Imaginal Exposure Scripts

Howard Hu

August 13, 2022

1 Exclusive Summary

This report aims to analyze the imaginal exposure scripts data given by Dr. Rachel Butler. We use McNemar's test to determine the reliability of multiple coders on the same categorical data. For category 1a, we find out all of the coders agree with the choice of coder 1. Therefore, coder 1 has good reliability on category 1a in time point 1. However, the p-value of coder 6 and coder 8 is 0.02, indicating they disagree on this category variable. By removing the least reliable coder 8, the reliability of category 1a is 69.8%. Moreover, since there is multiple sadistically significant p-value in our McNemar's test, we cannot choose one coder as our standard. Therefore, we would like to use a combination of coders to reduce the bias when calculating the reliability.

2 Introduction

2.1 General Background

According to the discussion with Dr. Rachel Butler, only 30% to 50% of people can recover their eating disorders from evidence-based treatment. Exposure therapy is one of the treatments that cure eating disorders. In imaginal therapy, people develop a script and repeatedly read/visualize the situation. It will help them to know which potential subject causes their eating disorders. In our study, we conducted four online sessions with each participant. A sizeable open trial of online imaginal exposure was conducted based on a parent study. Trained coders analyzed a total of 47 items to identify fears related to eating disorders. Specifically, the goal of this treatment is to determine the presence or absence of specific fears associated with eating disorders. Moreover, multiple coders may rate the same scripts. As a result, scripts had a different combination of coders, which didn't have the same coders.q

2.2 Objectives

The main question of our client is to calculate the reliability of coding. What is the best method?. Then the second question is, is it better to calculate the frequency of each item separately for every time point in the scripts? Is it better to use one coder's ratings or a combination of coders when calculating a frequency?

2.3 Exploratory Data Analysis

Our data is given by Dr. Rachel Butler, which contains 3543×128 subjects. We focus on the 47 binary variables, which means whether the specific topic is mentioned in the scripts. However, the category 45 variable is not a binary variable. Then we would like to drop that column and rename variables 46 and 47 to variables 45 and 46. Therefore, we only keep patient ID, coder name, and 46 binary variables in our data set.

Firstly, we would like to know the distribution of all 46 binary variables using the histogram. The result is shown in the following figure.

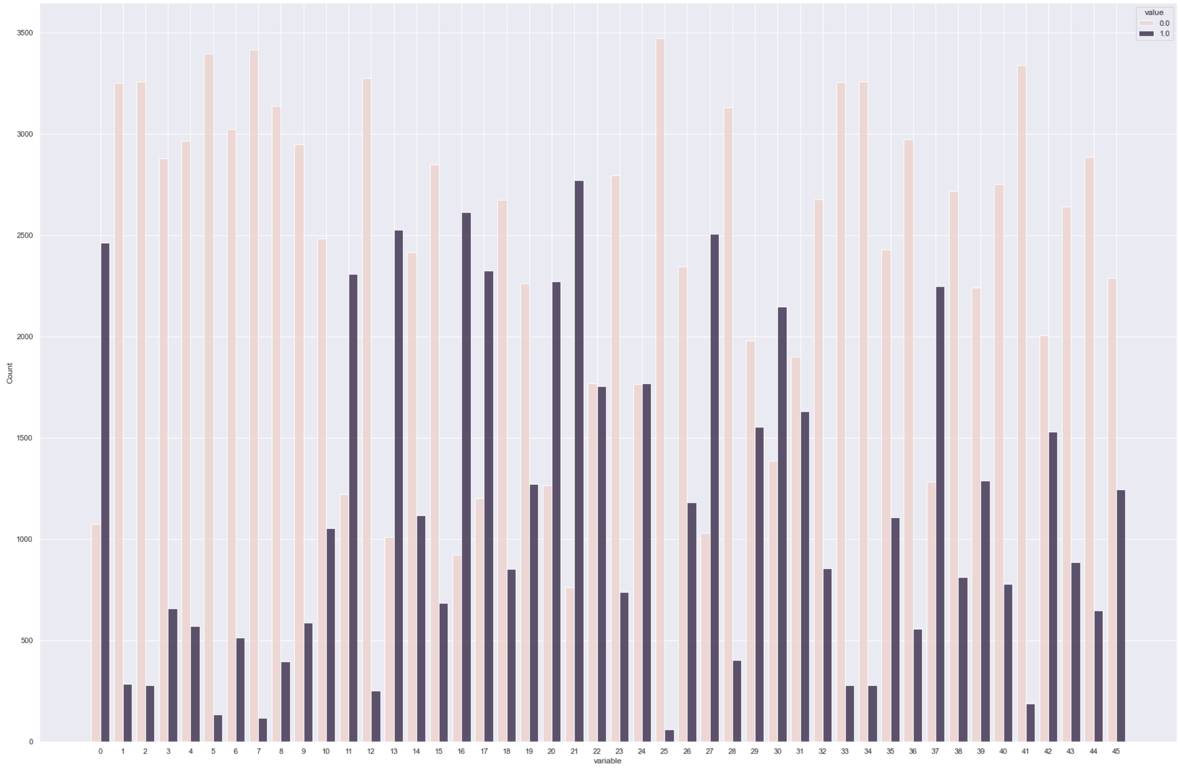

Figure 1: The distribution of all 46 binary variables

From the previous figure, we find that much more coders tend to grade 0 in the 25th categorical variable. According to the given coding template by Dr. Rachel Butler, we know the 25th categorical variable means "Eating More Food Than You Can Burn Via Exercise Mentioned." It shows that most patient coders think they did not eat excess food. Therefore, eating disorders may still be a problem for those patients.

Secondly, we would like to focus on each coder to see how they give grades for each binary variable. We drop the NA values of our data set because we have many observations. Then we average a mean grade of 46 variables for each coder and then build box plots. We only have 16 coders in our data set, and there are no coder 2, coder 9, and coder 14. Finally, We will have a 16 data set which means 16 coders and 46 category variables. The result is shown in the following figure.

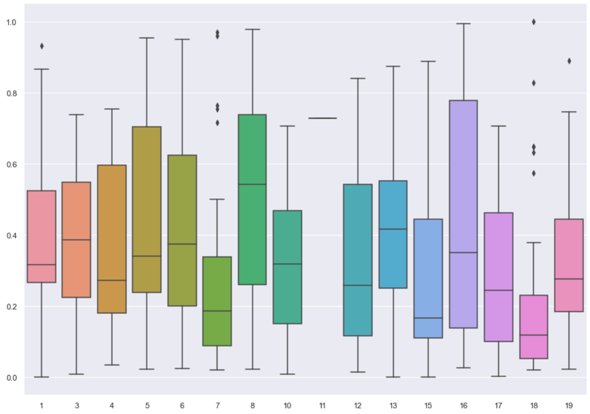

Figure 2: The distribution of all 16 coders

From the previous figure, we find out there is only one bar for code 11. It means the coder 11 only grade category 1a. Therefore, we definitely should not consider coder 11 when we calculate the reliability of the coding. Moreover, we find out there are five outliers for coder 18. It means coder 18 will tend to give grade 1 on some category variable. For example, for category 22a, the mean grade is 1. Therefore, for any patients' script, coder 18 will grade it as 1. Consequently, we should not consider coder 18 as a reliable coder.

3 Approach

This report aims to find a suitable way to calculate coding reliability. We want to use McNemar's test to determine the frequency of multiple coders giving the same grade on a category variable. If multiple coders provide the same rate on a category variable, we can consider those coders have good reliability on this category variable. Firstly, we would like to do the McNemar's test on our category 1a on time point 1. The result is shown in the following table.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		1.00	0.25	0.25	1.00	1.00	1.00	0.62	1.00	0.37	0.48	1.00	0.13	0.37	0.48	1.00
2			1.00	1.00	0.48	0.13		0.62	1.00	1.00	1.00	1.00	0.45	0.68	1.00	1.00
3				1.00	1.00	0.25	1.00		1.00	1.00	1.00	0.48	0.48		0.48	
4					0.48	0.13	1.00	1.00	0.48	0.62		0.13	1.00	1.00	0.13	
5						1.00		0.48	1.00	1.00		1.00	0.48	0.48	1.00	
6							1.00	0.02	0.62	0.13	0.48	1.00	0.13	0.13	0.13	
7								0.13	0.48	0.62	1.00	1.00	0.48	0.48	0.62	
8									0.00	0.00	1.00	0.00	0.00	0.50	0.00	1.00
9										1.00	1.00	0.68	0.23	0.42	1.00	1.00
10											1.00	0.04	0.04	0.34	1.00	
11												0.48	1.00	1.00	0.48	
12													0.00	0.07	0.15	1.00
13														0.55	0.02	1.00
14															0.58	1.00
15																1.00
16																

Null hypothesis H(0): The row and column marginal frequencies are equal (It means two coders agree on this category variable)
Alternative hypothesis H(1): The row and column marginal frequencies are not equal (It means two coders do not agree on this category variable)

From the previous table, we find out that no p-value of coder 1 is lower than our statistically significant level of 0.05. It means the rest of the 15 coders all agree with coder 1. However, the p-value for coder 6 and coder 8 is statistically significant,0.02, meaning coder 6 and coder 8 do not agree. We also can find the p-value of the following pairs are statistically significant ((6,8),(8,9),(8,10),(10,12),(8,13), (10,13),(12,13),(8,15),(13,15)). There is a lot of disagreement for coder 8, which is 60% of whole pairs. Therefore, for calculating the reliability of category 1a, we should not consider coder 8. Moreover, we should not only choose one coder because if we only choose coder 10, coder 13 and 15 will disagree with his choice. Therefore, we need to use a combination of coders to calculate the reliability of category 1a. Then we build the following equation to calculate the reliability(Frequency) of grading category 1a for time point 1.

$$\text{Reliability(Frequency) of grading category i for time j} = \frac{\sum coder's grade}{number of subjects}$$

hold for every i in$1 \leq i \leq 46$
hold for every j in$1 \leq j \leq 4$
coder's grade does not contain the least reliable coder

If we choose coder 1,2,3,4,5,6,7,9,10,11,12,13,14,15 and 16, the frequency of grading category 1a as 1 is 69.8%. We could apply this method on all 4 time points as long as 46 different category variables.

4 Conclusion

The purpose of this report is to analyze the imaginal exposure scripts given by Dr. Rachel Butler. To determine the reliability of multiple coders on the same categorical data, we use McNemar's test. All of the coders agree with the choice of coder 1 for category 1a. Therefore, coder 1 has good reliability on category 1a at time point 1. Coders 6 and 8 disagree on this category variable since their p-values are 0.02. Therefore, category 1a has 69.8% reliability after removing the least reliable coder 8. Furthermore, since our McNemar's test has multiple sadistically significant p-values, we cannot choose one coder as our standard. To reduce the bias when calculating reliability, we would like to use a combination of coders.

5 Appendix