Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis

Enhancing the power of permutation tests for positive serial dependence in binary data by using streak-breaking subgroups

Inge Clemens (535959)

Supervisor: Dr. N. Koning

Second Assessor: Prof. dr. C. Zhou

November 21, 2021

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

Recently, Koning and Hemerik (2021) have developed a method to enhance the power of group-invariance tests by using subgroups instead of uniform sampling from the group. For a location model, they find it possible to derive Oracle subgroups for which the distribution of the test statistic is equal under the null and alternative hypothesis. In this thesis, I examine the application of this method to permutation tests for positive serial dependence in binary sequences. These types of tests are extensively studied within the hot hand fallacy literature. Since derivations of Oracle subgroups is not feasible for the type of tests studied here, I have developed a heuristic approach to identifying subgroups that beat a Monte Carlo sample of permutations of the same size. I show that this heuristic is successful at identifying these types of subgroups for sequences of lengths 9 and 20. Furthermore, I present a systematic approach to gain insight into why some subgroups achieve a higher power than others in tests for positive serial dependence in binary sequences. This approach is successful in explaining differences in power for subgroups of $\mathcal{S}_6$.

Keywords: Group-Invariance Tests, Permutation Tests, Hot Hand Fallacy, Binary Sequences

Contents

1 Introduction

Traditional hypothesis testing relies on the derivation of the distribution of the test statistic under the null hypothesis. Given a pre-specified level α, which should fix the type I error rate, the critical value of the test can be determined from the derived null distribution of the test statistic. The disadvantages associated with this method can be attributed to the first step, deriving the distribution of the test statistic under the null. These derivations require a distributional assumption on the data generating process (DGP). Many conventional statistical tests assume that the data originate from a distribution that is approximately normal. If this assumption holds, the type I error rate is controlled, and therefore the test is referred to as exact. If this assumption does not hold, the size of the test is not guaranteed to be equal to α and thus the test is not exact. Moreover, when the test statistic becomes more complicated, analytical derivations of the null distribution might not be available at all.

Luckily, other methods of hypothesis testing exist that can circumvent these issues. One of these alternative methods are permutation tests, which were already described by Fisher in the 1930s (Fisher, 1937, Chapter 2). Fisher wrote about permutation tests 'the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method'. Given the computational power at hand today, these 'tedious' calculations have become available to all. In general, a permutation test is applicable in situations where the distribution of the test statistic under the null hypothesis does not change upon permutations of the data (Ernst et al., 2004). For example, under the null hypothesis that two samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ have equal means, interchanging some of the elements of the groups should not affect the distribution of the test statistic $\bar{Y} - \bar{X}$. The critical value in a permutation test is based on the permutation distribution of the test statistic. This permutation distribution is obtained by calculating the test statistic for all possible permutations of the data. The critical value is then given by the α-upper or -lower quantile of the permutation distribution of the test statistic. If the original value of the test statistic is more extreme than the critical value, the null hypothesis is rejected. (Lehmann and Romano, 2005, Chapter 15.2)

Recently, Koning and Hemerik (2021) have developed a new method to enhance the power of permutation tests. The authors illustrate their method using a location model. The aim of this thesis is to examine whether their method can be extended to other settings. More specifically, I investigate the possibility of increasing the power of permutation tests for positive serial dependence in binary sequences based on the methodology of Koning and Hemerik (2021).

Advantages of permutation tests over traditional hypothesis testing

The main advantage of permutation tests over traditional hypothesis testing is that the level of the test is assured under weaker assumptions. Therefore permutation tests are more reliable in various situations. This can be illustrated by two examples from different fields, namely treatment-control studies and the examination of the hot hand.

The benefits of permutation tests in treatment-control studies were already pointed out by Ludbrook and Dudley (1998) and Berger (2000). However, as the authors noted, traditional t- and F-tests remained the standard within biomedical research. Two decades after these articles were published, Young (2019) examined whether the results of 53 recent treatment-control studies could be reproduced with permutation tests. In the original studies, the authors based their results on conventional hypothesis testing. Young (2019) found that 22% of the results that were reported to be significant on a 1% level, were no longer significant when using permutation tests. This number even increased to 49% for joint test of significance. This reported discrepancy was assigned to the presence of high leverage points, which highly influence estimates of coefficients and standard errors and therefore make the results volatile. For a t-test, the presence of these high leverage points generates t-distributions with fatter tails than those of the *assumed* distribution of the test statistic under null. As a result, the size of the test is larger than desired. This is where the advantage of the permutation tests comes forward. In a treatment-control experiment, the act of random assignment of the treatment and control group is sufficient to perform exact permutation tests (Ernst et al., 2004; Hemerik and Goeman, 2021).

Another example that illustrates the benefits of permutation tests is within the hot hand literature. Here, it was shown by Miller and Sanjurjo (2018b) that the paired t-tests conducted by Gilovich et al. (1985) suffered from a substantial small sample bias. This could be rectified by applying a bias correction. However, such a bias correction could be omitted altogether by simply using permutation tests.

Approximating the permutation distribution of a test statistic

One limitation of permutation tests is that the total number of permutations of a sample increases rapidly with the size of the sample. Consequently, for larger samples, it is computationally infeasible to obtain the exact permutation distribution by calculating the test statistic *for all possible permutations*. An easy solution is to approximate the permutation distribution by Monte Carlo sampling from the permutation distribution. An approximate permutation distribution of the test statistic is then obtained by calculating the test static only for a random set of permutations, which is uniformly sampled from the entire group of permutations. If the permutation distribution is approximated in this way, the permutation test is still exact. Koning and Hemerik (2021) state that the number of random permutations required to obtain replicable results and adequate power is usually several times higher than α^{-1}. For example, if $\alpha = 0.05$, typically a set of 200-5000 random permutations is used.

Novel method that replaces uniform sampling from the permutation distribution

For some applications, using such a large set of random permutations is too computationally expensive. (Kofler and Schlötterer, 2012) A smaller number of random permutations can be used but this leads to a loss in power and less replicability. Recently, Koning and Hemerik (2021) have developed a method that provides a solution for both of these issues. Their method substitutes uniform sampling from the permutation distribution by using a subgroup of the permutation group. This new method has two advantages over the conventional practice. First, given a specific subgroup, the results of the test are fully replicable as they do not depend on random sampling from the permutation distribution. Second, Koning and Hemerik (2021) are able to identify subgroups for which the power of the permutation test is higher than it would be when an equally-sized random sample of permutations is used. Hence, there exist relatively small subgroups that can obtain a power that would require much more randomly sampled permutations. Using such a subgroup would decrease the computational burden of a permutation test compared to using a random set of permutations, without compromising on power.

Aim and overview of this thesis

Koning and Hemerik (2021) illustrate their new method to enhance the power of group-invariance tests with a location model. Permutation tests are a subclass of these group-invariance tests. The aim of this thesis is to investigate whether the methodology of Koning and Hemerik (2021) can be extended to another setting. More specifically, I will aim to identify subgroups of the permutation group that enhance the power of permutation tests for positive serial dependence in binary data. To summarize, the central question of this thesis is: ‘How to identify subgroups of the permutation group that obtain a higher power in a permutation test for positive serial dependence in binary sequences than an equally-sized random sample of permutations?’ To simplify notation, I will refer to tests of positive serial dependence in binary data as tests for ‘streakiness’ and to subgroups that outperform a random sample of permutations in such tests as ‘streak-breaking’ subgroups.

Permutation tests for streakiness in binary sequences are common within the ‘hot hand’ literature. Throughout this thesis, I will use the hot hand debate as a motivating example and examine an alternative hypothesis and test statistic often applied in this body of literature. In Chapter 2, I give a brief overview of the literature that examines the existence of the hot hand within basketball. However, with this thesis, I aim to explore new methodology rather than contribute to the hot hand debate. As such, the particular binary sequences studied in this thesis are not fully representative of those studied within the hot hand literature.

I will aim to identify streak-breaking subgroups for binary sequences of lengths 6, 9, and 20. However, due to the complex nature of the test statistic and DGP examined, I am not able to use the same approach as Koning and Hemerik (2021) and analytically derive properties of streak-breaking subgroups. Therefore, I have explored other methods to identify streak-breaking subgroups. In Chapter 5, I focus on short sequences of length 6 as this allows for calculating the power in a test for streakiness for all subgroups of a given order. For sequences of length 6 and a specific parametrization of the alternative hypothesis, I identify the best- and worst-performing subgroups of order 24 in a test for streakiness. I find that there exists a subgroup that not only outperforms an equally-sized random sample of permutations but even the entire group of permutations. In addition, I invest the relation between the performance and the structure of these subgroups which would make it possible to identify streak-breaking subgroups *a priori*. Unfortunately, the resulting relation is based on discreetness effects of the short sequence length studied and hence can not be generalized to longer sequence lengths. For longer sequences, it is computationally infeasible to identify streak-breaking subgroups by investigating all possible subgroups. Therefore, in Chapter 6 and Chapter 7, I formulate a heuristic that selects candidate streak-breaking subgroups for sequences of length 20 (Chapter 6) and length 9 (Chapter 7). I find that the heuristic is able to identify streak-breaking subgroups that outperform a random sample of permutations for various parameterizations of the alternative hypothesis. Moreover, this heuristic approach has the potential to be extended to longer sequences

2 Motivating example

I will use the statistical examination of the existence of the hot hand as a motivating example throughout this thesis. The hot hand literature aims to identify whether there is positive serial dependence in shot sequences of basketball players. In this section, I will give a quick overview of the origin and the current state of the hot hand debate. However, tests for streakiness in binary data also have other applications. For example, within empirical finance to assess serial dependence in asset returns. (Fama, 1965; Ritzwoller and Romano, 2021)

Initial study of the hot hand

The study of the hot hand was initiated by Gilovich et al. (1985). These authors noted that within the basketball community there is a firm belief that when a player is on a winning streak, he has a higher probability than usual to make the next shot. When a player is on such a winning streak, he is said to have a 'hot hand'. The authors compared basketball enthusiasts' belief in such a hot hand with the statistical evidence for its existence. To statistically examine the existence of the hot hand, GVT conducted a controlled shooting experiment to rule out external factors such as defensive pressure. The shooting experiment included 26 basketball players from Cornwell University's basketball team. Each player took 100 shots from a fixed distance for which it was estimated that the player would make 50% of the shots. If the hot hand were to exist, the shot sequences should contain more streaks than what would be expected from a random Bernoulli sequence with success probability=0.5. Hence, the null hypothesis that the sequences are generated by a random Bernoulli process should be tested against some streak generating alternative.

To test this null hypothesis, the conditional probability of a hit after a streak of l hits was compared with the conditional probability of a hit after a streak of l misses for $l = 1, 2, 3$. A paired t-test of the null hypothesis $\mathbb{E}[\hat{\mathbb{P}}(\text{hit } \mid l \text{ hits}) - \hat{\mathbb{P}}(\text{hit } \mid l \text{ misses})] = 0$ failed to reject for all but one player. Based on these results, the authors concluded the belief in the hot hand to be an illusion.

Follow-up studies to Gilovich et al. (1985)

Since the initial study of Gilovich et al. (1985) multiple replication studies have been conducted. For example, Avugos et al. (2013) performed a similar controlled shooting experiment with Olympian athletes that take 40 shots. Koehler and Conley (2003) examined the existence of the hot hand in four years of data from the NBA three-point shooting contest. The results of both of these studies are in agreement with the results of the initial study of Gilovich et al. (1985) and no evidence for the existence of the hot hand was found.

However, recently it has been pointed out that the experimental set-up and statistical analysis of these studies suffer from two flaws that, when corrected for, *reverse* the conclusion of these studies. (1) The analysis of the conditional hit probabilities with a paired t-tests suffers from a substantial small sample bias as was discovered by Miller and Sanjurjo (2018b). This can be adjusted for by using a bias correction or by simply using *permutation tests.* (2) The tests used in the above mentioned studies are underpowered for plausible specifications of the alternative hypothesis and longer shot sequences are required to obtain sufficiently powered tests. (Miller and Sanjurjo, 2018a; Ritzwoller and Romano, 2021).

After applying a bias correction, Miller and Sanjurjo (2018b) find evidence of hot hand shooting for both the study of Gilovich et al. (1985) and Avugos et al. (2013). Moreover, to obtain a higher power for these tests, Miller and Sanjurjo (2018a) and Miller and Sanjurjo (2021) examined longer shot sequences from, respectively, a controlled shooting experiment and the NBA three-point shooting contest. In both of these studies evidence of hot hand shooting was found.

Implications of the hot hand debate

Gilovich et al. (1985) concluded that the belief in the hot hand is a fallacy and attributed this misperception to the belief in the 'law of small numbers'. The law of small numbers was introduced by Tversky and Kahneman (1971) and refers to the fact that people 'tend to regard a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics.' Otherwise stated, the belief in the law of small numbers refers to the wrongful belief that the law of large numbers is applicable to small samples. Gilovich et al. (1985) state that due to the believe in the law of small numbers, basketball enthusiasts will perceive streaks of hits as overly representative of the dependence in the sequence. Hence, when they are confronted with a sequence of hits, they overestimate the serial dependence in the sequences and conclude that the sequence is not just generated by a random process but some other factors (the hot hand) must be in play.

Since the influential study of Gilovich et al. (1985), the hot hand 'fallacy' has been regarded as a convincing example of the belief in the law of small numbers. Since then, this human misperception of randomness has been incorporated in standard models used in behavioral finance and economics. (Barberis and Thaler, 2003; Barberis, 2018)

However, due to the newly found evidence of hot hand shooting, it is not certain anymore whether the believe in the hot hand actually is a fallacy. The question remains whether the deviation from randomness in basketball shooting is in accordance with people's perception of the hot hand and thus whether the belief in the hot hand is a fallacy or not.

3 Theory

In this chapter, I will give an overview of the theoretical framework on which this thesis is based. First, some key concepts from group theory will be discussed that are essential to the understanding of this research. Following, I will explain the principles of a group-invariance test and give a short overview of the method developed by Koning and Hemerik (2021) to enhance the power of these tests using subgroups. Although the methodological approach taken in this thesis differs from that of Koning and Hemerik (2021), the key concept of using subgroups in group-invariance tests finds its origin here.

3.1 Group theory

In his book on group theory, (Armstrong, 1988) Armstrong introduces groups with the statement 'Numbers measure size, groups measure symmetry'. Armstrong (1988) formally defines a group as '*a set $\mathcal{G}$ together with a multiplication on $\mathcal{G}$ which satisfies three axioms:*

1. *The multiplication is associative, that is to say $(xy)z = x(yz)$ for any three (not necessarily distinct) elements from $\mathcal{G}$.*
2. *There is an element e in $\mathcal{G}$, called the identity element, such that $xe = x = ex$ for every x in $\mathcal{G}$.*
3. *Each element x of $\mathcal{G}$ has a (so-called) inverse x^{-1} which belongs to the set $\mathcal{G}$ and satisfies $x^{-1}x = e = xx^{-1}$.'*

In this definition 'multiplication' should not be interpreted in the usual sense but as an operation that combines elements of a set.

Example. A simple example of a group is the set of all integers $\mathbb{Z}$ along with addition as the multiplication operator, annotated by $\mathcal{G} = (\mathbb{Z}, +)$. To verify this is a group, we can check whether $\mathcal{G}$ satisfies the above three axioms.

1. As $(x + y) + z = x + (y + z)$ for $x, y, z \in \mathbb{Z}$, the multiplication is associative.
2. The set contains an identity element, namely 0, as $x + 0 = x$ for all $x \in \mathbb{Z}$.
3. As $x + (-x) = 0$ for all $x \in \mathbb{Z}$, the inverse of x is equal to $-x$ and hence the inverse of x belongs to the set $\mathbb{Z}$.

Three other key properties common to all groups are: (1) the identity element of a group is unique, (2) the inverse of each element $x \in \mathcal{G}$ is unique, and (3) a group is closed under multiplication. The latter property means that, for every $x, y \in \mathcal{G}$ it holds that $xy \in \mathcal{G}$. The number of elements in a group is referred to as the order of the group and denoted by $|\,\mathcal{G}\,|$.

Subgroups

A subgroup of a group $\mathcal{G}$ is defined as *'a subset of $\mathcal{G}$ which itself forms a group under the multiplication of $\mathcal{G}$'.* A subset $\mathcal{H}$ of $\mathcal{G}$ forms a subgroup if, (1) the identity element of $\mathcal{G}$ belongs to $\mathcal{H}$; (2) for all $x, y \in \mathcal{H}$, the product xy is an element of $\mathcal{H}$; (3) the inverse of each element in $\mathcal{H}$, which by definition is a member of $\mathcal{G}$, is contained in $\mathcal{H}$.

Example. The even integers, denoted by $2\mathbb{Z}$, are a subset of $\mathbb{Z}$. Moreover, $2\mathbb{Z}$ also forms a subgroup of the group $\mathbb{Z}$ under addition. The latter statement can be verified by checking the three requirements for subgroups given above.

1. The identity element of $\mathbb{Z}$, namely 0, is an element of $2\mathbb{Z}$.

2. For all $x, y \in 2\mathbb{Z}$, $x + y \in 2\mathbb{Z}$. Thus the product xy in an element of the subgroup $2\mathbb{Z}$.

3. The inverse of $x \in 2\mathbb{Z}$ equals $-x$. As $-x$ is an even integer it is contained in $2\mathbb{Z}$.

Permutations

As this thesis focuses on permutation tests, I will discuss some essential properties of (sub)groups of permutations in more detail. A permutation on a set X is defined as a bijection of X onto itself. The collection of all permutations of X forms a group under composition of functions and is denoted by $\mathcal{S}_X$. 'Composition of functions' means that for two permutations σ and τ, $\sigma\tau(x) = \sigma(\tau(x))$. Verifying that $\mathcal{S}_X$ indeed forms a group under composition of functions can be done by checking the three requirements for a group using the properties of a bijection and composition of functions. (1) Composition of functions is associative; (2) there is a permutation ϵ which leaves all elements of X fixed and hence is the identity element of the group; (3) as each permutation σ is a bijection, it has an inverse σ^{-1}. Hence, $\mathcal{S}_X$ satisfies all three criteria. Finally, when the set X consists of the first n positive integers, $\mathcal{S}_X$ is referred to as the symmetric group of degree n and is denoted by $\mathcal{S}_n$.

Now I introduce two notations of permutations that will be used throughout this thesis.

Example. Consider the permutation $\sigma \in \mathcal{S}_6$ which is defined by $\sigma(1) = 5$, $\sigma(2) = 6$, $\sigma(3) = 1$, $\sigma(4) = 4$, $\sigma(5) = 3$ and $\sigma(6) = 2$.

Cyclic form notation. $\sigma = (153)(26)$. In a cyclic form notation, each integer within the brackets is mapped onto the integer on its right. The last integer within the brackets is mapped onto the first integer. Note that the integer 4 is not included in this notation as it is mapped onto itself.

Array form notation. $\sigma = [3,\ 6,\ 5,\ 4,\ 1,\ 2]$. In the array form notation of σ, the indices are places on their new positions.

To finish this section on permutations, I provide some more definitions that will be used in Chapter 6. A cyclic permutation is a permutation with a single pair of brackets *i.e.* $\tau = (162)$. The permutation $\sigma = (153)(26)$ is the product of two disjoint cyclic permutations. Every permutation in $\mathcal{S}_n$ can be written as either a cyclic permutation or the product of multiple disjoint cyclic permutations. A cyclic permutation of length k is called a k-cycle. Hence, σ consists of a 3-cycle and a 2-cycle.

Cyclic subgroups

In Chapter 6, I solely focus on finding *cyclic* streak-breaking subgroups, as cyclic subgroups have some useful properties. One of these properties is that the order of a cyclic subgroup can be determined easily. In this section, I will explain what cyclic subgroups are and give a formula for the order of cyclic subgroups.

A cyclic subgroup of $\mathcal{G}$ can be created by taking an element $x \in \mathcal{G}$ and the set of all powers of x. Such a cyclic subgroup is denoted by $< x >$ and x is referred to as the generator of the group. It can be verified that the set $< x >$ actually forms a subgroup by checking the requirements of a subgroup: (1) the identity element of $\mathcal{G}$ is equal to x^0 and hence the identity element of $\mathcal{G}$ belongs to $< x >$; (2) $x^i x^j = x^{i+j}$ and hence the product of x^i and x^j is an element of $< x >$; (3) the inverse of x^n is equal to x^{-n} which is an element of $< x >$.

Example. Given the permutation $\sigma = (1234) \in \mathcal{S}_4$, the cyclic subgroup $< \sigma >$ consists of the elements $\{\sigma = (1234),\ \sigma^2 = (13)(24),\ \sigma^3 = (1432),\ \sigma^4 = \epsilon\}$.

Subgroups can also be constructed by selecting multiple generators and subsequently closing the group. However, subgroups created in this way are not termed ‘cyclic’ subgroups. Note the difference between a cyclic *permutation* and a cyclic *subgroup*. A cyclic permutation is a permutation consisting of only one pair of brackets, a cyclic subgroup is a subgroup generated by just one permutation.

The order of the above described cyclic subgroup $< \sigma >$ is equal to 4. For all subgroups of $\mathcal{S}_n$ generated by a *cyclic* permutation - a permutation consisting of one k-cycle - the order will equal k. When a cyclic subgroup of $\mathcal{S}_n$ is generated by a permutation that is the product of disjoint cyclic permutations, the order can be determined with the following formula. Let σ be a product of r disjoint cyclic permutations of length $k_1, k_2, ..., k_r$. Then the order of $< \sigma >$ equals

$$| < \sigma > | = \text{lcm}(k_1, k_2, ..., k_r), \tag{1}$$

where lcm denotes the least common multiple. For a formal proof of this theorem refer to Proposition 9.8 of Humphreys (1996). For an intuitive explanation of this theorem consider the permutation $\sigma = (153)(26)$. The order of the cyclic subgroup $< \sigma >$ will be equal to the smallest integer s for which $\sigma^s = \epsilon$. In other words, the order of $< \sigma >$ is equal to the number of times the permutation σ has to be repeated until all elements will have ended up in their initial position. In this case, after 3 repetitions of σ elements 1, 5, 3 will be in their initial position but elements 2 and 6 will be interchanged. The smallest value of s for which $\sigma^s = \epsilon$ equals 6, which is the least common multiple of 2 and 3.

3.2 Group-invariance tests

Permutation tests are a subclass of group-invariance tests. In this section, I will give a brief explanation of the construction of group-invariance tests and describe the new methodology of Koning and Hemerik (2021) to enhance the power of these tests by using specific subgroups instead of uniform sampling from the group.

Group-invariance assumption

A group-invariance test is based on the assumption that under the null hypothesis, the distribution of the data is invariant under the transformations of a group. Hence, group-invariance tests do not rely on a parametric, distributional assumption.

If the group-invariance assumption holds, the size of the test is controlled. For a formal proof of this statement, refer to Theorem 1 of Hemerik and Goeman (2021). It is a crucial element of this proof that the set of transformations used actually forms a group. If the set of transformations used does not form a group, the test is not guaranteed to be exact.

I will now give two examples of a group-invariance assumption.

Example. For a permutation tests with data $\mathbf{x} = [x_1, x_2, ..., x_n]$, the group-invariance assumption is that under the null

$$\mathbf{x} \overset{d}{=} \tau(\mathbf{x}), \text{ for all } \tau \in \mathcal{S}_n. \tag{2}$$

Hence, under the null, the distribution of the data does not change upon permutations of the data.

Example. This second example illustrates the group-invariance assumption that is used in the location model as described by Koning and Hemerik (2021). Consider the following DGP,

$$\mathbf{x} = \boldsymbol{\iota}\mu + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are n-vectors and $\boldsymbol{\iota}$ is a n-vector of ones. The group-invariance assumption made in this example is,

$$\boldsymbol{\epsilon} \overset{d}{=} \mathbf{R}\boldsymbol{\epsilon}, \text{ for all } \mathbf{R} \in \mathcal{R}. \tag{4}$$

Here, $\mathbf{R}$ denotes a $(n \times n)$ sign-flipping matrix and $\mathcal{R}$ the group of all $(n \times n)$ sign-flipping matrices. A sign-flipping matrix is a diagonal matrix with diagonal entries equal to -1 or 1. Premultiplying a n-vector with a sign-flipping matrix, flips some of the signs of the elements of the vector. Using the axioms of a group defined in section 3.1, it can be verified that the set of sign-flipping matrices forms a group under matrix multiplication.

The group-invariance assumption on $\boldsymbol{\epsilon}$ holds if the error terms ϵ_i are independent and if the marginal distribution of ϵ_i is reflection symmetric about the y-axis. This is for example the case if the error terms are independent and normally distributed.

The remainder of this chapter will use the location model defined in Equation (3) and the group-invariance assumption from Equation (4) as an example to illustrate the construction of group-invariance tests.

Construction of group-invariance tests

I continue with the location model to illustrate the construction of a group-invariance test. I start by presenting a more intuitive approach to the construction of the group-invariance test and formalize this method afterward. As a group-invariance test is a superclass of the permutation tests described in the introduction, the procedure followed is quite similar.

Given the location model in Equation (3), we want to test the hypothesis $H_0 : \mu = 0$ against $H_a : \mu > 0$ with the test statistic $T(\mathbf{x}) = n^{-1/2}\boldsymbol{\iota}\mathbf{x}$. Note that under H_0, the group-invariance assumption that was made for $\boldsymbol{\epsilon}$ also holds for $\mathbf{x}$.

Under the alternative hypothesis, $T(\mathbf{x})$ is expected to be relatively large. Premultiplying $\mathbf{x}$ with a sign-flipping matrix $\mathbf{R}$ is likely to lower the value of the test statistic as some of the elements of $\mathbf{x}$ will get the opposite sign. The critical value of the test is determined by taking the α-upper quantile of the set $\{T(\mathbf{Rx}) \mid \mathbf{R} \in \mathcal{R}\}$. When $T(\mathbf{x})$ is larger than the critical value, the null hypothesis is rejected.

For a formalisation of this procedure, I will first introduce the concept of orbits. The group $\mathcal{R}$ introduces a partitioning $\mathcal{O}$ of the sample space into sets, which are called orbits. The orbit of an arbitrary vector $\mathbf{a} \in \mathbb{R}^n$ is $O_{\mathbf{a}} := \{\mathbf{Ra} \mid \mathbf{R} \in \mathcal{R}\} \in \mathcal{O}$, where $\mathbf{a}$ is used to represent the orbit.

For the location model, the group invariance assumption allows the disintegration of the distribution of $\boldsymbol{\epsilon}$ into a distribution over $\mathcal{O}$ and a distribution over each orbit. Because of the group-invariance assumption, there exists a conditional distribution $\boldsymbol{\epsilon} \mid \boldsymbol{\epsilon} \in O_{\boldsymbol{\epsilon}}^{\mathcal{R}}$ that is uniform over the elements of $O_{\boldsymbol{\epsilon}}^{\mathcal{R}}$. This is the key property to derive analytical results for these type of tests and will be applied later on.

Using the concept of orbits, the group-invariance test for the location model is constructed as follows. Given the observed data $\mathbf{x}$ the orbit $O_{\mathbf{x}}^{\mathcal{R}} = \{\mathbf{Rx} \mid \mathbf{R} \in \mathcal{R}\}$ and the set of test statistics $T(O_{\mathbf{x}}^{\mathcal{R}}) = \{T(\mathbf{Rx}) \mid \mathbf{R} \in \mathcal{R}\}$ can be determined. For a given level α, the α-upper quantile of the set $T(O_{\mathbf{x}}^{\mathcal{R}})$ is used as critical value. The null hypothesis is rejected if $T(\mathbf{x})$ is larger than the critical value.

Controlling the size of group-invariance tests

The advantage of a group-invariance test is that it controls the size of the test if the group-invariance assumption holds. However, depending on the number of elements in the multiset $T(O_{\mathbf{x}}^{\mathcal{R}})$, it is possible that the α-upper quantile of the set $T(O_{\mathbf{x}}^{\mathcal{R}})$ does not exists. Moreover, assuming that the α-upper qauntile does exists, if the critical value occurs multiple times in the multiset of test statistics, the test is no longer exact. To circumvent both of these issues, the following procedure can be used to determine the critical value and the probability with which the null hypothesis is rejected.

The critical value is chosen as the k-th value in the ordered set of test statistic, denoted with $T^{(k)}(O_{\mathbf{x}}^{\mathcal{R}})$. Setting $k = \lceil (1-\alpha) \,|\, \mathcal{R} \,|\, \rceil$ assures that the test has at most level α as shown by Hemerik and Goeman (2018). To make the test exact, randomization is used in the case that $T(\mathbf{x}) = T^{(k)}(O_{\mathbf{x}}^{\mathcal{R}})$. The null hypothesis is then rejected with a probability,

$$\mathbb{P}(\text{reject } H_0 \mid \mathbf{x},\, \mathcal{R}) = \frac{\alpha|\mathcal{R}| - |\{\mathbf{R} \in \mathcal{R} \mid T(\mathbf{Rx}) > T^{(k)}(O_{\mathbf{x}})\}|}{|\{\mathbf{R} \in \mathcal{R} \mid T(\mathbf{Rx}) = T^{(k)}(O_{\mathbf{x}})\}|} \tag{5}$$

Although this method assures the size of the test is controlled, the results might become less reproducible since rejection depends on a stochastic event.

Note that equation (5) is adapted for the location model discussed here, where a one-sided test is used and the critical value is taken as the α-upper quantile. For other tests, *e.g.* a two-sided test, this equation can easily be adjusted.

Enhancing the power of group-invariance tests using subgroups

In this section, I will explain the basic idea behind the enhancement of the power of group-invariance tests as described by Koning and Hemerik (2021). To this end, I first examine the source of power of a group-invariance test in the location model. To recall, in the location model, we want to test the hypothesis $H_0 : \mu = 0$ against $H_a : \mu > 0$ with the test statistic $T(\mathbf{x}) = n^{-1/2}\boldsymbol{\iota}\mathbf{x}$.

To study the power of the group-invariance test, results are derived conditional on $\boldsymbol{\epsilon} \in O_{\boldsymbol{\epsilon}}^{\mathcal{R}}$. This is convenient as under the group-invariance assumption the distribution of $\boldsymbol{\epsilon}$ is uniform over $O_{\boldsymbol{\epsilon}}^{\mathcal{R}}$. After deriving conditional results, the resulting expressions can be integrated over all orbits $O_{\boldsymbol{\epsilon}}^{\mathcal{R}} \in \mathcal{O}$, which yields general, unconditional properties of the group-invariance test (Chang and Pollard, 1997).

The power of a group-invariance test conditional on $\boldsymbol{\epsilon} \in O_{\boldsymbol{\epsilon}}^{\mathcal{R}}$, is given by

$$\begin{aligned} \text{power} \mid \boldsymbol{\epsilon} \in O_{\boldsymbol{\epsilon}}^{\mathcal{R}} &= \mathbb{P}\big[\text{reject } H_0 \mid H_a,\ \boldsymbol{\epsilon} \in O_{\boldsymbol{\epsilon}}^{\mathcal{R}}\big] && (6)\\ &= \mathbb{P}\big[T(\mathbf{x}) > \alpha\text{-upper quantile of } T(O_{\mathbf{x}}^{\mathcal{R}}) \mid H_a,\ \boldsymbol{\epsilon} \in O_{\boldsymbol{\epsilon}}^{\mathcal{R}}\big]. && (7)\end{aligned}$$

Hence, to study the power of this test it is required to know the distribution of the test statistic $T(\mathbf{x})$. Given that $\boldsymbol{\epsilon}$ is uniformly distributed over $O_{\boldsymbol{\epsilon}}^{\mathcal{R}}$, $T(\boldsymbol{\epsilon}) = n^{-1/2}\boldsymbol{\iota}'\boldsymbol{\epsilon}$ is uniformly distributed over $T(O_{\boldsymbol{\epsilon}}^{\mathcal{R}})$. Hence, the test statistic $T(\mathbf{x}) = T(\boldsymbol{\iota}\mu + \boldsymbol{\epsilon})$ is uniformly distributed over the set

$$\begin{aligned} \big\{T(\boldsymbol{\iota}\mu + \mathbf{R}\boldsymbol{\epsilon}) \mid \mathbf{R} \in \mathcal{R}\big\} &:= \big\{n^{-1/2}\boldsymbol{\iota}'\boldsymbol{\iota}\mu + n^{-1/2}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\epsilon} \mid \mathbf{R} \in \mathcal{R}\big\} \\ &= n^{1/2}\mu + \big\{n^{-1/2}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\epsilon} \mid \mathbf{R} \in \mathcal{R}\big\} \\ &= n^{1/2}\mu + T(O_{\boldsymbol{\epsilon}}^{\mathcal{R}}). \end{aligned} \tag{8}$$

The set $T(O_{\mathbf{x}}^{\mathcal{R}})$ is equal to

$$\begin{aligned} T(O_{\mathbf{x}}^{\mathcal{R}}) &:= \big\{T(\mathbf{R}\mathbf{x}) \mid \mathbf{R} \in \mathcal{R}\big\} \\ &= \big\{n^{1/2}\mu(n^{-1}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\iota}) + n^{-1/2}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\epsilon} \mid \mathbf{R} \in \mathcal{R}\big\}. \end{aligned} \tag{9}$$

Equation (8) and Equation (9) are quite similar. Comparing these equations gives insight into the source of power of this test. For sign-flipping matrices $\mathbf{R}$, it holds that $(n^{-1}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\iota}) \leq 1$. Therefore, for all $\mathbf{R} \in \mathcal{R}$ it follows that

$$n^{1/2}\mu(n^{-1}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\iota}) + n^{-1/2}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\epsilon} \leq n^{1/2}\mu + n^{-1/2}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\epsilon}. \tag{10}$$

Therefore, the power of a group-invariance test depends on the term $(n^{-1}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\iota})$ in Equation (9). If $(n^{-1}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\iota}) = 1$ for all $\mathbf{R} \in \mathcal{R}$ the group-invariance test has trivial power α. If $(n^{-1}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\iota}) < 1$ the test has non trivial power.

The smaller the value of $(n^{-1}\boldsymbol{\iota}'\mathbf{R}\boldsymbol{\iota})$, the smaller the critical value of the test, and hence the larger the power. This forms the core of the methodology of Koning and Hemerik (2021) to enhance the power of group-invariance tests. If it is possible to select a subgroup $\mathcal{Q} \subset \mathcal{R}$ such that $\boldsymbol{\iota}'\mathbf{Q}\boldsymbol{\iota} = 0$ for all $\mathbf{Q} \in \mathcal{Q}$, the power of the test will be larger than when an equally-sized set of random permutations is used.

Koning and Hemerik (2021) refer to such a subgroup as an oracle subgroup. The authors note that for the location model with sample mean test statistic and sample size n, the largest attainable order of an oracle subgroup of the sign-flipping group, is equal to 2^t. Here, t is the number of 2's in the prime factorization of n. In practice, this number will be quite low, which might make the estimate of the critical value less accurate. Therefore, they propose to use near-oracle subgroups $\mathcal{U} \subset \mathcal{R}$ for which $\boldsymbol{\iota}'\mathbf{U}\boldsymbol{\iota}$ is relatively small for all $\mathbf{U} \in \mathcal{U}$.

The benefit of using these oracle or near-oracle subgroups instead of uniform sampling from the group is that either the power of a group-invariance test can be enhanced while using the same number of transformations or the power of the test remains the same while using fewer transformations, which can greatly decrease the computational cost required.

4 Definition of alternative hypothesis and test statistic

In this thesis, I will examine whether the use of subgroups can enhance the power of permutation tests for positive serial dependence in binary data. To this end, I will use an alternative hypothesis and test statistic commonly used in the hot hand debate.

Definition of alternative hypothesis

In tests for positive serial dependence in binary sequences the null hypothesis assumes that the data is generated by a sequence of independent Bernoulli variables with success probability p. There are multiple possibilities for defining the alternative hypothesis based on the description of the hot hand by basketball enthusiasts. Within the basketball community, the hot hand is regarded as an increase in the probability of a hit when a player is on a winning streak. There are multiple ways to formalize this belief into a specification of the alternative hypothesis. Miller and Sanjurjo (2018a) distinguish three forms. First, a regime shift model that can be modeled as a hidden Markov chain. In this case, an external factor triggers the hot hand of a player. Second, a positive feedback model in which the probability of a hit increases after a streak of l hits. Third, a hit streak model, in which a player may enter a hot state with a given probability and remains there for a pre-specified number of throws.

In this thesis, I will opt for the positive feedback model for which the power has been derived by Ritzwoller and Romano (2021). The DGP under the alternative hypothesis can be defined via a Markov chain with 2^m states, where a state represents the outcome of the past m shots. I denote a shooting sequence of n shots with $\mathbf{x} = [x_1, x_2, ..., x_n]$, $x_i \in \{0, 1\}$. Each state is represented by a tuple $(x_1, x_2, ..., x_m) \in \{0, 1\}^m$, where x_1 denotes the most recent shot outcome. For instance, for $m = 3$ the state of a hit followed by two misses equals $(0, 0, 1)$. For most states the probabilities of transitioning from $(x_1, x_2, ..., x_m)$ to $(1, x_1, ..., x_{m-1})$ and $(0, x_1, ..., x_{m-1})$ are equal to, respectively, p and $(1 - p)$. The only exceptions are the states corresponding to a streak of m ones or zeros. For these states, the probability to remain in the same state, thus to elongate the streak, increases by η. Here, η is a positive number less than $\min(1 - p, p)$. For example, the transition probability matrix for $m = 2$ is given in Table 1. From now on I, will simply denote this alternative hypothesis as H_a.

	(0, 0)	(1, 0)	(0, 1)	(1, 1)
(0, 0)	$(1-p)+\eta$	$p-\eta$	0	0
(1, 0)	0	0	$1-p$	p
(0, 1)	$1-p$	p	0	0
(1, 1)	0	0	$(1-p)-\eta$	$p+\eta$

Table 1: Transition probability matrix under the alternative hypothesis for $m = 2$.

In this thesis, I will fix the parameter p at 0.5 as the power of tests for streakiness is highest for this parameterization. Furthermore, I examine sequences of length $n = 6, 9$, and 20. I have decided to focus on such relatively short sequences because these are computationally easier to examine. As noted by Ritzwoller and Romano (2021) these sequence lengths are too short to obtain sufficiently powered tests for probable values of η. The authors propose that a plausible upper bound for η is half of the interquartile range of the distribution of the hit percentages of NBA players. This equals an upper bound on η of 0.038. In this thesis I will examine $\eta \in \{0.2, 0.3, 0.4\}$. I have chosen these relatively high values of η because this results in a higher power of the permutation tests. As a consequence, this parameterization is not representative of the magnitude of a possible hot hand. However, this research focuses on exploring new methodology rather than contributing to the hot hand debate.

Definition of the test statistic

Commonly used test statistics in the hot hand literature are $\hat{\mathbb{P}}(\text{hit} \,|\, l \text{ hits}) - \hat{\mathbb{P}}(\text{hit } \,|\, l \text{ misses})$ and $\hat{\mathbb{P}}(\text{hit } \,|\, l \text{ hits}) - \hat{\mathbb{P}}(\text{hit})$. (Ritzwoller and Romano, 2021; Gilovich et al., 1985) The advantage of these test statistics is that it is also informative of the magnitude of the hot hand. A disadvantage is that these test statistics are not defined if a sequence does not contain a streak of l consecutive hits/misses. For longer sequences this rarely occurs but for short sequences, which I will study in this thesis, it occurs quite frequently. Hence, I have opted for another test statistic, namely the runs test statistic ($T_R(\mathbf{x})$) also employed by Gilovich et al. (1985) and Miller and Sanjurjo (2018a). $T_R(\mathbf{x})$ counts the number of runs in a sequence, *e.g.* $T_R([0, 0, 1, 0, 1, 1]) = 4$. Under the alternative hypothesis streaks occur more often and thus the number of runs will be lower. Therefore, the null hypothesis should be rejected if $T_R(\mathbf{x})$ is relatively small compared to the set $T_R(O^{\mathcal{G}}_{\mathbf{x}})$. Hence, the critical value of a permutation test with $T_R(\mathbf{x})$ test statistic is the α-lower quatile of the set $T_R(O^{\mathcal{G}}_{\mathbf{x}})$.

It is important to note that $T_R(\mathbf{x})$ is a discrete test statistic and hence the same values may occur frequently in the multiset $T_R(O_{\mathbf{x}}^{\mathcal{G}})$. For shorter sequences, this phenomenon is even more pronounced. One of the consequences of such a discrete test statistic is that the critical value might be encountered multiple times in the multiset $T_R(O_{\mathbf{x}}^{\mathcal{G}})$. As described in Section 3.2, this might impact the size of the test. This can be accounted for by rejecting the null hypothesis with a specific probability based on Equation (5), in the case that $T_R(\mathbf{x})$ equals the critical value. Since in this application, the α-lower quantile is taken as critical value the $>$ sign in Equation (5) should be replaced with a $<$ sign.

5 Understanding the difference in performance between subgroups in tests for streakiness

Koning and Hemerik (2021) are able to analytically derive the existence and properties of oracle subgroups in a location model. I have not been able to use a similar approach for permutation tests for streakiness in binary data. This is due to the complex nature of the test statistic and DGP examined. However, the underlying idea that the power of a permutation test can be enhanced by using a subgroup instead of an equally-sized random sample of permutations might still be applicable.

In the following three chapters, I explore two other methods to find streak-breaking subgroups. In this chapter, I present a systematic approach to examine why some subgroups outperform others in tests for streakiness given a specific parameterization of H_a. This approach, in short, focuses on the *most likely* sequences under H_a and aims to explain why $\mathbb{P}(\text{reject } H_0 \,|\, \mathbf{x},\ \mathcal{G})$ for a given subgroup $\mathcal{G}$ deviates from $\mathbb{P}(\text{reject } H_0 \,|\, \mathbf{x},\ \mathcal{S}_n)$, for these most likely sequences $\mathbf{x}$. As a case study, I apply this approach to the best- and worst-performing subgroups of order 24 of $\mathcal{S}_6$.

Given H_a with $m = 2$ and $\eta = 0.3$, I find that for $\mathcal{S}_6$ there exist subgroups that not only beat a random sample of permutations but even the entire symmetric group for many significance levels α. Moreover, I show that this systematic approach is successful in explaining differences in performance based on the structure of these subgroups. However, it seems that these differences in power arise from discreetness effects due to the small sequence lengths studied ($n = 6$). Therefore, I have not been able to extend the findings to longer sequences to identify streak-breaking subgroups for other values of n. For this reason, in the remaining two chapters, I use a heuristic approach to identify possible streak-breaking subgroups of $\mathcal{S}_{20}$ (Chapter 6) and $\mathcal{S}_9$ (Chapter 7).

5.1 Methodology

In this section, I will first discuss the approach I will take to relate the performance of subgroups in tests for streakiness to the structure of these subgroups. Then, I will introduce the case study that I use to demonstrate this approach with.

Relating the power of a subgroup in a test for streakiness to its structure

To understand differences in power between subgroups, I first present an analytical expression for the power of a permutation tests using a given subgroup. Using the law of total probability, the power of a permutation test, given a specific group $\mathcal{G}$, equals

$$\begin{aligned} \text{power} \mid H_a,\ \mathcal{G} &= \mathbb{P}(\text{reject } H_0 \mid H_a,\ \mathcal{G}) \\ &= \sum_{\mathbf{x} \in \mathcal{X}} \mathbb{P}(\text{reject } H_0 \mid \mathbf{x},\ \mathcal{G}) * \mathbb{P}(\mathbf{x} \mid H_a), \end{aligned} \tag{11}$$

where $\mathcal{X}$ denotes the sample space. Hence, the power of these tests is equal to the weighted average of the rejection probabilities of all sequences, where the weighing factor is the likelihood of the sequences. The likelihood of the sequences can vary strongly under H_a. For higher values of the parameters η and n, these differences in likelihood become more pronounced.

The analytical expression for the power given in Equation (11) shows that the sequences with the highest likelihood have the most influence on the overall power of the test. Since it is too complicated to examine the rejection probabilities of all feasible sequences, I propose to focus on the rejection probabilities of the most likely sequences to understand differences in performance between subgroups. These rejection probabilities for a given subgroup can be compared with the rejection probabilities of $\mathcal{S}_n$. Following, remarkable differences in the rejection probabilities of these most likely sequences can be related to the structure of the subgroup.

Case study for sequences of length 6

To illustrate the approach discussed above, I will examine the best- and worst-performing subgroups of order 24 of $\mathcal{S}_6$. I have decided to focus on short sequences of length 6 for two reasons:

1. The main reason is that for $\mathcal{S}_6$ the size of the sample space is only 64. In general, the size of the sample space for sequences of length n equals 2^n. The small sample space for $n = 6$ has two advantages. First, the number of sequences with a relatively high likelihood is limited, which simplifies the analysis described above. For example, if I want to investigate the top 10% sequences with the highest likelihood, I only have to consider 6 sequences. For sequences of length 20, this would correspond to examining 104,857 sequences. Second, since the sample space for sequences of length 6 is small, the power of the test can be calculated exactly with Equation (11). For longer sequences the power should be approximated with simulations.

2. The second reason for studying sequences of length 6 is that it is feasible to search all subgroups of $\mathcal{S}_6$ of a given order and select those with the highest and lowest power in a test for streakiness. The number of subgroups of $\mathcal{S}_n$ grows rapidly with n and hence it would not be possible to examine all subgroups for higher values of n.

For this case study, I decided to examine the best- and worst-performing subgroups of order 24 of $\mathcal{S}_6$. I will refer to these subgroups as $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$. The reasoning for focusing solely on subgroups of order 24 is twofold. First, subgroups of order 24 are not that small that a gain in power might be based on small sample properties. Second, the order of $\mathcal{S}_6$ is only 720. Therefore, larger subgroups would comprise a substantial amount of $\mathcal{S}_6$, which might smooth out differences between subgroups.

Finally, within this chapter the parameters of H_a are fixed at $\eta = 0.3$ and $m = 2$. Additionally, the significance level α is fixed at 3/24.

5.2 Results

In this section, I will first present the analytically determined power of a test for streakiness using the subgroups $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$. Then, I will examine the most likely sequences under H_a, to relate the power of $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ in a test for streakiness to the structure of these subgroups. Finally, I will discuss the power of these subgroups for other significance levels α.

Power of $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ in test for streakiness

Table 2 gives the analytically determined power for $\mathcal{S}_6$, $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ in a test for streakiness with α=3/24. For 24-rnd, the power was simulated using 100.000 Monte Carlo samples. The power of $\mathcal{G}^{low}$ is 3.8 percentage points lower than that of 24-rnd. However, $\mathcal{G}^{high}$ beats 24-rnd by 0.72 percentage points. Moreover, the power of the permutation test with $\mathcal{G}^{high}$ almost equals the power of the test with $\mathcal{S}_6$.

Later on in this section, I aim to relate the difference in power between $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ to the structure of these groups. To be able to understand that reasoning, it is necessary to know the structure of $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$. $\mathcal{G}^{high}$ is generated by the permutations $\sigma^{high} = (1, 2, 5)$ and $\tau^{high} = (1, 5, 2, 3)(4, 6)$ and $\mathcal{G}^{low}$ is generated by $\sigma^{low} = (2, 4, 5)$ and $\tau^{low} = (2, 4, 3, 5)(1, 6)$. Interestingly, these groups have a very similar structure. For both subgroups, all the permutations within the group permute 4 and 2 elements of the sequence separately. For $\mathcal{G}^{high}$, the elements 1, 2, 3, 5 and the elements 4, 6 are permuted separately. For $\mathcal{G}^{low}$, the elements 2, 3, 4, 5 and 1, 6 are permuted separately. All elements of the subgroups are given in Table A.1 in the Appendix.

	24-rnd	$\mathcal{S}_6$	$\mathcal{G}^{high}$	$\mathcal{G}^{low}$
power $\mid \mathcal{G}$	0.1802	0.1883	0.1874	0.1423

Table 2: Analytically determined power of test for streakiness with $T_R(\mathbf{x})$ test statistic, α=3/24 and alternative hypothesis H_a with $m = 2$ and η=0.3.

Examination of most likely sequences under H_a

Now, I will examine the 10 most likely sequences under H_a and describe for each of these sequences what properties of a subgroup will result in a high rejection probability. In the following section, I will relate these properties to the structure of $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ to explain the power of these groups in a test for streakiness.

Table 3 and Table A.2 enumerate all 64 unique binary sequences of length 6. For now, I will only focus on the first two columns $\mathbf{x}$ and $\mathbb{P}(\mathbf{x} \mid H_a)$. The other columns will be discussed later on. The sequences in Table 3 and Table A.2 are ordered from highest to lowest likelihood under H_a with m=2 and $\eta = 0.3$. The likelihood of the sequences vary strongly, from 0.102 for sequences 1 and 2 to 0.003 for sequences 57-64. Hence, based on Equation (11) the most likely sequences have 34 times as much influence on the overall power of the test as the least likely sequences. This justifies the approximation to solely examine the rejection probabilities for the most likely sequences to explain the difference in power between subgroups. Furthermore, the sequences in Table 3 are ordered in pairs that are each other's complement. Since the DGP and test statistic do not discriminate between sequences of 0's and 1's, the results for these pairs are the same.

For the 10 most likely sequences, I provide reasoning which properties of a subgroup will yield a high rejection probability in a test for streakiness. Rejection of the null hypothesis occurs if the test statistic $T_R(\mathbf{x})$ is low compared to the set of test statistics $T_R(O_{\mathbf{x}}^{\mathcal{G}}) = \{T_R(\tau(\mathbf{x})) \mid \tau \in \mathcal{G}\}$. The combination of short sequences with the discrete test statistic $T_R(\mathbf{x})$ will yield multisets $T_R(O_{\mathbf{x}}^{\mathcal{G}})$ in which the same values occur frequently, *e.g.* for $\mathbf{x} = [0, 1, 0, 0, 0, 0]$ the set $T_R(O_{\mathbf{x}}^{\mathcal{S}_6})$ contains only 2's and 3's. Therefore, rejection probabilities will usually not equal 0 or 1 but some intermediate value determined by Equation (5). (Since in this application, the α-lower quantile is taken as critical value the $>$ sign in Equation (5) should be replaced with a $<$ sign).

Sequence 1, 2: [0, 0, 0, 0, 0, 0] and [1, 1, 1, 1, 1, 1]

These sequences do not change upon permutation, hence the rejection probability will be the same across all groups. Based on Equation (5) the rejection probability will equal $\frac{\alpha \mid \mathcal{G} \mid - 0}{\mid \mathcal{G} \mid} = \alpha$.

			$\mathbb{P}(\text{reject } H_0 \mid \mathbf{x},\ \mathcal{G})$			
	$\mathbf{x}$	$\mathbb{P}(\mathbf{x} \mid H_a)$	24-rnd	$\mathcal{S}_6$	$\mathcal{G}^{high}$	$\mathcal{G}^{low}$
1	[0, 0, 0, 0, 0, 0]	0.102	0.125	0.125	0.125	0.125
2	[1, 1, 1, 1, 1, 1]	0.102	0.125	0.125	0.125	0.125
3	[1, 0, 0, 0, 0, 0]	0.064	0.374	0.375	0.5	0.125
4	[0, 1, 1, 1, 1, 1]	0.064	0.374	0.375	0.5	0.125
5	[0, 1, 0, 0, 0, 0]	0.040	0	0	0	0.125
6	[1, 0, 1, 1, 1, 1]	0.040	0	0	0	0.125
7	[1, 1, 1, 1, 1, 0]	0.026	0.375	0.375	0.25	0.125
8	[0, 0, 0, 0, 0, 1]	0.026	0.374	0.375	0.25	0.125
9	[1, 0, 1, 0, 0, 0]	0.025	0	0	0	0
10	[0, 1, 0, 1, 1, 1]	0.025	0	0	0	0
11	[0, 1, 1, 1, 1, 0]	0.016	0.085	0	0	0.125
12	[1, 0, 0, 0, 0, 1]	0.016	0.085	0	0	0.125
13	[0, 0, 0, 0, 1, 1]	0.016	0.767	0.938	1	0.5
14	[1, 1, 1, 1, 0, 0]	0.016	0.768	0.938	1	0.5
15	[1, 0, 1, 0, 1, 0]	0.016	0	0	0	0
16	[0, 1, 0, 1, 0, 1]	0.016	0	0	0	0
17	[0, 1, 0, 1, 0, 0]	0.016	0	0	0	0
18	[1, 0, 1, 0, 1, 1]	0.016	0	0	0	0
19	[1, 1, 0, 0, 0, 0]	0.016	0.766	0.938	0.75	0.5
20	[0, 0, 1, 1, 1, 1]	0.016	0.766	0.938	0.75	0.5
21	[1, 1, 1, 0, 0, 0]	0.016	0.861	1	0.5	0.75
22	[0, 0, 0, 1, 1, 1]	0.016	0.859	1	0.5	0.75
23	[1, 1, 1, 1, 0, 1]	0.016	0	0	0	0.125
24	[0, 0, 0, 0, 1, 0]	0.016	0	0	0	0.125

Table 3: The 24 most likely sequences under H_a with $m = 2$ and $\eta = 0.3$. For each sequence the likelihood and the rejection probability using 24-rnd, $\mathcal{S}_6$, $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ are given. The significance level is set at $\alpha = \frac{3}{24}$.

Sequence 3, 4: [1, 0, 0, 0, 0, 0] and [0, 1, 1, 1, 1, 1]

In this case $T_R(O_{\mathbf{x}}^{\mathcal{S}_6})$ contains only 2's and 3's and $T_R(\mathbf{x}) = 2$. Hence, $T_R(\mathbf{x}) = 2$ is the smallest values in the set $T_R(O_{\mathbf{x}}^{\mathcal{S}_6})$. The rejection probability will be based on Equation (5) and is highest for groups $\mathcal{G}$ for which the number of 2's in the multiset $T_R(O_{\mathbf{x}}^{\mathcal{G}})$ is small. If element 1 of sequence 3/4 is placed at position 1 or 6, the test statistic will equal 2. Hence, the group should scarcely place element 1 at the outer positions 1 and 6.

Sequence 5, 6: [0, 1, 0, 0, 0, 0] and [1, 0, 1, 1, 1, 1]

Here, $T_R(O_{\mathbf{x}}^{\mathcal{S}_6})$ again only contains 2's and 3's and $T_R(\mathbf{x}) = 3$. For a given level α and group $\mathcal{G}$, if $T_R(O_{\mathbf{x}}^{\mathcal{G}})$ contains at least $\alpha * |\mathcal{G}|$ 2's, the rejection probability will equal 0. If this is not the case, the rejection probability will still be relatively low. The highest rejection probability is attained if the multiset $T_R(O_{\mathbf{x}}^{\mathcal{G}})$ consists of only 3's. In this case the rejection probability will equal $\frac{\alpha|\,\mathcal{G}\,|-0}{|\,\mathcal{G}\,|} = \alpha$. Hence, the rejection probability is severely limited for sequences 5 and 6.

Sequence 7, 8: [1, 1, 1, 1, 1, 0] and [0, 0, 0, 0, 0, 1]

This case is similar to that of Sequence 3, 4, *i.e.* $T_R(O_{\mathbf{x}}^{\mathcal{S}_6})$ contains only 2's and 3's and $T_R(\mathbf{x}) = 2$. Again, the rejection probability is highest if the number of 2's in the set $T_R(O_{\mathbf{x}}^{\mathcal{G}})$ is small. This will be the case for a group that does not place element 6 at positions 1 or 6 often.

Sequence 9, 10: [1, 0, 1, 0, 0, 0] and [0, 1, 0, 1, 1, 1]

Here, $T_R(\tau(\mathbf{x})) \in \{2, 3, 4, 5\}$ and $T_R(\mathbf{x}) = 4$. Hence, for most subgroups the rejection probability will equal 0.

This analysis shows that focusing on sequences 1, 2, 5, 6, 9, and 10 is not a wise strategy for constructing subgroups with a high power since the rejection probability for these sequences is severely limited. Hence, a subgroup with high power should do well for sequences 3, 4, 7, and 8.

Rejection probability for sequences 3, 4, 7, 8 for $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$

Given that the rejection probabilities of sequences 3, 4, 7, 8 are most important to explain differences in power amongst subgroups, I examine these rejection probabilities for $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$. I compare the rejection probabilities for $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ with those of $\mathcal{S}_6$. Then, I relate the rejection probabilities for sequences 3, 4, 7, 8 to the structure of the subgroups and aim to identify characteristics of the subgroups that lead to a high or low power in a test for streakiness. The rejection probabilities for all feasible sequences of length 6 using $\mathcal{S}_6$, $\mathcal{G}^{high}$ and, $\mathcal{G}^{low}$ are given in the latter three columns of Table 3 and Table A.2.

$\underline{\mathcal{G}^{high}}$ Table 3 shows that the rejection probability for sequence 3, 4 is 0.125 higher for $\mathcal{G}^{high}$ than for $\mathcal{S}_6$. On the other hand the rejection probability for sequence 7, 8 is 0.125 lower for $\mathcal{G}^{high}$ than for $\mathcal{S}_6$. Because the likelihood of sequences 3, 4 is 2.5 times higher than that of sequences 7, 8, this results in a net positive effect on the power of $\mathcal{G}^{high}$ compared to $\mathcal{S}_6$.

The difference in rejection probabilities between $\mathcal{S}_6$ and $\mathcal{G}^{high}$, can be related to the structure of $\mathcal{G}^{high}$. Previously, I discussed that $\mathbb{P}(\text{reject } H_0 \,|\, \text{sequence } 3/4,\ \mathcal{G})$ will be high for a subgroup $\mathcal{G}$ that does not place element 1 at an outer position frequently. Similarly, $\mathbb{P}(\text{reject } H_0 \,|\, \text{sequence } 7/8,\ \mathcal{G})$ will be high for a subgroup $\mathcal{G}$ that does not place element 6 at an outer position frequently. $\mathcal{G}^{high}$ permutes the elements 1, 2, 3, 5 and 4, 6 of a sequence separately. Hence, element 1 will be placed at an outer position for 1/4 of the permutations in $\mathcal{G}^{high}$ and element 6 will be placed at an outer position for 1/2 of the permutations. The symmetric group $\mathcal{S}_6$ places the elements 1 and 6 at an outer position for 1/3 of the permutations. Thus, $\mathcal{G}^{high}$ obtains a higher rejection probability than $\mathcal{S}_6$ for sequences 3, 4 (for which the rejection probability is high if element 1 is not at an outer position frequently) and a lower rejection probability for sequences 7, 8 (for which the rejection probability is high if element 6 is not at an outer position frequently).

This might raise the question of why $\mathcal{G}^{high}$ is the best performing subgroup of order 24 as it does not obtain a high rejection probability for sequences 7, 8. The answer is that there is no subgroup of $\mathcal{S}_6$ that obtains a higher rejection probability than $\mathcal{S}_6$ for both sequences 3, 4 and sequences 7, 8. Since sequences 3, 4 have a higher likelihood than sequences 7, 8. The subgroup with the highest power should maximize the rejection probability of sequences 3, 4. This is exactly the case for $\mathcal{G}^{high}$.

$\underline{\mathcal{G}^{low}}$ The rejection probabilities of sequences 3, 4, 7, 8 are all equal to 0.125 for $\mathcal{G}^{low}$. These rejection probabilities are 0.250 lower than for $\mathcal{S}_6$. Hence, $\mathcal{G}^{low}$ performs poorly for all of the ‘influential’ sequences.

These low rejection probabilities can again be related to the structure of $\mathcal{G}^{low}$. $\mathcal{G}^{low}$ permutes the elements 2, 3, 4, 5, and 1, 6 of a sequence separately. Hence, elements 1 and 6 will *always* be at an outer position. This results in the low rejection probabilities for sequences 3, 4, 7, 8.

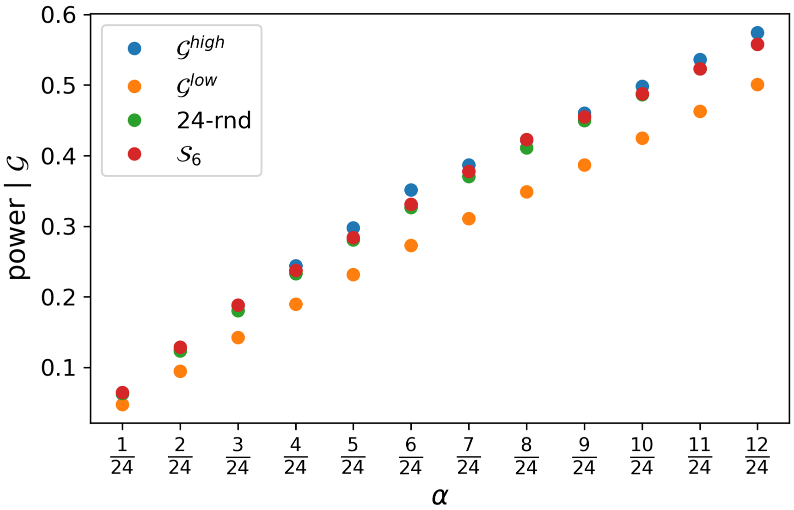

Figure 1: Power of the test for streakiness with alternative hypothesis H_a with parameters $m = 2$ and $\eta = 0.3$ for varying levels of α.

Power of $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ for other values of α

The forgoing analysis focused on a test with a fixed significance level of 3/24. In this section, I will discuss the power of $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ for other values of α.

Figure 1 shows the power of a test for streakiness using $\mathcal{G}^{high}$, $\mathcal{G}^{low}$, 24-rnd, and $\mathcal{S}_6$ for multiple significance levels. This plot shows that $\mathcal{G}^{low}$ performs worse than 24-rnd for all values of α. Surprisingly, $\mathcal{G}^{high}$ obtains a higher power than $\mathcal{S}_6$ for $\alpha > 3/24$, with the exception of $\alpha = 8/24$. The fact that a subgroup beats the entire symmetric group is very remarkable as a higher power is obtained under weaker assumptions, *i.e.* permutation invariance under a subgroup instead of permutation invariance under the entire symmetric group. Statistically, this seems counterintuitive as this would suggest there is a 'free lunch'. However, the higher power of $\mathcal{G}^{high}$ for $\alpha > 3/24$ comes at the cost of the lower power of $\mathcal{G}^{high}$ for smaller values of α.

To explain why $\mathcal{G}^{high}$ outperforms the entire symmetric group for $\alpha > 3/24$, I will again focus on the 'influential' sequences 3, 4. Based on Equation (5) the rejection probability of sequences 3, 4 given $\mathcal{G}^{high}$ or $\mathcal{S}_6$ can be expressed as a function of α.

$$\mathbb{P}(\text{reject } H_0 \,|\, \text{sequence } 3/4,\ \mathcal{S}_6) = \begin{cases} 3\alpha, & \text{if } \ \alpha \leq 7/24 \\ 1, & \text{otherwise} \end{cases} \tag{12}$$

$$\mathbb{P}(\text{reject } H_0 \,|\, \text{sequence } 3/4,\ \mathcal{G}^{high}) = \begin{cases} 4\alpha, & \text{if } \ \alpha \leq 5/24 \\ 1, & \text{otherwise} \end{cases} \tag{13}$$

Combining these expressions yields,

$$\mathbb{P}(\text{reject } H_0 \,|\, \text{sequence } 3/4,\ \mathcal{G}^{high}) - \mathbb{P}(\text{reject } H_0 \,|\, \text{sequence } 3/4,\ \mathcal{S}_6) = \begin{cases} \alpha, & \text{if } \ \alpha \leq 5/24 \\ 1 - 3\alpha, & \text{if } \ 5/24 < \alpha < 7/24 \\ 0, & \text{if } \ \alpha \geq 7/24 \end{cases} . \tag{14}$$

Hence, for $\alpha \leq 6/24$, the difference in rejection probability between $\mathcal{G}^{high}$ and $\mathcal{S}_6$ increases with increasing α. (More precise, the difference in rejection probability is increasing for $\alpha < 19/72$.) This gives an explanation why $\mathcal{G}^{high}$ only outperforms $\mathcal{S}_6$ for higher values of α. Moreover, the advantage of $\mathcal{G}^{high}$ with respect to sequence 3, 4 disappears for $\alpha \geq 7/24$. Thus, the fact that $\mathcal{G}^{high}$ also outperforms $\mathcal{S}_6$ for $\alpha \geq 9/24$ can not be attributed to a high rejection probability for sequences 3, 4 and must be due to differences in rejection probabilities for the less likely sequences.

5.3 Discussion

The aim of this chapter was to explain the differences in performance between subgroups based on the structure of those subgroups. The presented method that focuses on the most likely sequences under H_a was successful in explaining these differences. For the case studied in this chapter, the property that made a subgroup successful was that it should not contain many permutations that place element 1 of the sequence at one of the outer positions. This conclusion is very specific to this case and can be regarded as a discreetness effect of the short sequences studied here. Therefore, this conclusion can not be generalized to longer sequences.

However, the key insight that can be learned from this case study is that successful subgroups are those that have a high rejection probability for the *most likely* sequences. As shown in Table 3 and Table A.2 the likelihood of the most and least likely sequences differs by a factor of 34. Hence, not every sequence contributes equally to the overall power of the test, and subgroups that achieve a high power are those that have high rejection probabilities for the most likely sequences. This conclusion can also be generalized to other parameter settings. The extent to which the likelihood of the sequences differs depends on the sequence length and the parameters m and η of H_a. For example, for $n = 20$ and $m = 2$, the likelihood of the most and least likely sequences differ by a factor of $18 * 10^6$ for $\eta = 0.3$, and a factor 198 for $\eta = 0.1$.

The question remains whether the approach presented in this chapter is feasible for longer sequences. As the size of the sample space grows rapidly with increasing n, it is not feasible to apply the exact same approach for higher values of n. As an alternative approach for longer sequences, a fixed number of the most likely sequences could be examined. This could provide some insight into the structure of high-performing subgroups of $\mathcal{S}_n$.

6 Identifying cyclic streak-breaking subgroups with a heuristic for n=20

As it is not feasible to enumerate all subgroups for larger values of n, another approach is required to identify streak-breaking subgroups for longer sequences. In the following two chapters, I present a heuristic approach to select subgroups with possible streak-breaking properties. I apply this approach to sequences of length 20 and 9, in respectively Chapter 6 and Chapter 7. For both $n = 20$ and $n = 9$, the heuristic is able to identify subgroups that outperform a random sample of permutations for various specifications of H_a.

In this chapter, I limit the scope to cyclic subgroups, as these have some nice properties that help to guide the heuristic search. I apply the heuristic method to identify the most-promising subgroups of $\mathcal{S}_{20}$ of orders 30 and 99. For both of these orders, 4 out of the 15 subgroups that were selected by the heuristic, are streak-breaking. In contrast, when randomly selecting 15 subgroups, none of the subgroups outperform a random sample of permutations.

6.1 Methodology

The heuristic approach I developed to identify streak-breaking, *cyclic* subgroups of $\mathcal{S}_{20}$ consists of three steps.

Step 1: Selecting the most-promising class of subgroups.

Step 2: Selecting the most-promising subgroups within a class.

Step 3: Simulating the power of selected subgroups in tests for streakiness.

Before discussing these three steps, I will clarify why I limit the heuristic search to cyclic subgroups and how I define a 'class of subgroups'.

Advantages of cyclic subgroups

This method focuses on cyclic subgroups, *i.e.* subgroups that are generated by only one permutation. However, it is also possible to create subgroups by selecting multiple generators and combining these to form a group. The reason for focusing on cyclic subgroups is twofold, where both arguments are related to the order of the subgroups. First, for cyclic subgroups, the order of the group can easily be calculated based on the generator of the subgroup using Equation (1). For non-cyclic subgroups, there is no straightforward formula to determine the order of the subgroup based on the set of generators. Second, forming non-cyclic subgroups by combining two permutations of $\mathcal{S}_{20}$, typically results in huge subgroups. These subgroups are not suitable for permutation tests as this requires too much computational power. Moreover, the aim of this thesis is to explore whether it is possible to identify relatively small subgroups that beat an equally-sized random set of permutations or, equivalently, that obtain the same power as a larger set of randomly sampled permutations.

Throughout this chapter, I only focus on cyclic subgroups. For simplicity, I will sometimes drop the adjective 'cyclic' and just refer to cyclic subgroups as subgroups.

Classes of subgroups

In Step 1 of this heuristic method, I select the most-promising class of subgroups. In this section, I will provide a definition of such a class of subgroups. For this definition, I will use the fact that every permutation in $\mathcal{S}_n$ can either be written as a cyclic permutation or as the product of r disjoint cyclic permutations of length $k_1, k_2, ..., k_r$. This was described in more detail in Section 3.1.

Definition. Given that a permutation $\tau \in \mathcal{S}_n$ can be written as the product of r disjoint cyclic permutations of length $k_1, k_2, ..., k_r$, I define the 'cyclic-structure' of τ as $\{k_1, k_2, ..., k_r\}$.
Definition. I define the '$\{k_1, k_2, ..., k_r\}$-class' of subgroups as the set of all *cyclic* subgroups of $\mathcal{S}_n$ that are generated by a permutation with a $\{k_1, k_2, ..., k_r\}$ cyclic-structure.
Example. The permutations $\sigma = (26)(153)$ and $\tau = (15)(246)$ have the same cyclic-structure and the subgroups generated by σ and τ belong to the $\{2, 3\}$-class.

All subgroups within a class will have the same order. This follows from Equation (1).

Step 1: Selecting the most-promising class of subgroups

Now I will discuss the first step of the heuristic method to identify possible streak-breaking subgroups. In this step, the class of subgroups that is most likely to contain streak-breaking subgroups is selected. Some classes are more likely to contain streak-breaking subgroups than others. I introduce a new metric, the Least Permuted Index, to identify which classes are most likely to contain streak-breaking subgroups.

Definition The Least Permuted Index (LPI) of a subgroup $\mathcal{G}$ equals the least number of elements that are permuted by any of the permutations within $\mathcal{G}$ (excluding the identity permutation).

All subgroups within a class have the same LPI. To illustrate this, consider the following example.

Example Consider the permutations $\sigma = (26)(153)$ and $\tau = (15)(246)$. The subgroups generated by σ ($\mathcal{G}_\sigma$) and τ ($\mathcal{G}_\tau$) have order 6. The number of elements permuted by the permutations in $\mathcal{G}_\sigma$ and $\mathcal{G}_\tau$ is the same. This is shown in the enumeration below. The smallest number of permuted elements is 2. Hence, the LPI of both $\mathcal{G}_\sigma$ and $\mathcal{G}_\tau$ equals 2.

$\sigma^0 = \tau^0 = \epsilon$ - 0 elements permuted

σ^1, τ^1 - 5 elements permuted

σ^2, τ^2 - 3 elements permuted

σ^3, τ^3 - 2 elements permuted

σ^4, τ^4 - 3 elements permuted

σ^5, τ^5 - 5 elements permuted

Classes with a high LPI are most likely to contain streak-breaking subgroups. The reasoning behind this is as follows. Consider a subgroup that contains a permutation that only permutes 2 elements of a sequence. I call such a permutations that leaves the majority of the sequence in tact a 'conservative' permutation. Hence, the LPI of this subgroup equals 2. The presence of such a permutation can have a substantial impact on the power of a permutation test for streakiness. This is due to the fact that the null hypothesis in a permutation test is only rejected if $T_R(\mathbf{x})$ is smaller than the α-lower quantile of $T_R(O^{\mathcal{G}}_{\mathbf{x}})$. A few conservative permutations, for which the permuted test statistic is close to the actual test statistic, therefore can have a large impact on the power of the test. The presence of these conservative permutations is captured in the LPI. A high LPI, therefore, gives an indication that each permutation in the subgroup permutes the sequence relatively well.

Only subgroups from classes with a high LPI will be examined for streak-breaking properties. However, the number of subgroups within a class is still huge. Therefore, in Step 2 of this method, I introduce a heuristic to select the most promising subgroups within a class.

Step 2: Selecting the most-promising subgroups within a class

To select the most-promising subgroups from a class with cyclic-structure $\{k_1, k_2, ..., k_r\}$, I have developed the following heuristic procedure.

- Sample 100,000 generators with the cyclic-structure $\{k_1, k_2, ..., k_r\}$ and create 100,000 cyclic subgroups from these generators.
- Check these subgroups against the criterion that none of the permutations within the subgroup is allowed to contain 4 consecutive elements. As such, both the permutation [..., 7, 8, 9, 10, ...] and [..., 10, 9, 8, 7, ...] are not permitted.
- Score the remaining subgroups with a heuristic scoring function that can be used to rank the subgroups for their streak-breaking qualities.

In the last step, I have used the following scoring function (SF),

$$\mathrm{SF}(\sigma) = \sum_{i=1}^{n-1} |\sigma_{i+1} - \sigma_i|. \tag{15}$$

In this formula the array form representation of σ should be plugged in. For example the array form representation of $\sigma = (3, 6, 2)(1, 5)$ is [5, 6, 2, 4, 1, 3] and $\mathrm{SF}(\sigma) = 1 + 4 + 2 + 3 + 2 = 12$.

The scoring function is applied to each permutation within a subgroup $\mathcal{G}$ and the minimum and mean values of the set $\{SF(\sigma \mid \sigma \in \mathcal{G})\}$ are taken as metrics for the quality of the entire group. From now on, I will refer to these metrics as the mean score and the min score of a subgroup. The 15 subgroups with the highest mean score are selected for the simulation study in Step 3.

Step 3: Simulating the power of selected subgroups in tests for streakiness

In this step, I simulate the power of the selected subgroups in tests for streakiness and compare the results with the power of an equally-sized random set of permutations. I examine the power of the test for multiple specifications of the alternative hypothesis. The following parameters of H_a are considered: $m \in \{2, 3\}$ and $\eta \in \{0.2, 0.3, 0.4\}$. Additionally, the level of the test is set at $\alpha = 0.05$ and 100,000 Monte Carlo samples are used to approximate the power of the test.

To determine if the power obtained with a subgroup ($\hat{p}_g$) is statistically different from the power obtained with a random sample of permutations ($\hat{p}_r$), I apply a binomial hypothesis test for equal success probability (Bain and Engelhardt, 2014, Chapter 12.4). When performing l replications in a simulation study resulting in s rejections of the null hypothesis, this outcome can be regarded as drawn from a binomial distribution with l trials and a true rejection probability p. For testing the null hypothesis of equal power of the subgroup and the random sample of permutations ($H_0:\ p_g = p_r$) against the alternative $H_a: p_g \neq p_r$, the following test statistic can be applied,

$$Z = \frac{\frac{s_g}{l_g} - \frac{s_r}{l_r}}{\sqrt{\frac{s_g+s_r}{l_g+l_r} * (1 - \frac{s_g+s_r}{l_g+l_r}) * \frac{1}{l_g+l_r}}}. \tag{16}$$

The two-sided p-value is then given by,

$$\text{p-value} = 2 * \big(1 - \Phi(|Z|)\big), \tag{17}$$

where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution.

A subgroup is classified as 'streak-breaking' if (1) for at least one specification of H_a, the power of the subgroup is significantly higher than the power of the random sample and (2) the power of the subgroup is never significantly lower than the power of the random sample. Here, the significance level is set at 5%.

6.2 Results

In this section, I present the results of each of the three steps of the heuristic method. In Step 1, two classes of subgroups are chosen from which possible streak-breaking subgroups are selected in Step 2. In Step 3, I find that for both of the classes examined, 4 out of the 15 selected subgroups are streak-breaking.

Step 1: Selecting the most-promising class of subgroups

For this study, I want to investigate subgroups of $\mathcal{S}_{20}$ of two different orders. I decided to focus on subgroups of order 30 and on the largest attainable cyclic subgroups of $\mathcal{S}_{20}$. Given these order constraints, I select the most-promising classes of subgroups based on the LPI.

Subgroups of order 30. There are multiple classes that contain subgroups of order 30, *e.g.* {2, 3, 10}-class, {2, 15}-class or {2, 6, 10}-class. The LPI values of these classes equal 3, 2, and 9, respectively. Hence, out of these three options, the latter class is preferred. The {2, 6, 10}-class also has the highest LPI attainable for cyclic subgroups of order 30.

Largest attainable cyclic subgroups. For $n = 20$, the maximal order that can be attained with a cyclic subgroup is 210. The cyclic subgroups of $\mathcal{S}_{20}$ with an order of 210 are all members of the {3, 7, 10}-class. The LPI of this class equals 3. Therefore, this class is not very likely to contain streak-breaking subgroups. Hence, I decided to focus on slightly smaller subgroups. The subgroups of the {9, 11}-class are the largest subgroups of $\mathcal{S}_{20}$ that also attain a high LPI value. The subgroups of the {9, 11}-class have order 99 and a LPI value of 9. The {9, 11}-class is the only class that consists of subgroups of order 99.

The {2, 6, 10}-class and the {9, 11}-class both have a LPI value of 9 and therefore are selected as promising classes of subgroups. In the following step, the most-promising subgroups within these classes are selected with a heuristic.

Step 2: Selecting the most-promising subgroups within a class.

For both of these classes, 100,000 subgroups are generated. Of the 100,000 subgroups of the {3, 6, 10}-class (order=30), 23,057 satisfy the requirement that no permutation within the subgroup can contain 4 consecutive elements. For the subgroups of the {9, 11}-class (order=99), 25,987 pass this first criterion. The mean scores for the remaining subgroups are in the range [88.552, 143.138] for the subgroups of the {3, 6, 10}-class and in the range [105.235, 139.582] for the subgroups of the {9, 11}-class.

For both of these classes, the 15 subgroups with the highest mean scores are selected for a simulation study in Step 3. The mean scores, min scores, and generators of these subgroups are given in Table 4. The subscripts 30 and 99 refer to the order of the group. The underlined groups are those that perform significantly better than an equally-sized random subset of permutations and are classified as 'streak-breaking'. For both classes, 4 out of the 15 selected subgroups perform significantly better than an equally-sized random sample. The simulation results of these groups will be discussed in the results of Step 3.

Based on the results in Table 4, we can examine the ability of the heuristic score measure to select streak-breaking subgroups. The fact that, for both classes examined, only 4 out of the 15 selected subgroups are streak-breaking indicates that the mean score is not a fail-proof metric to select streak-breaking subgroups. Moreover, the actual streak-breaking subgroups do not have the highest mean scores out of the selected subgroups.

Additionally, it is remarkable that all streak-breaking subgroups have a relatively high min score, but not all groups with a high min score perform well. Hence, a high min score is a necessary

	mean score	min score	generator
$\mathcal{G}_{30}^{a}$	143.138	98	(3, 4, 19)(1, 11, 8, 18, 12, 16)(2, 7, 6, 13, 10, 9, 17, 15, 5, 20)
$\underline{\mathcal{G}_{30}^{b}}$	142.931	99	(8, 16, 9)(2, 5, 18, 4, 10, 13)(1, 14, 3, 20, 7, 6, 17, 12, 11, 15)
$\mathcal{G}_{30}^{c}$	142.31	90	(4, 6, 10)(1, 14, 17, 7, 2, 13)(3, 19, 15, 12, 9, 11, 5, 16, 20, 18)
$\mathcal{G}_{30}^{d}$	142.138	100	(2, 9, 11)(6, 19, 20, 10, 13, 15)(1, 4, 7, 5, 18, 3, 14, 12, 8, 16)
$\mathcal{G}_{30}^{e}$	141.759	92	(6, 18, 17)(1, 7, 3, 11, 14, 5)(2, 19, 16, 20, 4, 9, 8, 12, 10, 15)
$\mathcal{G}_{30}^{f}$	141.69	102	(3, 4, 17)(1, 9, 7, 12, 8, 15)(2, 6, 11, 14, 10, 5, 13, 19, 18, 20)
$\underline{\mathcal{G}_{30}^{g}}$	141.655	100	(7, 8, 16)(4, 18, 9, 11, 12, 20)(1, 13, 19, 3, 15, 17, 6, 5, 10, 14)
$\mathcal{G}_{30}^{h}$	141.517	102	(4, 19, 10)(1, 9, 5, 7, 12, 15)(3, 11, 16, 6, 13, 14, 17, 18, 8, 20)
$\mathcal{G}_{30}^{i}$	141	88	(1, 8, 12)(3, 17, 5, 14, 7, 4)(2, 20, 10, 19, 18, 6, 9, 13, 15, 16)
$\mathcal{G}_{30}^{j}$	140.793	76	(2, 6, 4)(1, 10, 7, 17, 14, 18)(3, 15, 19, 20, 16, 12, 5, 13, 11, 9)
$\mathcal{G}_{30}^{k}$	140.793	69	(6, 15, 16)(2, 18, 12, 8, 20, 14)(1, 9, 17, 5, 3, 11, 4, 19, 13, 10)
$\underline{\mathcal{G}_{30}^{l}}$	140.621	101	(2, 9, 20)(5, 18, 16, 19, 14, 12)(1, 7, 3, 15, 13, 11, 8, 17, 10, 6)
$\underline{\mathcal{G}_{30}^{m}}$	140.621	103	(12, 20, 15)(2, 14, 17, 7, 10, 6)(1, 19, 16, 3, 9, 13, 8, 4, 5, 11)
$\mathcal{G}_{30}^{n}$	140.448	105	(8, 20, 11)(1, 6, 17, 3, 12, 13)(2, 7, 15, 10, 14, 9, 5, 4, 18, 19)
$\mathcal{G}_{30}^{o}$	140.241	101	(3, 19, 14)(1, 9, 6, 17, 10, 11)(2, 12, 20, 13, 15, 7, 5, 8, 18, 4)
$\mathcal{G}_{99}^{a}$	139.582	85	(2, 9, 13, 6, 17, 19, 4, 12, 8)(1, 18, 7, 20, 3, 11, 10, 15, 16, 5, 14)
$\mathcal{G}_{99}^{b}$	139.378	93	(2, 6, 4, 7, 16, 19, 8, 11, 15)(1, 10, 20, 14, 3, 5, 13, 18, 12, 17, 9)
$\mathcal{G}_{99}^{c}$	139.296	91	(2, 14, 5, 18, 19, 10, 15, 4, 6)(1, 13, 7, 8, 20, 3, 16, 11, 12, 9, 17)
$\underline{\mathcal{G}_{99}^{d}}$	139.153	94	(3, 9, 20, 14, 12, 15, 6, 18, 17)(1, 19, 5, 2, 8, 10, 11, 16, 7, 4, 13)
$\underline{\mathcal{G}_{99}^{e}}$	138.969	95	(2, 10, 20, 8, 16, 12, 13, 9, 4)(1, 17, 3, 15, 18, 11, 6, 7, 14, 5, 19)
$\mathcal{G}_{99}^{f}$	138.888	69	(2, 12, 6, 18, 4, 5, 10, 11, 16)(1, 3, 20, 14, 17, 15, 13, 9, 7, 19, 8)
$\mathcal{G}_{99}^{g}$	138.745	72	(4, 14, 18, 16, 20, 10, 17, 6, 11)(1, 7, 9, 19, 3, 8, 2, 5, 13, 15, 12)
$\underline{\mathcal{G}_{99}^{h}}$	138.52	87	(3, 10, 16, 12, 14, 20, 9, 6, 13)(1, 11, 4, 5, 17, 2, 18, 19, 7, 8, 15)
$\mathcal{G}_{99}^{i}$	138.52	73	(4, 10, 18, 12, 16, 8, 20, 15, 14)(1, 19, 11, 17, 6, 9, 7, 2, 3, 13, 5)
$\mathcal{G}_{99}^{j}$	138.5	76	(4, 10, 13, 8, 12, 15, 20, 18, 16)(1, 2, 7, 3, 19, 14, 6, 9, 5, 11, 17)
$\mathcal{G}_{99}^{k}$	138.439	85	(3, 4, 18, 8, 16, 20, 12, 9, 15)(1, 10, 14, 17, 5, 6, 11, 19, 7, 2, 13)
$\mathcal{G}_{99}^{l}$	138.357	94	(2, 18, 19, 4, 13, 7, 16, 8, 9)(1, 10, 12, 6, 15, 17, 20, 11, 14, 3, 5)
$\underline{\mathcal{G}_{99}^{m}}$	138.337	91	(2, 11, 13, 20, 14, 10, 17, 18, 6)(1, 5, 12, 15, 8, 16, 19, 4, 3, 9, 7)
$\mathcal{G}_{99}^{n}$	138.316	86	(3, 18, 17, 7, 9, 15, 20, 14, 12)(1, 16, 8, 10, 4, 5, 13, 19, 6, 2, 11)
$\mathcal{G}_{99}^{o}$	138.296	71	(2, 4, 7, 9, 12, 19, 16, 11, 8)(1, 5, 17, 10, 14, 6, 13, 15, 18, 20, 3)

Table 4: Mean scores, min scores, and generators of the 15 selected cyclic subgroups of order 30 and 99. The streak-breaking subgroups are underlined.

but not a sufficient condition for a streak-breaking subgroup. Intuitively, it seems logical that a high min score is a necessary condition. A high min score indicates that all permutations in the subgroup permute the elements well. As discussed in the previous section, if a subgroup contains a few conservative permutations (which would result in a lower min score), the power of these tests can be seriously weakened. Finally, the reason as to why subgroups with comparable scores, *e.g.* $\mathcal{G}_{30}^{a}$ and $\mathcal{G}_{30}^{b}$, perform very differently in these tests, is not clear yet. However, finding an explanation for this could be interesting for further research.

Step 3: Simulating the power of selected subgroups in tests for streakiness

Table 5 shows the results of the simulation study for the streak-breaking subgroups. The subgroups $\mathcal{G}_{30}^{b}$ and $\mathcal{G}_{99}^{d}$ achieved a gain in power for 5 out of the 6 parameter settings studied and hence might be regarded the 'best' subgroups within their category. The surge in power was in the range of [0.41, 0.76] percentage points for $\mathcal{G}_{30}^{b}$ and in the range of [0.57, 1.02] for $\mathcal{G}_{99}^{d}$. The highest gain in power amounted to 1.48 percentage points and was achieved by $\mathcal{G}_{99}^{m}$ for $\eta = 0.4$ and $m = 3$.

Furthermore, it is interesting to note that whether a subgroup is better than a Monte Carlo sample of permutations strongly depends on the specific alternative hypothesis that is examined. None of the subgroups investigated outperformed the Monte Carlo sample for all six specifications of the alternative hypothesis examined. Moreover, the cases in which the subgroups achieve a gain in power differ across subgroups.

6.3 Discussion

The results showed that it is possible to identify streak-breaking subgroups of different orders for sequences of length $n = 20$. Especially the fact that streak-breaking subgroups of order 99 were found, gives an indication that it is possible to find such subgroups of higher-order and that this is not just an artifact of small groups are short sequences. Hence, based on these results, it seems that there is potential to apply this method to different sequences lengths.

With this approach, only cyclic subgroups were examined. For $\mathcal{S}_{20}$, the largest cyclic subgroup has order 210. This might be regarded as a limitation of this method as this might restrict the power that can be obtained with such a subgroup, especially when the power is compared to a test that uses a few thousand random permutations. However, for longer sequences, this limitation disappears. For example, in the hot hand literature often a sequence length of 100 throws is used. (Gilovich et al., 1985; Miller and Sanjurjo, 2021). For $\mathcal{S}_{100}$, there exist cyclic subgroups of a much higher order than

	$\mathcal{G}_{30}^{b}$				$\mathcal{G}_{30}^{g}$				$\mathcal{G}_{30}^{l}$			
η	m=2		m=3		m=2		m=3		m=2		m=3	
0	0.04931	-	0.04938	-	0.05010	-	0.04963	-	0.05050	-	0.05007	-
0.2	0.22505	(0.44)	0.13401	(0.41)	0.22464	(0.40)	0.13293	(0.30)	0.22336	-	0.13563	(0.57)
0.3	0.35870	(0.76)	0.19927	(0.74)	0.35470	-	0.19568	(0.38)	0.35603	(0.50)	0.19535	(0.35)
0.4	0.41994	-	0.25016	(0.64)	0.43023	(0.91)	0.24051	-	0.42288	-	0.24582	-
	$\mathcal{G}_{30}^{m}$				30-rnd							
0	0.05018	-	0.05018	-	0.05138	-	0.05014	-				
0.2	0.22303	-	0.13230	-	0.22065	-	0.12991	-				
0.3	0.35730	(0.62)	0.19486	-	0.35107	-	0.19184	-				
0.4	0.43167	(1.06)	0.25196	(0.82)	0.42108	-	0.24373	-				
	$\mathcal{G}_{99}^{d}$				$\mathcal{G}_{99}^{e}$				$\mathcal{G}_{99}^{h}$			
0	0.04799	-	0.05099	-	0.05123	-	0.04992	-	0.05036	-	0.04844	-
0.2	0.24113	-	0.14336	(0.57)	0.23979	-	0.14102	(0.34)	0.23580	-	0.13942	-
0.3	0.38737	(0.80)	0.21357	(0.91)	0.38645	(0.71)	0.20887	(0.44)	0.37925	-	0.20881	(0.44)
0.4	0.45462	(0.79)	0.26905	(1.02)	0.45755	(1.08)	0.26735	(0.85)	0.45198	(0.53)	0.26035	-
	$\mathcal{G}_{99}^{m}$				99-rnd							
0	0.04991	-	0.05086	-	0.05064	-	0.04864	-				
0.2	0.23948	-	0.14283	(0.52)	0.23927	-	0.13763	-				
0.3	0.38330	-	0.21132	(0.69)	0.37936	-	0.20443	-				
0.4	0.45724	(1.05)	0.27364	(1.48)	0.44671	-	0.25888	-				

Table 5: Power of test for streakiness with $T_R(\mathbf{x})$ test statistic, $\alpha = 0.05$ and alternative hypothesis with $m \in \{2, 3\}$ and $\eta \in \{0, 0.2, 0.3, 0.4\}$. The power of the test is given for the streak-breaking subgroups of order 30 and 99, 30-rnd and 99-rnd. For the cases that a subgroup outperformed the random sample on a 5% significance level, the difference in power in %-points is given between brackets.

for $\mathcal{S}_{20}$. For instance, subgroups of $\mathcal{S}_{100}$ with the cyclic-structure $\{20, 20, 59\}$ have an order of 1180 and LPI equal to 40. Therefore, this method has a lot of potential to be extended to longer sequences in future research.

The heuristic presented is not fail-proof as groups with similar scores, *e.g.* $\mathcal{G}_{30}^{a}$ and $\mathcal{G}_{30}^{b}$ can perform very differently in tests for streakiness. However, for both cyclic structures examined, 27% of the selected subgroups turned out to be streak-breaking. When 15 randomly selected subgroups of orders 30 and 99 are used in simulations, none of these subgroups turned out to be streak-breaking. This shows that the heuristic is able to identify streak-breaking subgroups. Additionally, the percentage of streak-breaking subgroups might even be increased if the subgroups used in the simulation study are selected on both a high mean and a high min score. In this research, the subgroups were only selected on a high mean score. However, based on the presented results, it seems as though a high min score is a necessary condition for a streak-breaking subgroup.

Finally, it would be an interesting objective for future research to identify why some subgroups are better than others. Determining this underlying reason might also provide a starting point to improve the heuristic selection criteria.

7 Identifying streak-breaking subgroup with a heuristic for n=9

The previous chapter focused on finding *cyclic* streak-breaking subgroups of $\mathcal{S}_{20}$ with a heuristic. For short sequences, however, the order of cyclic subgroups is severely limited. For example, the largest cyclic subgroup of $\mathcal{S}_9$ has order 20. Hence, focusing on merely cyclic subgroups seems too restrictive for such short sequences.

In this chapter, I focus on sequences of length $n = 9$ and extend the methodology discussed in the previous chapter to include all subgroups. To this end, I create subgroups by combining two generators. Following, I examine the largest subgroups of $\mathcal{S}_9$ that satisfy the constraints of the heuristic (order=216). I find that all ten subgroups of order 216 that satisfy the constraints, turn out to be streak-breaking.

7.1 Methodology

In Section 6.1 it was discussed that the order of cyclic subgroups can be determined easily. Hence, one can selectively construct subgroups of a given order. In this chapter, I will not examine cyclic subgroups but subgroups that are created by combining two generators. When creating subgroups in this way, it is not trivial to know upfront what the order of the group will be. Therefore, in this chapter, I am not able to generate subgroups of a specific order. To account for this, I have changed Step 1 of the heuristic procedure to generate subgroups of any order. Then, in Step 2 the most-promising subgroups of a given order are selected. Step 3 remains the same as in the previous chapter and hence will not be discussed here.

Step 1: Generating subgroups by randomly combining permutations

In this step, I generate subgroups by combining two generators to form a group. To this end, I randomly sample 5000 permutations from $\mathcal{S}_n$ and create groups by combining every possible pair of those 5000 permutations. Only the subgroups with an order between 25 and 3000 are selected to continue to Step 2.

Step 2: Selecting the most-promising subgroups of a given order

For the resulting set of subgroups, I apply the same heuristic criteria as in Step 2 of Chapter 6 to select the most-promising subgroups. The first criteria is that none of the permutations within a subgroup is allowed to contain 4 consecutive elements. The second criterion consists of the mean scores of the remaining subgroups. For a given order, the subgroups with the highest mean scores are selected for a simulation study.

Note that the first criterion is more restrictive for subgroups of $\mathcal{S}_9$ than for subgroups of $\mathcal{S}_{20}$. Additionally, permutations with 3 consecutive elements are also quite conservative for sequences of length 9, while the heuristic does not account for this. Therefore, I have introduced an additional metric that counts the number of permutations in a group with 3 consecutive elements. I will refer to this metric as '#consecutive(3)'.

7.2 Results

In this section, I present the results of the heuristic procedure for $n = 9$. With a simulation study, I examine the power in tests for streakiness for ten subgroups of order 216. I find that all of these ten subgroups are streak-breaking.

Results of Step 1 and Step 2

The combining of 5000 randomly generated permutations results in 1700 unique subgroups that do not contain any permutations with 4 consecutive elements and have an order between 25 and 3000. The order of these groups ranged from 27 to 216. The reason that no subgroups with an order larger than 216 are found, is that the constraint on the absence of permutations with 4 consecutive elements is too restrictive.

As discussed before, the search for streak-breaking subgroups of $\mathcal{S}_9$ is extended to include non-cyclic subgroups because the order of cyclic subgroups is too small. Since I am interested in finding somewhat larger subgroups that beat a Monte Carlo sample of permutations, I examine the largest subgroups found in this search, *i.e.* the subgroups of order 216. In total, there are 2364 unique subgroups of order 216 identified in this search. However, only 10 of these subgroups do not contain any permutations with 4 consecutive elements. Table 6 gives the metrics and generators of these subgroups. Interestingly, all of the groups have the same mean score. The underlying reason for this phenomenon is not clear yet.

Step 3: Simulating the power of selected subgroups in a test for streakiness

The power of these ten subgroups in tests for streakiness is given in Table 7. All of the ten subgroups of order 216 are streak-breaking meaning that they are never significantly worse than 216-rnd and in some cases significantly better.

	mean score	min score	#consec(3)	generators
$\mathcal{G}_{216}^{a}$	26.754	16	27	(1, 7, 3, 9, 5, 4)(2, 6) and (1, 3, 6)(2, 4, 8)(5, 9, 7)
$\mathcal{G}_{216}^{b}$	26.754	16	27	(1, 5, 7, 3, 4, 2)(8, 9) and (1, 7, 2)(3, 5, 8)(4, 6, 9)
$\mathcal{G}_{216}^{c}$	26.754	15	33	(1, 4)(2, 5, 3, 6, 7, 8) and (1, 3, 2, 4)(6, 7, 8, 9)
$\mathcal{G}_{216}^{d}$	26.754	15	33	(1, 7, 4, 5, 8, 6)(2, 9) and (1, 8, 3, 9)(2, 6, 4, 7)
$\mathcal{G}_{216}^{e}$	26.754	15	30	(1, 6, 3, 7, 9, 5) (2, 4) and (1, 3, 6, 7)(4, 5, 8, 9)
$\mathcal{G}_{216}^{f}$	26.754	14	30	(1, 7, 3, 6, 4, 8)(2, 9) and (1, 2, 8, 5)(3, 6, 7, 4)
$\mathcal{G}_{216}^{g}$	26.754	14	30	(1, 3, 4, 8, 6, 9)(2, 7) and (1, 6, 3, 2, 7, 8)(4, 5)
$\mathcal{G}_{216}^{h}$	26.754	13	35	(1, 4, 9, 5)(2, 6, 3, 8) and (1, 8, 6)(2, 9, 4)(3, 5, 7)
$\mathcal{G}_{216}^{i}$	26.754	12	28	(1, 5)(2, 6, 4, 3, 9, 8) and (1, 9, 2)(3, 7, 5)(4, 6, 8)
$\mathcal{G}_{216}^{j}$	26.754	12	28	(2, 4, 7)(3, 5, 9) and (1, 3, 9)(2, 4, 5)(6, 8, 7)

Table 6: Mean scores, min scores, #consec(3), and generators of the 10 groups of order 216 that were selected.

The groups $\mathcal{G}_{216}^{h}$ and $\mathcal{G}_{216}^{i}$ achieve a significant gain in power for all specifications of the alternative hypothesis examined. The increase in power ranges from 0.35 to 0.94 for $\mathcal{G}_{216}^{h}$ and from 0.37 to 1.06 for $\mathcal{G}_{216}^{i}$. For both groups, the largest increase in power is realized for $m = 2$ and $\eta = 0.4$, this amounted to 0.94 and 1.06 percentage points, respectively. Although these latter two groups might be considered the 'best' subgroups examined, this was not necessarily expected based on the metrics investigated. Table 6 shows that $\mathcal{G}_{216}^{h}$ and $\mathcal{G}_{216}^{i}$ have a relatively low min score and $\mathcal{G}_{216}^{h}$ even has the highest number of permutations with 3 consecutive elements. Hence, these metrics do not fully capture the streak-breaking qualities of a subgroup.

7.3 Discussion

The methodology presented is successful in identifying streak-breaking subgroups of a relatively large order for sequences of length 9. The largest streak-breaking subgroups found have an order of 216. Some of these subgroups even beat 216-rnd for all specifications of the alternative hypothesis examined. It is probably possible to find larger streak-breaking subgroups if the constraint on the presence of permutations with 4 consecutive elements is relaxed. This could be done by allowing that, for example, a maximum of 2.5% of the permutations may contain 4 consecutive elements.

Interestingly, all 10 subgroups of order 216 that did not contain 4 consecutive elements were streak-breaking. For subgroups of different orders, this is not generally the case. As mentioned in the previous chapter, it is not clear yet what property makes that some subgroups outperform others. This would be valuable to know and hence is an interesting topic for future research.

	$\mathcal{G}_{216}^a$				$\mathcal{G}_{216}^b$				$\mathcal{G}_{216}^c$			
η	m=1		m=2		m=1		m=2		m=1		m=2	
0	0.05109	-	0.05094	-	0.04955	-	0.04958	-	0.04954	-	0.05017	-
0.2	0.22638	-	0.12434	(0.31)	0.22707	(0.42)	0.12389	-	0.22841	(0.56)	0.12426	(0.30)
0.3	0.33935	(0.70)	0.15355	-	0.33623	-	0.15588	(0.32)	0.33735	(0.50)	0.15478	-
0.4	0.35371	(0.82)	0.15547	(0.39)	0.35578	(1.02)	0.15569	(0.41)	0.35733	(1.18)	0.15413	-
	$\mathcal{G}_{216}^d$				$\mathcal{G}_{216}^e$				$\mathcal{G}_{216}^f$			
0	0.05053	-	0.04969	-	0.04961	-	0.05073	-	0.04968	-	0.04857	-
0.2	0.22665	(0.38)	0.12114	-	0.22544	-	0.12256	-	0.22475	-	0.12387	-
0.3	0.33842	(0.60)	0.15109	-	0.33862	(0.62)	0.15400	-	0.33881	(0.64)	0.15515	-
0.4	0.35812	(1.26)	0.15411	-	0.35420	(0.87)	0.15476	-	0.35701	(1.15)	0.15466	-
	$\mathcal{G}_{216}^g$				$\mathcal{G}_{216}^h$				$\mathcal{G}_{216}^i$			
0	0.04920	-	0.04984	-	0.04990	-	0.0502	-	0.05055	-	0.05077	-
0.2	0.22839	(0.55)	0.12454	(0.33)	0.22699	(0.41)	0.12570	(0.45)	0.22802	(0.52)	0.12522	(0.40)
0.3	0.33912	(0.67)	0.15437	-	0.33897	(0.66)	0.15616	(0.35)	0.33773	(0.53)	0.15634	(0.37)
0.4	0.35550	(1.00)	0.15469	-	0.35490	(0.94)	0.15577	(0.42)	0.35614	(1.06)	0.15551	(0.39)
	$\mathcal{G}_{216}^j$				216-rnd							
0	0.04999	-	0.05089	-	0.05014	-	0.05038	-				
0.2	0.22645	-	0.12687	(0.56)	0.22285	-	0.12125	-				
0.3	0.33917	(0.68)	0.16045	(0.78)	0.33239	-	0.15268	-				
0.4	0.35514	(0.96)	0.16020	(0.86)	0.34555	-	0.15162	-				

Table 7: Power of test for streakiness with $T_R(\mathbf{x})$ test statistic, $\alpha = 0.05$ and alternative hypothesis with $m \in \{1, 2\}$ and $\eta \in \{0, 0.2, 0.3, 0.4\}$. The power of the test is given for the selected subgroups of order 216 and 216-rnd. For the cases that a subgroup outperformed the random sample on a 5% significance level, the difference in power in %-points is given between brackets.

Finally, this method is not suitable for long sequences. Even for sequences of length 20, randomly combining generators results in huge subgroups. Hence, for long sequences using cyclic subgroups, as described in the previous chapter, is recommended.

8 Conclusion

In this thesis, I have identified subgroups of the permutation group that obtain a higher power in a permutation test for positive serial dependence in binary sequences than an equally-sized Monte Carlo sample of permutations. I have identified such streak-breaking subgroups for sequences of lengths 6, 9, and 20.

Streak-breaking subgroups for sequences of lengths 6, 9, and 20

For sequences of length 6, I have found a subgroup of order 24 that not only outperforms an equally-sized random sample of permutations but even the entire symmetric group for many significance levels α. This is very remarkable as a higher power is obtained under weaker assumptions, *i.e.* permutation invariance under a subgroup instead of permutation invariance under the entire symmetric group.

For sequences of length 9 and 20, I have been able to identify streak-breaking subgroups based on a heuristic. I have presented two variations of this heuristic method. For long sequences, *e.g.* $n = 20$, this heuristic focuses solely on *cyclic* streak-breaking subgroups while for short sequences, *e.g.* $n = 9$, non-cyclic subgroups are examined. For $n = 20$, a cyclic subgroup of order 99 ($\mathcal{G}^d_{99}$) is found that achieves a gain in power ranging from 0.57 to 1.02 percentage points for various specifications of the alternative hypothesis. For $n = 9$, the best performing subgroup of order 216 ($\mathcal{G}^i_{216}$) achieves an increase in power between 0.37 and 1.06 percentage points.

Therefore, it can be concluded that the method developed by Koning and Hemerik (2021) to enhance the power of group-invariance tests can be extended to a completely different DGP and test statistic.

Understanding the differences in power between subgroups

In this research, I have not been able to fully explain differences in performance between subgroups based on the characteristics of these subgroups.

In Chapter 5, I have presented a systematic approach to examine why some subgroups perform better than others in permutation tests for positive serial dependence - given a particular specification of H_a. The key insight of this systematic approach is that streak-breaking subgroups are those that have a high rejection probability for the *most likely* sequences under the alternative hypothesis. For sequences of length 6, this method showed to be insightful to explain differences between subgroups. However, this result was based on discreetness effects of the short sequences studied and hence could not be generalized to identify streak-breaking subgroups for longer sequences.

Furthermore, in Chapter 6, subgroups were found with very similar scores on the heuristic measures but very different performances in tests for streakiness ($\mathcal{G}_{30}^{a}$ and $\mathcal{G}_{30}^{b}$). It is not clear yet what the underlying reason for this phenomenon is.

Suggestions for future research

Since this is the first research that examines the application of subgroups in permutation tests for streakiness, there are still some questions left answered. In particular, I have two suggestions that would be interesting to examine in future research.

First, it would be interesting to gain more insight into why some subgroups perform better than others in tests for streakiness. This could be done by applying the systematic approach of Chapter 5 to the subgroups of $\mathcal{S}_{20}$ that were selected by the heuristic in Chapter 6. Due to the large sample space for $n = 20$, it is not possible to enumerate all feasible sequences as I have done for $n = 6$. Therefore, I recommend focusing on the 50 most likely sequences under H_a and aim to find an explanation why subgroups with very similar scores on the heuristic measures, perform very differently in tests for streakiness. Determining the underlying reason for this difference in performance might also provide a starting point to improve the heuristic selection criteria used in Chapter 6 and Chapter 7.

Second, it would be interesting to extend the heuristic method, developed to identify candidate streak-breaking subgroups, to longer sequence lengths. In the hot hand literature, often shot sequences of at least 100 throws are studied. It would be interesting to see whether it is possible to identify streak-breaking subgroups of $\mathcal{S}_{100}$ with the heuristic approach. An additional advantage of studying longer sequence lengths is that the maximal order attainable for cyclic subgroups increases.

References

Armstrong, M. A. (1988). *Groups and symmetry.* Springer Science & Business Media.

Avugos, S., Bar-Eli, M., Ritov, I., and Sher, E. (2013). The elusive reality of efficacy–performance cycles in basketball shooting: an analysis of players' performance under invariant conditions. *International Journal of Sport and Exercise Psychology*, 11(2):184–202.

Bain, L. J. and Engelhardt, M. (2014). *Introduction to probability and mathematical statistics.* Brooks/Cole. (Custom Edition for the Econometric Institute in Rotterdam).

Barberis, N. (2018). Psychology-based models of asset prices and trading volume. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 79–175. Elsevier.

Barberis, N. and Thaler, R. (2003). A survey of behavioral finance. *Handbook of the Economics of Finance*, 1:1053–1128.

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in medicine*, 19(10):1319–1328.

Chang, J. T. and Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317.

Ernst, M. D. et al. (2004). Permutation methods: a basis for exact inference. *Statistical Science*, 19(4):676–685.

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, 38(1):34–105.

Fisher, R. A. (1937). *The design of experiments.* Oliver & Boyd, Edinburgh.

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

Hemerik, J. and Goeman, J. J. (2018). Exact testing with random permutations. *Test*, 27(4):811–825.

Hemerik, J. and Goeman, J. J. (2021). Another look at the lady tasting tea and differences between permutation tests and randomisation tests. *International Statistical Review*, 89(2):367–381.

Humphreys, J. F. (1996). *A course in group theory.* Oxford University Press.

Koehler, J. J. and Conley, C. A. (2003). The "hot hand" myth in professional basketball. *Journal of sport and exercise psychology*, 25(2):253–259.

Kofler, R. and Schlötterer, C. (2012). Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*, 28(15):2084–2085.

Koning, N. and Hemerik, J. (2021). Improving group-invariance tests through subgroups. *Unpublished manuscript.*

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses.* Springer Science & Business Media.

Ludbrook, J. and Dudley, H. (1998). Why permutation tests are superior to t and f tests in biomedical research. *The American Statistician*, 52(2):127–132.

Miller, J. B. and Sanjurjo, A. (2018a). A cold shower for the hot hand fallacy: Robust evidence that belief in the hot hand is justified. *University of Alicante mimeo.*

Miller, J. B. and Sanjurjo, A. (2018b). Surprised by the hot hand fallacy? a truth in the law of small numbers. *Econometrica*, 86(6):2019–2047.

Miller, J. B. and Sanjurjo, A. (2021). Is it a fallacy to believe in the hot hand in the nba three-point contest? *European Economic Review*, 138:103771.

Ritzwoller, D. M. and Romano, J. P. (2021). Uncertainty in the Hot Hand Fallacy: Detecting Streaky Alternatives to Random Bernoulli Sequences. *The Review of Economic Studies.*

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, 76(2):105.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–598.

Acknowledgements

I would have not been able to write this thesis without the help of my supervisor Nick Koning. Many thanks for all your creative and constructive input throughout this project. Next, I would like to thank the Dreamteam, Esmee Breeman, Joris Bentvelsen and Pepijn Thijssen. The pre-master and master would not have been the same without you. Lastly, I would like to thank my family and friends for their endless support throughout this journey.

A Appendix

	$\mathcal{G}^{high}$	$\mathcal{G}^{low}$
1	[1, 2, 3, 4, 5, 6]	[1, 2, 3, 4, 5, 6]
2	[2, 5, 3, 4, 1, 6]	[6, 5, 4, 2, 3, 1]
3	[3, 5, 2, 6, 1, 4]	[1, 3, 2, 5, 4, 6]
4	[5, 1, 3, 4, 2, 6]	[6, 4, 5, 3, 2, 1]
5	[1, 3, 2, 6, 5, 4]	[1, 4, 3, 5, 2, 6]
6	[2, 3, 5, 6, 1, 4]	[6, 2, 4, 3, 5, 1]
7	[3, 2, 5, 4, 1, 6]	[1, 5, 3, 2, 4, 6]
8	[5, 3, 1, 6, 2, 4]	[6, 3, 4, 5, 2, 1]
9	[1, 5, 2, 4, 3, 6]	[1, 2, 4, 5, 3, 6]
10	[2, 1, 5, 4, 3, 6]	[6, 5, 2, 3, 4, 1]
11	[3, 1, 5, 6, 2, 4]	[1, 3, 5, 4, 2, 6]
12	[5, 2, 1, 4, 3, 6]	[6, 4, 3, 2, 5, 1]
13	[1, 2, 5, 6, 3, 4]	[1, 4, 5, 2, 3, 6]
14	[2, 5, 1, 6, 3, 4]	[6, 2, 3, 5, 4, 1]
15	[3, 5, 1, 4, 2, 6]	[1, 5, 2, 4, 3, 6]
16	[5, 1, 2, 6, 3, 4]	[6, 3, 5, 2, 4, 1]
17	[1, 3, 5, 4, 2, 6]	[1, 2, 5, 3, 4, 6]
18	[2, 3, 1, 4, 5, 6]	[6, 5, 3, 4, 2, 1]
19	[3, 2, 1, 6, 5, 4]	[1, 3, 4, 2, 5, 6]
20	[5, 3, 2, 4, 1, 6]	[6, 4, 2, 5, 3, 1]
21	[1, 5, 3, 6, 2, 4]	[1, 4, 2, 3, 5, 6]
22	[2, 1, 3, 6, 5, 4]	[6, 2, 5, 4, 3, 1]
23	[3, 1, 2, 4, 5, 6]	[1, 5, 4, 3, 2, 6]
24	[5, 2, 3, 6, 1, 4]	[6, 3, 2, 4, 5, 1]

Table A.1: Array form representation of all permutations within $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$

			$\mathbb{P}(\text{reject } H_0 \,\vert\, \mathbf{x},\ \mathcal{G})$			
	$\mathbf{x}$	$\mathbb{P}(\mathbf{x} \,\vert\, H_a)$	24-rnd	$\mathcal{S}_6$	$\mathcal{G}^{high}$	$\mathcal{G}^{low}$
25	[0, 1, 1, 0, 0, 0]	0.010	0.084	0	0	0.25
26	[1, 0, 0, 1, 1, 1]	0.010	0.085	0	0	0.25
27	[0, 0, 0, 1, 0, 1]	0.010	0	0	0.125	0
28	[1, 1, 1, 0, 1, 0]	0.010	0	0	0.125	0
29	[1, 0, 1, 1, 1, 0]	0.010	0	0	0	0
30	[0, 1, 0, 0, 0, 1]	0.010	0	0	0	0
31	[0, 0, 1, 0, 0, 0]	0.010	0	0	0	0.125
32	[1, 1, 0, 1, 1, 1]	0.010	0	0	0	0.125
33	[1, 0, 0, 0, 1, 1]	0.010	0.193	0.125	0.375	0.25
34	[0, 1, 1, 1, 0, 0]	0.010	0.192	0.125	0.375	0.25
35	[1, 0, 0, 0, 1, 0]	0.010	0	0	0	0
36	[0, 1, 1, 1, 0, 1]	0.010	0	0	0	0
37	[0, 0, 0, 1, 0, 0]	0.010	0	0	0	0.125
38	[1, 1, 1, 0, 1, 1]	0.010	0	0	0	0.125
39	[0, 0, 1, 0, 1, 0]	0.006	0	0	0	0
40	[1, 1, 0, 1, 0, 1]	0.006	0	0	0	0
41	[1, 1, 0, 1, 0, 0]	0.006	0.002	0	0	0
42	[0, 0, 1, 0, 1, 1]	0.006	0.002	0	0	0
43	[1, 0, 1, 1, 0, 1]	0.006	0	0	0	0
44	[0, 1, 0, 0, 1, 0]	0.006	0	0	0	0
45	[1, 0, 1, 1, 0, 0]	0.006	0.002	0	0	0
46	[0, 1, 0, 0, 1, 1]	0.006	0.001	0	0	0
47	[0, 1, 0, 1, 1, 0]	0.006	0	0	0	0
48	[1, 0, 1, 0, 0, 1]	0.006	0	0	0	0
49	[1, 0, 0, 1, 0, 1]	0.006	0	0	0	0
50	[0, 1, 1, 0, 1, 0]	0.006	0	0	0	0
51	[0, 1, 1, 0, 1, 1]	0.006	0	0	0	0
52	[1, 0, 0, 1, 0, 0]	0.006	0	0	0	0
53	[1, 1, 0, 0, 0, 1]	0.004	0.192	0.125	0.375	0.25
54	[0, 0, 1, 1, 1, 0]	0.004	0.191	0.125	0.375	0.25
55	[1, 1, 1, 0, 0, 1]	0.004	0.085	0	0	0.25
56	[0, 0, 0, 1, 1, 0]	0.004	0.084	0	0	0.25
57	[1, 1, 0, 0, 1, 1]	0.003	0.085	0	0	0.25
58	[0, 0, 1, 1, 0, 0]	0.003	0.085	0	0	0.25
59	[1, 1, 0, 0, 1, 0]	0.003	0.002	0	0	0
60	[0, 0, 1, 1, 0, 1]	0.003	0.002	0	0	0
61	[1, 0, 0, 1, 1, 0]	0.003	0.002	0	0	0
62	[0, 1, 1, 0, 0, 1]	0.003	0.002	0	0	0
63	[0, 0, 1, 0, 0, 1]	0.003	0	0	0	0
64	[1, 1, 0, 1, 1, 0]	0.003	0	0	0	0

Table A.2: The 40 least likely sequences under H_a with $m = 2$ and $\eta = 0.3$. For each sequence the likelihood and the rejection probability using 24-rnd, $\mathcal{S}_6$, $\mathcal{G}^{high}$ and $\mathcal{G}^{low}$ are given.