

A META-ANALYSIS OF
RANDOMNESS
IN HUMAN BEHAVIORAL RESEARCH

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

In

The Department of Mathematics

By
Summer Ann Armstrong
B.S., Southeastern Louisiana University, 2000
August 2004

Acknowledgments

The idea for this thesis was suggested by Professor James J. Madden, due to my interest in probability. It is with much gratitude that I thank Dr. Madden for his tireless efforts, advice, and guidance. It is also a pleasure to thank Dr. George Cochran and Dr. Jason Hicks for their time and valuable input.

This thesis is dedicated to my husband, Dr. Michael Nuccitelli and to my parents, Kevin and Debra Armstrong for their steadfast encouragement and support.

Table of Contents

| | |
|---|-------------|
| ACKNOWLEDGMENTS..... | . ii |
| ABSTRACT..... | iv |
| CHAPTER 1. INTRODUCTION..... | 1 |
| CHAPTER 2. PROBABILITY THEORY..... | 4 |
| 2.1 The Sample Space and The Event Space..... | 5 |
| 2.2 Probability Measure..... | 6 |
| 2.3 The Discrete Random Variable and Its Distribution Function..... | 10 |
| CHAPTER 3. RANDOMNESS..... | 12 |
| 3.1 The Mathematical Approach..... | 12 |
| 3.2 The Statistical Approach..... | 15 |
| 3.3 The Psychological Approach..... | 16 |
| CHAPTER 4. CAN PEOPLE BEHAVE RANDOMLY? | 18 |
| 4.1 Budescu..... | 18 |
| 4.2 Neuringer..... | 20 |
| 4.3 Falk and Konold..... | 22 |
| CHAPTER 5. STATISTICAL TESTS FOR RANDOMNESS..... | 26 |
| 5.1 Parameters and Statistics..... | 26 |
| 5.2 An Example Relating to Hypothesis Testing..... | 26 |
| 5.3 Hypothesis Testing..... | 27 |
| 5.4 Frequency Test for Randomness..... | 29 |
| 5.5 Runs Test For Randomness..... | 30 |
| REFERENCES..... | 35 |
| APPENDIX: MATHEMATICA COMMANDS..... | 37 |
| VITA..... | 40 |

Abstract

This work analyzes the concept of randomness in binary sequences from three different perspectives: mathematically, statistically, and psychologically and examines the research on human perception of randomness and the question of whether or not humans can simulate random behavior. Generally, research shows that human subjects have great difficulty producing random sequences, even when they are instructed and motivated. We survey some of the literature and present some leading theoretical proposals. Finally, we present some basic statistical tests that can be used to evaluate randomness in a given binary sequence.

Chapter 1. Introduction

What is randomness? Most people have some intuitions about what it is. Many would cite examples of randomness, like the tossing of a coin or the occurrence of mutations in a gene. People behave and events in nature occur randomly on a regular basis. However, research has shown that when humans are instructed to act randomly, generally they are unsuccessful at this task. Usually, they are not capable of distinguishing accurately between random and non-random data—indeed, their perceptions appear to be biased in certain specific ways. Perhaps, this inability to randomize lies in misconceptions about randomness, or it could be that scientists have yet to provide an experiment engaging enough for humans to demonstrate some untapped randomization skills.

Much research has been done since 1953 on the perception that humans have of randomness and ability of human subjects to generate random sequences. In 1972, W. A. Wagenaar surveyed the literature available on studies involving generation of random sequences by human subjects. Of the thirteen studies that he examined, only one of the experiments reported that the human subjects were good randomizers.

A variety of measures have been taken and adjusted to broaden the experimental conditions in hopes of determining the precise extent to which humans are able to perform randomization tasks and the respects in which they regularly fail. Studies of randomization by humans can be divided into two categories: production and judgment. In experiments, participants were either instructed to produce a sequence that was later tested for randomness, or they were asked to judge sequences already available as random or non-random. As an example of a production study, Bakan (1960) had human subjects produce three hundred responses of “heads” and “tails.” They were told to try to produce a sequence that would resemble what would be produced by actually flipping a fair coin. As an example of a judgment study, Cook (1967) asked subjects to examine sequences of binary digits one hundred symbols long. They had to compare two at a time and then make judgments as to which was more patterned.

In such experiments, the kind of data subjects were asked to produce or examine varied. Often, the data was a sequence of symbols from some fixed alphabet. The size of the alphabet has ranged from two to twenty-six, and in different experiments it has consisted of letters, digits, head-tails, button pushing, or marked tokens (*e.g.*, cards marked with “X” or “O”). This means that the subjects, when constructing or judging a sequence, could have as little as two choices or as many as twenty-six. The two-choice alternative classically involved trying to mimic the tossing of a coin. In the case of twenty-six alternatives, subjects might have been instructed to pretend they were pulling letters of the alphabet from a box and recording the result. Despite the wide variety possible in choice of alphabet, most studies used a binary alphabet, and this paper will concentrate on this case.

Instructions to subjects have included asking them to produce symbols from the designated alphabet “as randomly as possible” or in imitation of some known random process. In judgment experiments, two or more sequences might be given simultaneously and subjects asked to determine which is most random, or subjects might observe a sequence as it is produced and be asked to decide periodically if it conforms to a random process. Some studies have used computer input or output devices as a means for subjects to generate or view data. Subjects have been required to press numeric or alphabetic keys, or they have reviewed sequences presented on a computer screen.

Some production experiments have been subject-paced, while others demand responses at intervals from 0.25 seconds to 4 seconds. Weiss (1964), used a one second and two second pilot light to signal his subjects to respond by pressing one of two buttons. Baddeley (1966) conducted an experiment with paced and un-paced conditions, finding that randomness decreased as pace increased.

An overwhelming number of experiments indicate that humans fail to produce random responses even when instructed and motivated to do so. Many studies report a tendency for subjects to produce sequences with negative recency, *i.e.*, too many alternations. Judgment studies are consistent in indicating that human subjects perceive sequences with too many alternations (THTHTH...) as random. Thus, some researchers have concluded that negative recency is the culprit explaining sub-optimal randomization in humans. However, a small number of studies have noticed “positive bias” (or positive recency) in data. This is a sequence with too many repetitions (TTTTT... or HHHHH...). Budescu (1987) proposed that human randomizing behavior could be modeled as a Markov process. This accommodates both negative and positive recency. Budescu claimed support for this model in experiments with 18 subjects.

Although the research mentioned so far strongly suggests that human subjects are sub-optimal randomizers, a few papers have supported human subjects in their ability to behave randomly or have attempted to explain human behavior in randomization tasks without denying the possibility that under the right circumstances humans might be more successful. One such case is Ross (1955). His subjects stamped cards with one of two symbols and were instructed to arrange them in a random order. His results did not support the idea that human subjects over-alternate. In another example, Cook (1967) concluded that the subjects were successful in recognizing the bias in certain sequences.

Diener and Thompson (1985) addressed the problem of how an observer decides whether a series was generated by a random process. In this study, subjects viewed sequences of “heads” and “tails” and were asked to judge which ones had been created by a random process similar to the tossing of a coin. Diener and Thompson concluded that rather than directly recognizing a sequence as representative of a random process, subjects decided that a sequence was random only after eliminating alternative nonrandom sequences.

Neuringer (1986) showed that subjects, who do not possess randomization skills at first, may learn to produce random-like behavior if provided with proper information training.

Following this introduction, we begin by providing a concise discussion of random experiments and related concepts including discrete random variables and their distributions. The subsequent chapter looks at a random sequence from three different perspectives: mathematical, statistical, and psychological. Next, we survey some of the literature involving human beings' concepts of randomness and whether or not they can mimic randomness successfully. The survey specifically deals with the works of Budescu, Neuringer, and Falk and Konold. Finally, we describe some statistical tests for randomness, and illustrate them through computer simulations using *Mathematica*.

Chapter 2. Probability Theory

This chapter is devoted to providing the reader with an overview of Probability Theory. Our primary interest involves determining how well humans can simulate random experiments. The foundation of the research conducted on randomness studies involving human subjects, is the concept of a random experiment. Thus, it is crucial to describe this concept and its characteristics, which is the primary focus of this chapter. Another aspect of this section is to recognize that standard randomness tests depend on using probability distributions, associated with certain judiciously selected random variables. We will specifically deal with the discrete random variable and its associated probability distribution.

We begin with the idea of a *random experiment*. Because a random experiment is a direct abstraction from situations with which most people have plenty of experience, expositions of probability theory often begin here. This is the connection between the formal calculus of probability theory and real-world applications. Basically, a random experiment is one in which the outcome cannot be predicted with certainty. An important assumption concerning a random experiment is that it can be repeated indefinitely under conditions that are essentially the same. The significance of this assumption stems from the idea that probability theory analyzes the long-term behavior as the experiment is replicated.

Examples of random experiments—or concrete realizations of the general concept—include:

- tossing a coin once
- tossing a coin 100 times,
- throwing two dice,
- selecting a card from a shuffled deck,
- selecting a random sample of people from a larger population,
- selecting 25 people and counting how many are left-handed.

Each of these examples has three important aspects: a sample space, an event space, and a probability measure, which we now define. Numerous examples of each concept will be given in the next section.

A **sample space**, commonly denoted by S , consists of the set of all possible outcomes. For example, when tossing a coin, the possible outcomes are head (H) or tail (T). (If anything else occurs, we have not successfully performed the experiment.)

An **event space** (E^*), contains all subsets of the sample space. In other words, an event E occurs if the observed outcome s is an element of E ($s \in E$). If a coin is tossed 100 times,

some examples of events are the following: 1) more heads than tails occur, 2) exactly 50 heads occur, 3) a head is obtained on the 5th toss.

The **probability** of an event is a measure of how likely it is for the event to occur.

Suppose we conduct an experiment where we throw two dice, and we are interested in the probability that both dice show an even number. When the two dice are thrown, there are 36 total possibilities, all with equal probability. If we consider the event that both dice show an even number, we have $\{(2,2),(2,4),(2,6), (4,2),(4,4),(4,6), (6,2), (6,4), \text{ and } (6,6)\}$ as the 9 ways that this could happen. Thus the probability of this occurring is $\frac{9}{36} = \frac{1}{4}$.

A **probability measure** (or distribution), P , is a real-valued function defined on the collection of events. It must satisfy certain criteria, but we will postpone describing them until Section 2.2.

2.1 The Sample Space and The Event Space

Recall that the sample space of a random experiment is a set S that includes all possible outcomes of the experiment. (A set is just a collection of objects, where the objects are referred to elements of the set.)

The sample space consists of exactly the set of possible outcomes. It serves as the universal set for all questions concerned with the experiment.

Example 2.1.1. Suppose one throws a standard die and records the outcome. Then, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$, the set of possible outcomes.

Example 2.1.2. Suppose someone tosses a coin in the air and observes the result. Thus, the sample space $S = \{Tail, Head\}$.

In creating a sample space, we need to decide on an appropriate level of detail. Whatever those things are that we choose to call the outcomes, they become the irreducible atoms of any further description of what can happen. If we think about shooting an arrow in the air, we might care about whose land it falls on, and we might never have any need to know more than that. In this case, the set of nearby estates could become the sample space. But we might have much more detailed concerns—to the extent, possibly, of needing to know, to the nearest inch, the distance of the landing point from three fixed markers.

Sample spaces can be finite or infinite. An experiment involving the tossing of a finite number of coins or the throwing of some fixed number of dice has a finite sample space. An infinite sample space occurs in an experiment where a real-number measure is taken: *e.g.*, measuring heights of people. Infinite sample spaces also occur where the outcomes are associated with an integer or a whole number, *e.g.*, tossing a coin until a head occurs and counting the number of tosses required.

Any subset of the sample space of an experiment is referred to as an “event”. This use of the word “event” is quite distant from the common meaning. The idea is that when we perform an experiment, we may not be interested in the particular outcome, but a particular property that an outcome may have. Now, each property of outcomes corresponds exactly to a subset of sample space—namely the set of all outcomes with the said property. Every time that the experiment is run, a given event E occurs provided the outcome of the experiment is an element of E . If the outcome of the experiment is not an element of E , then it does not occur. Note that the sample space S is an event because by definition it always occurs. At the other end, the empty set (\emptyset) or the set with no elements, is also an event. By the same reasoning that the sample space S is an event that always occurs, the empty set is an event that never occurs. The event space (E^*) is the set of all events that can result once an experiment is run. In the case of a finite sample space, event space is the set of all subsets of sample space.

Example 2.1.3. For instance, consider an experiment where a coin is tossed three successive times. Let $S = \{HHH, TTT, HTH, THT, HHT, TTH, HTT, THH\}$ be the associated sample space. Perhaps, the interest lies in the event “the number of heads exceeds the number of tails.” For any outcome of this experiment, we can easily determine whether this event does or does not occur by counting the number of heads. It is obvious that HHH, HTH, HHT, THH are the only elements of S corresponding to outcomes for which this event occurs.

The event, “the number of heads exceeds the number of tails,” occurs when the outcome is in the set $E = \{HHH, HTH, HHT, THH\} \subseteq S$. However, if the observations result in one of the other elements of S , then the event in question does not occur. Below are some other event possibilities given three successive coin tosses.

Description of Event and its Corresponding Subset of S

Second toss is heads – $\{HHH, HHT, THH, THT\}$

All tosses show the same face – $\{HHH, TTT\}$

Second toss is heads and the number of heads is exactly 2 – $\{HHT, THH\}$

Second toss is heads or the number of heads is exactly 2 – $\{HHH, HTH, THT, HHT, THH\}$

Example 2.1.4. Suppose we are presented with a deck of cards, instructed to shuffle them, select a card at random, and observe the selected card. Here, the associated sample space S is the set of all 52 cards—ace through king in all four suits. Suppose we are interested in the event “the card selected is either a diamond or a heart.” The elements of S in this event are the 26 diamonds and hearts. The probability of this event is $\frac{1}{2}$.

2.2 Probability Measure

Definition 2.2.1. A probability measure P for a random experiment is a real-valued function defined on the collection of events satisfying the following three axioms:

- i. $P(E) \geq 0$ for all E (where E represents an event)
- ii. $P(S) = 1$ (where S represents the sample space)
- iii. If E_1, \dots, E_j is a collection of disjoint events, then $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$

The first two axioms are straightforward. The first axiom just states the probability of each event is zero or greater. Axiom II states that the probability of the sample space S is 1. The third axiom, known as **countable additivity**, concerns a countable, disjoint collection of events. It states that the probability of a union of a finite or countably infinite collection of disjoint events is the sum of the corresponding probabilities.

Example 2.2.1. Suppose we toss a coin three successive times and record the results. Observe the possible outcomes:

$$S = \{HHH, TTT, HTH, THT, HHT, TTH, THH, HTT\}$$

We now describe a probability measure on this set.

Let $P(E) := \frac{1}{8} \bullet (\text{the number of elements of } E)$. We shall show that this P satisfies the axioms. It is clear that $P(E) \geq 0$ for all E . Also, $P(S) = \frac{1}{8} \bullet 8 = 1$. Finally, suppose that E_1, \dots, E_j are disjoint sets. Then the number of elements in the union of these events is equal to the sum of the number of elements in each individual set. This shows that Axiom 3 holds.

When would this be a reasonable probability measure? If the coin is fair, then we would expect each of the 8 one-element events to have equal probability. The second and third axioms then force the probability of each one-element event to be $\frac{1}{8}$. Finally, the third axiom forces us to use the definition above.

Of course, this is not the only probability measure we might want to consider. Suppose we had a biased coin that landed on heads only $\frac{1}{3}$ of the time. Then it would be reasonable to assign a probability of $\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)$ to the event $\{HHH\}$, $\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)$ to $\{THT\}$ and so on. In general, the probability of a one-element event is $\left(\frac{1}{27}\right)2^n$, where n is the number of occurrences of tails in the outcome in the event. The probability of an event with more than one element is found by adding together the probabilities of the one-element subsets of that event.

Example 2.2.2. To exhibit an experiment having an infinite sample space, suppose a coin is tossed until a head is observed. We designate the outcome of the experiment as the first time that a head turns up. Then, the sample space for this experiment is $S = \{\emptyset, 1, 2, 3, \dots, N, \dots\}$. Review the possibilities that follow.

| <u># of tosses before head is observed</u> | <u>Probability that the first head is observed on the given number of tosses</u> |
|--|--|
| 1(<i>H</i>) | $\frac{1}{2}$ |
| 2(<i>TH</i>) | $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ |
| 3(<i>TTH</i>) | $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$ |
| 4(<i>TTTH</i>) | $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$ |
| ⋮ | ⋮ |
| $N(\text{TTT} \dots \text{TN} = H)$ | $\left(\frac{1}{2}\right)^N$ |
| ⋮ | ⋮ |

Now we have defined the sample space— $\{1, 2, 3, \dots\}$, and we have partially defined the probability measure, in that we have specified a probability for each one-element event. From the above table, it is obvious that the first axiom is satisfied. To show Axiom 2 holds, we know a sum exists because the partial sums form an increasing sequence which is bounded above, so the partial sums have a limit. This sum is a convergent series, which we know converges to 1. Finally, the last axiom is also clearly satisfied.

Thus, we have the necessary ingredients for a random experiment: the sample space, the event space and a probability measure. Together, these three items define a probability space (S, E^*, P) .

Example 2.2.3. (Birthday Problem.) This is a classic problem in probability classes. The gist of the birthday problem is as follows: Given a group of people, what is the probability of at least two people in the group having the same birthday? The phrase “having the same birthday” means that they celebrate their birthday on the same day. Thus, the year of the birth is omitted. We will ignore leap years and assume that birthdays are uniformly distributed throughout the year.

Suppose n people are chosen at random, and their birthdays are noted. Then, the sample space consists of 365^n possible outcomes. Because the birthdays are uniformly distributed and the people are chosen randomly, each of the outcomes is equally likely.

Thus, as before, each one-element event has the same probability—this being $\frac{1}{365^n}$. Our sample space being finite, the probability of an arbitrary event E is the sum of the probabilities of the one-element events in it. Thus $P(E) = \frac{k}{365^n}$, where k is the number of outcomes in E .

To see the reasoning of this example most clearly, we examine the probabilities one step at a time. There are 365 distinct birthdays for one person. If two people are chosen, there are 364 different ways that the second could have a birthday without matching the first. If three people are chosen, there are 363 different birthdays that do not match the other two. So, if three people are chosen at random, the probability of their birthdays all being distinct would be determined by calculating the following:

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} = 0.9918$$

So, let p_n denote the event that the birthdays of the n people are distinct. Thus,

$$p_n = \frac{365 \cdot 364 \cdot 363 \cdot \dots \cdot (365 - n + 1)}{365^n}$$

In general for the standard birthday problem, $p_n = \frac{365!}{(365 - n)! \cdot 365^n}$. Since the above formula

calculates the probability that the birthdays of the n people are distinct, $1 - \frac{365!}{(365 - n)! \cdot 365^n}$ calculates the probability of at least two people having the same birthday. Now, as mentioned earlier, the results of this problem are surprising.

Examine the table below.

Table 2.2.1

| Number of People of Chosen (n) | Probability of at least One Match ($1-p_n$) |
|------------------------------------|---|
| 10 | 0.1169 |
| 20 | 0.4114 |
| 30 | 0.7063 |
| 40 | 0.8912 |
| 50 | 0.9704 |
| 60 | 0.9941 |

What is interesting is the fact that if 40 people are chosen at random, the probability that at least two of those people from that sample will share the same birthday is as high as 89.12%! Most people would not think that the probability for this sample size would be that high. A graphic illustration depicting the probability of at least one match versus the sample size is shown in Figure 2.2.1

Probability of at Least One Match Versus the Sample Size

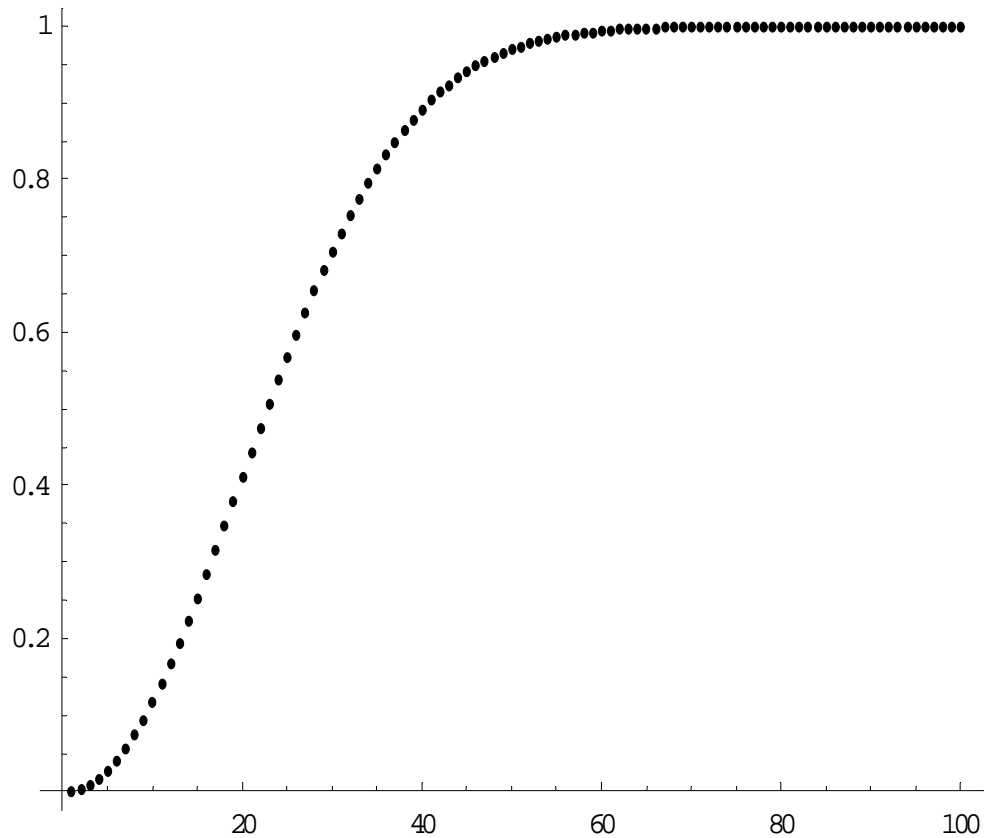


Figure 2.2.1

From the graph, if one selects 60 persons randomly, he/she can almost guarantee that at least two people will have the same birthday.

2.3 The Discrete Random Variable and Its Distribution Function

This section introduces the discrete random variable and its distribution function. It also presents some examples to illustrate these concepts.

Definition 2.3.1. Given a random experiment with a countable sample space S , a random variable, denoted by \mathbf{X} , is a function from S into another set \mathbf{W} [9, pg. 212].

The definition above allows a generalization to continuous sample spaces, but we will never need this level of generality in this thesis. Indeed, the only random experiments that are relevant to the problems we will address have finite sample spaces.

A random variable represents a measurement of interest in the context of the random experiment. A random variable \mathbf{X} is random in the sense that its value is dependent upon the outcome of the experiment. Moreover, it cannot be predicted with certainty before the experiment is run. Each time the experiment is run, an outcome $s \in S$ occurs, and a given random variable \mathbf{X} takes on the value $\mathbf{X}(s) \in W$.

Example 2.3.1. If an experiment is run where a coin is tossed until the first tail occurs, then the number of heads that results before the tail appears is a random variable. Call this variable X . If the sample space is $S = \{T, HT, HHT, HHHT, \dots\}$. Then, $\mathbf{X}(T) = 0$, $\mathbf{X}(HT) = 1$, $\mathbf{X}(HHT) = 2$, etc.

Example 2.3.2. Suppose the experiment is flipping a coin 100 times. The following are random variables:

- The number of heads (in an outcome)
- The number of alternations, *i.e.*, transitions from head to tail or tail to head
- The number of complete runs, where a complete run is a maximal uninterrupted sequence of the same face.
- The maximum length of a complete run.
- The average length of all the complete runs.

Definition 2.3.2. Let x be a number and let $\mathbf{X}: S \rightarrow W$ be a random variable on a sample space S . Then the event $\{s \in S \mid \mathbf{X}(s) = x\}$ is denoted $\{\mathbf{X} = x\}$. The function f defined by $f(x) := P(\{\mathbf{X} = x\})$ is called the probability distribution associated with the variable \mathbf{X} [9, pg. 213].

Chapter 3. Randomness

Randomness can be associated with the lawless, the pattern-less, the unpredictable, and the haphazard. These words are ambiguous and appeal more to feelings than to precise definitions. Randomness is a subtle concept that tends to elude attempts at scientific or mathematical precision.

This chapter will present and discuss three general ways that scientists have devised to sharpen the idea of randomness. Mathematical approaches attempt to define the property of randomness in infinite strings. As is common in mathematics, this removes the problem from the world of direct experience. Randomness becomes a property that an object in an abstract domain may have or may fail to have. We may reason about the objects, random and otherwise, that belong to that domain. However, none of our senses can interact with them in any direct manner at all.

The statistical approach, in contrast, skirts the issue of the essential nature of randomness. Its tactic is merely to ask, “What would suffice as good evidence of randomness?” This method focuses on the fact that though we may lack the ability to precisely define randomness, we nonetheless know enough about it to be able to recognize clues of its presence, when they are around.

The third approach is psychological. In this, we are even further from the question of the essential nature of randomness. Here, the question ceases to be “What is randomness?” and becomes, “What do people characterize as random? What do they accept as evidence of randomness?” There are two surprises here. The first is that people tend to have consistent ideas concerning randomness, classifying what they believe to be more random or less random roughly in the same way. The second is that their perceptions about randomness differ in subtle but uniform ways from the more rigorous ideas developed in the mathematical and statistical approaches.

3.1 The Mathematical Approach

One might suppose a random sequence could easily be defined as a pattern-less sequence, where all of the numbers in that sequence occur with equal frequency, and all subsequences pass statistical tests for randomness. There seem to be more difficulties involved though.

To further explain, suppose a binary sequence is constructed implementing the coin-tossing example, consisting of each possible outcome of successive tosses. The occurrence of “Heads” is represented by “1”; “tails” is represented by “0.” A reasonable question to ask would be, “When is a binary sequence random?” Is it possible to make a reasonable mathematical definition that really clarifies the meaning of “random”?

Suppose you perform three trials of tossing a coin twenty times. Examine the following outcomes:

11111111111111111111
01101010000010011110
00100001100010100000

The first result, twenty heads in twenty tosses, looks suspicious. The second and third appear rather random. However, the second sequence consists of the first twenty digits of the binary expansion $\sqrt{2} - 1$, and therefore was obtained by a process—root extraction. Thus, it fails to be random since it is clearly generated by a rule [16, pg. 48].

By probability theory, all three outcomes above have the same probability, $\left(\frac{1}{2}\right)^{20}$. So, why does it seem that the 2nd and 3rd sequences are random while the first sequence is not? (Note that the third string actually was obtained by tossing a coin.)

For these sequences, it is only possible to develop a perception of degrees of randomness, there being no abrupt distinction of the set of all sequences into random and nonrandom ones. Indeed, the idea that some finite sequences might be random, while others are not, leads to a problem. Given a finite binary sequence that is considered random, there is no cause to claim that the same binary sequence with one entry altered is not random. If in a finite sequence we can change one entry at a time without destroying randomness (if it is present), then we must also be able to change any two or any three or any number of entries whatsoever, still without making it non-random. But any sequence may be turned into any other by a finite number of changes. So every finite sequence is random, or none is.

Although real-world applications restrict one to finite encounters, the mathematical approach concerns infinite strings. The previous discussion suggests why this seems necessary. Every finite string is unique and special in some ways, and yet indistinguishable from other strings in probability.

Volchan (2002) reviews the most important mathematical approaches to randomness. He classified attempts to define randomness in three groups as follows:

- approaches based on stochastic-ness (or frequency stability)
- approaches based on incompressibility (or chaotic-ness) and
- approaches based on typicality.

All three proposals involved bringing to bear ideas related to algorithms and computability. The details become quite intricate. We will give an overview with only as much detail as we need for our purpose, which is to examine their bearing on psychological studies. We follow Volchan's exposition closely.

The notion of randomness as stochastic-ness originated with Richard von Mises. It was developed by Alonzo Church and Abraham Wald. Von Mises proposed that randomness could be modeled in a mathematically precise way by means of certain sequences of 0s and 1s, which he named “collectives.” He coined the terms *stochasticness* or *frequency stability*, which arose from the statistical regularity that he observed in random experiments such as coin-tossing, to name the defining property of a collective. Without attempting to provide full details, the condition for a sequence of 0s and 1s to be a collective is that the ratio of 0s to 1’s should approach the same limiting value in the sequence itself and in every subsequence of a certain, specific type.

Von Mises received many criticisms, and his arguments were considered inexact. Eventually, Kamke showed collectives do not exist unless appropriate restrictions (which von Mises failed to state) are placed on the subsequences. However, in 1937 Wald showed that collectives do exist if restrictions are placed on sets of admissible selections. By placing restrictions on the subsequences, this result opened the question of which subsequences should be required to meet the asymptotic condition on the ratio of 0s to 1s. In 1940, Alonzo Church proposed that in order to separate those sequences that are intuitively random, the set of admissible selections should consist of computable functions.

Church, and another mathematician Alan Turing, are known for the so-called *Church-Turing thesis*. The Church-Turing thesis is an idea in computer science stating that every effective computation or algorithm can be carried out by a Turing machine. Any computer program in any of the conventional programming languages can be translated into a Turing machine, and any Turing machine can be translated into most programming languages. Introduced in 1936 by Alan Turing, the Turing machine is an abstract model of computer execution and storage that gives a mathematically precise definition of an algorithm or other mechanical procedure. The concept of the Turing machine is based on the idea of a person executing a well-defined procedure by reading and writing symbols on a potentially infinite strip of paper. The person needs to remember one of a finite set of states, and the procedure is formulated in very basic steps.

Although the von Mises-Church-Wald proposal had some desirable traits, it revealed a severe flaw in 1939 when Jean Ville showed that collectives are not random enough by proving that there are collectives with a preference for 1s over 0s. In this sense, the cumulative number of 1s always exceeds the number of 0s. As a result, this first proposal to define a random sequence was unsatisfactory.

The second proposal originated with Ray Solomonoff, Andrei Kolmogorov, and Gregory Chaitin described randomness as incompressibility. The logic behind this idea is that a sequence is irregular or patternless if it cannot be described more efficiently than by giving the entire sequence itself. This is known as algorithmic complexity and declares a sequence random if no program other than the entire sequence itself can generate or describe it. Kolmogorov, Solomonoff, and Chaitin incorporated Turing machines in their definitions and theorems in support of their quest to define a random sequence. One of

Kolmogorov's propositions called an infinite sequence random if it had initial segments of high complexity, designating these sequences as incompressible. Unfortunately, Per Martin-Löf showed that no such sequence exists and proposed the third notion of randomness, typicality. In fairness to the second proposal, Chaitin along with several others were later able to prove that the incompressibility idea can be made consistent by placing suitable restrictions on the class of algorithms.

When we hear the word “typical,” we can intuitively think of something as common or ordinary. If we consider typical binary sequences, then we could say that these sequences are featureless. Through his research, Martin-Löf incorporates measure theory to develop his notion of a random sequence. Even though his theories contain some abstract ideas, they can be simplified in terms of the concept of an effective statistical sequential test for randomness. To further explain, it compares the sequence in question to an enumerable sequence. This enumerable test sequence looks for certain regularity properties that are considered incompatible with the sequence being random. Thus, a sequence is called Martin-Löf - random if it passes all of the effective sequential tests for randomness. In other words, Martin-Löf concluded that there is a universal sequential test that, if passed, defines a sequence as random [14, pg. 60].

To date, no serious flaw has been found with Martin-Löf's definition of a random sequence. Furthermore, Martin-Löf's notion of random sequences is mathematically consistent, making his argument the best candidate for the mathematical definition of a random sequence thus far.

3.2 The Statistical Approach

Bar-Hillel and Wagenaar suggest that we think of “randomness” as an unobservable property of a generating process [3, pg. 429]. A classic example would be the previously mentioned “coin toss”. This paragon fully demonstrates random behavior because we cannot predict the outcome of the individual coin tosses. If the coin is fair, the likelihood of obtaining a “head” is equal to that of obtaining a “tail,” and the occurrence of one outcome in a coin toss does not influence future outcomes. Thus, the proportion of heads (or tails) will converge to $\frac{1}{2}$ after a significantly large number of tosses.

This shows that though we may lack a precise definition of randomness that would be useful and interesting in an abstract domain, we nonetheless agree fairly well that a sample of output from a random generating process should have certain statistical properties. There should, for example, be about $\frac{1}{2}$ heads in a large sample. Similarly, there ought to be no predominance of either option among the flips that follow heads.

Suppose you ask a human subject to attempt to simulate a random experiment, say by providing a string of 100 0s and 1s each time you request it—for reasonable

compensation, of course. You might form the hypothesis that some abstract model of the production process provides an accurate account of what the subject is doing. For example, the abstract model might be a random experiment—*e.g.*, the flipping of a fair coin.

Now, given any random variable on the sample space, you can evaluate that variable at the string provided by the subject. If the model is not too complex, you can find the probability distribution for the variable that the model determines. From this, you can find the expected value of the random variable in the model. Also, you can determine the probability that the value of the variable at a random outcome departs from the expected value for the model by any given amount, say Δ . In other words, we can compute $P(|X - E_m| > \Delta)$, where E_m is the expected value in the model. By this computation we can judge whether the subject's data was likely to have been produced by a random experiment. Now, if this probability is very small when $\Delta = |X(\text{subject's sequence}) - E_m|$, then we take it as evidence against the model. This is the idea of hypothesis testing, which we will describe in more detail later.

3.3 The Psychological Approach

In general, people feel as though they understand what they mean when speaking of *randomness*. They communicate in daily affairs using their intuitive understanding of this term. Still, randomness, as we have seen, is a complex and subtle notion, and attempts to define randomness include complex philosophical or mathematical concepts. Is there some uniformity or agreement in the way people employ the notion? If so, does it agree with the technical conceptions we've described? This is a question of psychology – an important and non-trivial one.

Psychologists have been conducting studies for many years trying to pinpoint people's subjective sense of randomness. A majority of these studies examine generation of randomness by participants, particularly production involving two symbol types. A most prominent feature of their research shows that people identify randomness with an excess of alternation between symbol types. Participants do not perceive sequences that are typically random as random because to them, the runs appear to be too long, or they produce sequences containing too many short runs. This bias might be an expression of the "gambler's fallacy." This occurs when people assume that randomness self-corrects itself. In other words, if "black" has won six times in a row on the roulette wheel, people believe "red" is due. If the roulette wheel is unbiased, the six consecutive occurrences of "black" do not affect the incidental probability of "red."

Explanations have been offered for this behavior. One contends that the subjects produce or perceive sequences reflecting their own intuitive and subjective concept of randomness, which is incorrect (the gambler's fallacy). That is, it may not necessarily coincide with the probabilistic model. Another argues that humans are simply restricted by functional limitations such as the limited ability to ignore their own recent responses, limited ability to generate paced responses, or other limitations.

Allen Neuringer, a well-known psychologist involved in randomness studies, offers lack of skill as an explanation. He claims that people fail to comprehend all the requirements of randomness and have very little experience dealing with random series. Thus, they fail to generate such series when told to do so. Psychologist, David Budescu presented a stochastic model to explain human subjects' deviation from randomness, which concentrates on the process generating the sequence, and suggested that future work should consider the processes used by subjects in a randomization task rather than the series of outcomes that result. Falk and Konold claimed that people associate the randomness of a sequence with how difficult it is to mentally encode it. These will all be discussed in detail in the next section.

Chapter 4. Can People Behave Randomly?

This chapter focuses on three leading contemporary theoretical problems of randomness studies involving human subjects. Numerous studies have concluded that human subjects are suboptimal random generators. Those conducting the experiments suggested a variety of reasons as to why participants were not successful, such as attention span, boredom, or misconceptions of what true randomness really is. This section provides some of the leading contemporary theoretical problems and provides potential suggestions to improve future results.

David Budescu’s work entitled “A Markov Model for Generation of Random Binary Sequences” will be discussed. Budescu presented the idea that a subject’s performance in a randomization task can be described by a stochastic model, the Markov Chain, which is limited to the case of two response alternatives. This stochastic model considered probabilities of certain responses, given its direct predecessor response. Budescu claimed that his results fit this model fairly well.

Allen Neuringer received notable attention for his study, “Can People Behave Randomly?: The Role of Feedback.” He claimed that humans can learn random-like behavior provided they receive proper information and training. His results support this hypothesis, but some criticisms are possible.

Finally, Ruma Falk and Clifford Konold make an important point in their study, “Making Sense of Randomness: Implicit Encoding as a Basis for Judgment.” They incorporate the idea of algorithmic complexity in their hypothesis that human subjects judge the randomness of a stimulus in terms of how difficult it is for them to mentally encode it. Their findings support their theory.

4.1 Budescu

Budescu proposed that a subject’s performance in a randomization task can be described by a stochastic model, the Markov Chain, which is limited to the case of two response alternatives. A Markov Chain is described by the matrix of transition probabilities (*i.e.* probability of a certain response, given the response generated directly before), which is illustrated as follows:

| | | Trial <i>i</i> | |
|-------------------|--|----------------------------|---------------------------------|
| Trial <i>i</i> -1 | | 1 | 0 |
| 1 | | λ | $1 - \lambda$ |
| 0 | | $\frac{p(1 - \lambda)}{q}$ | $\frac{(1 - 2p + p\lambda)}{q}$ |

The matrix involved three parameters, probabilities $P(x = 1) = p$, $P(x = 0) = q = 1 - p$, and the conditional probability, λ , of a “1” following another “1” [4. pg. 27]. Besides

focusing on negative recency, there are three assumptions for any Markov chain. The responses in every trial depend on chance, the probability of a given response in every trial is independent of all previous responses except for its immediate predecessor, and the probability of any response, given the previous response, is independent of the trial's location in the sequence.

Three experiments were conducted to test the validity of this model [4, pg. 28]. The first experiment involved 18 subjects. A computer generated a random sequence of symbols (first two letters of the Hebrew alphabet). The word "choose" appeared on the screen to prompt the subject to press one of the two buttons on the keyboard. The responses were recorded and cleared from the screen after a certain number of responses were obtained. The subjects were then given a score ranging from zero to thirty (higher scores represented better performances). Their task was to imitate the random selection mechanism used by the computer.

The subjects in Experiment 1 generated a total of 36 sequences. These generated sequences were one of three lengths ($n = 20, 40, \text{ or } 60$), where n represents the length of sequence. However, the subjects did know the length of the sequence that they would generate each time. Also, the probability values for the two events (two Hebrew letters) varied on the following three levels: $p = 0.50, 0.70, \text{ and } 0.90$, where p represents the probability of the dominant event. The subjects were advised of the probability values each time they generated a sequence. The values of n , and the values of p were combined in order to consider all sequence possibilities. After generating nine sequences, the subjects repeated this activity three more times. Thus, each subject produced $4(3(20 + 40 + 60)) = 1440$ responses. During the procedure, the subjects were not given any feedback and were not allowed to write down their responses. The subjects did, however, receive monetary payment as an incentive to try to achieve a higher performance score.

Budescu claimed that his results supported the stochastic model. Ten of the eighteen participants consistently showed negative bias on all three probability levels, and one subject showed positive bias on all three probability levels. Four additional subjects displayed negative or positive bias for at least a subset of their generated sequences. Budescu also reported from his analyses that the recency bias was strongest when both events are equally probable, *i.e.* $p = 0.50$.

The purpose of the second experiment was to test the hypothesis that the performance in the first experiment was related to instructions and/or the performance score and monetary reward. Participants were assigned to one of six groups and completed one session of eighteen sequences. However, the groups' conditions varied by levels of instruction (two types: short and detailed) and feedback (three types: same as in Experiment 1, score provided but no payoff, and no feedback). No significant differences were found between the two levels of instruction or the three feedback conditions. The six groups achieved similar performance scores leading to a rejection of the former hypothesis.

The third experiment shared features of the first two, except all series were of length forty, and the number of previous responses displayed on the screen was 0,1,2,3, or all previous responses. (Recall that the former experiments did not allow the subjects to keep track of any responses.) Eight of the ten subjects produced too many alternations in their sequences. One subject produced too many repetitions (positive recency), and the other subject was inconsistent.

Budescu claimed that the Markov Model fit most of the subjects' responses fairly well. Fifty-two percent of the participants in all three experiments consistently showed negative bias, and eight percent showed positive bias. However, if only the first and third experiments are considered the percentages are considerably higher. In this case, seventy-one percent of the subjects were classified into negative bias and fourteen percent fell into the positive bias category.

Budescu's goal was to provide a reasonable descriptive model that showed consistency with the available empirical results. Although he did not test his data rigorously, he claimed success due to the following:

- The model used was the first one to clearly identify various independent styles in randomization, *i.e.* by classifying the participants into one of three groups: positive bias, negative bias, inconsistent.
- The model took into account the intensity of the individual bias (through the model's parameters).
- The model allowed for fairly precise forecasts of the individual patterns of responses.

Budescu suggested that future research should consider evaluating the processes that humans use in randomization, rather than the outcomes that result [4, pg. 38].

4.2 Neuringer

Neuringer introduced the idea that by providing feedback and statistical descriptors to subjects, random-like behavior can be learned. Two experiments having similar procedures supported this notion.

In Experiment 1, seven participants were told to press the "1" and "2" keys on the alpha numeric keypad of a computer as randomly as possible, with no feedback. No time limit existed, and the subjects completed sixty trials of 100 responses each (6000 responses). The subjects were also paid for participating. Next, the subjects repeated the above process, except after each trial, they were given the values of five statistical descriptors evaluated at the just-completed response. The subjects were asked to vary their responses to bring these five statistics as close as possible to values arising by applying the descriptors to output from a random number generator. As one would expect, all subjects

differed considerably from the random generator during the no feedback condition. However, during the sixty trials of the feedback condition, all subjects were eventually indistinguishable from the random generator on all five tests. Ultimately, it was determined that after less than an average of six hours of feedback training, the seven subjects behaved randomly according to the five statistical tests employed.

Experiment 2 duplicated the former, except ten statistical descriptors were used as opposed to five. Also, less advice and guidance were given to the four participants. The subjects were told to try to imitate the toss of a coin and enter the digits “1” or “0” as randomly as possible, with no feedback. Then, during the feedback condition, a table of numbers representing a different descriptive statistic appeared on the screen after each trial. Furthermore, toward the end of the session the subjects were told that they would receive two days off with pay if they attained a level of performance on all ten descriptors that showed them to be “random” over two successive sets of sixty trials each. The subjects again differed significantly from the random generator when denied feedback, but then learned to behave randomly as assessed by the tests.

Neuringer’s experiment shed new light on randomness studies. It was the first study to teach random behavior through feedback. Also, the success of his feedback condition argued against the theory that people are unable to behave randomly, suggesting that people can learn random-like behavior through a controlled feedback environment. Neuringer had previously done experiments with pigeons, where the animals learned to generate highly variable sequences when rewarded for doing so [13, pg. 72].

One might ask, did Neuringer’s study show human subjects could attain “true” randomness? Did subjects learn to satisfy requirements of his feedback condition or did they acquire a general ability? Neuringer took data from the last 60 trials in his second experiment and evaluated them using eight new statistics on which participants had never received any feedback. Using a 5% level of significance, half of the subjects were indistinguishable from the random generator on all eight tests and the other two were indistinguishable on six of the eight tests. These findings add further support to the effectiveness of the feedback procedure.

For the skeptics of Neuringer’s work, one cannot assert that Neuringer’s subjects behaved “truly randomly.” No individual set of statistics will prove randomness since it is possible for another statistic to show deviation from randomness. Furthermore, his study relied on a computer-based random number generator as the comparison norm, which in itself is not “truly random,” since this generator utilized an equation that passes most criteria for analyzing finite random sequences [13, pg. 73].

Combined data from the subjects failed some tests for randomness. The final 60 trials were put together to make a single sequence of 6,000 numbers. The concatenated 6,000 responses were examined using twelve tests including the previous eight statistics along with two chi-square tests and two additional autocorrelations. The data failed the

four additional tests. Of course, the subjects were not provided with any feedback from any of the twelve new tests nor from any test focused on a single set of 6,000 responses. Future research might demand more stringent requirements in order to determine whether or not human subjects can learn to be indistinguishable from random generators.

It is within the realm of possibility that subjects memorized long sequences of numbers that passed the tests. It is not clear if this could account for the results. Another possibility is that the subjects tuned into some random noise in the environment. Thus, the responses would pass tests for randomness, but the subject would not have been producing them.

4.3 Falk and Konold

Unlike the other authors, Falk and Konold feel that judging a sequence's randomness is more indicative of the subjective concept of randomness than is producing a random sequence. They use the analogy that a person may not be able to draw a scene, but may still be able to recognize the scene as one that he/she had observed. Thus, a person might be able to perceive randomness accurately, yet be unable to produce it. They used the algorithmic definition of randomness as the basis for their study. This approach relates to the 2nd mathematical proposal discussed earlier, which defines randomness in infinite sequences in terms of incompressibility. The following explains why. The algorithmic randomness of a binary sequence is the bit length of the shortest computer program that can reproduce the sequence. Falk and Konold suggest that a finite random sequence is "psychologically" random when it has maximum complexity in that it cannot be easily reproduced. They propose that their subjects attempt to make sense of a sequence in some way when asked to judge its randomness. For instance, they might try to encode the sequence before making a judgment on its randomness. They formed the hypothesis that human subjects will base their decision on the randomness in terms of how difficult it is to mentally encode the sequence. They conducted three experiments to test their claim.

In the first experiment, the participants were divided into three groups: the Judgment of Randomness group, the Memorization group, and the Assessed Difficulty of Memorization group. Each group worked with 10 sequences, each 21 bits in length. The probability of an alternation for the 10 sequences varied from 0.1 to 1.0 in steps of 0.1.

Note that a sequence would be "ideally random" when the probability of an alternation is 0.5. Proportions below this percentage indicate negative recency in a sequence, and those above indicate positive recency. The first group (Judgment of Randomness) consisted of 97 subjects. Each member of this group was instructed to rate each sequence on a scale of 0 to 10 according to his/her intuition of how likely it was that such a sequence was obtained by flipping a fair coin. These participants were advised to inspect all of the 10 sequences first before assigning any ratings. After examining all 10 sequences, they could then begin rating the sequences according to the scale above (where

a score of 10 represents a sequence most likely to have been obtained by flipping a coin, and a score of 0 represents a sequence least likely).

The second group (Memorization group) consisted of 80 participants. Their task was to study each sequence until they could reproduce it. An individual sequence was presented on a computer screen. Once the subject felt he/she could reproduce the given sequence, he/she pressed a key and the given sequence was masked. The subject then typed that sequence from memory on another line. If the subject reproduced the sequence correctly, then he/she proceeded to the next one. However, if the sequence were incorrect, then the computer displayed the sequence once again, and informed the participant of two things, where the first error had occurred and how many errors were made. The subject was then provided with another chance to view and to then type the sequence. This process was repeated until the sequence was reproduced accurately, although the subjects did have an option to skip to the following sequence if he/she had failed after 5 attempts. During this activity, the computer recorded the total time the given sequence was displayed.

The third group (Assessed Difficulty of Memorization group) consisted of 136 participants. They were presented with a given sequence and were required to assess its difficulty of memorization by rating the sequence with a 1, 2, or 3 – where 1 = easy, 2 = medium, and 3 = difficult.

Results from this experiment indicate that the difficulty of encoding predicts perceived randomness better than the sequence's degree of randomness does. This supports Falk and Konold's hypothesis that an assessment of the difficulty mediates the judgment of a sequence's randomness.

Falk and Konold speculated that their subjects would take advantage of any kind of a pattern in the stimuli in order to succeed at their individual tasks. In other words, they thought their subjects would be fairly capable of detecting patterns in the sequences that contained too many alternations relative to chance. Recall that a finite binary sequence would be "ideally random" if the probability of an alternation is 0.5. However, Falk and Konold reported that their subjects were oblivious to patterns with probabilities of alternations ranging from 0.1 to 0.4, leading them to conclude that the subjects were clueless to slight-to-moderate degrees of deviations from randomness in the negatively biased sequences. In fact, Falk and Konold claimed that these over-alternating sequences were even more difficult for the subjects to memorize than were the most random sequences.

In the second experiment, Falk and Konold compared the randomness ratings from 97 subjects in the first experiment to the difficulty other participants had in reproducing those sequences. The subjects in this second experiment were advised that copying a sequence efficiently can be accomplished in stages by breaking the sequences into chunks to simplify the encoding task. Twenty participants were instructed to look carefully at each sequence presented by a computer. Once subjects felt able to reproduce the sequence,

they pressed the return key to mask sequence. Subjects would then begin typing as much as remembered. If the subject reproduced the segment correctly, then another chunk would appear and the subject would proceed to learn and copy it. This process continued until the entire sequence was copied correctly. If a mistake were made, then a note appeared and the segment would reappear for additional viewing and retyping until copied correctly. The subjects were advised that the computer recorded the time spent in viewing the target sequence.

Experiment 2 showed high correlations between the difficulty of reproducing sequences and the perceived randomness of sequences. Thus, the findings in the second experiment corroborate those found in Experiment 1. These results provide evidence against the hypothesis that copying a given sequence in parts may help the subjects to decrease bias.

In Experiment 3, subjects assessed the difficulty of copying a sequence by breaking their assessments into stages. Falk and Konold used longer sequences (of length 41) and obtained randomness ratings for them also. Increasing the length of the sequences was done to extend the generality of their findings. The instructions for this experiment were identical to those in the first, except the scale ranged from 1 to 10 instead of 0 to 10. Specifically, subjects were presented with 5 sequences and were told to pretend that they had to copy the sequences on the back of the page. Since such sequences are usually copied in chunks, they were advised to try to divide each sequence as they would in order to reproduce in the fastest and most efficient way. They were also instructed to mark their divisions with lines to show how the sequence was partitioned. Once the subjects completed this task, they were instructed to return to their divided sequences and rate each segment from 1 to 3 representing how difficult it would be to copy (1=easy, 2=medium, and 3=difficult).

Data in this last experiment corroborated those results found in the previous experiments. The segments that the subjects rated as more difficult to reproduce were judged as most random. The results indicated that these segments, judged as most random by the subjects, were negatively biased.

Although the experiments varied in several ways such as the sequence length, rating scale, and format of sequence presentation, the results were uniform. Data from all three consistently attest to the intensity and prevalence of the bias towards alternations in the perception of randomness. Thus, the subjects in Falk and Konold's study seemed to have a more difficult time mentally encoding sequences with more alternations than would arise due to chance. They associated the degree of difficulty to encode a sequence with its degree of randomness. Thus, they judged these negatively biased sequences as "random."

From the results in Falk and Konold's study, it appears that the idea of randomness as maximal complexity accurately represents the intuitive concept. These results suggest that subjects equate randomness with the difficulty of encoding. They appear to attempt

some kind of direct encoding and use the difficulty of that attempt to judge the randomness of a sequence.

Chapter 5. Statistical Tests for Randomness

All of the tests in the present chapter are variants of statistical hypothesis testing. Therefore, we begin with an exposition of inferential statistics, which is a way to draw inferences about a population from a sample. Following the discussion of hypothesis testing, we will report on some of the statistical tests used in the randomness studies involving human subjects. Specifically, we look at the Frequency Test, the Runs Test, and tests involving the chi-squared statistic or variants of it.

In this chapter, we will explain how the information we need about probability distributions for testing the success of human “random number generators” can be estimated by computer simulations. Simulations may not be acceptable in sensitive tests of high-quality random number generators, but they are entirely adequate for testing the rather modest abilities of human randomizers. Basing our investigations on simulation permits great conceptual simplification.

5.1 Parameters and Statistics

A *parameter* is a numerical characteristic of a population. Examples include:

- the average (or mean) value of a numerical variable, *e.g.*, average height of a New Yorker,
- the largest (or smallest) value of a numerical variable.
- the standard deviation of a numerical variable, the root of the mean squared error in a “population” of observations,

Typically, a parameter is obtained from the distribution of some variable \mathbf{X} (on a population). For any population for which \mathbf{X} makes sense, the distribution of \mathbf{X} and the value of an associated parameter can be determined.

A sample is a subset of a given population, *i.e.*, a subpopulation. If \mathbf{X} is a variable on the population, then it is applicable to the subpopulation. We define a *statistic* as a numerical characteristic of a sample. The mean value of a numerical variable on a given sample would be a statistic, as would the sample maximum (or minimum). The standard deviation of the set of values the variable takes on a sample would also be a statistic.

5.2 An Example Relating to Hypothesis Testing

One very common form of hypothesis testing concerns the process whereby we obtain information about a population parameter from statistics on samples taken from that population. Often we are interested in determining how much confidence we can place in a claim that a population parameter lies in a particular range, when the only evidence available is the value of a related statistic.

Example 5.2.1. Suppose that an exit poll of 400 randomly selected voters shows that 220 people favor a certain candidate, say Judge Well. Should we accept with certainty that Judge Well has won the election? The answer is clearly “no” because the excess of favorable voters in the sample could have been due to pure chance.

Suppose that in actual fact fewer than 50% of all votes went for Judge Well. Then, obtaining a 400-voter sample with an excess of votes in favor of the above candidate becomes increasingly unlikely as the excess increases. Certainly, if a truly random sample of 400 voters included 300 who voted for Judge Well, then you would feel safe in acting on the assumption that Judge Well has won, even if actions based on error could be very damaging.

When we come to evaluate the evidence that Judge Well has lost his campaign, we recognize that certain exit poll statistics would support this notion unequivocally, *e.g.*, a huge preponderance of votes for Well. Other statistics from the exit poll would support the opposite conclusion, *e.g.*, only a few dozen votes for Well in a sample of hundreds. Finally, some kinds of polling results would support neither conclusion, *e.g.*, a near split. Can we be more precise and quantitative?

Suppose, as above, that 220 voters of the 400 in the exit poll vote for Judge Well. While this possibility exists even if the Judge has lost, it is not likely. We can compute the exact probability that there are at least 220 voters favoring the Judge in the sample despite the Judge losing. If Judge Well gets exactly half of the votes then by the binomial theorem, the probability that in a sample of 400 votes there are least 220 votes for the Judge is given by the following:

$$\sum_{i=220}^{400} \binom{400}{i} \left(\frac{1}{2}\right)^{400} \cong 0.0255$$

This probability is even smaller if the Judge gets less than half of the votes. This shows that if we bet that a candidate has won every time we see an excess of at least 20 more votes in an exit poll of 400 voters, then we can expect to win our bet at least 97% of the time.

On the other hand, it would be dangerous to assume either a win or a loss if the overage were only 2 or 3, since a bet on either a win or a loss would be wrong a substantial amount of the time. For example, suppose there is a near exact split in the actual voting. In roughly 40% of all samples of size 400 there will be 203 or more votes for Well and in another 40% there will be 203 votes against.

This example has illustrated the basic intuitions behind Hypothesis Testing.

5.3 Hypothesis Testing

Definition 5.3.1. A *null hypothesis* (usually denoted by H_0) is a hypothesis about a population parameter.

From collected data (representative of a sample) in an experiment, a test statistic can be determined, which helps to evaluate the competing hypotheses. If the statistic is very different from what would be expected if the null hypothesis were true, then the null hypothesis is rejected. However, if the data do not vary significantly from what would be expected under the assumption that the null hypothesis is true, then the null hypothesis is not rejected.

Definition 5.3.2. The criterion for rejecting the null hypothesis is known as the *significance level*.

A significance level is generally chosen by the user of the test, who, if rational, will select a level that is compatible with the risks involved in rejecting or not rejecting the null hypothesis. The use of significance levels in hypothesis testing occurs as a process of ordered steps. First, the distinction between the results of the experiment and the null hypothesis is made. Next, the probability of a obtaining a statistic as different or more different from the null hypothesis (again, assuming the null hypothesis is true) than the statistic obtained in the sample is calculated.

Definition 5.3.3. The *probability value* (denoted *p-value*) calculated in a hypothesis test represents the probability of obtaining data as extreme or more extreme than the current data (assuming is true).

Example 5.3.1. Suppose data from an experiment yielded a *p-value* of 0.005. This means that the probability of obtaining data as extreme or more extreme from the null hypothesis as those obtained in that experiment is 0.005. It is not the probability of the null hypothesis itself.

Last, this probability value is compared to the given significance level. In the event that the probability is less than or equal to the significance level, then the null hypothesis is rejected and the result is said to be statistically significant. Although subjective, 0.01 and 0.05 levels of significance are commonly used. As the significance levels decrease, data must diverge more from the null hypothesis in order to be significant. Thus, the 0.01 level is more conservative than the 0.05 significance level. In many instances, researchers designate the null hypothesis as the opposite of what the experimenter actually believes. They put forth a null hypothesis hoping that the data can discredit it. Also in hypothesis testing, an alternative hypothesis (usually denoted by H_1 or H_A) exists. In the case where data are sufficiently strong to reject the null hypothesis, then the null hypothesis is rejected in favor of an alternative hypothesis.

Another aspect of hypothesis testing concerns not accepting the null hypothesis. When results from data do not call for the rejection of the null hypothesis, this does not mean that the null hypothesis is accepted in the voting example. Some results may even indicate that the null hypothesis is false. However, evidence may lack the strength for a case convincing enough to say the null hypothesis is false.

We can, if we desire, to formulate randomness tests with hypotheses such as

- H_0 : Generated Data are representative of a random process
 H_1 : Generated Data are not representative of a random process

Data will be presented for each test, and the p -value will be calculated in each case. In contrast to practical uses of hypothesis testing, we do not need to act based on the truth or falsity of the null hypothesis. We simply use the p -value as a measure of “how random” a performance is.

5.4 Frequency Test For Randomness

The Frequency Test focuses on the proportion of zeros and ones for the entire given sequence. This test determines whether the number of zeros and the number of ones in a sequence are approximately the same. This is what would be expected in a truly random sequence. This test assesses the proximity of the fraction of ones to one-half, and shows how large an excess is required in order to reject the hypothesis of randomness at a given confidence level.

Example 5.4.1. Suppose σ is a sequence of 0s and 1s with length n , and suppose that the number of ones is w . (The quantity w is a fixed piece of data that we have concerning σ , not to be confused with the random variable W , which will be used below.) Now if $\frac{w}{n}$ is close to $\frac{1}{2}$ then we have no reason to reject the hypothesis that σ was generated by a process similar to the coin-flip experiment—or as we say for brevity, σ is “random.” (When we use “random” in this sense, we are referring to the third chapter. Recall, that since σ is finite, there is no absolute property of randomness in finite strings.) On the other hand, if $k := \left| \frac{n}{2} - w \right|$ is sufficiently large, then this provides evidence against the hypothesis of “randomness.”

How strong is the evidence? To evaluate this, we calculate the p -value as follows. Let $p = P\left(\left| \frac{n}{2} - W \right| \geq k\right)$, where W is a random variable representing the number of 1s that result from flipping a fair coin n times. If this p -value is very small, then we view it as

evidence against the hypothesis that σ is “random.” Why is this reasonable? Suppose $p = 0.01$. This value means that a sequence as deviant as σ occurs only 1% of the time, when a truly random process is the source. So, if our policy is to reject σ as random when $p = 0.01$, we will eliminate a sequence from a truly random source only once in 100 tests in the long run.

The p -value above can be computer from the binomial theorem. Let E be the set of all integers $i \in \{1, 2, 3, 4, \dots, n\}$ such that $\left| \frac{n}{2} - i \right| \geq k$. Then, the probability that we have

exactly i heads is $\frac{\binom{400}{i}}{2^{400}}$ and hence $\frac{\sum_{i \in E} \binom{400}{i}}{2^{400}}$.

5.5 Runs Test For Randomness

This final section begins by discussing the concepts of runs and lengths of runs in a binary sequence. It describes the Runs Test, a basic statistical test that can be used to evaluate randomness in a given sequence. Although the Runs Test has variations, we focused on three simple versions. One version examines the number of runs of some selected length in a given sequence. The second examines the total number of runs in a given sequence. The third examines the run-count vector, and examines the distance from this to the vector of expected run counts.

Definition 5.5.1. A run in a symbol sequence is a series of identical consecutive symbols bounded before and after by different symbols. The length of a run is simply the number of times the repeated symbol occurs.

Example 5.5.1. Suppose a coin is tossed 20 times producing the following result:

THHTTHHHHHHTHHTTHHTT

We mark the sequence in order to display the runs:

| T | HH | TT | HHHHHH | T | HH | TT | HH | TT |

Thus, this sequence has a total of nine runs. The run lengths are as follows:

1, 2, 2, 6, 1, 2, 2, 2, 2

The results indicate 2 runs of length one, 6 runs of length two, 0 runs of lengths three, four and five and 1 run of length six. The run-count vector is $\{2, 6, 0, 0, 0, 1\}$.

Proposition 5.5.1. In a truly random binary sequence of length n (where n is very large), you would expect very nearly $\frac{n}{4}$ runs of length 1, $\frac{n}{8}$ runs of length 2, $\frac{n}{16}$ runs of length 3, $\frac{n}{32}$ runs of length 4, and so forth.

Proof. In order for a symbol to occupy a run of length 1, two conditions must be met. They are the following: 1) the symbol preceding the symbol in question must differ from it, and 2) the same holds for the symbol following the symbol in question. Both of these conditions have probability $\frac{1}{2}$. They are independent (since we are assuming that the sequence is random), so the probability of them occurring together is $\frac{1}{4}$. Similarly, for a symbol to be the first in a run of length 2, three independent conditions must be met, each having probability $\frac{1}{2}$. In general, for a symbol to be first in a run of length k , $k+1$ conditions must be met. Note that the first and last symbols in the sequence are exceptions to this argument. However, if n is large, the effect is small. Now, the number of runs of length k in a sequence of length n is exactly the same as the number of symbols that occur *first* in a run of length k . As we have seen that number is very nearly $n/2^{k+1}$. \square

The simplest version of the runs test for randomness examines whether the total number of runs of some selected length in the sequence to be tested is within the range that would be expected if the sequence had been generated by a truly random process. There are several variants that build on this essential theme. For example, the investigator might ask whether the total number of runs is within the expected range. Other variants might test whether the total number of runs of several different lengths is reasonable, or whether some function of several run counts is in an expected range.

Example 5.5.2. In this example, we apply the variant of runs test in which we examine the runs of only certain lengths. We ask whether or not the sequence to be tested has the appropriate number of runs of one particular length.

To create an example, the author of this paper manually generated a sequence of 1024 0s and 1s, attempting to act as randomly as possible. This sequence was translated into a list of 0s and 1s and entered into *Mathematica* for analysis. The run count vector was {535, 38, 71, 26, 11, 4, 0, 1, 1}. The test sequence immediately appears to have inappropriate run counts for a truly random sequence, especially for runs of length 1 and runs of length 2. According to Proposition 5.5.1, in a truly random sequence of length 1024, we would expect to have about 256 runs of length 1, 128 runs of length 2, 64 runs of length 3, 32 runs of length 4, 16 runs of length 5, 8 runs of length 6, 4 runs of length 7, 2 runs of length 8, and 1 run of length 9.

One simulation of a sequence of 1024 0s and 1s in *Mathematica* produced a run count vector of {249,122,64,32,15,12,3,2,3}. This single simulation lies reasonably close to the expected run count vector, which tends to confirm our suspicion that the manually produced sequence is non-random. To obtain more detailed information, we programmed

Mathematica to generate 100,000 trial sequences, each consisting of 1024 pseudo-random “coin flips”. We then computed and examined the run count vectors of the 100,000 trials for 1024 “coin tosses.” Table 5.5.1 shows the range of the run counts that occurred in our data.

Table 5.5.1

| Number of Runs (Range) | Length of Run |
|------------------------|---------------|
| [178, 330] | 1 |
| [79, 180] | 2 |
| [34, 97] | 3 |
| [11, 57] | 4 |
| [2, 33] | 5 |
| [0, 22] | 6 |
| [0, 15] | 7 |
| [0, 11] | 8 |
| [0, 7] | 9 |
| [0, 6] | 10 |
| [0, 5] | 11 |
| [0, 3] | 12, 13 |
| [0, 2] | 14, 15 |
| [0, 1] | 16—25, 27 |

From the table, we see that among the 100,000 simulations, there were some with as few as 178 runs of length 1 and some with as many as 330 runs of length 1, some with as few as 79 runs of length 2 and some with as many as 180 runs of length 2, and so forth. There were sequences with as many as 3 different runs of length 12, and there were sequences with as many as 3 different runs of length 13. Runs of length 16 through 25 were observed. (The table does not tell us whether there were any trials that contained more than one run longer than 16.)

Reconsidering the run count vector $\{535,38,71,26,11,4,0,1,1\}$ from our manually generated sequence, it is clear that the number of runs of length 1 far exceed the expected number, and the number of runs of length 2 is way too small. So, we certainly would not declare that this sequence passes the Runs Test for randomness.

Example 5.5.3. In this case, we study the total number of runs. We determine whether or not the sequence to be tested has the appropriate run total.

The author of this paper manually generated a sequence of one thousand 0s and 1s, attempting to be as random as possible. After generating the sequence, we translated it into a list of 0s and 1s and entered it into *Mathematica* in order to analyze it. The run count vector for the data was found to be $\{407,47,59,30,14,6,6,3,1,1,1\}$. In other words, in our sequence of length 1000, there were 407 runs of length 1, 47 runs of length 2, 59 runs of

length 3, etc. All together, there were 575 runs of lengths one through 11, these accounting for all 1000 entries.

The pseudo-random number generator in *Mathematica* was used to simulate a sequence of 1000 “coin flips.” This generator performed 10,000 complete simulations, where each output consisted of a list of 1000 0s and 1s. In these 10,000 trials, the smallest total run count was 446 and the largest was 561. The table below gives the distribution of the variable “total number of runs” in the population of 10,000 simulations.

Table 5.5.2

| Run Count Range (RCR) | Number of Simulations in RCR |
|-----------------------|------------------------------|
| (441,450] | 9 |
| (451,460] | 59 |
| (461,470] | 222 |
| (471,480] | 717 |
| (481,490] | 1598 |
| (491,500] | 2369 |
| (501,510] | 2293 |
| (511,520] | 1670 |
| (521,530] | 782 |
| (531,540] | 225 |
| (541,550] | 45 |
| (551,560] | 10 |
| (561,570] | 1 |

The mean run count was 501 and the median was approximately the same. Run counts, between 461 and 540, accounted for more than 98% of all the data. We concluded that run counts outside this range have a p -value of roughly 2%. The total run count of 575 in the string that we produced manually was inconsistent with what would be expected in a random sequence, as the data from the simulation shows. In fact, in a second simulation we generated roughly 500,000 trials, in each of which a random string of 1000 0s and 1s was produced. There was only 1 case of a run count higher than 575. If the sequence that we generated were part of an experiment (where we were testing its randomness), we would conclude that it fails the runs test.

We extended our investigation with regard to the run count vector (from Example 5.5.2) even further by defining a function that takes into account the run counts of all lengths up to 10. This function is an approximate chi-square test statistic. We say approximate because the statistic that we compute does not have an exact chi-square distribution, since the distributions it comes from are neither normal nor independent. The chi-squared test statistic is calculated by comparing expected outcomes to observed outcomes and is defined in more detail below.

Definition 5.5.2. Let C_1, C_2, \dots, C_n be the observed counts of runs of each length up to n , *i.e.*, C_1 = observed number of runs of length 1, C_2 = observed number of runs of length 2, *etc.* The C_1, C_2, \dots, C_n are not independent since $C_1 + 2 C_2 + 3 C_3 + \dots + n C_n = n$. Let E_1, E_2, \dots, E_n be the expected count of runs of length j , *i.e.*, E_j = expected count of runs of length j , E_2 = expected count of runs of length 2). Then, we defined the chi-squared test statistic (for a run count vector) as follows:

$$\sum_{j=1}^n \frac{(C_j - E_j)^2}{E_j}$$

We can create a modified statistic by restricting to run lengths of only certain sizes, *e.g.*, from 1 to 10. We entered this restricted statistic into *Mathematica* and evaluated it at each of 100,000 simulations of 1024 “coin flips.” The following example describes the results of that simulation.

Example 5.5.4. Recall that the run count vector for our generated sequence of 1024 “coin flips” was $\{535, 38, 71, 26, 11, 4, 0, 1, 1, 0\}$. The chi-squared test statistic (restricted to runs of length 1 through 10) for this run count vector was 337.8. An initial sequence of 1024 “coin flips” was simulated for comparison producing a run count vector of $\{257, 109, 73, 37, 12, 12, 2, 2, 1, 0\}$ and a test statistic value of 15.9. After this initial simulation, we simulated 100,000 trials of 1024 “coin flips.” In these 100,000 trials, the smallest observed value of the statistic was 0.719, and the largest observed value of the statistic was 61.2. Our statistical value of 377.8 is extremely far from the range of simulated chi-square test statistic values. This unreasonably large value indicates a major departure from randomness in the manually generated sequence.

References

- [1] Baddeley, A.D. 1966. The Capacity for Generating Information by Randomization. *Quarterly Journal of Experimental Psychology*. 18:119-129.
- [2] Bakan, Paul. 1960. Response-Tendencies in Attempts to Generate Random Binary Series. *American Journal of Psychology*, 73: 127-131.
- [3] Bar-Hillel, Maya and Wagenaar, W.A. 1991. The Perception of Randomness. *Advances in Applied Mathematics*. 12: 428-454.
- [4] Budescu, David V. 1987. A Markov Model for Generation of Random Binary Sequences. *Journal of Experimental Psychology*. 13: 25-39.
- [5] Chaitin, G. J. 1975. Randomness and Mathematical Proof. *Scientific American*. 232(5): 47-52.
- [6] Cook, Alex. 1967. Recognition of Bias in Strings of Binary Digits. *Perceptual and Motor Skills*. 24: 1003-1006.
- [7] Diener, Don and W. Burt Thompson. 1985. Recognizing Randomness. *American Journal of Psychology*. 98: 433-447.
- [8] Falk, R and Konold, C. 1997. Making Sense of Randomness: Implicit Encoding as a Bias for Judgment. *Psychological Review*. 104: 301-318.
- [9] Feller, William. 1950. *An Introduction to Probability Theory and Its Applications*, Volume 1, Second Edition. New York: John Wiley & Sons, Inc.
- [10] Goldberg, Samuel. 1960. *Probability: An Introduction*. New Jersey: Prentice-Hall, Inc.
- [11] Knuth, Donald E. 1997. *The Art of Computer Programming: Semi-numerical Algorithms*, Volume 2, Third Edition. California: Addison-Wesley.
- [12] Lopes, Lola L. 1982. Doing the Impossible: A Note on Induction and the Experience of Randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 8: 626-636.
- [13] Neuringer, Allen. 1986. Can People Behave Randomly?: The Role of Feedback. *Journal of Experimental Psychology: General*. 115: 62-75.

- [14] Rath, Gustave. 1966. Randomization by Humans. *American Journal of Psychology*. 79: 97-103.
- [15] Ross, Bruce. 1955. Randomization of a Binary Series. *American Journal of Psychology*. 68: 136-138.
- [16] Volchan, Sergio B. January 2002. What is a Random Sequence? *American Mathematical Monthly*. 46-63.
- [17] Wagenaar, W.A. 1970. Appreciation of Conditional Probabilities in Binary Sequences. *Acta Psychologica*. 34: 348-356.
- [18] Wagenaar, W.A. 1972. Generation of Random Sequences by Human Subjects: A Critical Survey of Literature. *Psychological Bulletin*. 77: 65-72.
- [19] Weiss, Robert. 1964. On Producing Random Responses. *Psychological Reports*. 14: 931-941.

Appendix: Mathematica Commands

This appendix provides the reader with some of the input commands used in *Mathematica* to obtain data in the Runs Test for Randomness simulations (Chapter 5, Section 5). The steps below show how we obtained the results.

Suppose \mathbf{s} is a list of 0s and 1s (e.g., let us input $\mathbf{s}=\{0,1,0,0,1,1,1\}$). The function `Split[]` takes the list of 0s and 1s (that we call \mathbf{s}) and group the runs together. So applying this command to \mathbf{s} , we have a run of length 1, followed by another run of length 1, a run of length 2, and a run of length 3. Thus, if the input is `Split[s]`, then the output is `{{0},{1},{0,0},{1,1,1}}`.

If the input is `Length[Split[s]]`, then the output is 4. Since the `Split[]` command groups the runs together, `Length[]` counts the total number of runs for us.

If the input is `Random[Integer, {0,1}]` then the output is either a 1 or 0. This command gives us a random integer in the range [0, 1]; it behaves like a coin toss. `Table[Random[Integer, {0,1}], {1000}]` returns a list giving the result of 1000 “coin tosses”.

If the input is

```
Length[Split[Table[Random[Integer, {0,1}], {1000}]]],
```

then *Mathematica* returns the total number of runs in a sequence of 1000 “coin flips” simulated by the computer.

It requires more work to determine the run count vector for our simulations. We are interested in the number (possibly 0) of runs of a given length. Therefore, we define a function that will produce the run count vector for our sequence. We input the function as follows:

```
rcv[x_]:=
  Take[
    Map[Length,
      Split[Sort[Join[
        Range[10],
        Map[Length,Split[x]]
      ]]]
    ],
    10]-Table[1, {10}]
```

This function returns a list with the first 10 entries of the run count vector for a given sequence. It will display the individual number of runs of each length from 1 to 10 (will display 0, if for a certain length, there are no runs in that category). In table 5.5.1, we altered this function by replacing the number 10 by the number 30. (By making this change, the function would be able to consider the possibility of having a run with a length as high as 30 during the simulations.).

Example. In our work, we created a sequence of 1024 0s and 1s by hand. We entered it with the name `stringrep`. When we input `rcv[stringrep]`, the output was `{535, 38, 71, 26, 11, 4, 0, 1, 1, 0}`, telling us that we have 535 runs of length 1, 38 runs of length 2, 71 runs of length 3, etc. in our generated sequence.

The simulations required obtaining the run count vectors of 1024 “coin tosses” 100,000 times. We defined a function that generates a sequence of 1024 “coin flips.”

```
t1024[]:=Table[Random[Integer,{0,1}},{1024}]
```

With input `rcv[t1024[]]`, *Mathematica* returns an individual run count vector for the simulated sequence of 1024 “coin flips”, e.g., `{257, 109, 73, 37, 12, 12, 2, 2, 1, 0}`.

To perform the 100,000 trials, which are described in the Table 5.5.1, the `rcv[x_]` function was adjusted to allow possible (although unlikely) runs of length 30. The input `Table[rcv[t1024[]],{100000}]` lists the run-count vectors of 100,000 trials of 1024 “coin tosses”.

The chi-square statistic is programmed as follows.

```
chistat[s_]:=  
((s- $\{256,128,64,32,16,8,4,2,1,1/2\}$ )^2). $(1/\{256,$   
 $128,64,32,16,8,4,2,1,1/2\})$ 
```

Example. `chistat[{a,b,c,d,e,f,g,h,i,j}]` returns the following:

$$\frac{1}{256}(-256+a)^2 + \frac{1}{128}(-128+b)^2 + \frac{1}{64}(-64+c)^2 + \frac{1}{32}(-32+d)^2 + \frac{1}{16}(-16+e)^2$$

$$+ \frac{1}{8}(-8+f)^2 + \frac{1}{4}(-4+g)^2 + \frac{1}{2}(-2+h)^2 + (-1+j)^2 + 2\left(-\frac{1}{2}+j\right)^2$$

Here, `{a,b,c,d,e,f,g,h,i,j}` represents an observed run count vector. Recall that the run count vector for our generated sequence of 1024 “coin flips” was `{535, 38, 71, 26, 11, 4, 0, 1, 1, 0}`.

We calculated the approximate chi-square statistic using this data by the following:

```
chistat[{535,38,71,26,11,4,0,1,1,0}]
```

This gives the test statistic value of 377.8.

To perform the 100,000 trials of the above, we used

```
Sort@Table[chistat[rcv[t1024[ ]]],{100000}].
```

This calculates the approximate chi-square test statistic for 100,000 simulated toss-sequences and sorts them in ascending order.

Vita

Summer Ann Armstrong was born on August 29, 1976, and grew up in Ponchatoula, Louisiana. She now resides in the state of New York in the Hudson Valley with her husband, Dr. Michael Nuccitelli. After graduating from Louisiana State University in the summer of 2004, she plans to teach at a local university near her home.