate and deflate always can be found in Appendix B.0.1.

3.3.2 Cases 3 & 4: Inflation, then Deflation and Deflation, then Inflation

Now that we have discussed the mathematical possibilities where any change in the mixing proportion would not result in a change in the sign of the bias, we now discuss the case in which there are values of the mixing proportion for which the bias would change from positive to negative, or vice versa. Mathematically, these cases occur when $\exists p_0 \in (0, 1)$ such that either $Bias(p) \geq 0$ for $p \in (0, p_0]$ and $Bias(p) < 0$ for $p \in (p_0, 1)$ or $Bias(p) \leq 0$ for $p \in (0, p_0]$ and $Bias(p) > 0$ for $p \in (p_0, 1)$. We shall denote each of these practical cases as "Inflate/Deflate" and "Deflate/Inflate" going forward, as they imply that the mixed Spearman's Rho is larger, then smaller (or vice versa) than the Spearman's Rho of the target population, V.

First note that, since there is a point p_0 for which the sign of our bias changes, this implies that our mathematical scenarios for each of these cases will involve one root that is in $(0, 1)$. First, we know that since our cubic has one real zero root, that the remaining two roots must both either be real, or complex conjugates of each other. Since we assume

that there is a root in $(0, 1)$, we cannot have complex roots in this case. This significantly reduces the number of possible mathematical scenarios for each of these cases. Thus, for Inflate/Deflate and Deflate/Inflate there are only two mathematical possibilities. These include: One non-zero root in $(0, 1)$ and one non-zero root greater than one; and one non-zero root in $(0, 1)$ and one non-zero root less than zero. The mathematical possibilities that would result in inflate, then deflate are as follows:

3(a). Both Positive Roots: $a > 0,\ b \leq -a,\ c > 0,\ |b| > 2\sqrt{ac},\ a + b + c \leq 0$

3(b). One Positive, One Negative Root: $a < 0,\ c > 0,\ a + b + c \leq 0$

For deflate, then inflate, the mathematical possibilities are:

4(a). Both Positive Roots: $a < 0,\ b \geq -a,\ c < 0,\ |b| > 2\sqrt{ac},\ a + b + c \geq 0$

4(b). One Positive, One Negative Root: $a > 0,\ c < 0,\ a + b + c \geq 0$

The derivations of these sets of inequalities with respect to inflate, then deflate and deflate, then inflate can be found in Appendix B.0.2.

3.3.3 Cases 5 & 6: Inflate, Deflate, then Inflate, and Deflate, Inflate, then Deflate

We now discuss the last two practical possibilities of interest, namely when smaller values of the mixing proportion can result in a change in the sign of the bias, yet larger values of the mixing proportion will result in the same sign for the bias. Mathematically, this occurs when either $\exists p_0, p_1 \in (0, 1)$ such that $Bias(p) \geq 0$ for $p \in (0, p_0]$, $Bias(p) \leq 0$ for $p \in (p_0, p_1]$, and $Bias(p) > 0$ for $p \in (p_1, 1]$ or $Bias(p) \leq 0$ for $p \in (0, p_0]$, $Bias(p) \geq 0$ for $p \in (p_0, p_1]$, and $Bias(p) \leq 0$ for $p \in (p_1, 1]$. We shall denote each of these cases as "Inflate/Deflate/Inflate" and "Deflate/Inflate/Deflate" going forward, as they imply that

the mixed Spearman's Rho is larger, then smaller, then larger (or vice versa) than the value under the target population as one increases the mixing proportion.

First note that since there are points p_0 and p_1 for which the sign of our bias changes, this implies that our cubic has two roots in $(0, 1)$. Thus, there is only one mathematical possibility for each of these two practical scenarios. For inflate, then deflate, then inflate, the mathematical possibility is:

5. Both Roots in $(0, 1)$: $a > 0,\ -2a \leq b \leq 0,\ c \leq a,\ |b| > 2\sqrt{ac},\ a + b + c \geq 0$

As for deflate, then inflate, then deflate, the mathematical possibility is:

6. Both Roots in $(0, 1)$: $a < 0,\ -2a \geq b \geq 0,\ c \geq a,\ |b| > 2\sqrt{ac},\ a + b + c \leq 0$

The derivations of these sets of inequalities is presented in Appendix B.0.3. This completes all mathematical possibilities for a cubic function as described by our bias.

3.4 Existence of Mathematical Possibilities

Of particular interest to the practitioner is if each of the mathematical possibilities described in the previous section can actually exist for the given expressions of a, b, and c. In addition to this, given the mathematical complexity of the expressions for a, b, and c, it would also be of interest to the practitioner that if these mathematical possibilities do exist, if they can also be described by sets of means, variances, and covariances on the underlying distributions. If this were the case, a practitioner could then estimate these quantities to establish the possibility of a particular form of the bias in their data set. The answer to these questions is in the affirmative, as we shall demonstrate in this section through simulation results.

In order to demonstrate that each of these cases exist, we begin by randomly generating mean vectors and covariance matrices for the valid and contaminating classes V and C, respectively. The mean vectors are generated component wise from a discrete uniform

distribution on the interval $(-5, 5)$. For the covariance matrices, we first generate variances from a continuous uniform distribution on the interval $(1, 3)$, then generate the covariance for this matrix by sampling from a uniform distribution where the lower and upper bounds are derived from the Cauchy-Schwarz Inequality [25]. Finally, once we have generated these variances and covariances, we check to determine that the resulting matrix is positive semidefinite, to ensure that it is a valid covariance matrix.

Once we have generated the mean vectors and covariance matrices for the valid and contaminating classes, we use these as parameters in the multivariate normal cumulative distribution functions and univariate normal density functions. We then use these functions to calculate, via numerical integration, the values of a, b, and c using the expressions derived in Theorem 3.2. We then determine if these calculated values of a, b, and c satisfy one of the given sets of inequalities presented in Section 3.3. If these values do not satisfy the given set of inequalities, the process starts over again, until the set of inequalities is satisfied. We performed this criteria for each of the 14 cases presented in Section 3.3, and generated sets of mean vectors and covariance matrices that will generate each of the 14 mathematical possibilities. The sets of mean vectors and covariance matrices that generate the 14 cases is presented above, where Cases 1(a) through 1(d) correspond to mathematical possibilities for Inflate Always, Cases 2(a) through 2(d) to Deflate Always, Cases 3(a) and 3(b) to Inflate/Deflate, Cases 4(a) and 4(b) to Deflate/Inflate, Case 5 to Inflate/Deflate/Inflate, and Case 6 to Deflate/Inflate/Deflate.

Of additional interest would be to determine the sampling behavior of Spearman's Rho under the bivariate mixture model relative to the theoretical derivations provided in Theorem 3.2. To this end, we partition values of p into increments of .01. For each value of p, we generate 10,000 observations for Cases 1 through 4, 1,000,000 observations for Case 5, and 100,000 observations for Case 6, each of which is from either a multivariate normal distribution with parameters from the valid or contaminating classes depending on whether

Case	μ_V	μ_C	σ_V	σ_C	$\sigma^2_{V_1}$	$\sigma^2_{C_1}$	$\sigma^2_{V_2}$	$\sigma^2_{C_2}$
1(a)	(5,3)	(-1,5)	-2.010	1.240	1.880	2.850	2.180	2.270
1(b)	(-5,-1)	(0,4)	0.092	0.459	2.913	2.646	2.186	2.246
1(c)	(1,-3)	(0, 0)	-1.000	0.362	2.170	1.555	1.200	1.594
1(d)	(-2,-4)	(0,-4)	-1.460	0.651	2.010	1.629	2.530	2.170
2(a)	(-5,5)	(-4,3)	1.260	-2.390	1.920	2.790	2.960	2.210
2(b)	(3,-5)	(-1,-2)	1.890	-0.365	2.360	2.926	1.740	2.938
2(c)	(-5,1)	(4,0)	-1.140	-1.790	2.950	2.830	1.050	1.340
2(d)	(1,-5)	(-5,-5)	1.270	0.028	2.290	1.852	2.900	1.680
3(a)	(-4,3)	(0,3)	-1.060	-1.550	1.690	1.020	1.020	2.390
3(b)	(-1,3)	(2,4)	-1.060	-0.647	2.840	1.172	2.610	1.949
4(a)	(-2,-1)	(-1,-4)	-0.525	-0.507	2.743	1.714	1.523	2.626
4(b)	(-1,1)	(-4,2)	-0.091	0.057	1.449	2.707	2.029	1.993
5	(0,-2)	(-2,4)	-1.540	-1.470	2.150	2.860	1.630	1.190
6	(5,-4)	(-4,-2)	-2.130	-1.570	2.120	2.750	2.840	1.120

the given value of p is greater than or less than an observation from a uniform distribution on $(0, 1)$. We then compute Spearman's Rho for this mixed sample. Following this, we generate an equivalent number of observations from a multivariate normal distribution for only the valid class for each case, and calculate Spearman's Rho. The resulting difference of Spearman's Rho then serves as an estimate for our expression in Definition 3 for each value of p.

After completing the above simulations, we then plot these against the plots of Bias(p) derived by our estimates of a, b, and c for each of the 14 mathematical possibilities. A table of these graphs is presented below. Note that the estimated values of the bias for each p closely match those of the calculated analytical value at each p. This completes Chapter 3.

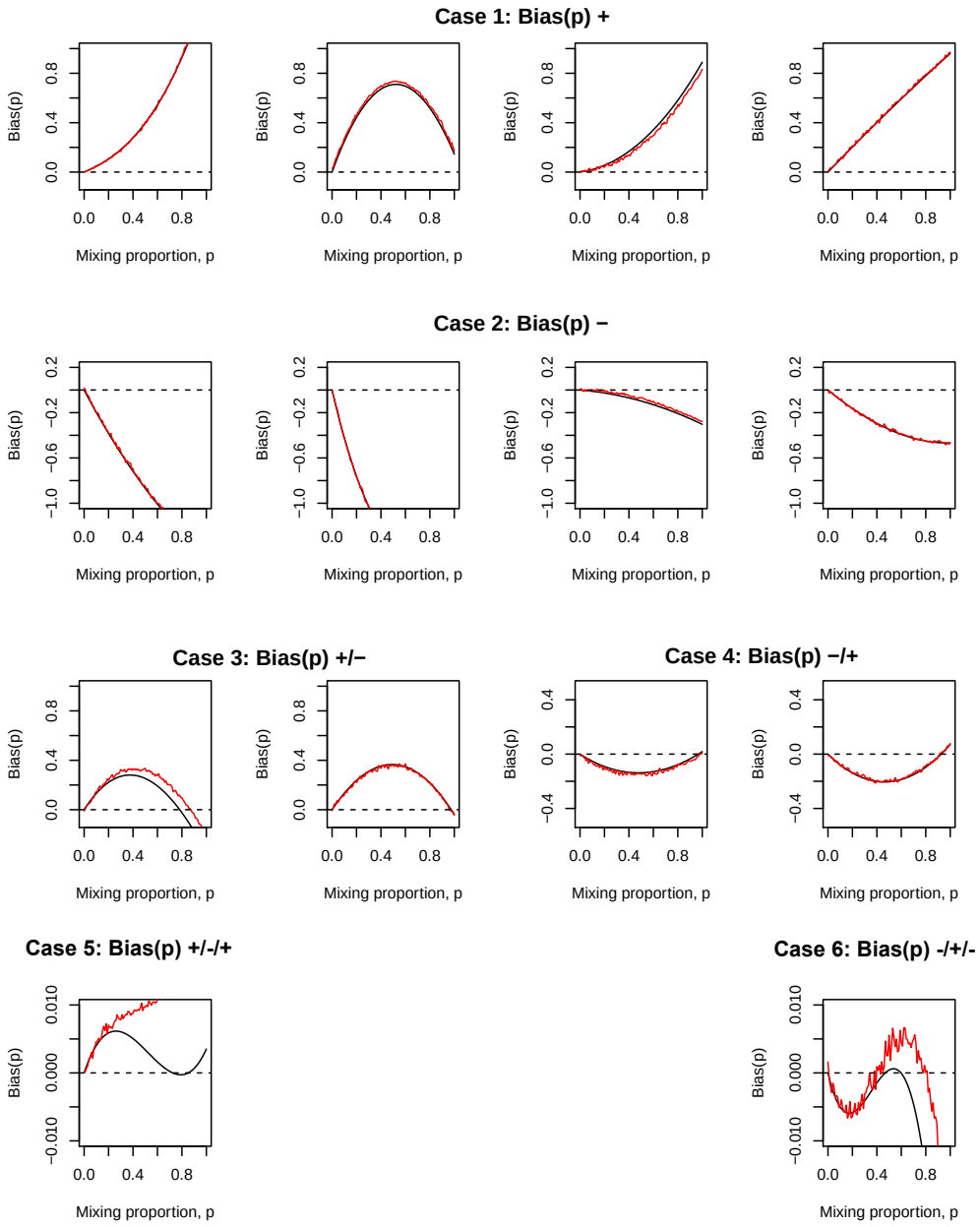

Figure 3.2: Graphs showing the general behavior of cases one through six for Bias$(p) = ap^3 + bp^2 + cp$, where p is the mixing proportion representing the probability of seeing a contaminating response. In red, simulated estimates of the bias are plotted.

REFERENCES

[1] Osborne, J.W. Best Practices in Data Cleaning; SAGE Publications: Newcastle upon Tyne, UK, 2012.

[2] Huang, J.L.; Curran, P.G.; Keeney, J.; Poposki, E.M.; DeShon, R.P. Detecting and Deterring Insufficient Effort Responding to Surveys. J. Bus. Psychol. 2012, 27, 99114.

[3] McGrath, R.E.; Mitchell, M.; Kim, B.H.; Hough, L. Evidence for response bias as a source of error variance in applied assessment. Psychol. Bull. 2010, 136, 450470.

[4] Huang, J.L.; Liu, M.; Bowling, N.A. Insufficient effort responding: examining an insidious confound in survey data. J. Appl. Psychol. 2015, 100, 828845.

[5] DeSimone, J.A.; DeSimone, A.J.; Harms, P.D.; Wood, D. The differential impacts of two forms of insufficient effort responding. Appl. Psychol. 2018, 67, 309338.

[6] Cred, M. Random responding as a threat to the validity of effect size estimates in correlational research. Educ. Psychol. Meas. 2010, 70, 596 612.

[7] Holtzman, N.S.; Donnellan, M.B. A simulator of the degree to which random responding leads to biases in the correlations between two individual differences. Pers. Individ. Differ. 2017, 114, 187 192.

[8] Kam, C.C.S.; Meyer, J.P. How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. Organ. Res. Methods 2015, 18, 512541.

[9] Schmitt, N.; Stuits, D.M. Factors defined by negatively keyed items: the result of careless respondents? Appl. Psychol. Meas. 1985, 9, 367373.

[10] Cronbach, L.J. Coefficient Alpha and the Internal Structure of Tests. Psychometrika 1951, 16, 293334.

[11] Fong, D.Y.; Ho, S.Y.; Lam, T.H. Evaluation of internal reliability in the presence of inconsistent responses. Health Qual. Life Outcomes 2010, 8, 110.

[12] Sijtsma, K. On the use, the misuse, and the very limited usefulness of Cronbachs alpha. Psychometrika 2009, 74, 107120.

[13] McNeish, D.M. Thanks Coefficient Alpha, Well Take It From Here. Psychol. Methods 2017, 23, 412433.

[14] Wertheimer, M.E. Identifying the types of insufficient effort responders. Masters Thesis, Middle Tennessee State University, Murfreesboro, TN, USA, 2017.

[15] Attali, Y. Reliability of speeded number-right multiple-choice tests. Appl. Psychol. Meas. 2005, 29, 357368.

[16] Waller, N.G. Commingled samples: A neglected source of bias in reliability analysis. Appl. Psychol. Meas. 2008, 32, 211223.

[17] Kuder, G.F.; Richardson, M.W. The theory of the estimation of test reliability. Psychometrika 1937, 2, 151160.

[18] Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika 1941, 6, 153160.

[19] Guttman, L. A basis for analyzing test-retest reliability. Psychometrika 1945, 10, 255282.

[20] Lord, F.M.; Novick, M.R. Statistical Theories of Mental Test Scores; Addison-Wesley: Reading, MA, USA, 1968.

[21] Pinsoneault, T.B. Detecting random, partially random, and nonrandom Minnesota Multiphasic Personality Inventory-2 protocols. Psychol. Assess. 2007, 19, 159164.

[22] Berry, D.T.R.; Wetter, M.W.; Baer, R.A.; Larsen, L.; Clark, C.; Monroe, K. MMPI-2 random responding indices: Validation using a self-report methodology. Psychol. Assess. 1992, 4, 340345. [CrossRef]

[23] Meehl, P.E. Psychodiagnosis: Selected Papers; University of Minnesota Press: Minneapolis, MN, USA, 1973; pp. 200224.

[24] Waller, N.G.; Meehl, P.E. Multivariate Taxometric Procedures: Distinguishing Types from Continua; Sage: Thousand Oaks, CA, USA, 1998.

[25] Kreyszig, E. Introductory Functional Analysis with Applications; Wiley: Hoboken, NJ, USA, 1989.

[26] Revelle, W. Psych: Procedures for Psychological, Psychometric, and Personality Research; R Package Version 1.8.10; Northwestern University: Evanston, IL, USA, 2018.

[27] R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2017.

[28] RStudio, Inc. Easy Web Applications in R., 2013. Available online: http://www.rstudio.com/shiny/ (accessed on 3 December 2018)

[29] Meijer, R.R.; Sijtsma, K. Methodology Review: Evaluating Person Fit. Appl. Psychol. Meas. 2001, 25, 107135.

[30] Felt, J.M.; Castenada, R.; Tiemensma, J.; Depaoli, S. Using person fit statistics to detect outliers in survey research. Front. Psychol. 2017, 8.

[31] Embretson, S.E.; Reise, S.P. Item Response Theory for Psychologists; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2000.

[32] Curran, P.G. Methods for the detection of carelessly invalid responses in survey data. J. Exp. Soc. Psychol. 2016, 66, 419.

[33] Meade, A.W.; Craig, S.B. Identifying careless responses in survey data. Psychol. Methods 2012, 17, 437455.

[34] Johnson, J.A. Ascertaining the validity of individual protocols from web-based personality inventories. J. Res. Personal. 2005, 39, 103129.

[35] DeSimone, J.A.; Harms, P.D.; DeSimone, A.J. Best practice recommendations for data screening. J. Organ. Behav. 2015, 36, 171181.

[36] Revelle, W. Hierarchical cluster analysis and the internal structure of tests. Multivar. Behav. Res. 1979, 14, 5774.

[37] McDonald, R.P. Test Theory: A Unified Treatment; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1999.

[38] Casella, G., Berger, R. L. (2017). Statistical inference. Belmont, CA: Brooks/COle Cengage Learning.

[39] Spearman, C. (1904). The Proof and Measurement of Association between Two Things. The American Journal of Psychology, 15(1), 72. doi:10.2307/1412159

[40] Kruskal, WH. (1958). Ordinal measures of association. J. Amer. Statist. Assoc. 53: 814-861.

[41] Lehmann, EL. (1966). Some concepts of dependence. Ann. Math. Statist. 37: 1137-1153.

[42] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris. 8: 229-231.

[43] Carden, S., Camper, T., Holtzman, N. (2018). Cronbachs Alpha under Insufficient Effort Responding: An Analytic Approach. Stats, 2(1), 1-14. doi:10.3390/stats2010001

[44] Nelsen, R. B. (2006). An introduction to copulas. New York: Springer.

[45] Rudin, Walter (1976). Principles of mathematical analysis. International Series in Pure and Applied Mathematics (Third ed.). New York: McGraw-Hill Book Co. pp. 221223.

Appendix A

APPENDIX TO CHAPTER 2

A.1 Proof of Lemma

Proof. Begin by applying the Cauchy-Schwarz Inequality to covariances:

$$\begin{aligned} \text{cov}(V_i, V_j) &\leq \sqrt{\text{var}(V_i)\text{var}(V_j)} \\ \implies \sum_{i \neq j}\sum \text{cov}(V_i, V_j) &\leq \sum_{i \neq j}\sum \sqrt{\text{var}(V_i)\text{var}(V_j)} \\ &\leq \sum_{i \neq j}\sum \frac{1}{2}(\text{var}(V_i) + \text{var}(V_j)) \end{aligned}$$

where the last line invokes the arithmetic-geometric mean inequality. Thus,

$$\begin{aligned} \implies \sum_{i \neq j}\sum \text{cov}(V_i, V_j) &\leq \frac{1}{2}\left(2(k-1)\sum_{i=1}^{k} \text{var}(V_i)\right) \\ \implies \frac{1}{k-1}\sum_{i \neq j}\sum \text{cov}(V_i, V_j) &\leq \sum_{i=1}^{k} \text{var}(V_i) \\ \implies \overline{\sigma_{ijV}} &\leq \overline{\sigma^2_{iV}}. \end{aligned}$$

□

Appendix B

APPENDIX TO CHAPTER 3

B.0.1 MATHEMATICS OF CASES 1& 2

First, consider Inflate Always. For this case, none of the roots of our cubic equation must occur in $(0, 1)$. In addition to this, since our cubic equation does not have a constant term, we know that there will be a root at zero. Thus, to necessitate the Bias being strictly positive on $(0, 1)$, the Bias must be increasing away from zero. Thus,

$$0 < [\text{Bias}(p)]'\Big|_0 = (3ap^2 + 2bp + c)\Big|_0 = c.$$

Thus, the first condition to necessitate Inflate Always is that $c > 0$.

In order to proceed, we must now consider the placement of the two nonzero roots with respect to $(0, 1)$. The first possible scenario is that our cubic equation has only a singular root in $\mathbb{R}$, and thus it's two nonzero roots take values in $\mathbb{C}$. Since our Bias is a cubic equation without a constant term, the two nonzero roots are those of the quadratic variety. Thus, these two roots are complex only when the discriminant is negative, or when

$$b^2 - 4ac < 0 \iff |b| < 2\sqrt{ac}.$$

In addition to this, for the cubic to be strictly positive in $(0, 1)$ with two complex roots, $\text{Bias}(p) \to \infty$ as $p \to \infty$. This only occurs when $a > 0$. Thus, the first mathematical possibility for Inflate Always can be characterized by the following set of inequalities:

$$a > 0, \quad c > 0, \quad |b| < 2\sqrt{ac}.$$

The second possible mathematical scenario is when both of our nonzero roots are real numbers and strictly greater than or equal to one. For this, we now choose to represent our cubic in a different manner. Let r_1 and r_2 represent our nonzero roots. By the Fundamental

Theorem of Algebra, we have that our cubic takes the form:

$$\begin{aligned}\text{Bias}(p) =& ap(p-r_1)(p-r_2)\\ =& ap^3 - a(r_1+r_2)p^2 + ar_1r_2p\\ \implies& b = -a(r_1+r_2), \quad c = ar_1r_2\end{aligned}$$

Thus, we can discuss bounding the coefficients of our cubic by discussing the placement of the two nonzero roots. First note that for the Bias to be positive in $(0,1)$ when the two nonzero roots are greater than 1, we must have thatBias$(p) \to \infty$ as $p \to \infty$. This only occurs when $a > 0$. Now, if $r_1, r_2 \geq 1$, we have that:

$$b = -a(r_1+r_2) \leq -a(1+1) = -2a$$

and

$$c = ar_1r_2 \geq a(1)(1) = a.$$

In addition to this, the above inequalities do not preserve the positivity of the cubic at $p = 1$. In order to preserve this, we add the following inequality:

$$\text{Bias}(1) = a + b + c \geq 0.$$

Finally, to ensure that r_1 and r_2 are real, we ensure that the discriminant is real, or:

$$|b| > 2\sqrt{ac}.$$

Thus, the second mathematical possibility for Inflate Always can be characterized by the following set of inequalities:

$$a > 0, \quad c \geq a, \quad a+b+c \geq 0, \quad |b| > 2\sqrt{ac}.$$

The third mathematical possibility is when our two nonzero roots are strictly less than zero. First, this implies that there are no roots in $\mathbb{R}^+$, and thus for the cubic to be positive in $(0,1)$,

we must have Bias$(p) \to \infty$ as $p \to \infty$, and thus $a > 0$. Now, using the expressions for b and c provided above by the Fundamental Theorem of Algebra, we have that for $r_1, r_2 < 0$:

$$b = -a(r_1 + r_2) > -a(0 + 0) = 0$$

and

$$c = ar_1r_2 = a(-|r_1|)(-|r_2|) = a|r_1||r_2| > 0.$$

Finally, to ensure that r_1 and r_2 are real, we include that $|b| > 2\sqrt{ac}$. Thus, the set of inequalities that generates the third mathematical possibility for Inflate Always is:

$$a > 0, \quad b > 0, \quad c > 0, \quad |b| > 2\sqrt{ac}.$$

The final mathematical possibility for Inflate Always is when one of our nonzero roots is negative, while the other one is greater than or equal to 1. First, note that for our Bias to be positive in $(0, 1)$ and for there to be a root past 1, we must have that Bias$(p) \to -\infty$ as $p \to \infty$, and thus $a < 0$. Using the expression for c provided by the Fundamental Theorem of Algebra, and without loss of generality suppose $r_1 < 0$, we get that:

$$c = ar_1r_2 = -|a|(-|r_1|)r_2 = |a||r_1|r_2 > 0.$$

However, we do not get a similar inequality on b, as the sign of b could be positive or negative, depending on the maginitude of the positive and negative roots with respect to each other. In addition to this, we must ensure that our Bias is non-negative at 1, and thus $a + b + c \geq 0$. Finally, to ensure that our roots are real numbers, we must ensure that our discriminant is non-negative, or:

$$b^2 - 4ac \geq 0 \iff b^2 > 4ac \iff c > \frac{b^2}{4a}.$$

Thus, the set of inequalities that generates the final mathematical possibility for Inflate always is as follows:

$$a < 0, \quad c > 0, \quad a + b + c \geq 0, \quad c > \frac{b^2}{4a}.$$

For the case where Bias$(p) < 0 \; \forall p \in (0,1)$, it is sufficient to consider the function $g(p) = -\text{Bias}(p)$. Then performing a similar analysis as above, will result in different signs to each of the above sets of inequalities.

B.0.2 MATHEMATICS OF CASES 3& 4

The first mathematical scenario that would result in Inflate/Deflate is when the second nonzero root is greater than or equal to one. First note that for the mathematical conditions of Inflate/Deflate to occur with one root being greater than or equal to one, we must have that Bias$(p) \to \infty$ as $p \to \infty$. Thus, we must have that $a > 0$. Going forward, assume that $0 < r_1 < 1$. Using the same formulation for b and c as in the previous section, we have that

$$b = -a(r_1 + r_2) \leq -a(0 + 1) = -a$$

and

$$c = ar_1r_2 > 0.$$

Furthermore, to ensure that deflation occurs for all values of p to the right of r_1, we must ensure that our Bias is non-positive at $p = 1$. Thus, we include the inequality $a + b + c \leq 0$. Finally, to ensure that both roots are real valued, we include the following inequality: $|b| > \sqrt{ac}$. Thus, the set of inequalities that generates the first mathematical possibility for Inflate/Deflate are as follows:

$$a > 0, \quad c > 0, \quad b \leq -a, \quad a + b + c \leq 0, \quad |b| > \sqrt{ac}.$$

The second mathematical scenario that would result in Inflate/Deflate is when the second nonzero root is negative. First note that under this scenario that Bias$(p) \to -\infty$ as $p \to \infty$, and thus $a < 0$. Using the expression for c from the previous section, we have that

$$c = ar_1r_2 = -|a|r_1(-|r_2|) = |a|r_1|r_2| > 0.$$

However, there does not need to be a bound on b, as the magnitude and sign of b depends on the magnitude of the positive and negative roots with respect to each other. In addition to this, we must ensure that the negativity of the Bias holds at $p = 1$. Thus, the following inequality is included: $a + b + c \leq 0$. Finally, to ensure that both roots take real values, we must ensure that the discriminant is positive, or:

$$b^2 - 4ac > 0 \iff b^2 > 4ac \iff c > \frac{b^2}{4a}.$$

Thus, the set of inequalities that generates the second mathematical possibility for Inflate/Deflate is as follows:

$$a < 0, \quad c > 0, \quad a + b + c \leq 0, \quad c > \frac{b^2}{4a}.$$

For the case of Deflation/Inflation, it is sufficient to consider the function $g(p) = -\text{Bias}(p)$. Then, performing a similar analysis as above will result in different signs to each of the above sets of inequalities.

B.0.3 Mathematics of Cases 5& 6

The only mathematical scenario that would result in Inflate/Deflate/Inflate is when both nonzero roots are strictly positive, and less than one. First note that for the mathematical conditions of this root behavior, we must have that $\text{Bias}(p) \to \infty$ as $p \to \infty$. Thus, we must have that $a > 0$. Going forward, assume that $0 < r_1, r_2 < 1$. Using the same formulations for b and c as in section B.0.1, we have that

$$b = -a(r_1 + r_2) > -a(1 + 1) = -2a$$

and

$$b = -a(r_1 + r_2) < -a(0 + 0) = 0$$

. Thus, we have that $-2a < b < 0$. For c, we have

$$c = ar_1r_2 < a(1)(1) = a$$

and

$$c = ar_1r_2 > a(0)(0) = 0$$

. Thus, we have that $0 < c < a$. In addition to this, to ensure that inflation occurs for all values of p to the right of r_2 (assuming $r_2 \geq r_1$), we include the following inequality: $a + b + c \geq 0$. Finally, to ensure that our two nonzero roots are real, we ensure that the discriminant is nonnegative, or when $|b| > 2\sqrt{ac}$. Thus, the set of inequalities that generates the mathematical possibility for Inflate/Deflate/Inflate is as follows:

$$a > 0, \quad 0 < c < a, \quad -2a < b < 0, \quad a + b + c \leq 0, \quad |b| > \sqrt{ac}.$$

For the case of Deflation/Inflation/Deflation, it is sufficient to consider the function $g(p) = -\text{Bias}(p)$. Then, performing a similar analysis as above will result in different signs to each of the above sets of inequalities.