SUBJECTIVE JUDGEMENTS OF RANDOMNESS AND THE SIGNIFICANCE OF THE SUBJECTIVE THRESHOLD

BRADLEY C. SMITH

A Thesis
Submitted to the Department of Psychology
and the Department of Mathematics & Computer Science
of the University of Lethbridge
in Partial Fulfillment of the
Requirements for the Degree

BACHELOR OF SCIENCE

Department of Psychology
Department of Mathematics & Computer Science
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

To my Mother,
for always telling me that I can accomplish anything I put my mind to.

Abstract

Amidst the controversy surrounding the use of the 0.05 threshold as a standard for statistical significance, one common complaint is that it is arbitrary. Is it? Or is 0.05 a reasonable approximation for the threshold at which participants and investigators would naturally attribute results to a cause, rather than to random chance? Participants were shown series of simulated coin flips or spinner spins with varying degrees of probability and then were asked to judge whether the coin or spinner series was fair. The results suggest that their subjective threshold is reasonably close to the 0.05 criterion. Additional research investigated whether there was a bias to attribute non-randomness to clustering. Results showed a significant clustering bias, but one that was less influential in decision-making than previously reported. Finally, a redundancy analysis based on information theory showed a significant effect only in conditions where the coin proportions were approximately even.

Acknowledgements

Thank you to my supervisor Dr. John Vokey for originally telling me this experiment wouldn't work so that I had more motivation to get it going. Also for being willing to let me barge into his office unannounced when I needed help. To Dr. Scott Allen for talking through questions with me and being the sanity check on my crazy ideas. To my second reader, Dr. John Sheriff, for consulting with me on my statistics and being open to an honours thesis such as this. And to my friends in the Vallen lab for grinding through my confusing presentations and explanations.

This work was presented at the University of Lethbridge, as a course requirement for completing an honours thesis. Portions of this work were also presented as a talk called "What makes 0.05 significant?" at the P-PACT 2016[1] conference. Portions of this work will be presented as a poster at the BASICS 2017[2] conference and CSBBCS 2017[3] conference.

[1]Prairie Perception Action Cognition Team, Canmore, Alberta November 2016
[2]Banff Annual Seminar in Cognitive Science, Banff, Alberta, May 2017
[3]Canadian Society for Brain, Behaviour, and Cognitive Sciences, Regina, Saskatchewan, June 2017

Contents

List of Figures

List of Tables

Chapter 1

An Abitrary Introduction

1.1 The 0.05 Threshold

"There are no routine statistical questions, only questionable statistical routines." -D.R. Cox (Chatfield, 1991)

It is generally accepted, with notable grumbling, that it is important to have some near universal standards to use as a measuring stick when reporting research (e.g., Bross, 1971). However, along with almost any universal standard there follows heated debate about that standard. Statistical significance testing, especially at $\alpha = 0.05$, has become an important measure of reliability in scientific research across a wide range of disciplines. Along with the use of this type of statistical testing has come a barrage of complaints and arguments over the use of significance testing and the 0.05 threshold. Many notable scholars, such as Cohen (1994), have attacked significance testing and the use of the 0.05 threshold. Recently even the ASA[1] released a statement clarifying the proper definition and use of *p*-values where they discourage their use as threshold standards (Wasserstein & Lazar, 2016). It is a heated and multi-disciplinary debate over the use of the significance testing in research. There is however at least one topic that is fairly well agreed upon, and that is the arbitrary origin of the 0.05 criterion.

1.1.1 A Significant Threshold?

Researchers such as Cohen (1990), Connelly (2014), Greenland et al. (2016), and Salsburg (1985) claim that the 0.05 threshold is arbitrary and suggest that its arbitrary nature makes it less important or valuable. A quieter, but large population of researchers readily agree that the

[1] American Statistical Association

threshold is arbitrary, but believe that being arbitrary does not take away from its importance as a standard. Then, finally, there is a small population of researchers such as Cowles and Davis (1982a) who believe that regardless of its origins, the 0.05 criterion is not actually arbitrary. A more complete explanation of how 0.05 became the common statistical threshold is given by Cowles and Davis (1982b). Their research demonstrates that the origins of the criterion are rooted before the time of R. A. Fisher, but that the widespread use of 0.05 traces back to statements made by Fisher. In particular Fisher wrote in his classic book, *Statistical Methods for Research Workers*,

> "The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant." (R. A. Fisher Sir, 1970, p. 44)

Later in the same book, and in subsequent papers, Fisher varies his approach to decisions of significance and also suggests that the use of other values are appropriate by saying,

> "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level." (R. A. Fisher, 1926)

From there, the use of that "convenient" or "preferred" threshold took root and spread into most research.

With this arbitrary beginning, one can understand the frustrations of those researchers who advocate for discontinuing the use of the 0.05 threshold. However the weight of experience cannot be ignored. Since the general acceptance of the 0.05 threshold, it has arguably served as a relatively good safeguard against spurious conclusions. Admittedly, this is a more complicated issue, but 0.05 has been a decent balancing point between making type I (false positive) and type II (false negative) errors (in many situations) despite its arbitrary origins. So we have a "convenient" value that works relatively well for making judgements and is still arbitrarily defined. Despite this arbitrary origin, the criterion is usually accepted as a reasonable one for most decisions. Why is it that such an arbitrary number is so rarely questioned for being unreasonably small (or large)? Why is it that such an arbitrary number is still so widely used? Why is it that such an arbitrary number often produces decisions similar to the researchers when faced with the collected data? This line of thought leads us to the same question that Cowles and Davis (1982a) had; *"Is the .05 level subjectively reasonable?"* Or put another way, are events that have less than a 5 percent chance of happening generally regarded as non-random while those more likely than 5 percent are generally regarded as random?

1.1.2 The Subjective Threshold

Testing subjective views of probability was studied by Alberoni (1962) as he searched to come up with axioms of subjective probability. His work included testing participants' "threshold of dismissal of the idea of chance" (Alberoni, 1962). However, his research was more interested in the participants intellectual processes rather than the actual threshold. Cowles and Davis (1982a) then took the same idea and applied it to test the common statistical threshold in an interesting, and more important, empirical way. They used a rigged gambling game to test when participants would begin to be suspicious of the game, and when they would decide to quit playing the game. Then, using a binomial probability distribution, they were able to attribute how probable (or improbable) the sequence of events was before participants reached their "threshold of dismissal of the idea of chance". Results from that experiment were that participants expressed suspicion when the events had a probability of about 10% and quit the game when the events had a probability of about 1%. The conclusion was that this experiment suggested that the 0.05 threshold may be close to the subjective threshold.

Although Cowles and Davis (1982a) proposed, and followed through with, a very interesting way of approaching the debate over the statistical criterion, there were a few issues with the study to consider. For one, the study was done with a gambling game involving real money. Even though the money was provided by the experimenters, gambling research has shown major differences in behavior when real money is involved, so that could have influenced the results by causing participants to play longer than they would have otherwise. For another issue, the design of their experiment was sequential such that the event became ever more improbable over time, meaning that it could be the case that participants generally became more suspicious as the experiment progressed. They could have simply had a general limit of how long the experiment could continue before reaching their threshold rather than their threshold having any bearing on the probability of the events. Also, the game used in the experiment had a discrete distribution with a low number of trials before participants reached their decisions. That low number of trials limited the estimation of the subjective threshold to a small possible subset of probabilities. With these considerations in mind it was decided to replicate the idea of Cowles and Davis (1982a) to get a more satisfying answer.

1.2 What is Random?

"Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin." -John von Neumann

It is important to take a moment to consider what it means to be random. Randomness is a concept that does not have a very satisfactory definition no matter how much it has been studied. A simplistic definition of randomness is unpredictability. If events are unpredictable then they are random. Events are unpredictable if we are unable to understand how they are produced or at least the pattern of production. For example, the process of rolling dice is only random because we are unable to fathom the precise physics involved[2] so we can predict the specific outcome. If one were able to understand how the dice would interact with each other and the surface they are rolled on then perhaps each roll could be predicted and the outcome would cease to be random. Therefore, randomness must be a property of the event generating process. The problem is, once we understand the event generating process, the events cease to be random. For example, random number generation in a computer is based upon very simple mathematics. So when someone knows the equation, and the seed, the output becomes completely predictable and non-random. This creates a paradox of randomness: when someone presents a supposedly random sequence to a naive observer, and one who understands the generating process the sequence is then both random and non-random depending on who is being asked. For example, consider the following sequences of numbers between 0-9; 8979323846, 2643383279, or 5028841971. For most readers these three sequences of ten digits are completely random but some may recognize them as the $11^{th} - 40^{th}$ digits in π. Now consider the binary sequences; XOOOOXOXXX, XXXOOXOXOO, or OXXXXXOOOO. Again for most readers these may seem random (or possibly not because of the clustering bias) but they are simply the re-coding of the previous sequences with odd numbers being represented as "O" and even numbers being represented as "X". Randomness is then an unobservable property, because when the generating procedure is understood it ceases to be random. Randomness can only be inferred indirectly from a generating procedure's output. Given these considerations, it is accepted that although the random number generator used in the experiments (described in the following chapters) is objectively not random once understood, it does pass all the standard statistical tests of randomness and would most likely be considered truly random by any participant.

[2]At the moment of event production. It may be possible to predict/fathom what will happen or has happened given all parameters.

1.3 The Clustering Bias

"If you torture the data enough, nature will always confess." -Ronald Coase

(Coase, 1982)

One of the most common topics when discussing subjective judgements of randomness is the clustering bias. The clustering bias occurs when sequences of events, especially binary alternations, have an element that repeats several times in a row then it is more likely to be perceived as non-random. For example, if there were four tails in a row in a sequence of ten coin flips, such as HTTTTHTHHH, that sequence is perceived as not being random even though it is reasonably probable. The bias is discussed by many researchers including: Bar-Hillel and Wagenaar (1991), Falk and Konold (1997), Gilovich, Vallone, and Tversky (1985),Green and Afima (1982), Lopes and Oden (1987), Nickerson (2002), Sanderson (2009), Sun and Wang (2010), Wagenaar (1970a), and Wagenaar (1970b). The clustering bias is usually discussed alongside a related bias called the over-alternation bias. The over-alternation bias occurs when sequences of events, especially binary alternations, have their elements alternate more than would be randomly expected. That sequence is then more likely to be perceived as random. For example, a sequence of coin flips that alternates nearly perfectly such as this, HTHHTHTHTT, is perceived as random even if that sequence is unlikely[3].

The extensive literature on the clustering bias gives the impression that clustering is the most influential factor in judgements of randomness. The idea is that the amount of clustering overwhelms the other judgements that could be made, such as proportion differences present in the sequence. For example, the sequence TTHHHHTTHT may be judged non-random because of the cluster of four "H" despite having equal proportions of "H" and "T". The literature makes the case that even expected amounts of clustering are enough that perception will be swayed to believe the sequence is non-random. This idea is present in Gilovich et al. (1985) and Sanderson (2009), just to name two. There is however, notably, very little research on the

[3]Unlikely by the runs test or similar tests.

clustering bias across levels of proportion biases. How are judgements swayed by clustering at differing proportions of elements? Is the clustering bias still important when there is obviously too many of one item and not enough of another for equal proportions? Does forcing a sequence to have fewer clusters than would be expected by chance, the over-alternation bias, decrease the number of times the sequence is judged non-random?

1.4 Redundancy

"Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital."

-Aaron Levenstein

Information theory analysis is used with some regularity in other fields but seldom considered in research on judgements of randomness. One such analysis compares different levels of redundancy. Informally, redundancy is a measure of patterns and pattern stability in the sequence. Jamieson, Nevzorova, Lee, and Mewhort (2016)[4] gives the example that the sequence TVXTVX is more regular and redundant than the sequence XTVTXV. Similarly, the sequence XXXXXOOOOO is more redundant than OXOOXOXXXO. Sequences that are more redundant have reoccurring patterns, are more predictable, and therefore less random.

Redundancy analysis is able to detect whether the judgements of randomness are due to pattern recognition and is able to direct the researcher towards the types of patterns that are being responded to. This specificity is because the analysis is done at different orders of redundancy. Zero-order redundancy refers to the frequency of individual characters in the sequences being judged. First-order redundancy refers to the frequency of bi-grams, such as TV, VX, XT, XX, OX, or XO, in the sequences being judged. Second-order redundancy refers to the frequency of tri-grams, such as TVX, VXT, XXO,XOO, or OXO in the sequences being judged. Higher order redundancies can be computed up to one order below the number of items

[4]Their paper gave the basis for the redundancy analysis presented in this work.

in each sequence; however, this process quickly becomes very computationally heavy. The judgements of randomness can then be compared with levels of redundancy to see whether/what levels of pattern recognition are being responded to. A more complete explanation of the orders of redundancy and calculations is given in Chapter 7.

1.5 Production Vs. Perception

"Do not trust any statistics you did not fake yourself." - Winston Churchill

In researching subjective judgements of randomness, there are generally two types of experiments: production experiments and perception experiments.

1.5.1 Production Experiments

Production experiments occur when participants are asked to produce random sequences. These experiments compare participants output to truly random sequences to pinpoint the differences. With this line of thought, experimenters have been able to show many differences between a participants production of "randomness" and truly random sequences. One example is that participants produce more local representativeness, meaning a lack of clustering and an equal proportion over subsections of the whole sequence, than would be expected. Production experiments follow an interesting line of thought but it is not clear what causes the systematic biases. It could be that they are true reflections of the participants biased notions of randomness, or it could be that they are biased reflections of the participants accurate notions of randomness. Bar-Hillel and Wagenaar (1991) gives the analogies that people with linguistic competence produce ungrammatical sentences while speaking and those with good musical pitch perception may fail to produce good pitch. Just because people fail to produce randomness does not mean they cannot perceive it.

1.5.2 *Perception Experiments*

Perception experiments occur when participants are presented with a sequence of events and asked to judge for randomness. Sometimes the task is presenting participants with several different sequences of events and asking them to make judgements of randomness. For example, when asked which of these two sequences is random: (1) OOOOOOXXXXOOOXXOOOOO, or (2) OOXOXOOXXOOXOXXXOOXO most participants would say (2) is the random sequence (Sanderson, 2009). Other times the task is presenting participants with one sequence of events and asking whether it was random. For example, participants would be asked if OOOOOOXXXXOOOXXOOOOO is random. These experiments have been able to show many of the same biases that production experiments were able to, including very large clustering biases, as discussed above. However, most perception experiments present the entire sequence all at once; as is demonstrated above when all 20 elements of the sequences are presented together. Very little work has been done experimentally on judgements of randomness where each element is presented one at a time. Consider a binary option presentation, like coin flipping: presenting all 20 "coin flips" at once is a very artificial experience. Admittedly, presenting all elements one at a time on a computer is also artificial, but it at least more closely approximates real world random experiences such as coin flipping or dice rolling. This method of presentation may limit the mental "clumping" of clusters of elements.

1.6 Summary

"In God we trust. All others must bring data." -W. Edwards Deming

(Lynch & Stuckler, 2012)

There has been a lot of research examining subjective judgements of randomness in general, but, very little to find a threshold of randomness. The remaining chapters will describe a series of experiments examining: (1) where the subjective threshold of randomness is and generalizing it to different situations, (2) the effect of the clustering bias vs the proportion bias, and (3) a redundancy analysis based on information theory examining stimuli and responses used in the previous experiments. These experiments will present sequences of simulated coin flips or spinner spins one event at a time to approximate real life situations. Certain implications of the results will follow in discussion.

Chapter 2

Flipping Coins

"...surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989)

2.1 Introduction

As with Cowles and Davis (1982a), the objective of this research was not to make an argument for or against using the 0.05 significance level but to test whether it has any relationship with the subjective threshold of participants. An experiment similar to the one done by Cowles and Davis (1982a) was designed, but specifically avoiding a few elements present in their study. First, the experiment avoided any gambling context. Second, the experiment provided opportunity for participants to see both fair and non-fair sequences in random order (instead of just a non-fair and increasingly improbable sequence) to rule out time and order factors. And finally, the experiment increased the number of possible probabilities that could be declared the subjective threshold and had more trials for each participant.

2.2 Method and Procedure

2.2.1 Participants

Thirty-two University of Lethbridge undergraduate students volunteered for this experiment and 29 of them were compensated with course credit for their efforts. The other 3 willingly donated their time without compensation.

2.2.2 Materials

LiveCode 8.0.1 Community Edition was used to create the program for the experiment. iMac computers were used but no features were used that are specific to any type or brand of computer.

2.2.3 Procedure

Participants were seated in front of a computer where the test would take place. No two participants were seated next to each other while the study was in progress. Participants were instructed that they would be observing the results of several series of coin flips with each series coming from a different coin. They were told that "X" would represent heads and "O" would represent tails. They were instructed that their role would be to decide whether the series of coin flips reasonably came from a fair coin or not. The participants were also told that they were free to take breaks to rest their eyes and look away from the screen after any response and before starting the next series of coin flips.

Sixteen of the participants were shown three series of fifty truly random and fair coin flips as a training phase. These coin flips were truly random in the sense that every single flip had a 50% chance of being a head or a tail on each individual trial. The participants were told explicitly that these were truly fair coin flips and to pay attention in order to familiarize themselves with the procedure of flipping coins. They were not asked for any response to these "coin flips". The sixteen participants who did not receive the training phase simply skipped these series of "coin flips" and went straight to the testing sequences.

All participants were then shown forty-five series of coin flips; there were nine different testing conditions with five replications for each condition. The order of presentation of these series of "coin flips" was randomized for every participant. The series of flips were created by making lists of fifty items and controlling how many of those items were "X"s with the rest being "O"s. The nine conditions are identified by the number of "X"s in the list. The different numbers of “X”s in the conditions were 15, 18, 20, 22, 25, 28, 30, 32, 35.

Once the list with a certain number of “X”s was created, the order was randomized and then presented to the subject one item at a time, very quickly. For example, one series of coin flips from the 15 “X”s condition was: O, O, O, O, O, O, O, O, O, O, X, O, X, O, X, O, O, X, O, X, O, O, O, O, O, X, O, O, X, O, O, X, O, O, O, X, X, O, X, X, X, O, X, O, O, O, X, O, O, O. The item, single letter, or individual “coin flip” would remain on the screen for 0.3 seconds and then be removed. The screen would then remain blank for 0.1 seconds to allow the subject to know that a new item or “coin flip” was about to be shown. No subjects ever voiced a concern that the coin flips were happening too quickly to be able to distinguish when a new one was taking place even when questioned following the experiment.

After all fifty “coin flips” had been presented to the subject, the program would then ask whether the series had been created by a “fair coin” with the two possible responses being “Yes” or “No.” After one-half of the “coins” or series of “coin flips” had passed, a message would be presented to the participants that they were one-half done and should take a break and look away from the computer screen. This break at the halfway point was not enforced.

The decision to have fifty “coin flips” per “coin” was made as it seemed to be an appropriate number where there was neither too few coins to limit analysis to a narrow range of p-values nor so many “flips” that the subjects would not be able to concentrate for the entire “coin”. The choice of which conditions to use was made to have the condition with the p-value just below the 0.05 threshold involved, to have symmetry in the conditions, and to limit ourselves to a reasonable number of conditions to ensure the participants could remain attentive. The time that each “coin flip” remained on and off the screen was chosen so that the series of “flips” would be too fast to count but not too fast that the flips blurred together. No variations on these parameters were tested in this set of experiments.

Response time was recorded for each response given by the participants. The amount of time was calculated from the moment the question was presented to the participant to the moment the participant selected one of the responses.

After all forty-five "coins" or series of "coin flips" were shown and responded to, a short survey was presented to the participants. It asked for participants' age, sex, university major, whether the participant had taken different types of statistics classes, and a scale of 1-100 response on how well they thought they understood what the 0.05 significance level meant.

2.2.4 Analysis

LibreOffice and R (R Core Team, 2015) within the RStudio (RStudio Team, 2015) environment were used for all of the analysis.

A binomial distribution was used to calculate the probability of the series of coin flips in each condition. Each condition is listed in Table 2.1 with its associated p-value. The p-values shown are the sum of the probabilities of that particular condition and any more extreme condition assuming that the coin was completely fair. The p-values listed are all one-tailed so only the more extreme values in one direction are considered in the calculation. The one-tailed p-value was used because it was assumed that participants would be able to tell whether one of the elements was happening more often than the other, so there would be a directionality to their decision criterion.

Condition	P-value
15	0.0033002
18	0.0324543
20	0.1013194
22	0.2399438
25	0.5561376
28	0.2399438
30	0.1013194
32	0.0324543
35	0.0033002

Table 2.1: One-Tailed P-Values for Conditions in Flipping Coins

The subjective threshold is measured by the probability between where the subjects on average called the condition random and non-random. When more than one-half of the trials in a certain condition were called non-random then that condition is on average non-random. With the discrete nature of the binomial distribution this measure will give a range of values in which the subjective threshold could lie; however, it could also be approximated by the middle value between the two conditions.

2.3 Results

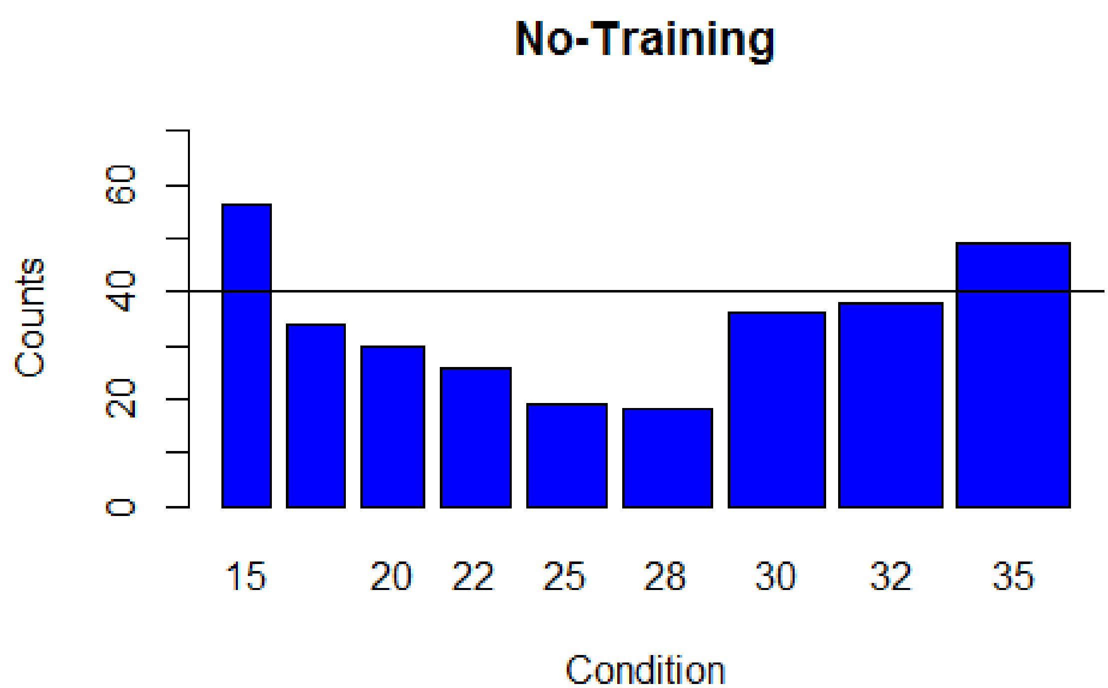

Figure 2.1: Flipping Coins: No Training Counts

Figures 2.1, 2.2, and 2.3 are three graphs showing the number of times each condition was called non-random summing across all participants. Figure 2.1 is a graph of all participants who did not receive the initial training phase, Figure 2.2 is a graph of all participants who did

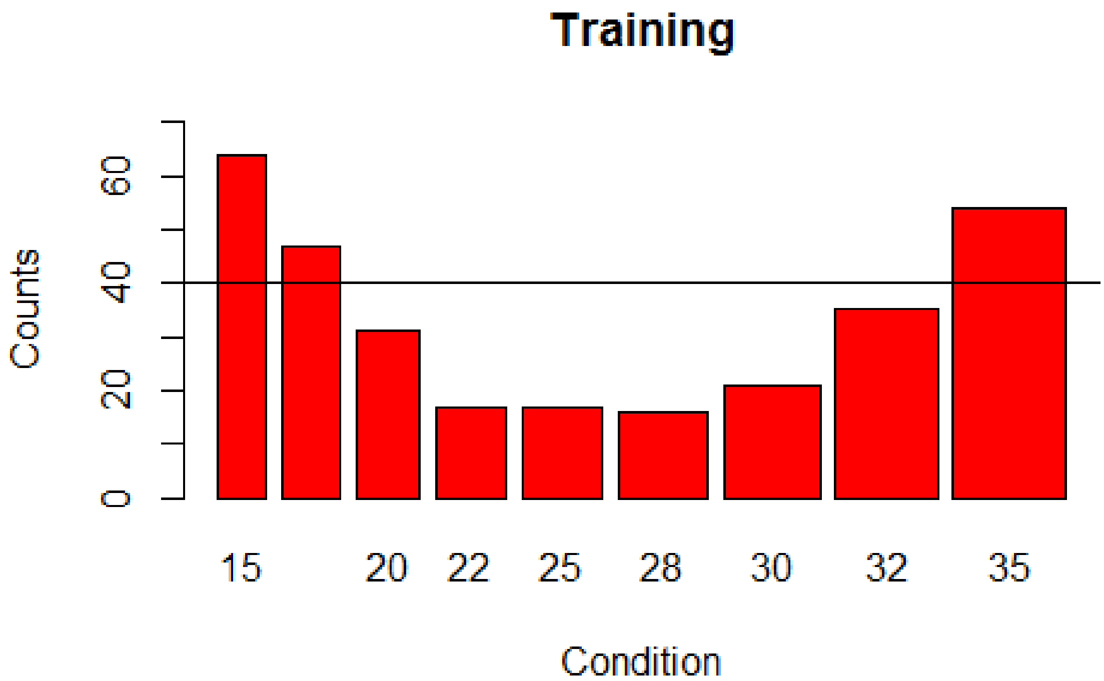

Figure 2.2: Flipping Coins: Training Counts

receive the initial training phase, and Figure 2.3 is a graph of all participants in both the training and non-training conditions. All participants were grouped together for Figure 2.3 because it was determined that training had no notable impact. Figure 2.3 denotes the condition by the numbers on the bars and the p-values for each condition on the x-axis. The horizontal line is drawn at half the total number of coins in each condition; any bars that extend above that line were called non-fair more often than they were called fair, so they were on average non-fair. Figure 2.4 is the same at Figure 2.3 except with the complementary conditions (those that have equal p-values) summed together and the y-axis is the proportion of coins called non-random instead of counts.

As can be seen in Figure 2.3 the measure of the subjective threshold puts it at a value between 0.0325 and 0.1013 (with 0.0669 being the mean value) for those conditions with fewer "X"s involved and between 0.0325 and 0.0033 (with 0.0179 being the mean value) for those

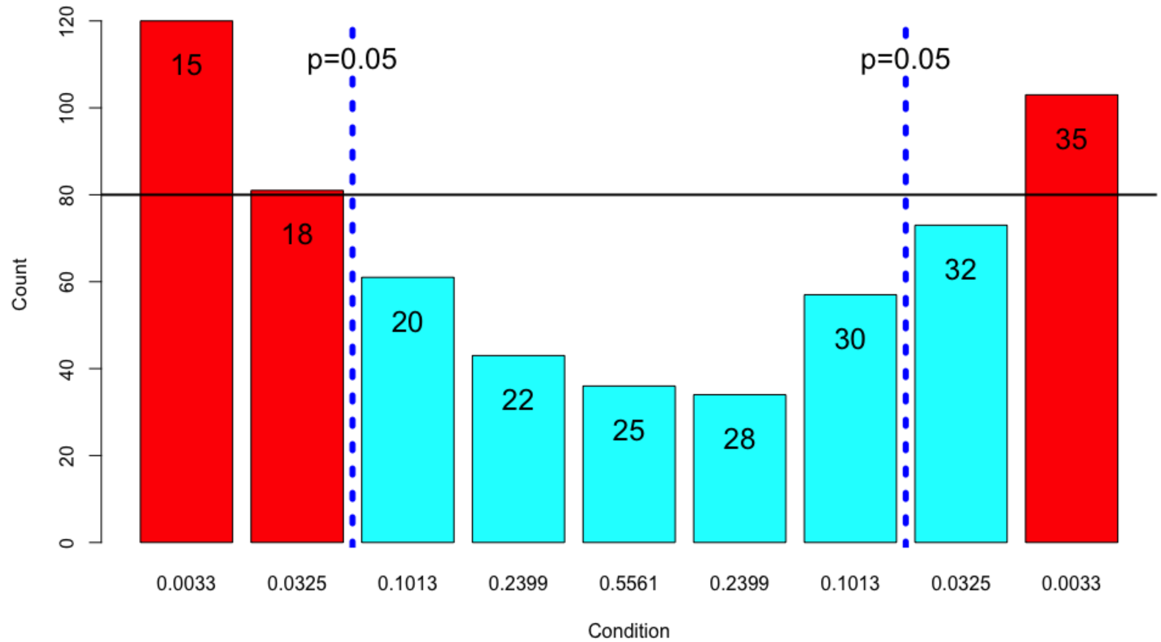

Figure 2.3: Flipping Coins: All Participants Counts

conditions with more “X”s involved.

Figure 2.5 is a barplot of the the average differences in response time for each condition. The differences were in fractions of a second and did not show any interesting patterns. Similarly, responses to the questionnaire at the end of the experiment showed no meaningful relationship between age, sex, university major, or statistical background, to judgements of randomness.

Figure 2.6 is a plot set up the same way as Figure 2.3 but with only a single subject’s data included in the plot.

A runs test for randomness was done on each “coin” but was not found to be influential in decisions of randomness; when those that failed[1] were excluded from the analysis the results did not change.

[1]Were in the most extreme 10% of theoretical coins.

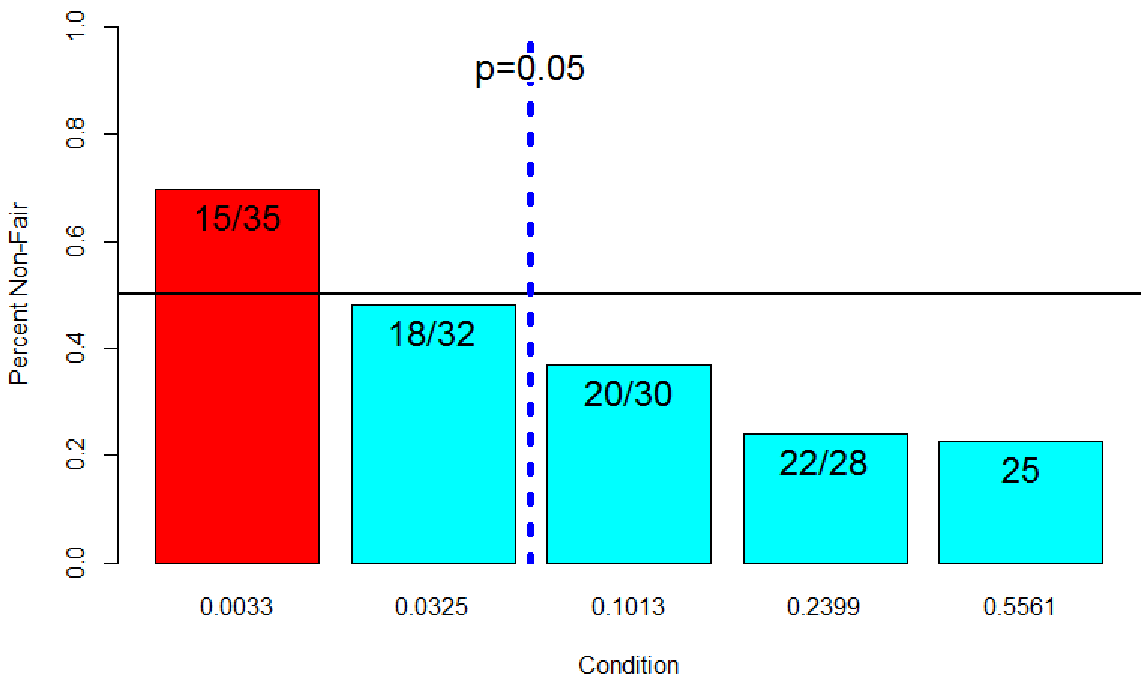

Figure 2.4: Flipping Coins: Proportion Called Non-Random with Complementary Conditions Summed Together

2.4 Discussion

An examination of Figure 2.3 leads one to believe that the answer to Cowles and Davis (1982a) question, "Is the .05 level subjectively reasonable?" is yes. If the proposed measurement of the subjective threshold is accepted then it is reasonable that we conclude that the subjective threshold is close to 0.0325 because it appears in both measurements of the threshold; or to conclude that it is close to 0.0424 which is the mean p-value of all conditions involved in the measure of the subjective threshold. Both of these measurements are reasonably close to 0.05, and would round to 0.05, so it could closely approximate the subjective threshold.

This experiment was able to replicate both the idea behind and the results of Cowles and Davis (1982a). It also avoided using any type of gambling device and it had participants making judgements of both randomness and non-randomness instead of having them stop the experiment

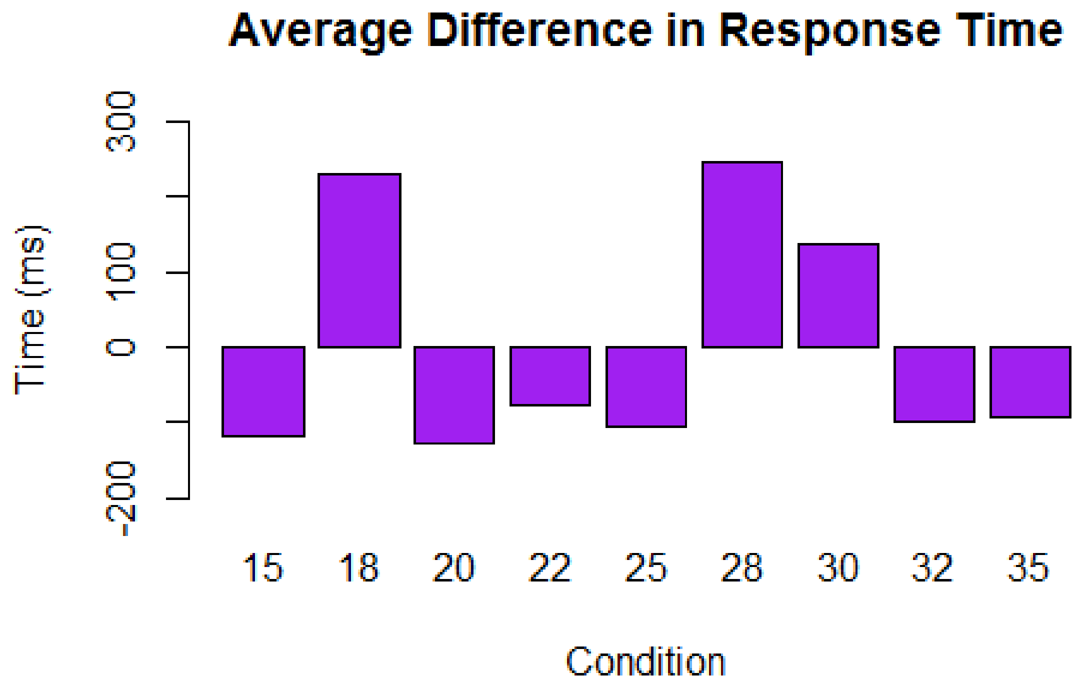

Figure 2.5: Flipping Coins: Average Difference in Response Time by Condition

at a certain point. These changes were made to avoid issues associated with gambling and sequential effects. This experiment was also able to narrow the subjective threshold a bit more closely than did Cowles and Davis (1982a) but is still limited by the discrete range of values associated with the binomial distribution.

It is interesting to note that the fairly clean results for a single subject seen in Figure 2.6 were relatively common across all of the subjects. Although each subject may have varied quite a bit on the value of their individual threshold, some preferring to only call the most extreme conditions non-random and others taking a more liberal approach, almost all of them had a fairly well defined personal cut-off point that was consistent between the high and low conditions. This is evidence that the participants were: not especially sensitive to either “X”s or “O”s, use some kind of threshold decision making, and that the threshold may vary from person to person.

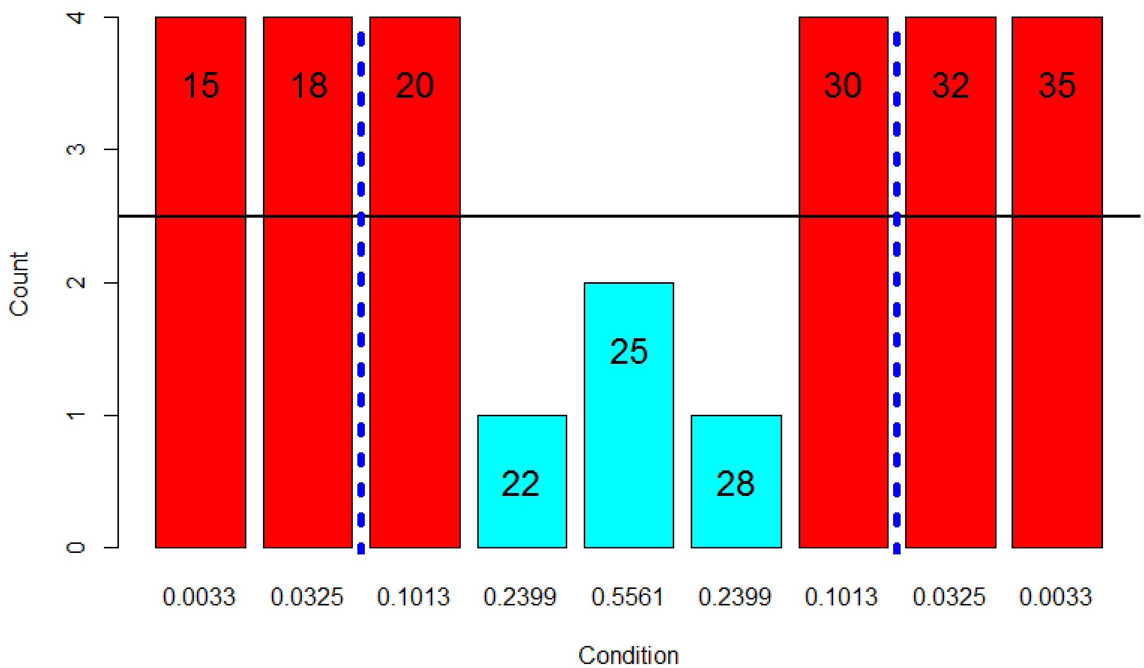

Figure 2.6: Flipping Coins: A Single Participant's Counts

Chapter 3

Flipping Coins 2.0

”If your experiment needs statistics, you ought to have done a better experiment.”

– Ernest Rutherford

3.1 Introduction

After successful completion of Flipping Coins, described in Chapter 2, it was decided to attempt to generalize the results and get a more exact estimate of the subjective threshold by targeting a broader range of possible probabilities. The same experiment was run again but with all 21 conditions between 15-35 being presented.

3.2 Method and Procedure

3.2.1 Participants

Thirty-four University of Lethbridge undergraduate students volunteered for this experiment and were compensated with course credit for their efforts.

3.2.2 Materials

LiveCode 8.0.1 Community Edition was used to create the program for the experiment. iMac computers were used but no features were used that are specific to any type or brand of computer.

3.2.3 Procedure

The same general procedure as was described in Section 2.2.3 was used with a few differences:

- Seventeen participants received the training phase and seventeen did not receive it.
- All participants were tested on 42 series of coin flips; there were twenty-one different testing conditions with two replicates for each condition.
- The twenty-one different conditions were identified by the number of "X"s in the list and included all integers from 15 to 35 inclusive.

3.2.4 Analysis

The same general analysis as was described in Section 2.2.4 was used except replacing Table 2.1 with Table 3.1.

3.3 Results

Once again there was no notable difference between participants who received the training phase and those that did not, so the data were summed together. Figure 3.1 is a graph showing the number of times each condition was called non-random summing across all participants. Figure 3.1 denotes the condition by the numbers on the bars and the p-values for each condition on the x-axis. The horizontal line is drawn at half the total number of coins in each condition; any bars that extend above that line were called non-fair more often than they were called fair, so they were on average non-fair. Figure 3.2 is the same as Figure 3.1 except with the complementary conditions (those that have equal p-values) summed together and the y-axis is the proportion of coins called non-random instead of counts.

Condition	P-value
15	0.0033002
16	0.0076733
17	0.0164196
18	0.0324543
19	0.0594602
20	0.1013194
21	0.1611182
22	0.2399438
23	0.3359055
24	0.4438624
25	0.5561376
26	0.4438624
27	0.3359055
28	0.2399438
29	0.1611182
30	0.1013194
31	0.0594602
32	0.0324543
33	0.0164196
34	0.0076733
35	0.0033002

Table 3.1: One-Tailed P-Values for Conditions in Flipping Coins 2.0

Figure 3.3 is a barplot of the the average differences in response time for each condition. The differences were in fractions of a second and did not show any interesting patterns. Similarly, responses to the questionnaire at the end of the experiment showed no meaningful relationship between age, sex, university major, or statistical background, to judgements of randomness

As can be seen in Figure 3.1 the measure of the subjective threshold puts it at a value between 0.016 and 0.032 (with 0.024 being the mean value) for both those conditions with fewer “X”s involved and those conditions with more “X”s involved.

A runs test for randomness was done on each “coin” but was not found to be influential in decisions of randomness; when those that failed[1] were excluded from the analysis the results did not change.

[1]Were in the most extreme 10% of theoretical coins.

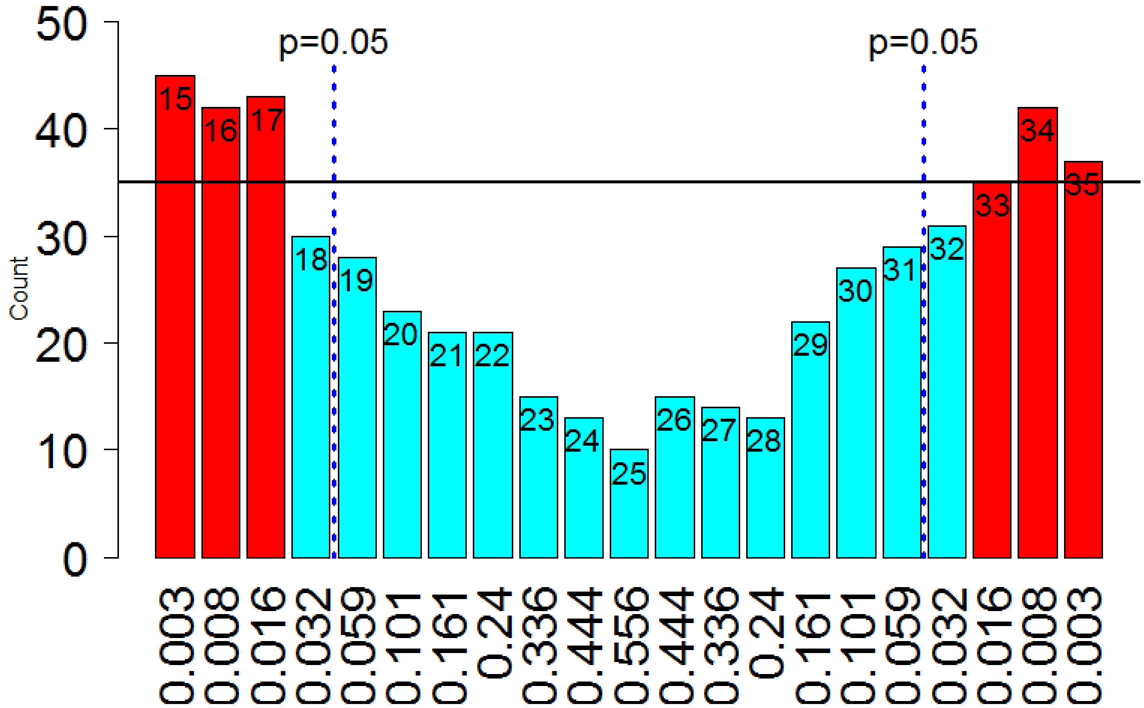

Figure 3.1: Flipping Coins 2.0: All Participants Counts

3.4 Discussion

This experiment replicated the results from the experiment in Chapter 2 and further confirmed the idea that the subjective threshold is reasonably close to 0.05. In fact, it suggests that the subjective threshold may be lower than 0.05 but still within the same general range.

After completing this experiment there was some concern that the results could simply be an artifact of the procedure of flipping coins and may not be generalizable outside of that context. To demonstrate that that was not the case, we confirmed the conclusions with the experiments described in Chapter 4 and Chapter 5.

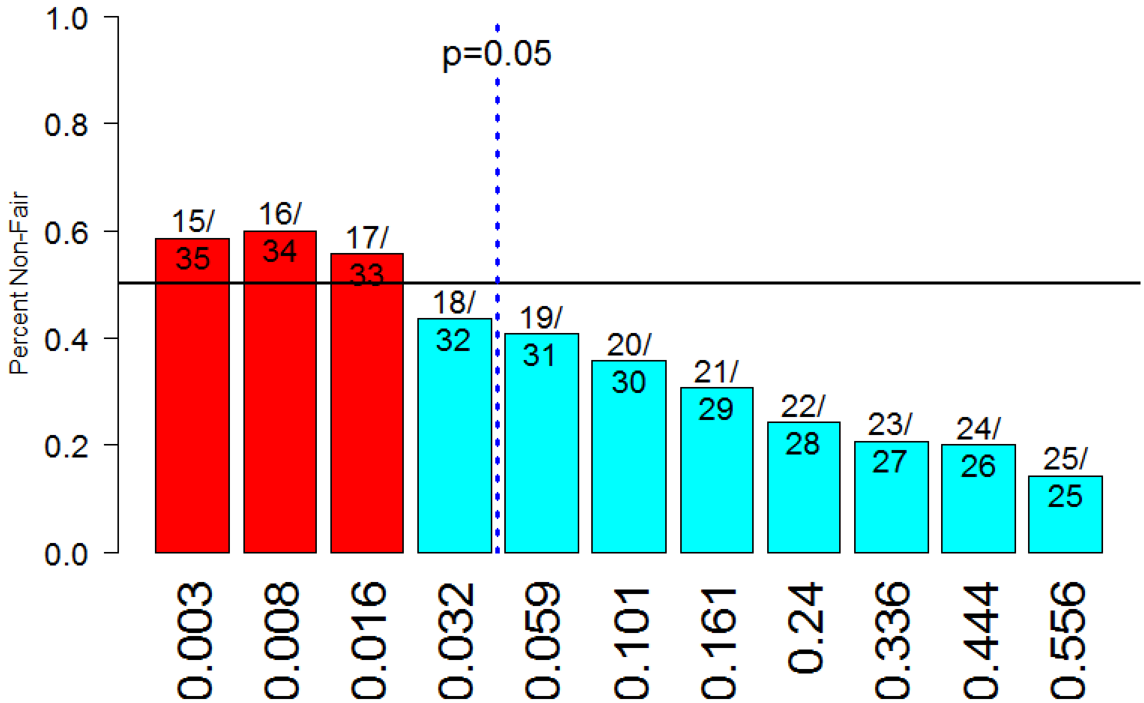

Figure 3.2: Flipping Coins 2.0: Proportion Called Non-Random with Complementary Conditions Summed Together

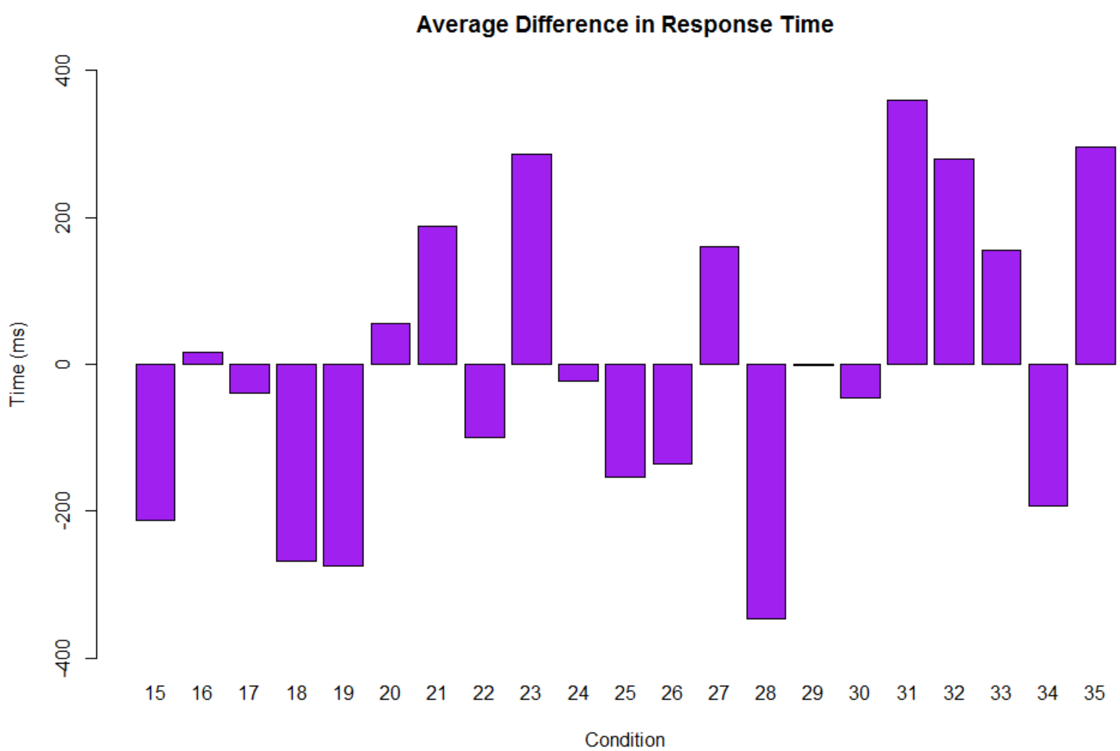

Figure 3.3: Flipping Coins 2.0: Average Difference in Response Time by Condition

Chapter 4

Spinning Spinners (3 Choices)

"All models are wrong, but some are useful." (Box, 1976)

4.1 Introduction

Because of the concern about the generalizability of the Flipping Coins experiment it was decided to do a similar experiment but extend it outside a binary task. A simple extension of the task was made to relate it to spinning spinners to give three possible results instead of just two. The goal of the experiment was the same as in Chapters 2 and 3, to find the subjective threshold, but now using the multinomial distribution of probabilities instead of the binomial.

4.2 Multinomial P-value Justification

4.2.1 Multinomial Distribution

The multinomial probability density function (PDF) is given by:

$$P(x) = \frac{n!}{\prod_{i=1}^{k}(x_i!)} \prod_{i=1}^{k} p_i^{x_i}; \tag{4.1}$$

$$x_i \in 0, ..., n, \sum(x_i) = n, \sum(p_i) = 1$$

Where n = the number of items in the sequence, x_i = the number of replications of item i, and p_i = the probability of item i

For the following experiment the null hypothesis, H_0, would have a multinomial distribution with three equally probable items and fifty-one items in every sequence so the experiments specific PDF, under H_0, is given by:

$$P(x) = \frac{51!}{(x_1!) * (x_2!) * (x_3!)} (1/3)^{x_1} * (1/3)^{x_2} * (1/3)^{x_3}; \tag{4.2}$$

$$x_i \in 0, ..., 51, \sum(x_i) = 51$$

4.2.2 Multinomial P-Value

The problem with the multinomial distribution is that it does not have a defined cumulative distribution function (CDF) or p-value. The p-value, in general, is defined as

> "...the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis H_0 is true." (Montgomery, 2013)

It was thought to be reasonable to define the multinomial p-value as the sum of the probabilities of all possible events that are equally or less probable.

A visual representation of this calculation is given in Figure 4.1. The same visual representation, but cutting out the extremely improbable values, is given by Figure 4.2 and a contour plot of the distribution is given by Figure 4.3. In all of these Figures, the x-axis and y-axis are representations of x_1 and x_2 from Equation 4.2 and x_3 is not represented because it is simply $51 - x_1 - x_2$.

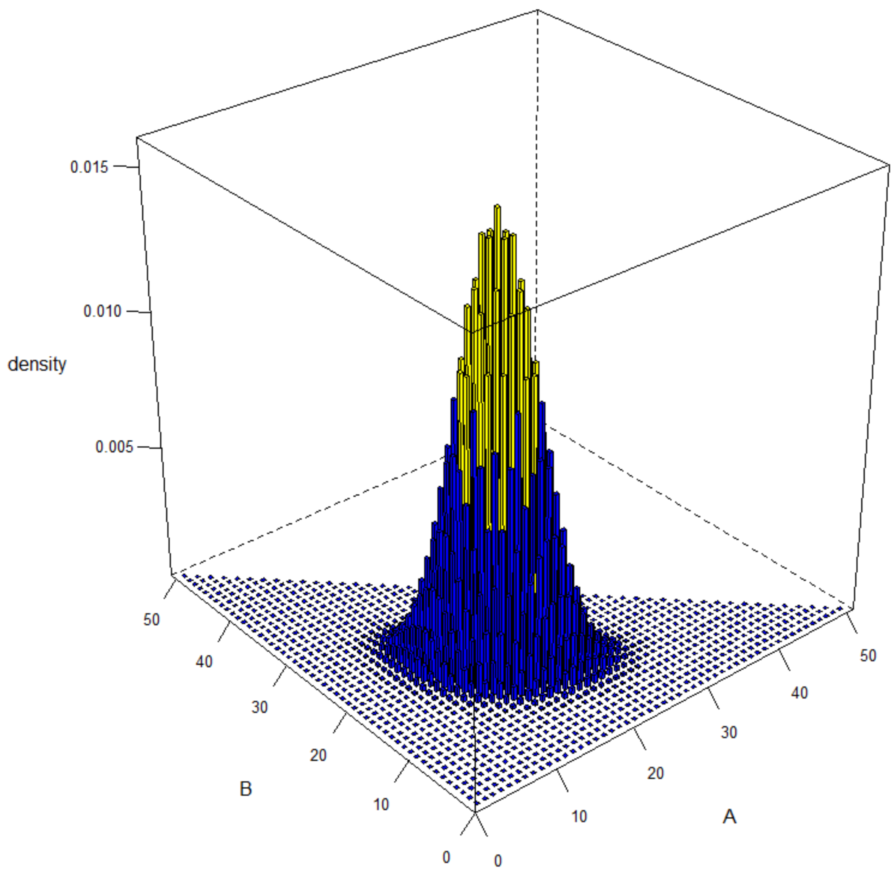

Figure 4.1: Visual Representation of Multinomial P-value With All Dark Bars Having a Density ≤ 0.01

4.2.3 One-Tailed Test

It was also thought reasonable to approximate a one-tailed test in the multinomial distribution by dividing the p-value by the number of unique possible items. In the binomial distribution a one-tailed test could have been done by summing all of the equally or less likely possible

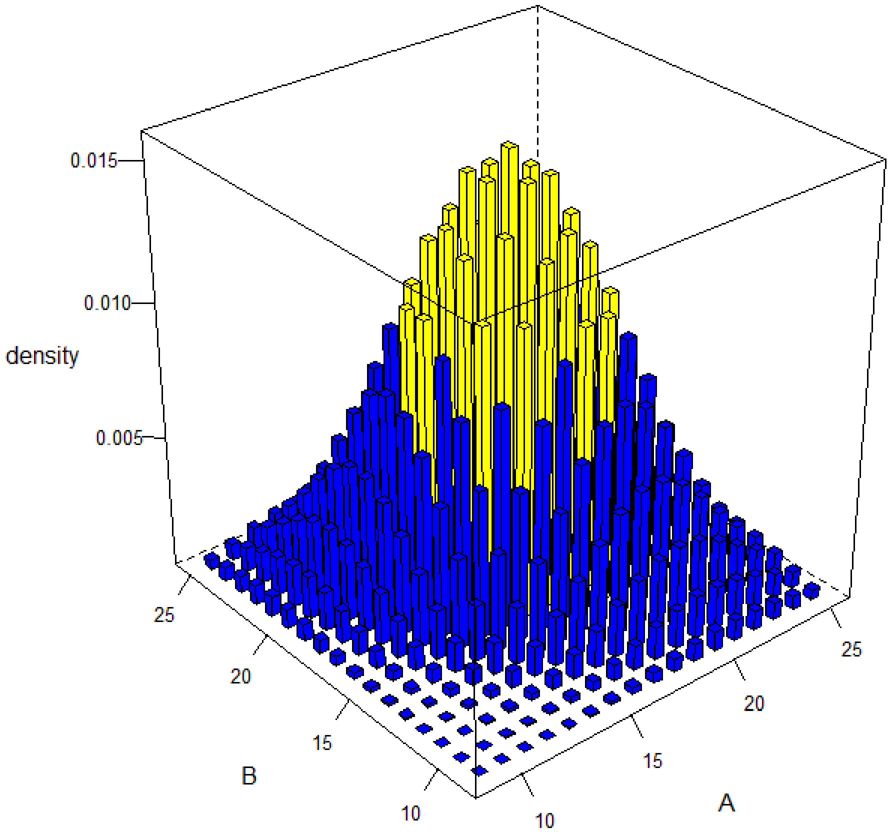

Figure 4.2: Close up Visual Representation of Multinomial P-value With All Dark Bars Having a Density ≤ 0.01

events' probabilities and then dividing by two.[1] In the multinomial distribution, calculating the one-tailed test in this way should follow the same logic but it is accepted to be an approximation; not an exact answer.

[1]With the exception of the perfectly even split this method would have given the exact same p-values as calculated in Section 2.2.4 and Section 3.2.4

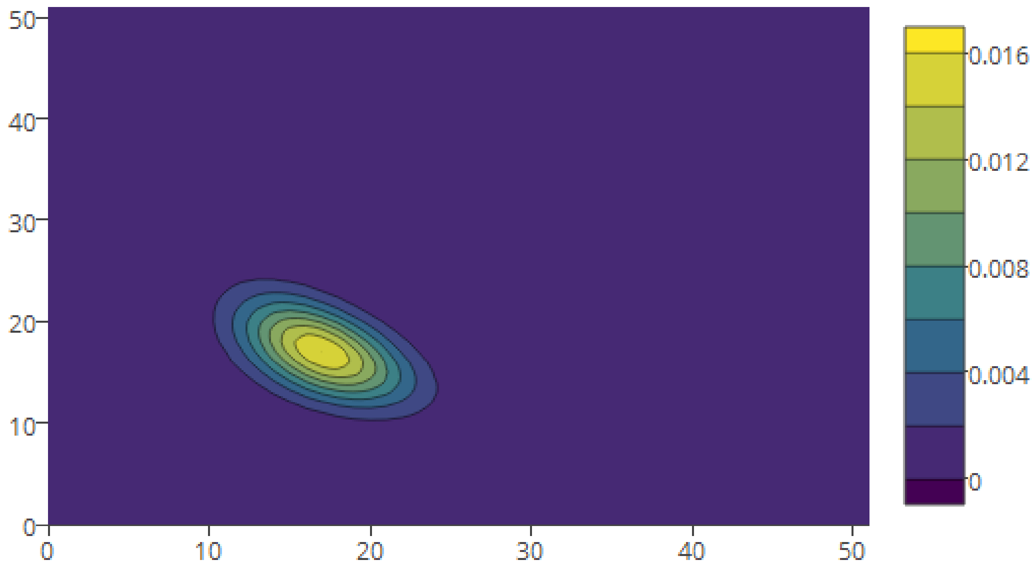

Figure 4.3: Contour Plot of 3 Choice Multinomial Distribution

4.3 Method and Procedure

4.3.1 Participants

Twenty University of Lethbridge undergraduate students volunteered for this experiment and all of them were compensated with course credit for their efforts.

4.3.2 Materials

LiveCode 8.0.1 Community Edition was used to create the program for the experiment. iMac computers were used but no features were used that are specific to any type or brand of computer.

4.3.3 Procedure

The same general procedure as was described in Section 2.2.3 was used with key differences:

- Instead of series of "X"s and "O"s the participants were presented either series of "A"s, "B"s, and "C"s (the letter condition) or blocks of blue, green and red (the colour condition).
- All participants were shown a virtual toy spinner that they could spin to understand the supposed background of the task. This also allowed participants to instantly see all possible values that the spinner could take before beginning the task.
- All twenty participants received the customary training sequences from completely fair spinners.
- All participants were tested on 45 series of spinner spins; there were nine different testing conditions with five replicates for each condition.
- The series of spinner spins were created by making a list of fifty-one items, randomly selecting one of the three possible results, and putting seventeen replicates[2] into the list. Then, randomly selecting a different possible result and assigning it a certain number of replications, and finally filling the rest of the list with the last possible result. The number of replicates for the items that were variable was decided by the condition the spinner was in. The conditions were 7, 9, 12, 15, 17, 19, 22, 25, and 27. After the list was generated its order was then randomized. For example, one series of spinner spins from the letter:7 condition (One item will appear 17 times, one 7 times, and one 27 times) was: A, A, A, A, A, B, C, A, A, A, C, A, A, A, C, C, A, C, A, A, A, B, C, C, A, C, C, A, C, A, A, B, C, C, C, A, A, B, A, B, A, A, C, C, A, B, C, A, B, A, C. Presentation of these items followed the same timing and procedure as in Section 2.2.3.

[2]The expected number of replicates in a fair spinner

4.3.4 Analysis

The same general analysis as was described in Section 2.2.4 was used except replacing Table 2.1 with Table 4.1 that is derived from the one-tailed multinomial p-value described in Section 4.2.

Condition	P-value
7	0.0008014
9	0.0075587
12	0.0838712
15	0.2708057
17	0.3333333
19	0.2708057
22	0.0838712
25	0.0075587
27	0.0008014

Table 4.1: One-Tailed P-Values for Conditions in Spinning Spinners (3 Choices)

4.4 Results

Figures 4.4, 4.5, and 4.6 are three graphs showing the number of times each condition was called non-random summing across all participants. Figure 4.4 is a graph of all participants who were in the colour condition, Figure 4.5 is a graph of all participants who were in the letter condition, and Figure 4.6 is a graph of all participants in both the colour and letter conditions. There was no notable difference between participants' results for those who received the colour condition and those that received the letter condition so the data were summed together. Figure 4.7 is the same as Figure 4.6 except with the complementary conditions (those that have equal p-values) summed together and the y-axis is the proportion of spinners called non-random instead of counts.

Figure 4.8 is a barplot of the the average differences in response time for each condition. The differences were in fractions of a second and did not show any interesting patterns.

As can be seen in Figure 4.7 the measure of the subjective threshold puts it at a value

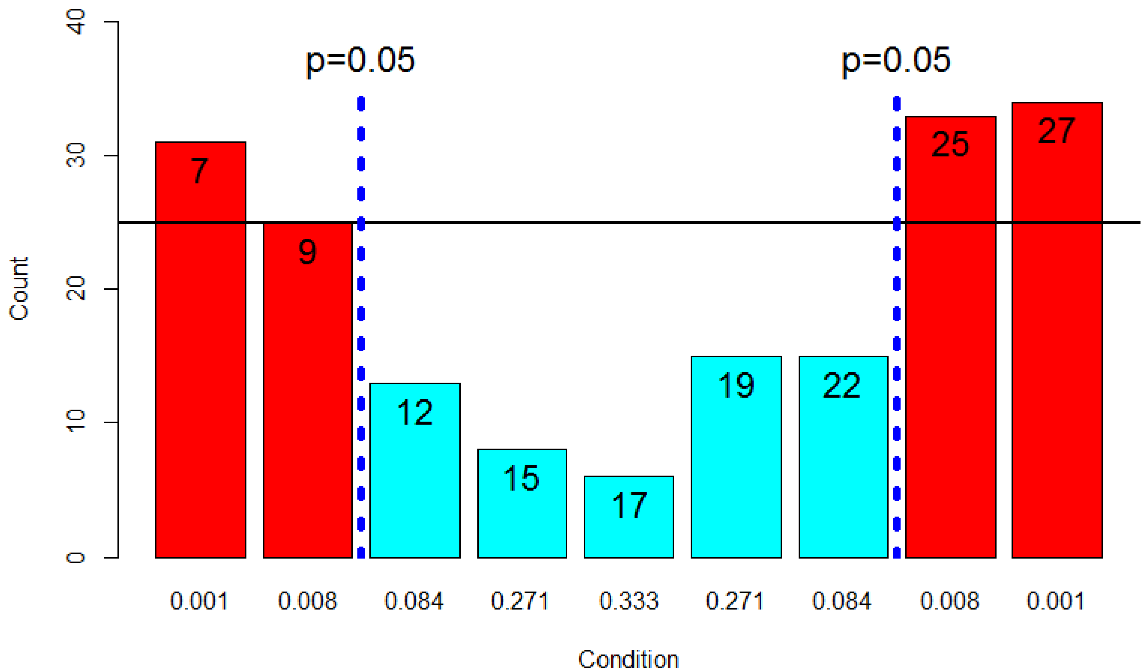

Figure 4.4: Spinning Spinners: Colour Spinner Counts

between 0.008 and 0.084 (with 0.046 being the mean value).

4.5 Discussion

If the assumptions made in this experiment are believed then it has given further evidence for the subjective threshold to be about 0.05. The tests in Chapters 2 and 3 have demonstrated that participants begin to judge binary alternation events as non-random at a level just below 0.05 and now that result is extended to events beyond binary tasks within a multinomial distribution. This is evidence that the result is not an artifact of the Flipping Coins task but somewhat more generalizable. To further that idea, we extended the testing into Spinning Spinners 2.0 with 5 possible results.

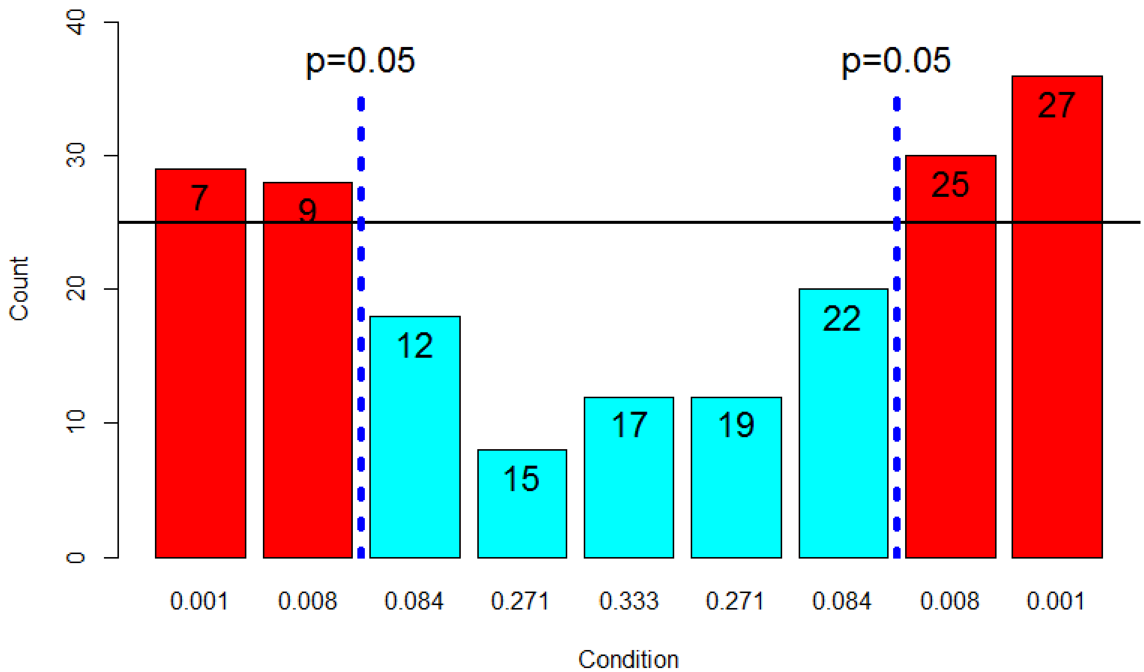

Figure 4.5: Spinning Spinners: Letter Spinner Counts

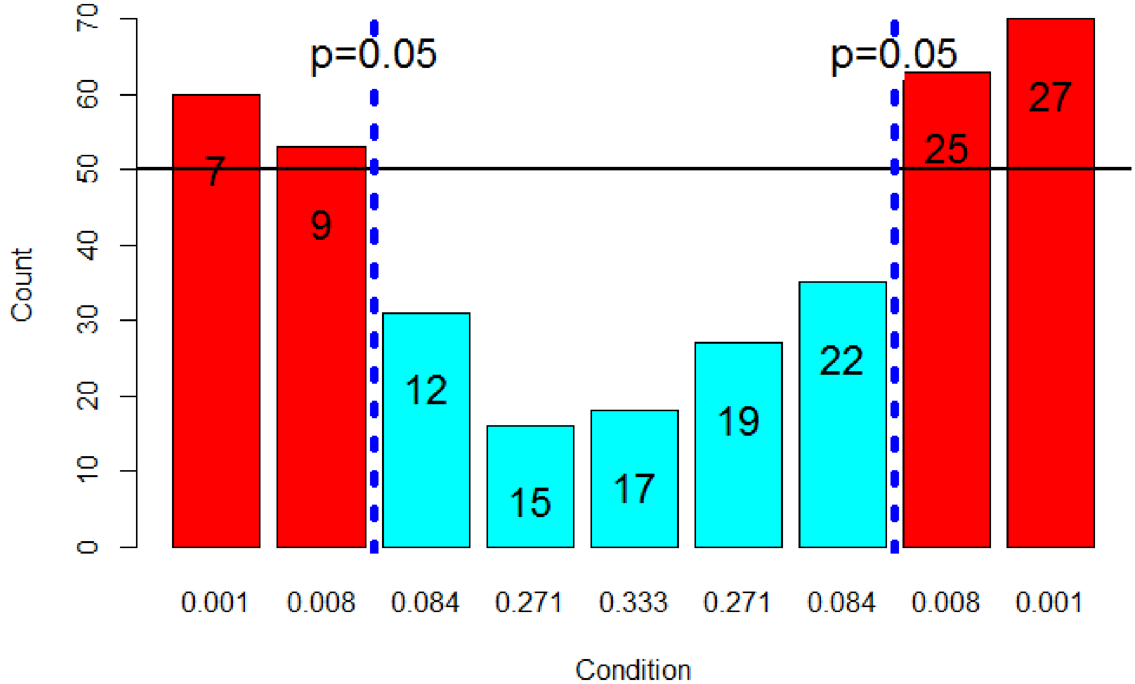

Figure 4.6: Spinning Spinners: All Participants Counts

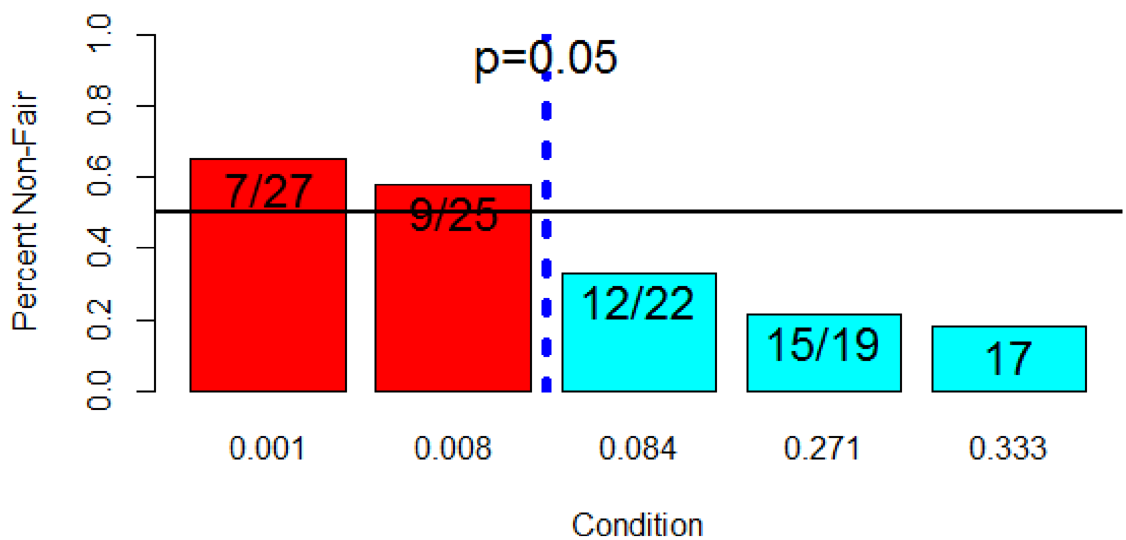

Figure 4.7: Spinning Spinners: Proportion Called Non-Random with Complementary Conditions Summed Together

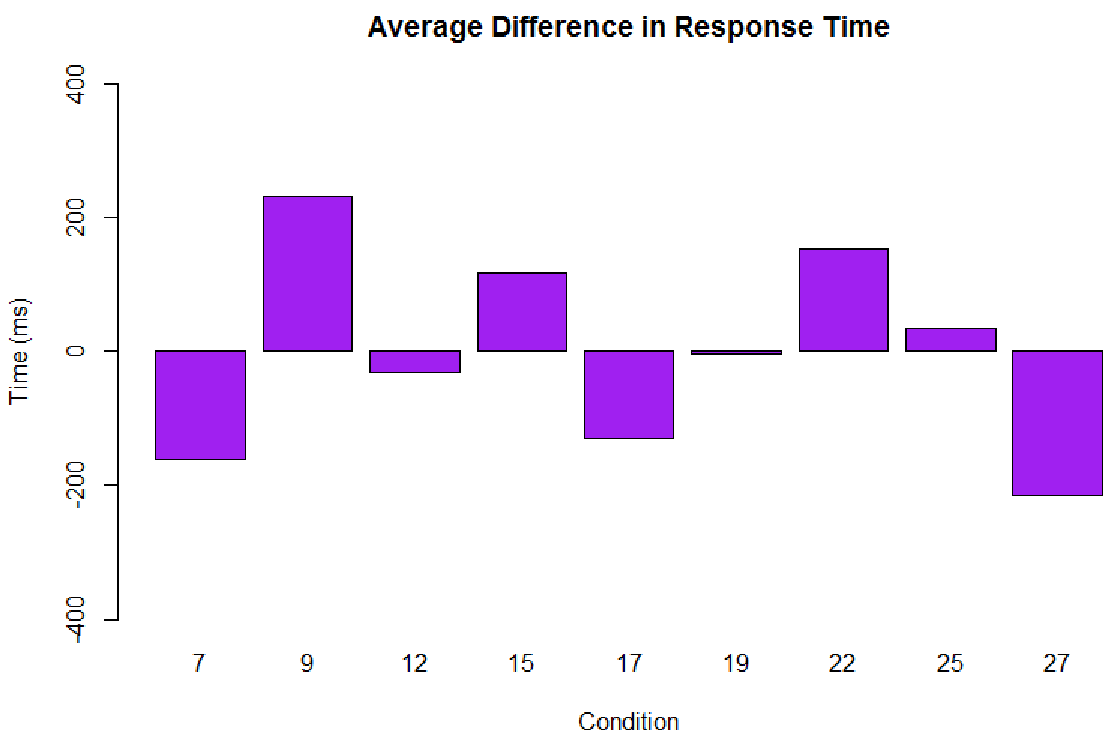

Figure 4.8: Spinning Spinners: Average Difference in Response Time by Condition

Chapter 5

Spinning Spinners 2.0 (5 Choices)

"Anyone who can do solid statistical programming will never miss a meal." - David Banks

5.1 Introduction

To extend the ideas presented in the last three chapters, a new experiment was completed that followed the same ideas as Spinning Spinners in Chapter 4 but with more than 3 possible options.

5.2 Multinomial P-value for 5 Options

With an experiment looking at a spinner with five options the null distribution would be a multinomial distribution with five equally probable items, similar in concept to Section 4.2.1. This time each sequence would involve fifty items in every sequence so the experiments specific PDF, under H_0, is given by:

$$P(x) = \frac{50!}{(x_1!) * (x_2!) * (x_3!) * (x_4!) * (x_5!)} (1/5)^{x_1} * (1/5)^{x_2} * (1/5)^{x_3} * (1/5)^{x_5} * (1/5)^{x_5}; \quad (5.1)$$

$$x_i \in 0, \ldots, 50, \sum(x_i) = 50$$

The same logic from Section 4.2.2 was used to calculate the p-value for this experiment. An approximation of a one-tailed test was done by dividing the p-value by five, the method discussed in Section 4.2.3.

5.3 Method and Procedure

5.3.1 *Participants*

Nineteen University of Lethbridge undergraduate students volunteered for this experiment and all of them were compensated with course credit for their efforts.

5.3.2 *Materials*

LiveCode 8.0.1 Community Edition was used to create the program for the experiment. iMac computers were used but no features were used that are specific to any type or brand of computer.

5.3.3 *Procedure*

The same general procedure as was described in Section 4.3.3 was used with a few differences:

- There was no colour condition. All participants saw a letter condition that presented a series of "A"s, "B"s, "C"s, "D"s, and "E"s.
- Nine of the participants were tested on 50 series of spinner spins; they were shown ten different conditions with five replicates for each condition. They were labeled the "odd condition group" because they were shown the odd numbered conditions. The other ten participants were tested on 55 series of spinner spins; they were shown eleven different conditions with five replicates for each condition. They were labeled the "even condition group" because they were shown the even numbered conditions.
- The series of spinner spins were created by making a list of fifty items, randomly selecting three of the five possible results, and putting ten replicates[1] into the list. Then, randomly

[1]The expected number of replicates in a fair spinner

selecting a different possible result and assigning it a certain number of replications, and finally filling the rest of the list with the last possible result. The number of replicates for the items that were variable was decided by the condition the spinner was in. The conditions were all integers from 0 to 20 inclusive. The "odd condition group" received all of the odd numbered conditions and the "even condition group" received all the even numbered conditions. After the list was generated, its order was then randomized. For example, one series of spinner spins from the 19 condition (three items will appear 10 times, one 19 times, and one 1 time) was: B, B, B, B, A, B, D, E, A, B, A, D, B, B, B, E, D, D, A, A, D, E, E, A, E, D, E, B, D, C, A, D, B, A, E, E, E, D, D, B, B, B, A, E, B, A, B, B, B, B. Presentation of these items followed the same timing and procedure as in Section 2.2.3.

5.3.4 Analysis

The same general analysis as was described in Section 2.2.4 was used except replacing Table 2.1 with Table 5.1 that is derived from the one-tailed multinomial p-value described in Section 5.2.

5.4 Results

Figure 5.1 is a graph showing the proportion of times each condition was called non-random summing across all participants. Figure 5.2 is the same as Figure 5.1 except with the complementary conditions (those that have equal p-values) summed together.

Figure 5.3 is a barplot of the average differences in response time for each condition. The differences were much larger than seen before but still with no interesting patterns.

As can be seen in Figure 5.2 the measure of the subjective threshold puts it at a value between 0.006 and 0.023 (with 0.0145 being the mean value).

Condition	P-value
0	0.0000090
1	0.0001628
2	0.0013094
3	0.0064540
4	0.0229574
5	0.0555212
6	0.1056208
7	0.1549310
8	0.1888122
9	0.1998912
10	0.2000000
11	0.1998912
12	0.1888122
13	0.1549310
14	0.1056208
15	0.0555212
16	0.0229574
17	0.0064540
18	0.0013094
19	0.0001628
20	0.0000090

Table 5.1: One-Tailed P-Values for Conditions in Spinning Spinners 2.0

5.5 Discussion

Once again the results demonstrate that the subjective threshold is reasonably close to or below 0.05. Although the way the spinners in this experiment (and those in Chapter 4) were produced made them not truly multinomial, the instructions and examples given to the participants were such that they should have expected multinomial patterns. Therefore, from the perspective of the participant, this task should appear multinomial. So having demonstrated that the subjective threshold is reasonably close to 0.05 with two multinomial tasks suggests that the idea is relatively robust and generalizable. Implications of 0.05 being close to the subjective threshold will be discussed in Chapter 8.

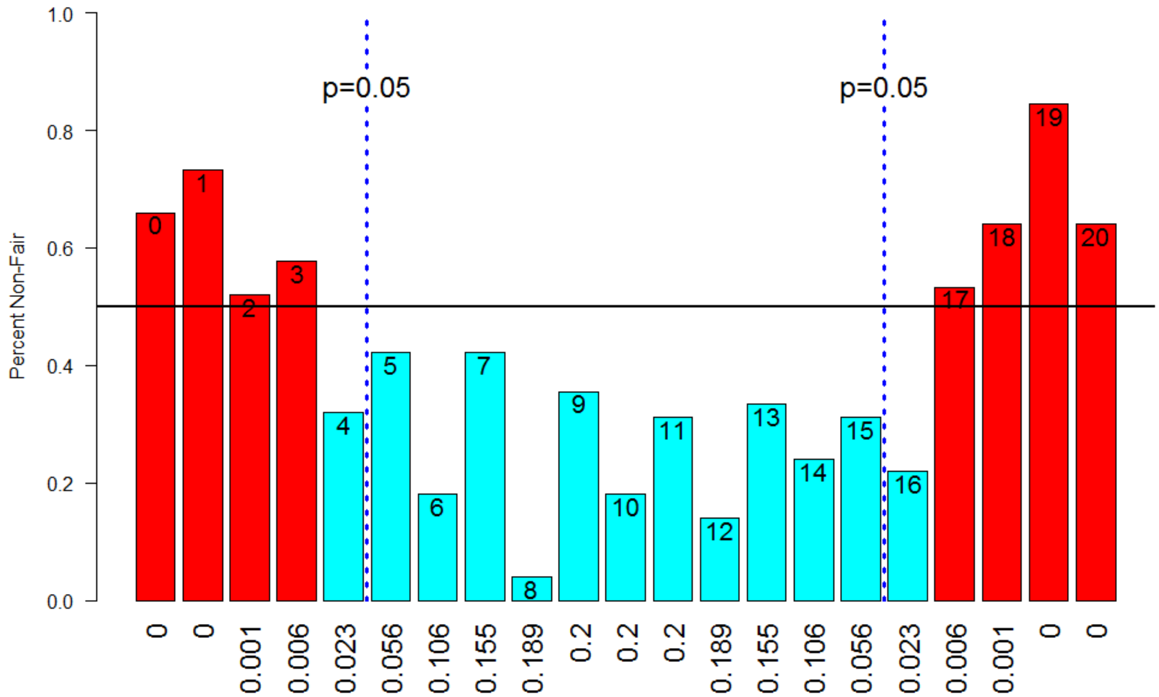

Figure 5.1: Spinning Spinners 2.0: All Participants, Proportion Non-Fair

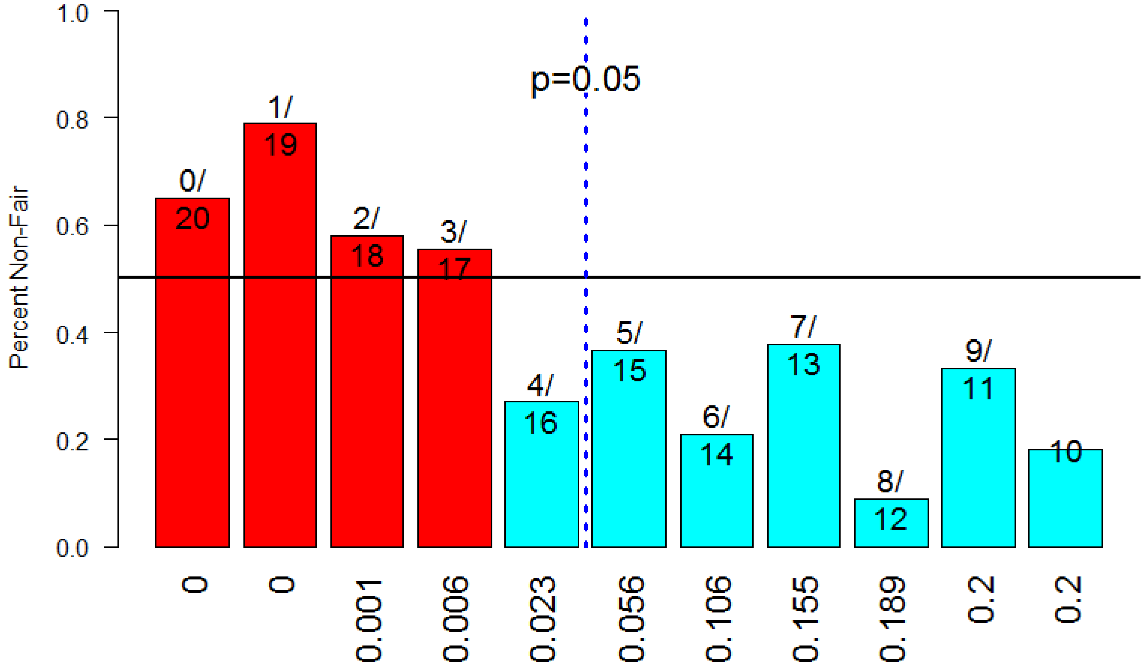

Figure 5.2: Spinning Spinners 2.0: Proportion Called Non-Random with Complementary Conditions Summed Together

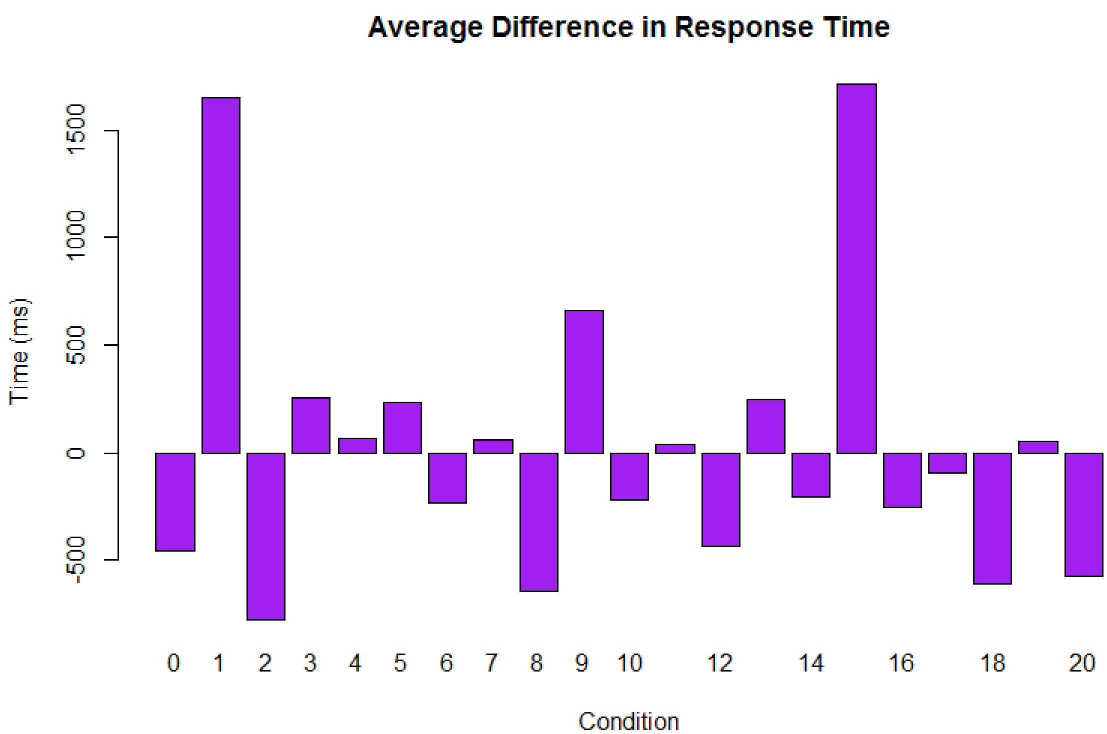

Figure 5.3: Spinning Spinners 2.0: Average Difference in Response Time by Condition

Chapter 6

Clustering Vs. Proportional Bias

"Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable." -Bobby Bragan

6.1 Introduction

After having established participants subjective threshold, and in doing so showing their sensitivity to proportion differences, it was decided to look at the effect of the clustering bias. It is hard to compare the effect of proportion differences and the clustering bias from the Flipping Coins or Spinning Spinners experiments because clustering correlates with proportion differences. In other words, when there is a large proportion difference, there tends to be large clusters.

In order to study clustering you have to be able to measure it. The measurement of clustering that was chosen for use was the number of runs. A run is a sequence of items sharing a common value. In the case of Flipping Coins, a run is a sequence of "X"s or "O"s. An alternative way of describing it is that every time the coin flip result alternates a new run is counted. This definition implies that if there are more runs there is a lot of alternation; also, if there are fewer runs then there is less alternation and more clustering. For example, this sequence: TTHHHHTTHT has five runs and this sequence: THTHTHTHTH has ten runs even though they both contain the same number of heads and tails. From that example, we can see that fewer runs means more clustering. In the extreme, a coin could have only two runs such as this: TTTTTHHHHH.

Initial analysis of the clustering bias was done by examining the Flipping Coins data from Chapter 2 for the influence of runs on judgements of randomness while accounting for the proportion differences. This analysis found a very modest effect of clustering and no effect of the over-alternation bias.[1] It was decided to design an experiment to control for runs and proportions to test the effect of clustering biases on judgements of randomness. The experiment used was a manipulation on the Flipping Coins experiment. This experiment allowed the comparison of the proportion differences and the clustering bias by using two levels of proportion and three levels of runs. It was also chosen because individual item presentation is also a relatively novel means of presentation when studying the clustering bias.

6.2 Method and Procedure

6.2.1 Participants

Twenty-four University of Lethbridge undergraduate students volunteered for this experiment and all of them were compensated with course credit for their efforts. One participant's data was removed from analysis because he responded yes to every sequence and said that he did so because he "couldn't find any palindromes". There was no mention of palindromes in any of the instructions.

6.2.2 Materials

LiveCode 8.0.1 Community Edition was used to create the program for the experiment. iMac computers were used but no features were used that are specific to any type or brand of computer.

[1] Discussed in Section 1.3

6.2.3 Procedure

The same general procedure as was described in Section 2.2.3 was used with a few differences:

- All participants received the three training phases at the beginning of the test.
- All participants were tested on forty-eight series of coin flips; there were six different testing conditions with eight replicates for each condition.
- The six different testing conditions were formed by two crossed factors, runs and "X" count. Runs had three levels; high (With 31-37 runs), medium (with 24-28 runs), and low (with 13-20 runs). Both the high and low condition would fail the runs test of an average coin. "X" count had two levels: 18 and 25.
- The conditions were denoted by *runs:"X" count*; so condition *low:18* had a low number of runs and 18 "X"s.[2]
- Sequences of coins were produced in the same way as Section 2.2.3 for the "X" count factor and then randomized or re-randomized until the runs factor was satisfied.

6.2.4 Analysis

LibreOffice and R (R Core Team, 2015) within the RStudio (RStudio Team, 2015) environment were used for all of the analysis.

[2]In reality, it was not always 18 "X"s in the 18 "X" count level. At the time the coin was produced it was randomly determined whether "X" or "O" would be used to fill up to the "X" count level. This means that about one-half of the time in the 18 "X" level there would be 18 "O"s and 32 "X"s.

6.3 Results

Figure 6.1 is an interaction plot showing the number of coins called non-random as a function of level of runs with separate lines for proportion differences. We can see major effects from proportion differences, a definite clustering bias, but no over-alternation bias. We can also see that there is no interaction between proportion differences and runs.

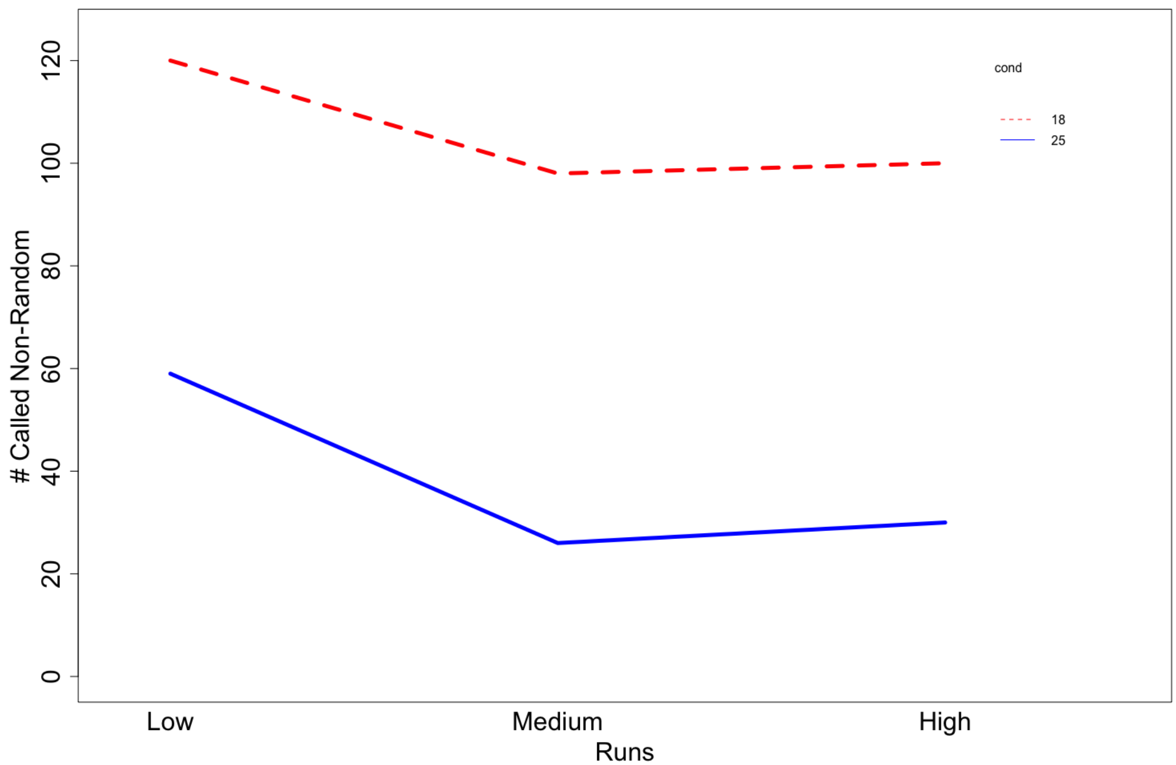

Figure 6.1: Clustering Vs. Proportion Interaction Plot

Tables 6.1, 6.2, and 6.3 are all pairwise comparisons across all conditions using a Pooled SD and a Bonferroni Correction. There is a significant proportion effect and a significant difference between the low runs condition and the other two runs conditions. There is a significant difference between 25L:25M but noticeably no significant difference between 18L:18M[3]. This indicates a clustering bias in the 25 "X" condition but not in the 18 "X" condition; or at least the

[3]When using other correction methods besides Bonferroni sometimes this comparison is significant. The Bonferroni correction was used in order to be conservative with multiple comparisons.

bias is less pronounced.

	18
25	<0.00001

Table 6.1: Two-Tailed P-Values for "X" Conditions Pair-Wise Comparisons with Bonferroni Correction and Pooled SD

	H	L
L	6e-04	
M	1	9e-05

Table 6.2: Two-Tailed P-Values for Runs Conditions Pair-Wise Comparisons with Bonferroni Correction and Pooled SD

	18H	18L	18M	25H	25L
18L	0.3035				
18M	1	0.15995			
25H	<0.00001	<0.00001	<0.00001		
25L	3e-05	<0.00001	1e-04	0.01159	
25M	<0.00001	<0.00001	<0.00001	1	0.00198

Table 6.3: Two-Tailed P-Values for X:Runs Interaction Conditions Pair-Wise Comparisons with Bonferroni Correction and Pooled SD

6.4 Discussion

The clustering bias found in this experiment is significant but smaller than previously found. Figure 6.1 shows the clustering bias to have nearly the same effect at two very different levels of proportion differences; however, pairwise comparisons show a slight difference. If the clustering bias was as strong as has been reported in the literature, then the low level of runs in the Low:25 condition should have resulted in many more judgements of non-randomness than was found.

Also, if expected levels of clustering[4] should produce the clustering bias, as described by Gilovich et al. (1985) and others, then the medium levels of runs (those with the expected

[4]Assuming truly random sequences

amount of runs/clustering) should be judged non-random more often than the high levels of runs (those with too little clustering). In these results there is no significant difference between medium and high levels of runs and any minimal difference is to judge those with high levels of runs non-random more often. It can then be concluded that when the coins had the expected levels of runs there was no discernible clustering bias. It can also be seen that over-alternation did not cause the participants to more favourably judge coins as random, as would be expected by the over-alternation bias. It is, however, notable that participants were not sensitive to the non-randomness in high runs conditions, although they were sensitive to the non-randomness in the low runs condition. This discrepancy could be termed an over-alternation bias.

A possible explanation for these results is that the seeing the entire sequence of events all at once, as is usually done in past experiments, draws the participants attention to the clusters where as if a participant sees them one at a time then clusters are less influential because they are harder to “clump” together. Additionally, it seems as though participants are making judgements of randomness based on a proportion difference unless there is clearly too much clustering. The idea that there is this two step decision process would explain why the lines on the interaction plot are so parallel and that the lines only significantly change when there is a lot of clustering compared to the other levels of clustering. It also explains why there was a significant clustering bias in the 25 “X” condition but not in the 18 “X” condition, because the proportion differences would account for most of those decisions before considering clustering. It is also an idea that was verbally expressed by a participant following his participation in the experiment.

These results are clearer, but similar to the results from the initial clustering bias analysis done on the data from the Flipping Coins experiment in Chapter 2. They also align with some of the conclusions that will be made in Chapter 7 and described in Chapter 8.

Chapter 7

Redundancy Analysis

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

(R. A. Fisher, 1938)

7.1 Introduction

As stated in Section 1.4, a redundancy analysis is able to detect patterns forming at different levels or orders in the stimuli that the participants would be responding too. Redundancy ranges from 0 to 1. If redundancy is equal to 0 then the sequence is perfectly random and completely unpredictable. If redundancy is equal to 1 then the sequence is perfectly predictable and non-random. How redundant a stimulus is can drastically change between orders of redundancy so many levels will be analyzed.

7.2 Redundancy Equations

In the context of the Flipping Coins experiment, zero order redundancy is given by:

$$R_0 = 1 - \frac{-\sum_{i=1}^{m} p_i log_2(p_i)}{log_2(n)} \tag{7.1}$$

where m is the number of unique symbols in the sequence, n is the total number of possible unique symbols, p_i is the proportion of symbols in the sequence that are symbol i, and a symbol is one character long.

With the Flipping Coins data the zero order redundancy will reduce to:

$$R_0 = 1 - \frac{-(p_{(X)}log_2(p_{(X)}) + p_{(O)}log_2(p_{(O)}))}{log_2(2)} \tag{7.2}$$

because there are only two possible symbols “X” and “O”.

First-order redundancy is the same as Equation 7.1 except symbols are two characters long so for the Coin Flipping data the first order redundancy is given by:

$$R_1 = 1 - \frac{-(p_{(XX)}log_2(p_{(XX)}) + p_{(XO)}log_2(p_{(XO)}) + p_{(OX)}log_2(p_{(OX)}) + p_{(OO)}log_2(p_{(OO)}))}{log_2(4)} \tag{7.3}$$

7.3 Examples

To illustrate the point consider the sequence XOXOXOXOXO: It has 5 “X”s and 5 “O”s at the zero-order so:

$$R_0 = 1 - \frac{-((5/10)log_2(5/10) + (5/10)log_2(5/10))}{log_2(2)} = 1 - \frac{-(0.5*(-1) + 0.5*(-1))}{1} = 0$$

It also has 0 “XX”s, 5 “XO”s, 4 “OX”s, and 0 “OO”s at the first order so:

$$R_1 = 1 - \frac{-((0/9)log_2(0/9) + (5/9)log_2(5/9) + (4/9)log_2(4/9) + (0/9)log_2(0/9))}{log_2(4)} = 0.504$$

So we can see that the sequence is considered perfectly random at the zero-order because it had equal proportions but very non-random at the first order. The logic is relatively easy to extend to the second-order and beyond but the computations involved become heavy in long sequences.

Here is the calculation for the second-order of the example sequence: it has 0 “XXX”, 0 “XXO”, 4 “XOX”, 0 “XOO”, 0 “OXX”, 4 “OXO”, 0 “OOX”, and 0 “OOO”.

$$R_2 = 1 - \frac{-((4/8)log_2(4/8) + (4/8)log_2(4/8))}{log_2(8)} = 0.667$$

As another example, this sequence O, O, O, O, X, O, X, O, X, X, X, O, X, X, X, X, O, X, X, X, X, X, O, O, O, O, X, X, O, O, O, O, O, O, X, X, O, O, X, O, O, O, X, X, X, X, O, X, O, X has $R_0 = 0$, $R_1 = 0.008$, $R_2 = 0.018$, and $R_3 = 0.026986$.

7.4 Analysis

A redundancy analysis was done on all of the data in previous chapters but only the analysis done on the original Flipping Coins experiment, in Chapter 2, will be presented. This choice was made because the results are clear, easy to see, and fairly representative of the other results.

For every sequence of coin flips presented, the 0-6th order redundancies were calculated. After the 6th order redundancy calculations, the computational demand was too great for the calculation to be completed. The data were sorted according to proportion differences (putting complimentary conditions together) and two-tailed unprotected Welch’s T-tests were preformed comparing the redundancies of those coins that were declared non-random by participants vs those coins that were declared random by participants. The zero-order redundancy was excluded from the t-tests because the control for proportion differences made that t-test unnecessary. The p-values of these t-tests are in Table 7.1. One tailed tests could have been considered because it is almost universally true that the coins considered non-random have higher scores on all orders of redundancy than those coins considered random but two-tailed tests were used to be at least somewhat conservative.

Tables 7.2 and 7.3 report the test statistics and degrees of freedom, respectively, for each of the Welch’s t-tests that were run.

Condition	1^{rst} Order	2^{nd} Order	3^{rd} Order	4^{th} Order	5^{th} Order	6^{th} Order
35/15	0.2715883	0.5676168	0.8345108	0.6679317	0.4577437	0.7369681
32/18	0.7529124	0.9844284	0.9026671	0.7584967	0.4860531	0.4463413
30/20	0.5278546	0.4066721	0.9853116	0.7467406	0.9647907	0.6821652
28/22	**0.0098926**	**0.0025635**	**0.0038670**	**0.0343632**	0.0898670	0.1585878
25	0.6513840	0.2795118	0.1024935	**0.0306350**	**0.0214978**	**0.0265368**

Table 7.1: Two-Tailed P-Values for Welch's T-Tests in Redundancy Analysis

Condition	1^{rst} Order	2^{nd} Order	3^{rd} Order	4^{th} Order	5^{th} Order	6^{th} Order
35/15	1.10340	0.572710	0.209230	-0.42963	-0.744000	-0.33633
32/18	-0.31508	0.019533	-0.122390	0.30772	0.697420	0.76248
30/20	-0.63208	-0.831090	-0.018429	0.32331	-0.044187	-0.40997
28/22	-2.62720	-3.082800	-2.948700	-2.14170	-1.710500	-1.41940
25	-0.45482	-1.094300	-1.664500	-2.22360	-2.373300	-2.28970

Table 7.2: T-Statistics for Welch's T-Tests in Redundancy Analysis

Condition	1^{rst} Order	2^{nd} Order	3^{rd} Order	4^{th} Order	5^{th} Order	6^{th} Order
35/15	154.020	165.500	178.250	197.110	201.780	202.71
32/18	311.220	317.850	317.260	319.990	316.220	317.33
30/20	277.050	265.580	250.380	242.430	247.510	258.62
28/22	105.470	115.820	114.790	113.110	114.930	111.42
25	45.833	46.004	48.291	50.941	50.201	47.37

Table 7.3: Degrees of Freedom for Welch's T-Tests in Redundancy Analysis

7.5 Discussion

Accepting that the t-tests preformed were unprotected against multiple comparisons, it is at least suggestive that the significant differences in redundancies between subjectively fair and non-fair coins were only within the approximately equal proportion conditions. This result suggests that the participants were in part basing their judgements of randomness on pattern recognition but only significantly so for the conditions that were approximately equal in element proportions. One possible explanation is that participants had at least a two-fold decision process where they would check whether the proportions were reasonable; then, if no major differences were found consider implicit pattern recognition. This suggestion seems reasonable given the results found by Jamieson et al. (2016). Their analysis of artificial grammar experiments found that when participants were given a discrimination task, and they could not consciously find reasonable criteria to base judgements on, they relied upon the implicit pattern recognition tested for by redundancy analysis. In the Flipping Coins experiments, proportions are a reasonable criterion on which to base decisions but only when they are obviously different.

Chapter 8

General Discussion

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." (Tukey, 1962)

8.1 The 0.05 Threshold

8.1.1 A Lack of Statistics

The reader may or may not have noticed a general lack of inferential statistics being used in the analyses of Chapters 2 to 5 and this was a conscious decision on the part of the author. Having done this work as part of the requirements for an Applied Statistics degree (Concentrating in Psychology) there is an understanding of the role of statistics but also a concern over testing a standard by its own merits. It was thought better to relay the data in a more basic form to the reader and allow them to make their own decisions of its merit. Many of the possible inferential tests have been completed and they lead to the same basic conclusions, so nothing seemed to be lost by excluding some inferential assumptions.

8.1.2 What Does it Mean?

No matter what the origins are of the 0.05 threshold, this research suggests that it is not an arbitrary standard. Participants were clearly able to discriminate, to some extent, the probabilities of the coin flips or spinner spins and made judgements of randomness based on those discriminations. There is also no reason to believe the participants responded to any element (such as "X"s or "O"s) in the presented sequences different from the other elements because of the clear symmetry in responses across all variations of the test. Because there was no bias toward any specific elements, summing complimentary conditions together is appropriate. When the arithmetic mean, of all the threshold probabilities shown in the figures that summed complimentary conditions together (Figures 2.4, 3.2, 4.7, and 5.2), is calculated we find that the value comes out to 0.026 (or if the median is preferred that comes out to 0.020). That approximation seems relatively close to the 0.05 threshold.

It seems possible that the 0.05 threshold has been able to stand up to its opponents so well because it has the merit of generally making subjectively reasonable decisions. It is possible that 0.05 is an approximation of the value where participants, students, and researchers alike become suspicious of supposedly random events. If true, this idea gives new meaning to inferential testing. It means that results that are deemed statistically significant would generally be noticed by human observers if they were able to attend directly to the phenomena under examination. It also means that inferential testing is also liable for many of the same errors that are attributed to human error, particularially the rate of errors made. It means inferential testing is only about as skeptical as people are and that is often not skeptical enough.

The idea that 0.05 is close to the subjective threshold could also make sense in an evolutionary perspective. Experience has shown that 0.05 is a relatively good balancing point between making type I and type II errors for many situations. It would give fitness advantages to be able to limit the number of errors in decision making. It could be that people have tended to learn to make decisions based on a threshold that limits errors, such as perhaps the 0.05 threshold. This line of reasoning is of course extremely speculative.

So to those who get caught up about the use of the 0.05 significance level because it's "arbitrary" and say there is no reason to use it – I say, here's a reason.

8.2 The Clustering Bias

In Chapter 6 a significant clustering bias was found but only once there was more clustering than would be expected by random chance. This result was different than what has been published in much of the literature. So the basic conclusion, as laid out in Section 6.4, is that there was a significant clustering bias but one that was smaller and less influential than previously found. It is believed this difference is because of the "one-at-a-time" method of presentation employed. If that is true, then the large clustering biases reported in past literature may be more a result of stimulus presentation and perception, rather than them being a true bias towards calling clusters non-random.

The bias was also more pronounced in the 25 "X" condition leading to the idea that proportion differences may affect decisions of randomness more than clustering. This conclusion also falls in line with the discussion from Section 7.5 where it was hypothesized that participants were attending to proportion differences; and then when no decision could be reached, considering pattern recognition as a secondary criterion. Clustering is a form of patterning so, based on the findings in Chapter 7 and this hypothesis, it makes sense that there would be a significant clustering effect in the 25 "X" condition but not in the 18 "X" condition.

The important conclusions to draw here are that more research is needed to understand peoples' perceptions of randomness and that the clustering bias is not nearly as clear cut as it has been made out to be.

8.3 Redundancy

Information theory, and specifically redundancy analysis, has been a useful tool to psychologists for many years but it is often under utilized. Many phenomena and tasks can be simplified and equated to one another by an application of information theory. For example, the tasks examined by Jamieson et al. (2016) can be compared to the tasks presented here. A set of seemingly random stimuli is presented and then the participant is asked to give a judgement. In the studies presented in this work, the requested judgement was of sequence randomness; in Jamieson et al. (2016) the judgement was of artificial grammaticality. In both studies the problem essentially comes down to whether the participant feels as though the sequence fits their expectations of what it should look like or does not. In order to have expectations to compare against, the participant must have some belief of patterning, or non-patterning, in the sequences. Redundancy analysis is sensitive to those patterns and is able to find what type of patterns are being responded to in both of those situations.

"I keep saying that the sexy job in the next 10 years will be statisticians." - Hal Varian

References

Alberoni, F. (1962, April). Contribution to the study of subjective probability. I. *Journal of General Psychology*, *66*, 241–264.

Bar-Hillel, M., & Wagenaar, W. A. (1991, December). The perception of randomness. *Advances in Applied Mathematics*, *12*(4), 428–454.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Bross, I. (1971). Critical Levels, Statistical Language and Scientific Inference. In *Foundations of Statistical Inference* (Godambe VP and Sprott ed.). Toronto: Holt, Rinehart & Winston of Canada, Ltd.

Chatfield, C. (1991). Avoiding Statistical Pitfalls. *Statistical Science*, *6*(3), 240–252.

Coase, R. H. (1982). *How should economists choose?* American Enterprise Institute.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Cohen, J. (1994). The earth is round ($p<.05$). *The American psychologist*, *49*(12), 997–1003.

Connelly, L. M. (2014, April). Statistical and Clinical Significance. *Medsurg Nursing*, *23*(2), 118–9.

Cowles, M., & Davis, C. (1982a, July). Is the .05 level subjectively reasonable? *Journal of Behavioural Science*, *14*(3), 248–252.

Cowles, M., & Davis, C. (1982b, May). On the origins of the .05 level of statistical significance. *American Psychologist*, *37*(5), 553–558.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301–318.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*(33), 503–513.

Fisher, R. A. (1938). Presidential Address. *Sankhy: The Indian Journal of Statistics (1933-1960)*, *4*(1), 14–17.

Fisher, R. A., Sir. (1970). *Statistical methods for research workers* (14th, rev. and enl. – ed.) (Nos. Book, Whole). New York: Hafner.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314.

Green, D. R., & Afima. (1982, January). Testing Randomness. *Teaching Mathematics and its Applications*, *1*(3), 95–100.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350.

Jamieson, R. K., Nevzorova, U., Lee, G., & Mewhort, D. J. K. (2016, March). Information theory and artificial grammar learning: inferring grammaticality from redundancy. *Psychological Research; Heidelberg*, *80*(2), 195–211.

Lopes, L. L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 392–400.

Lynch, J., & Stuckler, D. (2012, December). In God we trust, all others (must) bring data. *International Journal of Epidemiology*, *41*(6), 1503–1506.

Montgomery, D. C. (2013). *Design and analysis of experiments* (Eighth edition ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*(2), 330–357.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276.

RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc.

Salsburg, D. S. (1985). The Religion of Statistics as Practiced in Medical Journals. *The American Statistician*, *39*(3), 220–223.

Sanderson, Y. B. (2009, July). Effective generation of subjectively random binary sequences. *Advances in Applied Mathematics*, *43*(1), 1–11.

Sun, Y., & Wang, H. (2010, December). Perception of randomness: On the time of streaks. *Cognitive Psychology*, *61*(4), 333–342.

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

Wagenaar, W. (1970a). Appreciation of conditional probabilities in binary sequences. *Acta Psychologica*, *34*, 348–356.

Wagenaar, W. (1970b). Subjective randomness and the capacity to generate information. *Acta Psychologica*, *33*, 233–242.

Wasserstein, R. L., & Lazar, N. A. (2016, April). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133.