# Fat Chance

Charles M. Grinstead
Swarthmore College

J. Laurie Snell
Dartmouth College

# Chapter 1

# Fingerprints

## 1.1 Introduction

On January 7, 2002, in the case U.S. v. Llera-Plaza, Louis H. Pollack, a federal judge in the United States District Court in Philadelphia, barred any expert testimony on fingerprinting that asserted that a particular print gathered at the scene of a crime is or is not the print of a particular person. As might be imagined, this decision was met with much interest, since it seemed to call into question whether fingerprinting can be used to help prove the guilt or innocence of an accused person.

In this chapter, we will consider the ways in which fingerprints have been used by society and show how the current quandary was reached. We will also consider what probability and statistics have to say about certain questions concerning fingerprints.

## 1.2 History of Fingerprinting

It seems that the first use of fingerprints in human society was to give evidence of authenticity to certain documents in seventh-century China, although it is possible that they were used even earlier than this. Fingerprints were used in a similar way in Japan, Tibet, and India. In Simon Cole's excellent book on the history of fingerprinting, the Persian historian Rashid-eddin is quoted as having declared in 1303 that "Experience shows that no two individuals have fingers exactly alike." [1] This statement is one with which the reader is no doubt familiar. A little thought will show that unless all the fingerprints in the world are observed, it is impossible to verify this statement. Thus, one might turn to a probability model to help understand how likely it is that this statement is true. We will consider such models below.

In the Western World, fingerprints were not discussed in any written work until 1685, when an illustration of the papillary ridges of a thumb was placed in an anatomy book written by the Dutch scientist Govard Bidloo. A century later, the

[1] Cole, Simon A., "Suspect Identities: A History of Fingerprinting and Criminal Identification," Harvard University Press, Cambridge, Massachusetts, 2001, pgs. 60-61.

statement that fingerprints are unique appeared in a book by the German anatomist J. C. A. Mayer.

In 1857, a group of Indian conscripts rebelled against the British. After this rebellion had been put down, the British government decided that it needed to be more strict in its enforcements of laws in its colonies. William Herschel, the grandson of the discoverer of the planet Uranus, was the chief administrator of a district in Bengal. Herschel noted that the unrest in his district had given rise to a great amount of perjury and fraud. For example, it was believed that many people were impersonating deceased officers to collect their pensions. Such impersonation was hard to prove, since there was no method that could be used to decide whether a person was who he or she claimed to be.

In 1858, Herschel asked a road contractor for a handprint, to deter the contractor from trying to contest, at a later date, the authenticity of a certain contract. A few years subsequent to this, Herschel began using fingerprints. It is interesting to note that as with the Chinese, the first use of fingerprints was in civil, not criminal, identification.

At about the same time, the British were increasingly concerned about crime in India. One of the main problems was to determine whether a person arrested and tried for a crime was a habitual offender. Of course, to determine this required that some method be used to identify people who had been convicted of crimes. Presumably, a list would be created by the authorities, and if a person was arrested, this list would be consulted to determine whether the person in question was on the list or not. In order for such a method to be useful, it would have to possess two properties. First, there would have to be a way to store, in written form, enough information about a person so as to uniquely identify that person. Second, the list containing this information would have to be in a form that would allow quick and accurate searches.

Although, in hindsight, it might seem obvious that one should use fingerprints to help with the formation of such a list, this method was not the first to be used. Instead, a system based on anthropometry was developed. Anthropometry is the study and measurement of the size and proportions of the human body. It was naturally thought that once adulthood is reached, the lengths of bones do not change. In the 1880's Alphonse Bertillon, a French police official, developed a system in which eleven different measurements were taken and recorded. In addition to these measurements, a detailed physical description, including information on such things as eyes, ears, hair color, general demeanor, and many other attributes, was recorded. Finally, descriptions of any 'peculiar marks' were recorded. This system was called Bertillonage, and was widely used in Europe, India, and the United States, as well as other locations, for several decades.

One of the main problems encountered in the use of Bertillonage was inconsistency in measurement. Many measurements of each person were taken, and the 'operators,' as the measurers were called, were trained. Nonetheless, if a criminal suspect was measured in custody, and the suspect's measurements were already in the list, the two sets of measurements might vary enough so that no match would be made.

Another problem was the amount of time required to search the list of known offenders, in order to determine whether a person in custody had been arrested before. In some places in India, the lists grew to contain many thousands of records. Although these records were certainly stored in a logical way, the variations in measurements made it necessary to look at many records that were 'near' the place that the searched-for record should be.

The chief problem at that time with the use of fingerprints for identification was that no good classification system had been developed. In this regard, fingerprints were not thought to be as useful as Bertillonage, since the latter method did involve numerical records that could be sorted. In the 1880's, Henry Faulds, a British physician who was serving in a Tokyo hospital at the time, devised a method for classifying fingerprints. This method consisted of identifying each major type of print with a certain written syllable, followed by other syllables representing different features in the print. Once a set of syllables for a given print was determined, the set was added to a alphabetical list of stored sets of syllables representing other prints.

Faulds wrote to Charles Darwin about his ideas, and Darwin forwarded them to his cousin, Francis Galton. Galton was one of the giants among British scientists in the late 19th century. His interests included meteorology, statistics, psychology, genetics, and geography. Early in his adulthood, he spent two years exploring southwest Africa. He was also a promoter of eugenics; in fact, this word is due to Galton.

Galton became interested in fingerprints for several reasons. He was interested in the heritability of certain traits, and one such trait that could easily be tested were fingerprint patterns. He was concerned with ethnology, and sought to compare the various races. One question that he considered in this vein was whether the proportions of the various types of fingerprints differed among the races. He also tried to determine whether any other traits were related to fingerprints. Finally, he understood the value that such a system would have in helping the police and the courts identify recidivists.

To carry out such research, it was necessary for him to have access to many fingerprints. By the early 1890's, he had amassed a collection of thousands of prints. This collection contained prints from people belonging to many different ethnic groups. He also collected fingerprints from certain types of people, such as criminals. He was able to show that fingerprints are partially controlled by heredity. For example, it was found that a peculiarity in a pattern in a fingerprint of a parent might pass to the same finger of a child, or, with less probability, to another finger of that child. Nonetheless, it must be stated that his work in this area did not lead to any discoveries of great import.

One of Galton's most fundamental contributions to the study of fingerprints consisted of his publishing of material, much of which was due to William Herschel, that fully established the fact that fingerprint patterns persist over the lifetime of an individual. Of at least equal importance was his development of a method to classify fingerprints. His method had the important attribute that it could be quickly searched to determine if it contained a given fingerprint.

Very shortly thereafter, a committee consisting of various high officials in British law enforcement was formed to compare Bertillonage and the Galton fingerprint method, with the goal being to decide which method to adopt (although Bertillonage was in use in continental Europe, India, and elsewhere, it had not yet been used in Britain). The committee also considered whether it might be still better to use both methods at once.

In their deliberations, the committee noted that the taking of fingerprints is a much easier process than the one that is used by Bertillonage operators. In addition, a fingerprint, if it is properly taken (i.e. if the resulting impression is legible), is a true and accurate rendition of the patterns on the finger. Both of these statements lead to the conclusion that this method is more accurate than Bertillonage.

Given these remarks, it might seem strange that the committee did not recommend that fingerprints be the method of choice. However, there was still some concern about the accuracy of the indexing used in the method. It was recommended that identification be made by fingerprints, but indexing be carried out by Bertillonage. The committee did foresee that the problems with fingerprint indexing could be overcome, and that in this case, the fingerprint method might be the sole system in use.

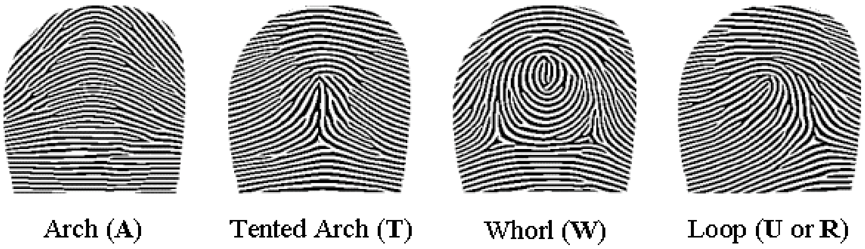
Galton continued to work on his method of classification, and in 1895, he published a system that greatly improved his previous attempts. Edward Henry, a magistrate of a district in India, worked on and modified Galton's indexing method between 1898 and 1900. This modification was adobted by Scotland Yard. Regarding credit for the method, a letter from Sir George Darwin to the London Times had this to say: "Sir Edward Henry undoubtedly deserves great credit in recognising the merits of the system and in organising its use in a practical manner in India, the Cape and England, but it would seem that the yet greater credit is due to Mr. Francis Galton."[2]

In 1902, Galton published a letter in the journal Nature, entitled "Finger-Print Evidence," in which he discusses a new aspect (for him, at any rate) of fingerprints. Scotland Yard had sent him two enlarged photographs of thumbprints. The first came from the scene of a burglary, and the second came from the fingerprint files at Scotland Yard. Galton discusses how the use of his system allows the prosecution to explain the similarities in the two prints. The question of accuracy in matching prints obtained from a crime scene with those in a database is one that is still being considered today. Before turning to this question, we will describe Galton's method.

Galton begins by noting that in the center of most fingerprints there is a 'core,' which consists of patterns that he calls loops and whorls (see Figure 1.1[3].) If no such core exists, the pattern is said to be an arch. Next, he defines a delta as the region where the parallel ridges begin to diverge to form the core. Loops have one delta, and whorls have two. These deltas serve as axes of reference for the rest of the classification. By tracing the ridges as they leave the delta(s) and cross the core, one can partition fingerprints into ten classes. Since each finger would be in

[2] George Darwin, quoted in Karl Pearson, "Life and Letters of Francis Galton,"

[3] Keogh, E.,An Overview of the Science of Fingerprints. Anil Aggrawal's Internet Journal of Forensic Medicine and Toxicology, 2001; Vol. 2, No. 1 (January-June 2001)



Arch (A) Tented Arch (T) Whorl (W) Loop (U or R)

Figure 1.1: Four examples of fingerprints.

one of the ten classes, there are $10^{10}$ possible sets of ten classes. Even though the ten classes do not occur with equal frequency among all fingerprints, this first level of classification already serves to distinguish between most pairs of people.

Of the ten classes, only two correspond to loops, as opposed to arches and whorls. However, about half of all fingerprints are loops, which suggests that the scheme is not yet precise enough. Galton was aware of this, and added two other types of information to the process. The first involved using the axes of reference arising from the deltas to count ridges in certain directions. The second involved the counting and classification of what he termed 'minutiae.' This term refers to places in the print where a ridge bifurcates or ends. The idea of minutiae is still in use today, although they are now sometimes referred to as 'Galton points' or 'points.'

There are many different types of points, and the places that they occur in a given fingerprint seems to be somewhat random. In addition, a typical fingerprint has many such points. These observations imply that if one can accurately write down where the points occur and which types of points occur, then one has a very powerful way to distinguish two fingerprints. The method is even more powerful when comparing sets of ten fingerprints from two people.

## 1.3 Models of Fingerprints

We shall investigate some probabilistic models for fingerprints that incorporate the idea of points. The two most basic questions that one might use such models to help answer are as follows. First, in a given model, what is the probability that no two fingerprints, among all people who are now alive, are exactly alike? Second, suppose that we have a partial fingerprint, such as one that might have been recovered from a crime scene (such partial prints are called latent prints). What is the probability that this latent print exactly matches more than one fingerprint, among all fingerprints in the world? The reason that we are interested in whether the latent print matches more than one fingerprint is that it clearly matches one print, namely the one belonging to the person who left the latent print. It is typically the case that the latent print, if it is to be of any use, will identify a suspect, i.e.

someone who has a fingerprint that matches the latent print. It is obviously of great interest in a court of law as to how likely it is that someone other than the suspect has a fingerprint that matches the latent print. We will see that this second question is of central importance in the discussions going on today about the accuracy of fingerprinting as a crimefighting tool.

Galton seems to have been the first person to consider a probabilistic model that might shed some light on the answer to the first question. He began by imagining a fingerprint as a random set of ridges, with roughly 24 ridge intervals across the finger and 36 ridge intervals along the finger. Next, he imagined covering up an $n$ by $n$ ridge interval square on a fingerprint, and attempting to recreate the ridge pattern in the area that was covered. Galton maintained that if $n$ were small, say at most 4, then most of the time, the pattern could be recreated by using the information in the rest of the fingerprint. However, if $n$ were 6, he found that he was wrong more often than right when he carried out this experiment.

He then let $n = 5$, and claimed that he would be right about one-half of the time in reconstructing the fingerprint. This led him to consider the fingerprint as consisting of a set of non-overlapping $n$ x $n$ squares, which he considered to be independent random variables. In Pearson's account, Galton used $n = 6$, although his argument is more understandable had he used $n = 5$. Galton claimed that any of the reconstructions, both the correct and incorrect ones, might have occurred in nature, so each random variable has two possible values,*given* the way that the ridges leave and enter the square, and given how many ridges leave and enter. Pearson says that Galton 'proceeds to gived a rough approximation to two other chances, which he considers to be involved: the first concerns guessing correctly the general course of the ridges adjacent to each square, and the second of guessing rightly the number of ridges that enter and issue from the square. He takes these in round numbers to be $1/2^4$ and $1/2^8$... .'[4] Finally, Galton multiplies all of these probabilities together, under the assumption of independence, and arrives at the number 64 billion which, at the time, was 4 times the number of fingerprints in the world. (Galton claims that the odds are roughly 39 to 1 against any particular fingerprint occurring anywhere in the world. It seems to us that the odds should be 3 to 1 against.)

We will soon see other models of fingerprints that arrive at much different answers. However, it should be remembered that we are trying to estimate the probability that no two fingerprints, among all people who are now alive, are exactly alike. Suppose, as Galton did, that there are 16 billion fingerprints among the people of the world, and there are 64 billion possible fingerprints. Does the reader think that these assumptions make it very likely or very unlikely that there are two fingerprints that are the same? To answer this question, we can proceed as follows. Consider an urn with 64 billion labeled balls in it. We choose, one at a time, 16 billion balls from the urn, replacing the balls after each choice. We are asking for the probability that we never choose the same ball more than once. This is the celebrated birthday problem, on a world where there are 64 billion days in a year,

[4] Pearson, ibid., pg. 182.

and 16 billion people. The birthday problem asks what is the probability that at least two people share a birthday. The answer is

$$\left(1-\frac{0}{n}\right)\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) ,$$

where $n = 64$ billion and $k = 16$ billion. This can be seen by considering the people one at a time. If 6 people, say, have already been considered, and if they all have different birthdays, then the probability that the seventh person has a birthday that is different than all of the first 6 people equals

$$\left(1-\frac{6}{n}\right) .$$

It is relatively straightforward to estimate the above product in terms of $k$ and $n$. For the values given by Galton, the product is less than

$$\frac{1}{10^{10^9}} .$$

This means that in Galton's model, with his estimates, it is extremely likely that there are two fingerprints that are the same.

In fact, to our knowledge, no two fingerprints from different people have ever been found that are identical. Of course, it is not the case that all fingerprints on Earth have been recorded or compared, but the FBI has a database with more than 10 million fingerprints in it, and we presume that no two fingerprints in it are exactly the same. (It must be said that it is not clear to us that all pairs of fingerprints in this database have actually been compared. In addition, one wonders whether the FBI, if it found a pair of identical fingerprints, would announce this to the world.) In any case, if we use Galton's estimate for the number of possible fingerprints, and let $k = 10$ million, the probability that no two are alike is still very small; it is less than

$$\frac{1}{10^{339}} .$$

We can turn the above question around and ask the following question. Suppose that there are 60 billion fingerprints in the world, and suppose that we imagine they are chosen from a set of $n$ possible fingerprints. How large would $n$ have to be in order that the probability that all of the chosen fingerprints are different exceeds .999? An approximate answer to this question is that it would suffice for $n$ to be at least $10^{25}$. Although this is quite a bit larger than Galton's estimate, there have been other, more sophisticated models of fingerprints, some of which we will now describe, have come up with estimates for $n$ that are much larger than $10^{25}$. Thus, if these models are at all accurate, it is extremely unlikely that there exist two fingerprints in the world that are exactly alike.

In 1933, T. Roxburgh described a model for fingerprint classification that is much more intricate than Galton's model. This model, and many others, are described and compared in an article in the Journal of Forensic Sciences, written by D. A.

Stoney and J. I. Thornton.[5] In Roxburgh's model, a vertical ray is drawn upwards from the center of the fingerprint (this idea must be accurately defined, but for our purposes, we can take it to mean the center of the loop or whorl, or the top of the arch). This ray is defined to be 0 degrees. Another ray, with endpoint at the center, is revolved clockwise from the first ray. As this ray passes over minutiae, the types of the minutiae are recorded, along with the ridge numbers on which the minutiae lie. If a fingerprint has $R$ concentric ridges, $n$ minutiae, and there are $T$ minutia types, then the number of possible patterns equals

$$(RT)^n \ ,$$

since as the second ray revolves clockwise, the next minutia encountered could be on any of the $R$ ridges and be of any of the $T$ minutia types. Roxburgh also introduces a factor of $P$ that corresponds to the number of different overall patterns and core types that might be encountered. Thus, he estimates the number of possible fingerprints to be

$$P(RT)^n \ .$$

He takes $P = 1000$, $R = 10$, $T = 4$, and $n = 35$; this last value is Galton's estimate for the typical number of minutia in a fingerprint. If we calculate the above expression with these values, we obtain the number

$$1.18 \times 10^{59} \ .$$

Roxburgh modified the above expression for the number of possible fingerprints to attempt to account for ambiguities between various types of minutiae. For example, it is possible that a fork in a ridge might be seen as a ridge ending, depending upon whether the ridges in question meet each other or not. Roxburgh suggested using a number $Q$ which would vary depending upon the quality of the fingerprint under examination. The value of $Q$ ranges from 1.5 to 3, with the smaller value corresponding to a higher quality fingerprint. For each minutia, Roxburgh replaced the factor $RT$ by the factor $RT/Q$. This leads to the expression

$$P((RT)/Q)^n$$

as an estimate for the number of discernable types of fingerprints, assuming their quality corresponds to a particular value of $Q$. Note that even if $Q = 3$, so that $RT/Q = 1.33R$, the number of discernable types of fingerprints in this model is

$$2.16 \times 10^{42} \ .$$

Stoney and Thornton note that although this is a very interesting, sophisticated model, it has been "totally ignored by the forensic science community."[6]

[5] Stoney, D. A. and J. I. Thornton, "A Critical Analysis of Quantitative Fingerprint Individuality Models," Journal of Forensic Sciences, v. 31, no. 4 (1986), pgs. 1187-1216.

[6] ibid., pg. 1192

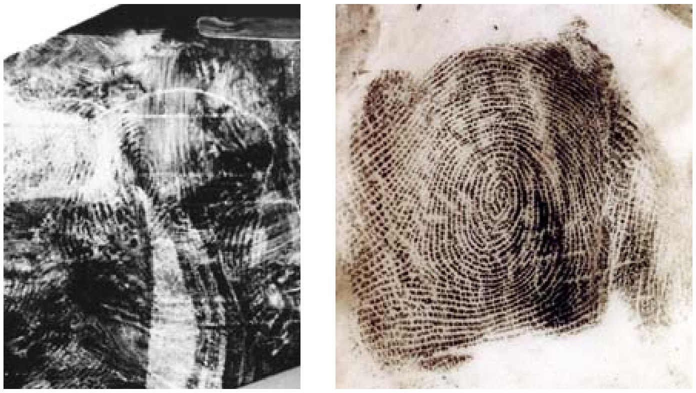


Figure 1.2: Examples of latent and rolled prints.

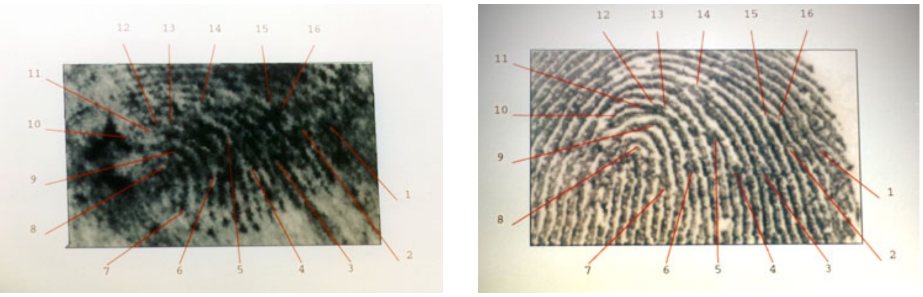


Figure 1.3: Minutiae matches.

## 1.4 Latent Fingerprints

According to a government expert who testified at a recent trial, the average size of a latent fingerprint fragment is about one-fifth the size of a full fingerprint. Since a typical fingerprint contains between 75 and 175 minutiae[7], this means that a typical latent print has between 15 and 35 minutiae. In addition, the latent print recovered from a crime scene is frequently of poor quality, which tends to increase the likelihood of mistaking the types of minutiae being observed.

In a criminal case, the latent print is compared with a high quality print taken from the hand of the accused or from a database of fingerprints. Figure 1.2 shows a latent print and the corresponding rolled print to which the latent print was matched. Figure 1.3 shows another pair of prints, one latent and one rolled, from the same case. The figure also shows the claimed matching minutiae in the two prints.

The person making the comparison states that there is a match if he or she believes that there are a sufficient number of common minutiae, both in type and

[7] 'An Analysis of Standards in Fingerprint Identification 1,' Federal Bureau of Investigation, Department of Justice, Law Enforcement Bulletin, vol. 1 (June 1972).

location, in the two prints. There have been many criminal cases in which an identification was made with fewer than fifteen matching minutiae[8]. There is no general agreement among various law enforcement agencies or among various countries, on the number of matching minutiae that must exist in order for a match to be declared. In fact, according to Robert Epstein[9], “many examiners ... including those at the FBI, currently believe that there should be no minimum standard whatsoever and that the determination of whether there is a sufficient basis for an identification should be left to the subjective judgment of the individual examiner.” It is quite understandable that a law enforcement agency might object to constraints on its ability to claim matches between fingerprints, as this could only serve to decrease the number of matches obtained.

In some countries, fingerprint matches can be declared with as few as eight minutiae matches (such minutiae matches are sometimes called ‘points.’) However, there are examples of fingerprints from different people that have seven matching minutiae. In a California bank robbery trial, *U. S. v. Parks*, in 1991, the prosecution introduced evidence that showed that the suspect’s fingerprint and the latent print had ten points. The trial judge, Spencer Letts, asked the prosecution expert what the minimum standard was for points in order to declare a match. The expert announced that the minimum was eight. Judge Letts had seen fingerprint evidence entered in other trials. He said “If you only have ten points, you’re comfortable with eight; if you have twelve, you’re comfortable with ten; if you have fifty, you’re comfortable with twenty.”[10] Later in the same trial, the following exchange occurred between Judge Letts and another prosecution fingerprint expert:

“The Witness: ‘The thing you have there is that each department has their own goals or their own rules as far as the number of points being a make [an identification]. ...that number really just varies from department to department.’

The Court: ‘I don’t think I’m ever going to use fingerprint testimony again; that simply won’t do...’

The Witness: ‘That just may be one of the problems of the field, but I think if there was [a] survey taken, you would probably get a different number from every department that has a fingerprint section as to their lowest number of points for a comparison and make.’

The Court: ‘That’s the most incredible thing I’ve ever heard of.’ ”[11]

According to Simon Cole, no scientific study has been carried out to estimate the probability of two different prints sharing a given number of minutiae. David Stoney and John Thornton claim that none of the fingerprint models proposed during the past century “even approaches theoretical accuracy ..., and none has been subjected to empirical validations.”[12] In fact, latent print examiners are prohibited by their

[8] see footnote 25 in Epstein, Robert, ‘Fingerprints Meet Daubert: The Myth of Fingerprint “Science” is Revealed,’ Southern California Law Review, vol. 75 (2002), pgs. 605-658.

[9] ibid., pg. 610

[10] Cole, op. cit., pg. 272.

[11] ibid., pgs 272-273.

[12] Stoney and Thornton, op. cit., pg. 1187.

primary professional association, the International Association for Identification ("IAI"), from offering opinions of identification using probabilistic terminology. A resolution, passed by the IAI at one of its meetings, states that "any member, officer, or certified latent print examiner who provides oral or written reports, or gives testimony of possible, probable, or likely friction ridge identification shall be deemed to be engaged in [unbecoming] conduct... and charges may be brought."[13]

In 1993, the Supreme Court rendered a decision in the case Daubert v. Merrell Dow Pharmaceuticals, Inc.[14] The Court described certain factors that courts needed to consider when deciding whether to admit expert testimony. In this decision, the Court concentrated on scientific expert testimony; it considered the issue of expert testimony of a non-scientific nature in the case Kumho Tire Co. v. Carmichael, a few years later.[15] In the first decision, the Court interpreted the Federal Rule of Evidence 702, which defines the term 'expert witness' and states when such witness are allowed, as requiring trial judges to determine whether the opinion of an expert witness lacks sufficient reliability, and if so, to exclude this testimony. The Daubert decision listed five factors that could be considered when determining whether scientific expert testimony should be retained or excluded. These factors are as follows:

1. "A preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue."[16]

2. "The court ordinarily should consider the known or potential rate of error... ."[17]

3. The court should consider "the existence and maintenance of standards controlling the technique's operation... ."[18]

4. "'General acceptance' can ... have a bearing on the inquiry. A "reliability assessment does not require, although it does permit, explicit identification of a relevant scientific community and an express determination of a particular degree of acceptance within that community."'[19]

5. "A pertinent consideration is whether the theory or technique has been subjected to peer review and publication... ."[20]

In the Kumho case, the Court held that a trial court's obligation to decide whether to admit expert testimony applies to all experts, not just scientific experts. The Court also held that the factors listed above may be used by a court in assessing nonscientific expert testimony.

In the case (U.S. v. Llera-Plaza) mentioned at the beginning of the chapter, the presiding judge, Louis Pollack, applied the Daubert criteria to the fingerprint identification process, as he was instructed to do by the Kumho case. In particular,

[13] Epstein, op. cit., pg. 611, footnote 32.
[14] 509 U.S. 579 (1993)
[15] 526 U.S. 137 (1999).
[16] 509 U.S. 579 (1993), note 593.
[17] ibid., note 594
[18] ibid.
[19] ibid., quoted from United States v. Downing, 753 F.2d 1224, 1238 (3d Cir. 1985).
[20] ibid., note 593.

he discussed the problem with the current process employed by the FBI (and other law enforcement agencies), which is called the ACE-V system. This name is an acronym, and stands for analysis, comparison, evaluation, and verification. Judge Pollack ruled that the third part of this process, in which a fingerprint expert states his or her opinion that the latent print and the comparison print (either a rolled print from a suspect or a print from a database) either match or do not match, did not measure up to several of the Daubert criteria.

With regard to the first of the criteria, the government (the plaintiff in the case) argued that the method of fingerprint matching had been tested empirically over a period of 100 years. It also argued that in any particular case, the method can be tested through the testimony of a fingerprint expert other than the one whose testimony is being heard. The judge rejected this argument, saying that neither of these actions could be considered as scientific tests of the method. He further noted that in the second case, the strength of the second examiner's 'test' of a claimed match is diluted by the fact that in many cases, the second examiner has been advised of the first examiner's claims in advance.

On the point of testing, it is interesting to note that in 2000, the National Institute of Justice (NIJ), which is an arm of the Department of Justice, had solicited proposals for research projects to study the reliability of fingerprinting. This solicitation was mentioned by the judge in his ruling, and was also taken as evidence by the defense that the government did not know whether fingerprinting was reliable.

The second Daubert criterion concerns the 'known or potential rate of error' of the method. In their arguments before the court, the government contended that there were two types of error - methodology error and practitioner error. One of the government's witnesses, when asked to explain methodology error, stated that 'an error rate is a wispy thing like smoke, it changes over time...'[21] The judge said that 'the full import of [this] testimony is not easy to grasp.' He summarizes this testimony as saying that if a method, together with its limitations, has been defined, then there is no methodology error. All of the error is practitioner error. The other government witness, Stephen Meagher, a supervisory fingerprint specialist with the FBI, also testified that if the scientific method is followed, then the methodology error rate will be zero, i.e. all of the error is practitioner error. We will have more to say about practitioner error below.

Judge Pollack also found problems concerning the third Daubert criterion, which deals with standards controlling a technique's operation. There are three types of standards discussed in the judge's ruling. The first is whether there is a minimum number of Galton points that must be matched before an overall match is declared. In the ACE-V process, no minimum number is prescribed, and in fact, in some jurisdictions, there is no minimum. The second type of standard concerns the evaluation of whether a match exists. The government and defense witnesses agreed that this decision is subjective. The judge concluded that 'it is difficult to see how fingerprint identification–the matching of a latent print to a known fingerprint–is controlled by any clearly describable set of standards to which most examiners

[21] U.S. v. Llera-Plaza, January 7, 2002, at 47.

subsribe.'[22] Finally, there is the issue of the qualifications of examiners. There are no mandatory qualification standards that must be attained in order for someone to become a fingerprint examiner, nor are there any uniform certification processes.

Regarding the fourth Daubert criterion, the judge had this to say:

> 'General acceptance by the fingerprint examiner community does not ... meet the standard... . First, there is the difficulty that fingerprint examiners, while respected professionals, do not constitute a 'scientific community' in the Daubert sense... . Second, the Court cautioned in Kumho Tire that general acceptance does not 'help show that an expert's testimony is reliable where the discipline itself lacks reliability. The failure of fingerprint identifications fully to satisfy the first three Daubert factors militates against heavy reliance on the general acceptance factor. Thus, while fingerprint examinations conducted under the general ACE-V rubric are generally accepted as reliable by fingerprint examiners, this by itself cannot sustain the government's burden in making the case for the admissibility of fingerprint testimony under Federal Rule of Evidence 702.[23]

The conclusion of the judge's ruling was as follows:

> For the foregoing reasons:
>
> A. This court will take judicial notice of the uniqueness and permanence of fingerprints.
>
> B. The parties will be able to present expert fingerprint testimony (1) describing how any latent and rolled prints at issue in this case were obtained, (2) identifying, and placing before the jury, such fingerprints and any necessary magnifications, and (3) pointing out any observed similarities and differences between a particular latent print and a particular rolled print alleged by the government to be attributable to the same persons. But the parties will not be permitted to present testimony expressing an opinion of an expert witness that a particular latent print matches, or does not match, the rolled print of a particular person and hence is, or is not, the fingerprint of that person.'[24]

The government asked for a reconsideration of this ruling. Not surprisingly, it felt that its effectiveness in both the trial at hand and in future trials would be seriously compromised if witnesses were not allowed to express an opinion on whether or not a latent print matches a rolled print. The government asked to be allowed to submit evidence that shows the accuracy of FBI fingerprint examiners. The defendants argued that the judge should decline to reconsider his ruling, and Judge Pollack stated that their argument was solid; 'neither of the circumstances conventionally justifying rconsideration – new, or hitherto unavailable facts or new controlling law – was present here.'[25] Nonetheless, the judge decided to grant a reconsideration

[22] ibid. at 58.

[23] ibid. at 61.

[24] ibid. at 69.

[25] U. S. v. Llera Plaza, March 13,2002, at 11.

hearing, arguing that the record on which he made his previous ruling was testimony presented two years earlier in another courtroom. 'It seemed prudent to hear such live witnesses as the government wished to present, together with any rebuttal witnesses the defense would elect to present.'[26]

At this point in our narrative, it makes sense to consider the various attempts to measure error rates in the field of fingerprint analysis. Lyn and Ralph Haber, who are consultants at a private company in California, and are also adjuncts at the University of California at Santa Cruz, have obtained and analyzed relevant data from many sources.[27] These data include both results on crime laboratories and individual practitioners. We will summarize some of their findings here.

The American Society of Crime Laboratory Directors (ASCLD) is an organization that provides leadership in the management of forensic science. It is in their interest to evaluate and improve the quality of operations of crime laboratories. In 1977, the ASCLD began developing an accreditation program for crime laboratories. By 1999, 182 labs had been accredited. One requirement for a lab to be accredited is that the examiners working in the lab must pass an externally administered proficiency test. We note that since it is the lab, and not the individual examiners, that is being tested, these proficiency tests are taken by all of the examiners as a group in a given lab.

Beginning in 1983, the ASCLD began administering such a test in the area of fingerprint identification. The test, which is given each year to all labs requesting accreditation, consists of pictures of 12 or more latent prints and a set of ten-print (rolled print) cards. The set of latent prints contains a range of quality, and is supposed to be representative of what is actually seen in practice. For each latent print, the lab was asked to decide whether it is 'scorable,' i.e. whether it is of sufficient quality to attempt to match it with a rolled print. If it is judged to be scorable, then the lab is asked to decide whether or not it matches one of the prints on the ten-print cards. There are 'correct' answers for each latent print on the test, i.e. the ASCLD has decided, in each case, whether or not a latent print is scorable, and if so, whether or not it matches any of the rolled prints.

The Habers report on results from 1983 to 1991. During this time, the number of labs that took the exam increased from 24 to 88; many labs took the tests more than once (a new test was constructed each year). Assuming that in many cases, the labs have more than one fingerprint expert, this means that hundreds of these experts took this test at least once during this period.

Each lab returned one answer for each question. There are four types of error that can be made on each question of each test. A scorable print can be ruled unscorable, or vice versa. If a print is scorable, it can be erroneously matched to a rolled print, or it can fail to be matched at all, even though a match exists. Of these four types of errors, the second and third are more serious than the others, assuming that we take the point of view that erroneous evidence against an innocent

[26] ibid.

[27] Haber, Lynn, and Ralph Norman Haber, "Error Rates for Human Latent Fingerprint Examiners," in Advances in Automatic Fingerprint Recognition, Nalini K. Ratha, ed., New York, Springer-Verlag, 2003.

person should be strenuously guarded against.

The percentage of answers with errors of each of the four types were 8%, 2%, 2%, and 8%, respectively. What should we make of these error rates? We see that the more serious types of errors had lower rates, but we must remember that these answers are consensus answers of the experts in a given lab. For purposes of illustration, suppose that there are two experts in a given lab, and they agree on an answer that turns out to be incorrect. Presumably they consulted each other on their answers, so we cannot multiply their individual error rates to obtain their group error rate, since their answers were not independent events. However, we can certainly say that if the lab error rate is 2%, say, then the individual error rates of the experts at the lab who took the test are all at least 2%.

In 1994, the ASCLD asked the IAI for assistance in creating and reviewing future tests. The IAI asked a company called Collaborative Testing Services (CTS) to design and administer these tests. The format of these tests is similar to the earlier ones, but all of the latent prints are scorable, so there are only two possible types of errors for each question. In addition, individual fingerprint examiners who wish to do so may take the exam by themselves. The Habers report on the error rates for the examinations given from 1995 through 2001. Of the 1685 tests that were graded by CTS, 95 of them, or more than 5%, had at least one erroneous identification, and 502 of the tests, or more than 29%, had at least one missed identification.

Since 1995, the FBI has administered its own examinations to all of its fingerprint examiners. These examinations are similar in nature to the ones described above, but there are a few differences worthy of note. These differences were described in Judge Pollack's reconsideration ruling, in the testimony of Allan Bayle, a fingerprint examiner for 25 years at Scotland Yard.[28]

> Mr. Bayle had reviewed copies of the internal FBI proficiency tests before taking the stand. He found the latent prints utilized in those tests to be, on the whole, markedly unrepresentative of the latent prints that would be lifted at a crime scene. In general, Mr. Bayle found the test latent prints to be far clearer than the prints an examiner would routinely deal with. The prints were too clear – they were, according to Mr. Bayle, lacking in the 'background noise' and 'distortion' one would expect in latent prints that were not identifyable; according to Mr. Bayle, at a typical crime scene only about ten per cent of the lifted latent prints will turn out to be matched. In Mr. Bayle's view the paucity of non-identifyable prints: 'makes the test too easy. It's not testing their ability. It doesn't test their expertise. I mean I've set these tests to trainees and advanced technicians. And if I gave my experts these tests, they'd fall about laughing.'

Approximately 60 FBI fingerprint examiners took the FBI test each year in the period from 1995 to 2001. On these tests, virtually all of the latent prints had matches among the rolled prints. Since many of the examiners took the tests most

[28] U. S. v. Llera Plaza, March 13,2002, at 24.

or all of these years, it is reasonable to suppose that they knew this fact, and hence would hardly ever claim that a latent print had no match. The results of these tests are as follows: there were no erroneous matches, and only three cases where an examiner claimed there was no match when there was one. Thus, the error rates for the two types of error were 0% and 1%.

It seems clear that the error rates of the crime labs for the various types of error are small, but not negligible, and the FBI's rates are suspect for the reasons given above. Given that in many criminal cases, fingerprint evidence forms a crucial part of the prosecution's case, it is reasonable to ask whether the above data, were it to be submitted to a jury, would make it difficult for the jury to find the defendant guilty 'beyond a reasonable doubt,' which is the standard that must be met in such cases.

The question of what this last phrase means is a fascinating one. The U.S. Supreme Court recently weighed in on this issue, and the majority opinion is thorough in its attempt to explicate the history of the usage of this phrase. The Court agreed to review two cases involving instructions given to juries by judges. Standard instructions to juries state that 'guilt beyond a reasonable doubt' means that the jurors need to be convinced 'to a moral certainty' of the defendant's guilt. In one case, 'California defended the use of the moral-certainty language as a 'commonsense and natural' phrase that conveys an 'extraordinarily high degree of certainty.'[29] In the second case, a judge in Nebraska 'included not only the moral-certainty language but also a definition of reasonable doubt as 'an actual and substantial doubt.' The jurors were instructed that 'you may find an accused guilty upon the strong probabilities of the case, provided such probabilities are strong enough to exclude any doubt of his guilt that is reasonable.'[30] The Supreme Court upheld both sets of instructions. The decision regarding the first set was unanimous, while in the second case, two justices dissented, noting that 'the jury was likely to have interpreted the phrase 'substantial doubt' to mean that 'a large as opposed to a merely reasonable doubt is required to acquit a defendant."[31]

The Court went on to note that the meaning of the phrase 'moral certainty' has changed over time. In the mid-19th century, the phrase generally meant a high degree of certainty, whereas today, some dictionaries define the phrase to mean 'based on strong likelihood or firm conviction, rather than on the actual evidence.'[32] Although the Court upheld both sets of instructions, the majority opinion stated that the Court did not condone the use of the phrase 'moral certainty.'

In a concurring opinion, Justice Ruth Bader Ginsburg noted that some Federal appellate circuit courts have instructed trial judges not to provide any definition of the phrase 'beyond a reasonable doubt.' Justice Ginsburg said that it would be better to construct a better definition than the one used in the instructions in the cases under review. She 'cited one suggested in 1987 by the Federal Judicial Center, a research arm of the Federal judiciary. Making no reference to moral certainty, that

[29] Linda Greenhouse, 'High Court Warns About Test for Reasonable Doubt,' The New York Times, March 22, 1994

[30] ibid.

[31] ibid.

[32] American Heritage Dictionary of the English Language, 1992.

definition says in part, ‘Proof beyond a reasonable doubt is proof that leaves you firmly convinced of the defendant’s guilt.”[33]

It may very well be the case that after wading through the above verbiage, the reader has no clearer an idea (and perhaps even has a less clear idea) than before of what the phrase ‘beyond a reasonable doubt’ means. However, juries are given this phrase as part of their instructions, and in the case of fingerprint evidence, they deserve to be educated about error rates involved. We leave it to the reader to ponder whether evidence produced by a technique whose error rate seems to be at least 2% is strong enough to be beyond a reasonable doubt.

On March 13, 2002, Judge Pollack filed his second decision in the Llera-Plaza case. The judge’s ruling was a partial reversal of the original one. His ruling allowed FBI fingerprint examiners to state in court whether there is a match between a latent and a rolled print, but nothing was said in the ruling about examiners not in the employ of the FBI. The judge’s mind was changed primarily because of the testimony of Mr. Bayle who, ironically, was a witness for the defense. Although, as noted above and in the judge’s decision, there are shortcomings in the FBI’s proficiency testing of its examiners, the judge was convinced by the facts that the ACE-V system used by the FBI is essentially the same as the system used in Great Britain and that Mr. Bayle believes in this system without reservation.

As an interesting footnote to this case, after Judge Pollack announced his second ruling, the NIJ cancelled its original solicitation, described above, and replaced it by a ‘General Forensic Research and Development’ solicitation. In the guidelines for this proposal under ‘what will not be funded,’ we find the phrase ’proposals to evaluate, validate, or implement existing forensic technologies.’ This is a somewhat strange way to respond to the judge’s worries about whether the method has been adequately tested in a scientific manner.

## 1.5 The 50K Study

At the beginning of Section 1.3, we stated that in order to decide whether fingerprints are useful in forensics, it is of central importance to be able to estimate how likely it is that a latent print will be incorrectly matched to a rolled print. In 1999, the FBI asked the Lockheed Martin Company to carry out a study of fingerprints. In a pre-trial hearing in the case U.S. v. Mitchell[34], Stephen Meagher, whom we have introduced earlier, explained why he commissioned the study. The primary reason for carrying out this study, he said, was to use the FBI database of over 34 million sets of 10 rolled prints to see how well the automatic fingerprint recognition computer programs distinguished between prints of different fingers. The results of the study could also be used, he reasoned, to strengthen the claim that no two fingerprints are alike. Thus, this study was not originally conceived as a test of the accuracy of matching latent and rolled prints. Nonetheless, as we shall see, this study touched on this second issue.

[33] Greenhouse, loc. cit.

[34] U.S. v. Mitchell, July 7, 1999.

Together with Bruce Budlowe, a statistician who works for the FBI, Meagher came up with the following design for the experiment. The overall idea was to compare every pair of rolled prints in the database, to see if the computer algorithms could distinguish among different prints with high accuracy. It was decided that the number of comparisons needed to carry this out for the whole database was not a reasonable number of comparisons to attempt to carry out (the number is about $5.8 \times 10^{16}$), so they instead chose 50,000 rolled fingerprints from the FBI's master file. These prints were not chosen at random; rather, they were the first 50,000 that were of the pattern 'left loop' from white males. It was decided to restrict the fingerprints in this way because according to Meagher, race and gender have some affect on the size and types of fingerprints. By restricting in this way, the resulting set of fingerprints are probably more homogeneous than a set of randomly chosen fingerprints would be, thereby making it harder to distinguish between pairs from the set. If the study could show that each pair could be distinguished, then the result is more impressive than a similar result accomplished using a set of randomly chosen prints.

At this point, Meagher turned the problem of design over to the Lockheed group. The design and implementation of the study were carried out by Donald Zeisig, an applied mathematician and software designer, and James O'Sullivan, a statistician (both at Lockheed). Much of what follows comes from testimony that Zeisig gave at the pre-trial hearing in U.S. v. Mitchell.

Two experiments with this data were performed. The first began by using two different software programs that each generated a measure of similarity between two fingerprints based on their minutiae patterns. A third program was used to merge these two measures. In a paper on this study, David Kaye[35] delved into various difficulties presented by this study. Information about this study was also provided by the fascinating transcripts of the pre-trial hearing mentioned above[36].

We follow Kaye in denoting the measure of similarity between fingerprints $f_i$ and $f_j$ by $x(f_i, f_j)$. Each of the fingerprints was compared with itself, and the function $x$ was normalized. Although this normalization is not explicitly defined in either the court testimony or the Lockheed summary of the test, we will proceed as best we can. It seems that the values of $x(f_i, f_j)$ were all multiplied by a constant, so that $x(f_i, f_i) \leq 1$ for all $i$, and there is an $i$ such that $x(f_i, f_i) = 1$. One would expect that a measure of similarity would be symmetric, i.e. that $x(f_i, f_j) = x(f_j, f_i)$, but this is never mentioned in the report, and in fact there is evidence that this is not true for this measure.

The value of $x(f_i, f_j)$ is then computed for all $2.5 \times 10^9$ ordered pairs of fingerprints. If this measure of similarity is of any value, it should be very small for all pairs of non-identical fingerprints, and large (i.e. close to 1) for all pairs of identical fingerprints.

Next, for each rolled print $f_i$, the 500 largest values of $x(f_i, f_j)$ are recorded. One of these values, namely when $j = i$, will presumably be very close to 1, but the

[35] Kaye, David, "Questioning a Courtroom Proof of the Uniqueness of Fingerprints," International Statistical Review, Vol. 71, No. 3 (2003), pgs 521-533.

[36] Daubert Hearing Transcripts, at www.clpex.com/Mitchell.htm

other 499 values will probably be very close to 0. At this point, the Lockheed group calculated the mean and standard deviation of this set of 500 values (for a fixed value of $i$). Presumably, the mean and the standard deviation are both positive and very close to 0 (since all but one of the values is very small and positive).

At this point, Zeisig and O'Sullivan assume that the distribution, for each $i$, is normal, with the calculated mean and standard deviation. No reason is given for making this assumption, and we shall see that it gives rise to some amazing probabilities. Under this assumption, one can change the values of $x(f_i, f_j)$ into values of a standard normal distribution, by subtracting the mean and dividing by the standard deviation. The Lockheed group calls these normalized values $Z$ scores. The reader can see that if this is done for a typical set of 500 values of $x(f_i, f_j)$, with $i$ fixed, one should obtain 499 $Z$ scores that are fairly close to 0 and one $Z$ score, corresponding to $x(f_i, f_i)$, that is quite large.

It is then pointed out that if one takes 500 values from the standard normal distribution, the expected value of the largest value obtained should be about 3. This corresponds to the fact that for a standard normal distribution, the probability that a sample value is greater than 3 is about .002 (= 1/500). Thus, Zeisig and O'Sullivan would be worried if any of the non-mate $Z$ scores (i.e. $Z$ scores corresponding to pairs $(f_i, f_j)$ with $i \neq j$) were greater than 3. In fact, except for three cases, which will be discussed below, all of the non-mate $Z$ scores were less than 1.83. This fact should make a statistician worry; if one takes 500 non-negative values from a standard normal distribution, and repeats this experiment 50,000 times, then one should expect to see many maximum values exceeding 3. In fact, when we carried out this experiment 100 times, we obtained a maximum value that was larger than 3 in 71 cases. Thus, the fact that the largest value was 1.83 casts much doubt on whether the distribution in question is normal.

The three non-mate $Z$ scores that were larger than 3 corresponded to the $(i, j)$-pairs $(48541, 48543)$, $(48543, 48541)$, and $(18372, 18373)$. The scores in these cases were 6.98, 6.95, and 3.41. When Zeisig and O'Sullivan found these high $Z$ values, they discovered that in all three cases, the pairs were different rolled prints of the same finger. In other words, the sample of 50,000 fingerprints were from 49,998 different people. It is interesting to note that the pair $(18373, 18372)$ must have had a $Z$ score of less than 1.83, even though the pair corresponds to two prints of the same finger. We'll have more to say about this below. This shows that it is possible for two different prints of the same finger to generate a $Z$ score which is in the same range as do two prints of different fingers.

Now things get murky. The smallest $Z$ score of any fingerprint paired with itself was stated to be 21.7. This high value is to be expected; the reader will recall that for any fingerprint $f_i$, the 500 values correspond to 499 small $Z$ scores and one very large $Z$ score. However, the conclusion drawn from this statement is far from clear. If one calculates the probability that a standard normal random variable will take on a value greater than 21.0, one obtains a value of less than $10^{-97}$. The Lockheed group states its conclusion as follows[37]: 'The probability of a non-mate

[37] Kaye, op. cit. , pg. 530

rolled fingerprint being identical to any particular fingerprint is less than $10^{-97}$.'

David Kaye points out that the real question is not whether a computer program can detect copies of photographs of rolled prints, as is done in this study when a rolled print is compared with itself. Rather, it is whether such a program can, for each finger in the world, put all rolled prints of that finger in one category, and make sure that no rolled prints from any other finger fall into that same category. Kaye notes that although there were so few repeated fingers in the study that one cannot determine the answer to this question with any great degree of certainty, one of the three pairs noted above, of different rolled prints of the same finger, produced a $Z$ score that would occur about once in every 3000 comparisons, assuming the comparisons generate scores that are normally distributed. This means that if one were to make millions of comparisons between pairs of rolled prints of different fingers, one would find thousands of $Z$ scores as high as the one corresponding to the pair $(18372, 18373)$. This would put the computer programmer in a difficult situation. To satisfy Kaye, the program would have to be assigned a number $Z^*$ with the property that if a $Z$ score were generated that was above this value, the program would state that the prints were of the same finger, while if the generated $Z$ score were below this value, the program would state that the two prints were of different fingers. But we can see that there can be no such $Z^*$ value that will always be right. If $Z^* > 3.41$ (the value corresponding to the pair $(18372, 18373)$) then the program would declare that the thousands of the pairs of prints of different fingers mentioned above are in fact prints of the same finger. If $Z^* < 3.41$, then the program would declare that the pair $(18372, 18373)$ are prints of different fingers.

As we noted above, the pair $(18373, 18372)$ was not flagged as having a large $Z$ score. The reason for this is that when the three non-mate pairings mentioned above were flagged, it was not yet known that they corresponded to the same fingers. However, one does wonder whether the Lockheed group looked at the $Z$ score of this pair, once the reversed pair was discovered to have a high $Z$ score. In any event, the $Z$ score of this pair is not given in the summary of the experiments. Robert Epstein, an attorney for the defense in U. S. v. Mitchell, noticed this fact as well, and asked Donald Zeisig, during cross-examination, what the $Z$ score of this pair was. It turns out that the $Z$ score was 1.79. This makes things still worse for the matching algorithm. First, there were other non-mate pairs with larger $Z$ scores. Second, one might expect that the $Z$ score of a pair would be roughly the same in either order (although it isn't clear that this should be so). In any event, a $Z$ score of 1.79 does not correspond to an extremely unlikely event; thus, the algorithm might fail, with some not-so-small probability, to detect an identification between two fingerprints (or else might, with some not-so-small probability, make false identifications). In fact Epstein, in his cross-examination, noted that the pair $(12640, 21111)$ had the $Z$ values 1.83 and 1.47 (depending upon the order), even though it was later discovered that both of this prints were of the same finger. When asked by Epstein, Zeisig agreed that there could possibly have been other pairs of different prints of the same finger (which must have had low $Z$ values, since they were not flagged).

The second experiment that the Lockheed group performed was an attempt to find out how well their computer algorithms dealt with latent fingerprints. To that end, a set of 'pseudo' latent fingerprints was made up, by taking the central 21.7% of each of the 50,000 rolled prints in the original data set. This percentage was arrived at by taking the average size of 300 latent prints from crime scenes versus the size of the corresponding rolled prints.

At this point, the experiment was carried out in essentially the same way as the first experiment. Each pseudo latent $l_i$ was compared with all 50,000 rolled prints, and a score $y(l_i, f_j)$ was determined. For each latent $l_i$, the largest 500 values of $y(l_i, f_j)$ were used to construct $Z$ scores. As before, the $Z$ score corresponding to the pair $(l_i, f_i)$ was expected to be the largest of these by far. Any non-mate $Z$ scores that were high were a cause for concern.

The two pairs $(48541, 48543)$ and $(18372, 18373)$ did give high $Z$ scores, but it was already known at this point that these pairs corresponded to different rolled images of the same finger. There were three other pairs whose $Z$ scores were above 3.6. One pair, $(21852, 21853)$ gave a $Z$ score of 3.64. The latent and the rolled prints were of fingers 7 and 8 of the same person. Further examination of this pair determined that part of finger 8 had intruded into the box containing the rolled print of finger 7. The computer algorithm had found this intrusion, when the pseudo latent for finger 8 was compared with the rolled print of finger 7. This is a somewhat impressive achievement.

One other pair, $(12640, 21111)$, generated large $Z$ scores in both orders. At the time the summary was written, it had not yet been determined whether these two prints were of the same finger. The Lockheed group compared all 20 fingerprints (taken from the two sets of 10 rolled prints corresponding to this pair) with each other. Not surprisingly, the largest scores were generated by prints being compared with themselves. The second highest score for each print was generated when that print was compared with the corresponding print in the other set of 10 rolled prints, and these second-highest scores were quite a bit higher than any of the remaining scores. This is certainly strong evidence that the two sets of 10 rolled prints corresponded to the same person.

The second experiment does not get at one of the central issues concerning latent prints, namely how the quality of the latent print affects the ability of the fingerprint examiner (or a computer algorithm) to match this latent print with a rolled one. Figures 1.2 and 1.3 show that latent prints do not look much like the central 21.7% of a rolled print. Yet it is just these types of comparisons that are used as evidence in court. It would be interesting to conduct a third experiment with the Lockheed data set, in which care was taken to create a more realistic set of latent prints.

## Exercises

1. By the middle of the 20th century, the FBI had compiled a set of more than 10 million fingerprints. Suppose that there are $n$ fingerprint patterns among all of the people on Earth. Thus, $n$ is some number that does not exceed 10 times the number of people on Earth, and it equals this value if and only if

no two fingerprints are exactly alike.

(a) Suppose that all $n$ fingerprint patterns are equally likely. Estimate the number $f(n)$ of random fingerprints that must be observed in order that the probability that two of the same pattern are observed exceeds .5. Hint: To do this using a computer, try different values of $n$ and guess an approximate relationship between $n$ and $f(n)$.

(b) Under the supposition in part a), given that $f(n) = 10$ million, estimate $n$. Note that it is possible to show that if not all $n$ fingerprint patterns are assumed to be equally likely, then the value of $f(n)$ decreases.

(c) Suppose that $n < 60$ billion (so that at least two fingerprints are alike). Estimate $f(n)$.

(d) Suppose that $n = 30$ billion, so that, on the average, every pattern appears twice among the people who are presently alive. Using a computer, choose 10 million patterns at random, thereby simulating the set compiled by the FBI. Was any pattern chosen more than once? Repeat this process many times, keeping track of whether or not at least one pattern is chosen twice. What percentage of the time was at least one pattern chosen at least twice?

(e) Do the above calculations convince you that no fingerprint pattern appears more than once among the people who are alive today?

# Chapter 2

# Streaks

## 2.1 Introduction

Most people who watch or participate in sports think that hot and cold streaks occur. Such streaks may be at the individual or the team level, and may occur in the course of one contest or over many consecutive contests. As we will see in the next section, there are different probability models that might explain such observations. Statistics can be used to help us decide which model does the best job of describing (and therefore predicting) the observations.

As an example of a streak, suppose that a professional basketball player has a lifetime free throw percentage of .850. This means that in her career, she has made 85% of her free throw attempts. We assume that over the course of her career, this probability has remained roughly the same.

Now suppose that over the course of several games, she makes 20 free throws in a row. Even though she shoots free throws quite well, most sports fans would say that she is 'hot.' Most fans would also say that because she is 'hot,' the probability of making her next free throw is higher than her career free throw percentage. Some fans might say that she is 'due' for a miss. The most basic question we look at in this chapter is whether the data show that in such situations, a significant number of players make the next shot (or get a hit in baseball, etc.) with a higher probability than might be expected, given the players' lifetime (or season) percentages. In other words, is the player streaky? One can also ask whether the opposite is true, namely that many players make the next shot with a lower probability than might be expected. We might call such behavior 'anti-streaky.' Both types of behavior are examples of non-independence between consecutive trials.

An argument in favor of dependence might run something like this. Suppose that the player's shot attempts are modeled by a sequence of Bernoulli trials, i.e. on each shot, she has an 85% chance of making the shot, and this percentage is not affected by the outcomes of her previous shot attempts. Under this model, the probability that she makes 20 shots in a row, starting at a particular shot attempt, is $(.85)^{20}$, which is approximately .0388. This is a highly improbable event under the assumption of independence, so this model does not do a good job of explaining the observation.

An argument in favor of the Bernoulli trials model might run as follows. It can be shown that in a sequence of 200 independent trials, where the probability of a success on a given trial is .85, the average length of the longest run of successes is about 20.5. Since many players shoot 200 or more free throws in a given season, it is not surprising that this player has a success run of 20. We will have more to say about the length of the longest run under this model.

There are models that might be used to model streaky behavior. We will consider two of these models in this chapter. The first uses Markov chains. In this model, the probability of a success in a given trial is $p_1$, if the preceding trial resulted in a success, and $p_2$, if the preceding trial resulted in a failure. If $p_1 > p_2$ in the model, then one might expect to see streaky behavior in the model. The model is a Markov chain regardless of whether $p_1 > p_2$, $p_1 = p_2$, or $p_1 < p_2$. If $p_1 = p_2$, the model is the same as the Bernoulli model.

For example, suppose that in the basketball example given above, the player has a 95% chance of making a free throw if she has made the previous free throw. It is possible to show that in order for her overall success rate to be 85%, the probability that she makes a free throw after missing a free throw is 28.3%. It should be clear that in this model, once she makes a free throw, she will usually make quite a few more before she misses, and once she misses, she may miss several in a row. In fact, in this model, once she has made a free throw, the number of free throws until her first miss is a geometric random variable, and has expected value 19, so including the first free throw that she made, she will have streaks of made free throws of average length 20. In the Bernoulli trials model, the average length of her streaks will be only 11.8. Since these two average lengths are so different, the data should allow us to say which model is closer to the truth.

A second possible meaning of streakiness, which we call block-Bernoulli, refers to the possibility that in a long sequence of independent trials, the probability of success varies in some way. For example, there may be blocks of consecutive trials, perhaps of varying lengths, such that in each block, the success probability is constant, but the success probabilities in different blocks might be unequal. As an example, suppose the basketball player has a season free throw percentage of 85%, and assume, in addition, that during the season, there are some blocks of consecutive free throws of length between 20 and 40 in which her probability of success is 95%. This means there must be other blocks in which her success probability is less than 95%. If we compute the observed probability of success over all blocks of length 30, say, we should see, under these assumptions, a wide variation in these probabilities. The question we need to answer is how much wider this variation will be than in the Bernoulli model with constant probability of success. The greater the difference between the two variation sizes, the easier it will be to say which model fits the data better.

It is natural to look at success and failure runs under the various models described above, since the ideas of runs and streakiness are closely related. We will describe some statistical tests that have been used in attempts to decide which model most accurately reflects the data. We will then look at data from various sports and from other sectors of human endeavor.

### Exercises

1. Would you be more likely to say that the basketball player in the example given above was 'hot' if she made 20 free throws in a row during one game, rather than over a stretch of several games?

## 2.2 Models for Repeated Trials

The simplest probability model for a sequence of repeated trials is the Bernoulli trials model. In this model, each trial is assumed to be independent of all of the others, and the probability of a success on any given trial is a constant, usually denoted by $p$. This means, in particular, that the probability of a success following one success, or even ten successes, is unchanged; it equals $p$. It is this last statement that causes many people to doubt that this model can explain what actually happens in sports. However, as we shall see in the next section, even in this model, there are 'hot' and 'cold' streaks. The question is whether the observed numbers and durations of 'hot' and 'cold' streaks exceed the predicted numbers under this model.

This model is a simple one, so one can certainly give reasons why it should not be expected to apply very well in certain situations. For example, in a set of consecutive at-bats in baseball, a batter will face a variety of different pitchers, in a variety of game situations, and at different times of the day. It is reasonable to assert that some of these variables will affect the probability that the batter gets a hit. In other situations, such as free throw shooting in basketball, or horseshoes, the conditions that prevail in any given trial probably do not change very much. Some baseball models have been proposed that have many such situational, or explanatory, variables.

Another relatively simple model is the Markov chain model. It is not necessary for us to define Markov chains here; the interested reader should consult [7]. For our purposes, this model has two parameters, $p_1$ and $p_2$, which correspond to the probability of a success in a trial immediately following a success or a failure, respectively. (The first trial can be defined to have any probability without affecting the large-scale behavior of this model.) It is, of course, possible to define similar, but more complicated, models where the probability of success on a given trial depends upon the outcomes in the preceding $k$ trials, where $k$ is some integer greater than 1. In the case of baseball, some statisticians have considered such models for various values of $k$, and have also assumed that the dependence weakens with the distance between the trials under consideration.

Another model, which has been used to study tennis, is called the odds model. Under this model, the probability $p_{(0,0)}$ that player A wins the first set in a match against player B might depend upon their rankings, if they are professionals, or on their past head-to-head history. If the set score is $(i, j)$ (meaning that player A has won $i$ sets and player B has won $j$ sets), the probability that player A wins the next set is denoted by $p_{(i,j)}$. In the particular model we will consider, odds, instead of probabilities, are used. The odds $O_{ij}$ that player A wins a set, if the set score is $(i, j)$, is defined by the equation

$$O_{ij} = k^{i-j} O_{00} \ ,$$

where $k$ is a parameter that is estimated from the data. If $k > 1$, then this means for example, that a player does better as the set score becomes more and more favorable to him; in other words, he has momentum. The relatively simple form of the above equation is the reason that odds, and not probabilities, are used in this model (the corresponding equation involving probabilities is more complicated).

Finally, models have been proposed that add in 'random effects' to one of the above models, i.e. at each trial, or possibly after a set of trials of a given length, a random real number is added to the probability of success.

## 2.3 Runs in Repeated Trials

Suppose that we have an experiment which has several possible outcomes, and we repeat the experiment many times. For example, we might roll a die and record the face that comes up. Suppose that the sequence of rolls is

$$1, 4, 4, 3, 5, 6, 2, 2, 2, 3, 3, 5, 6, 5, 6, 1, 1, 2, 2 \ .$$

We define a run to be a consecutive set of trials with the same outcome that is not contained in any longer such set. So, in the sequence above, there is one run of length 3, namely the set of three consecutive 2's, and there are four runs of length 2.

If we wish to compare various models of repeated trials, with two possible outcomes, we might look at the length of the longest success run (or failure run), or the number of runs (here it makes little difference whether one looks at the number of success runs or the total number of runs, since the second is within one of being twice the first in any sequence). One might also look at the average length of the success runs. When considering whether a process is Markovian, one might look at the observed success probabilities following successes and failures (or perhaps following sequences of consecutive successes or failures). When considering the block-Bernoulli model, one might compute observed values of the probability of success over blocks of consecutive trials.

In order to use statistical tests on any of the above parameters, one needs to compute or estimate the theoretical distributions of the parameters under the models being considered. For example, suppose that we have a set of data that consists of many strings of 0's and 1's, with each string being of length around 500. For each string, we can determine the length of the longest run of 1's. At this point, we could compare the observations with the theoretical distribution of the longest success run in a Bernoulli trial, where the parameters are $n = 500$ and $p$ equaling the observed probability of a 1 in the strings. This comparison between the data and the theoretical distribution would yield, for each string, a p-value. The reader will recall that if we are testing a hypothesis, the p-value of an observation is the probability, assuming the hypothesis is true, that we would observe an observation that is as extreme or more extreme than the actual observation.

For example, suppose that in a sequence of 500 0's and 1's, we observe 245 0's and 255 1's, and we observe that the longest run of 1's is of length 11. One can show that if $n = 500$ and $p = .51$, then about 90% of the time, the longest success run of 1's in a Bernoulli trials sequence with these parameters is between 5 and 11. Thus, the p-value of this observation is about .1, which means that we might be somewhat skeptical that the model does a good job of explaining the data.

### Exercises

**1** (Classroom exercise) Half of the class should create sequences of 200 coin flips. The other half should write down sequences of length 200 that they think look like typical sequences of coin flips. These data should be labelled only with the person's name. In the next few exercises, we will use some statistics to see if we can determine which are the actual coin flip sequences, and which are made up.

**2** Suppose that we have a Bernoulli trials process in which the probability of a success equals $p$. If there are $n$ trials, what is the expected number of runs? Hint: In any outcome sequence, the number of runs is equal to one more than the number of two consecutive unequal trials. For example, if the outcome sequence is $SFFFSSFSSS$, then there are four pairs of consecutive unequal trials (these pairs correspond to the trials numbered $(1, 2)$, $(4, 5)$, $(6, 7)$, and $(7, 8)$). There are five runs in this sequence. If we let $X_i$ be a random variable which equals 1 if the outcomes of trials $i$ and $i + 1$ are different, then the number of runs $R$ is given by

$$R = \sum_{i=1}^{n-1} X_i \ .$$

Thus, to find the expected value of $R$, it suffices to find the expected value of the right-hand sum; this equals

$$\sum_{i=1}^{n-1} E(X_i) \ .$$

## 2.4 Statistical Tests on Repeated Trials

We now proceed to describe the distributions for the parameters that were mentioned above. We begin by noting that if we let $p_1 = p_2$ in the Markov chain model, we obtain the Bernoulli model. If we are trying to test whether there is evidence of streakiness in a set of data, we might set the null hypothesis to be the statement that $p_1 = p_2$, and the alternative hypothesis to be the statement that $p_1 > p_2$, since one definition of streakiness is that $p_1 > p_2$. If we are testing whether the

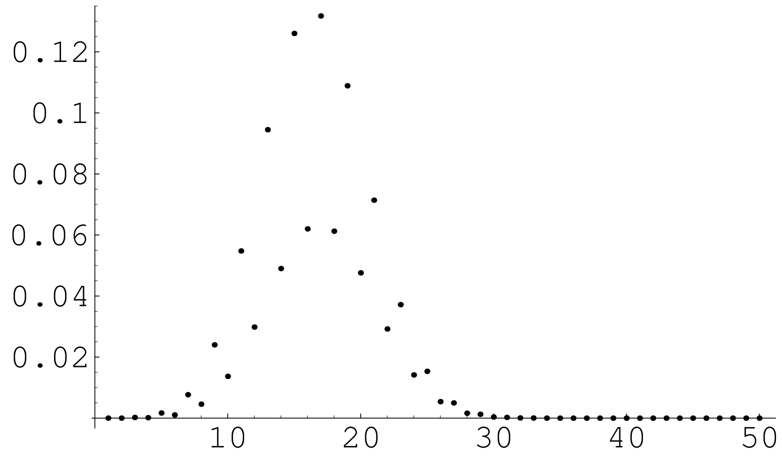


Figure 2.1: Distribution of number of runs in the Bernoulli model

Bernoulli trials model fits the data, without any prejudice towards streakiness as an alternative, we might set the alternative hypothesis to $p_1 \neq p_2$.

In both the Bernoulli and the Markov models, the observed value of $p_1$ is closely related to the observed value of $p$ and the number of pairs of consecutive successes in a sequence. In fact, if we denote the number of such pairs by $d$, then in Exercise 1 the reader is asked to derive the relationship between $n$, $p$, $p_1$, and $d$. This means that if we wish to use $p_1$ as a parameter to compare the two models, we can instead use $d$. In the Appendix, the distribution of $d$ is derived.

The distribution for the number of runs in the Markov chain model is derived in the appendix of this chapter. Plots of both distributions for the case $n = 50$, $p = .2$, $p_1 = .3$, and $p_2 = .175$ are shown in Figures 2.1 and 2.2. The reason for the choices of $p_1$ and $p_2$ is that if $p_1 = .3$, then in order to make the long-range percentage of successes equal .2, as it is in the first model, we must choose $p_2 = .175$. The most obvious difference between the distributions in the Markov chain model and the Bernoulli model is that the number of runs in the former model is generally smaller than in the latter model. This means that one might use an interval of the form $[a+1, \infty)$ as the acceptance region and the interval $[1, a]$ as the critical region in this hypothesis test. One can also compute a p-value of an observation under the Bernoulli model.

Another parameter that we will consider is the length of the longest success run. The distribution of this parameter in the case of Bernoulli trials was computed in [13]. In the appendix, we derive the corresponding distribution for the Markov chain model. In Figure 2.3 we show the two distributions with parameters $n = 100$, $p = .5$, $p_1 = .7$, and $p_2 = .3$. The Markov chain distribution is the one with the right-hand peak. The means of the two distributions for these values of the parameters are 5.99 and 9.57. Again, one can test the hypothesis that $p_1 = p_2$ against the hypothesis that $p_1 > p_2$, using this parameter.

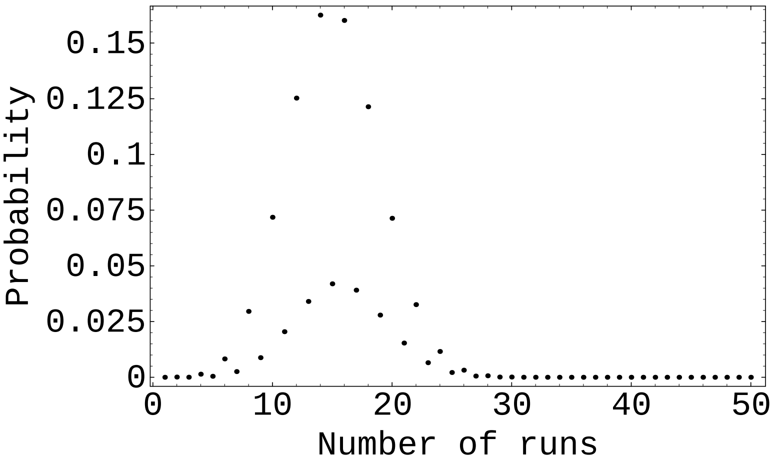


Figure 2.2: Distribution of the number of runs in the Markov model

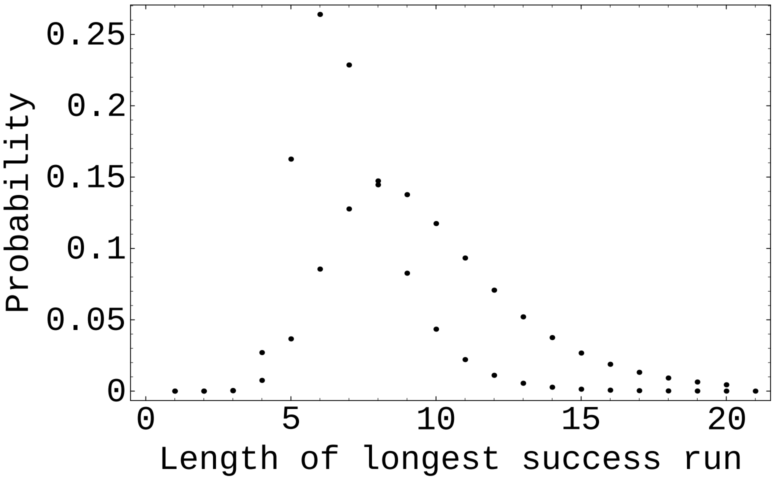


Figure 2.3: Distributions of longest success run for Markov and Bernoulli models

In [2], Albright used a $\chi^2$-test to test for dependence of consecutive at-bats for baseball players; we will now give a general description of this test. Given a sequence of 0's and 1's, define $n$ to be the length of the sequence, $n_0$ to be the number of 0's in the sequence, $n_1$ the number of 1's, and $n_{ij}$ to be the number of consecutive pairs of terms in the sequence with the first equal to $i$ and the second equal to $j$. Then the statistic

$$\chi^2 = \frac{n(n_{00}n_{11} - n_{10}n_{01})^2}{n_0^2 n_1^2}$$

is, under the assumption that the sequence has been generated by a Bernoulli trials process, approximately $\chi^2$-distributed with 1 degree of freedom. So, to apply this test, we compute the value of $\chi^2$ and see if the value exceeds the critical value of the $\chi^2$-distribution at a given level of significance, or we report the p-value corresponding to the observed value.

When trying to decide whether $p \neq p_1$ another parameter that is useful is the number of success doubletons (hereafter called doubletons), i.e. consecutive pairs of successes. If we let $\hat{d}$ denote the number of doubletons in a sequence of length $n$, then Exercise 1 shows that $\hat{d}$ is within one of $n\hat{p}\hat{p}_1$, where $\hat{p}$ and $\hat{p}_1$ are the observed values of $p$ and $p_1$. Thus, using the number of doubletons to study a sequence is very similar to using the value of the parameter $p_1$.

In the appendix, we explain why the number of doubletons is asymptotically normally distributed in the Bernoulli trials model, and give expressions for the asymptotic mean and standard deviation. We now go through an example to show how we can use this parameter. We assume that the sequence is of length $n$, and that the observed value of $p$ (which we will take as the value of $p$) is .3. In Figure 2.4, we show the normal distributions corresponding to values of $p_1 = .3$ (the Bernoulli case) and $p_1 = .4$; the first of these is on the left. The vertical line in the graph marks the right-hand endpoint of a 95% confidence interval. The horizontal coordinate of this line is 57.7. Although this distribution and the next one are discrete, we have drawn them as continuous to make it easier to see the vertical line.

In this case, we are testing the hypothesis that $p_1 = .3$ (the null hypothesis) against the alternative hypothesis $p_1 = .4$. If we have an observed sequence of length 500, we count the number of doubletons. If the null hypothesis were true, the number of doubletons would be greater than 57.7 only 5% of the time. Ths, if the observed number of doubletons is greater than 57.7, we reject the null hypothesis at the 95% level of significance.

When one is carrying out a test of a hypothesis, there are two types of errors that can occur. A type I error occurs when the null hypothesis is rejected, even though it is true. The probability of a type I error is typically denoted by $\alpha$. We see that in the present case, $\alpha = .05$. A type II error occurs if the null hypothesis is accepted, even though it is false. The probability of such an error is denoted by $\beta$. In order to estimate $\beta$, one needs an alternative value of the parameter being estimated. In the present case, if we take this alternative value to be $p_1 = .4$, then one can calculate that $\beta = .404$. It is possible to visualize both $\alpha$ and $\beta$. In Figure 2.4, $\alpha$ is the area under the left-hand curve to the right of the vertical line, and $\beta$ is the area under the right-hand curve to the left of the vertical line.

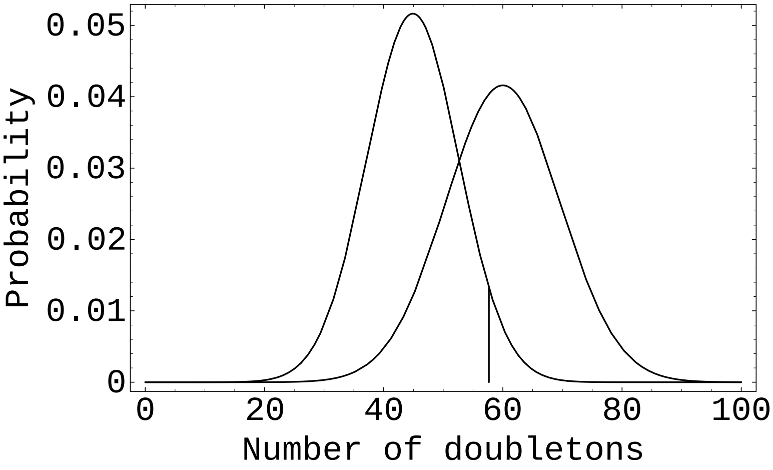


Figure 2.4: Number of Doubletons for $n = 500$, $p = .3$ and either $p_1 = .3$ or $p_1 = .4$

The power of a hypothesis test is defined to be $1-\beta$; it is the probability that the null hypothesis will be accepted at the 95% level of significance when the alternative hypothesis is true. In this case, the power is .596. The higher the power of a test, the better one feels about accepting the results of the test. If we change the value of $n$ to 2000, and graph the same distributions again, we obtain Figure 2.5. In this case, the power of the test is .964.

The reader will recall that we defined a process that generates sequences of successes and failures to be a block-Bernoulli process if the process has success probabilities that change over time. For example, over the course of a baseball season, a batter might have periods in which he has different probabilities of getting a hit.

We will take as our model one that consists of blocks (intervals) of consecutive trials, such that in each block, the individual trials are mutually independent and the success probability is constant. Of course, for such a model to be interesting, the lengths of the blocks in which the success probability is constant must be fairly long, for otherwise one could not hope to differentiate this process from a Bernoulli process. For example, suppose that we define a process that, for each block of size 1, has a success probability that is picked uniformly at random from the interval $[0, 1]$. This process will be completely indistinguishable from a sequence of coin flips of a fair coin. The reason for this is that on each trial, the probability that the coin will come up heads is exactly $1/2$, by symmetry considerations.

We will assume that the lengths of the blocks vary uniformly between $a$ and $b$, and the success probabilities on these blocks vary uniformly between $p_{min}$ and $p_{max}$. It is possible to imagine other assumptions one might make; for example, perhaps the success probabilities are more likely to be near the middle of the interval $[p_{min}, p_{max}]$ than near the end.

What parameter might we use to distinguish block-Bernoulli trials processes

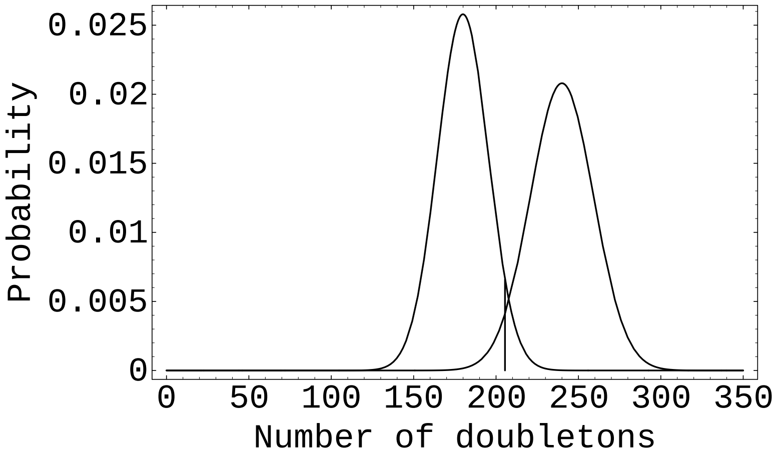


Figure 2.5: Number of Doubletons for $n = 2000$, $p = .3$ and either $p_1 = .3$ or $p_1 = .4$

from Bernoulli ones? One fairly obvious parameter can be obtained in the following way. Suppose we think that the blocks in a block-Bernoulli trials process are of average length 40. In a sequence of length $n$, generated by this process, there will be $n - 39$ blocks of length 40 (these blocks are allowed to overlap; they must start between position 1 and position $n - 39$). For each of these blocks, we compute the observed probability of success by dividing the number of successes in the block by 40, the length of the block. We then take the difference between the maximum and minimum observed probabilities.

In the Bernoulli trials model, if the success probability is $p$, then the observed success probability in a block of length $m$ is, for moderate-sized $m$, approximately normally distributed with mean $p$ and variance $p(1-p)/m$. If we have a sequence of length $n$, we wish to know the distribution of the difference between the maximum and minimum observed success probabilities over all blocks of length $m$.

We do not know whether this distribution has been calculated. However, it can certainly be simulated. In Figures 2.6 and 2.7, we show simulations of the success probabilities for blocks of length 40 in both a block-Bernoulli and a Bernoulli trials sequence. In both simulations, $n = 200$. In the Bernoulli case, $p = .3$, and in the block-Bernoulli case, we let $a = 30$, $b = 50$, $p_{min} = .25$, and $p_{max} = .35$. In the sequence corresponding to Figure 2.6, the block sizes are 38, 40, 38, 39, and 30, and the success probabilities in these blocks are .250, .293, .279, .319, and .288.

The reader will notice that the two figures look similar in the amount of variation in the success probabilities. This similarity should make the reader wonder whether the parameter described above, the difference between the maximum and minimum success probabilities over the blocks, does a very good job at distinguishing between the two models. We will call this parameter the windowed difference of the sequence.

One way to answer this question is to compute the power of a test. In the present case, we let the null hypothesis be the statement that the success probability

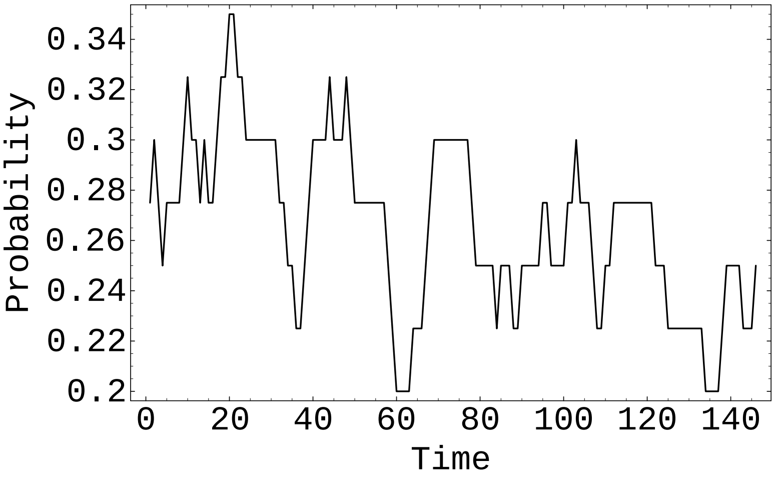


Figure 2.6: Running Success Probabilities for Block-Bernoulli Model

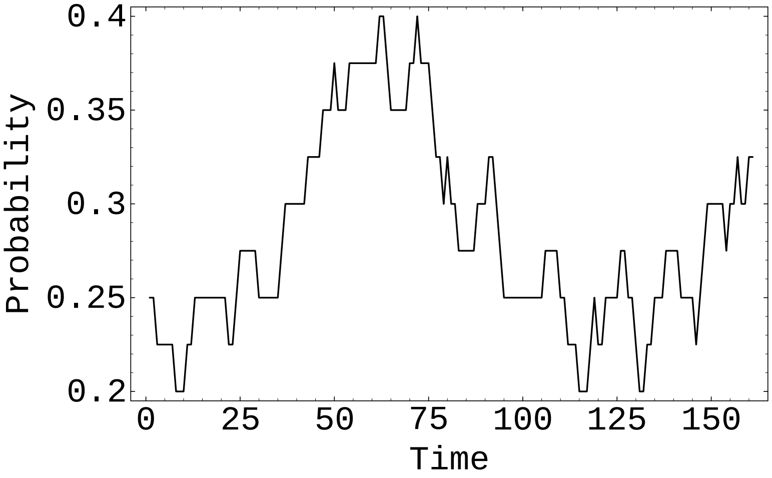


Figure 2.7: Running Success Probabilities for Bernoulli Model

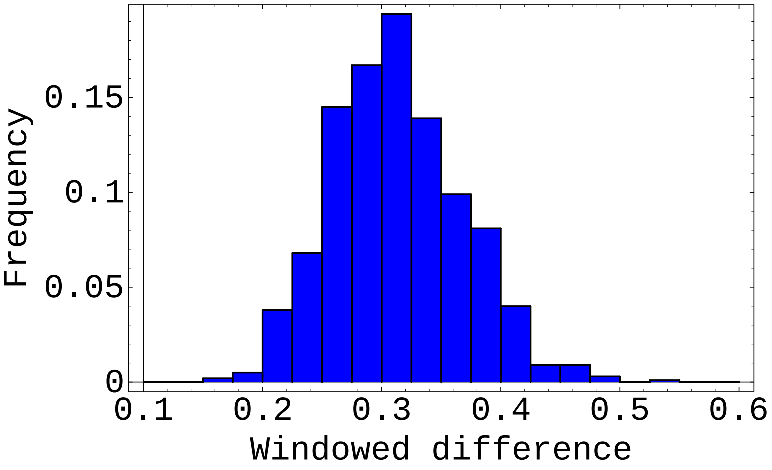


Figure 2.8: Simulated Values of Windowed Difference for Bernoulli Process

is constant over the entire sequence, i.e. that our process is a Bernoulli process. We will take $n = 500$ and $p = .3$. The alternative hypothesis deals with the block-Bernoulli process. Here we assume that $a = 30$ and $b = 50$, although one can certainly imagine varying these values. The specific parameter whose value we will let vary is $p_{max}$, the maximum allowable success probability. We then let $p_{min} = 2p - p_{max}$, so that $p$ is the midpoint of the interval $[p_{min}, p_{max}]$. Note that if we let $p_{max} = .3$, then the block-Bernoulli process reduces to the Bernoulli process.

We now wish to find a critical region, at the .05 level of significance, for our hypothesis test. To do this, we simulate the Bernoulli process 1000 times, and determine a value, called the critical value, below which we find 95% of the values of the windowed difference. In Figure 2.8, we show a histogram of the values of the windowed difference for the Bernoulli process. The critical value is .425, so the region $[.425, 1]$ is the critical region, meaning that if we obtain a value of the parameter that exceeds .425, we reject the null hypothesis.

In Figure 2.9, we show the corresponding histogram for the block-Bernoulli process for the parameter value $p_{max} = .4$. It can be seen that the values in this case are slightly larger than in the Bernoulli case, but the two distributions overlap quite a bit. The power of this test for the specific value of $p_{max} = .4$ is one minus the probability that we will accept the null hypothesis when, in fact, the alternative hypothesis is true. In this case, this means that we want to estimate the probability that the windowed difference falls in the interval $[0, .425]$ for the block-Bernoulli process. Our simulation gives us a value of .84 for this probability, so the power of this test is .16, which isn't very good. One way to increase the power is to increase the size of $n$, but this may or may not be feasible, depending upon how we come upon the data.

When using a parameter to help decide whether a particular model fits the data well, an alternative to testing the null hypothesis against an alternative is to

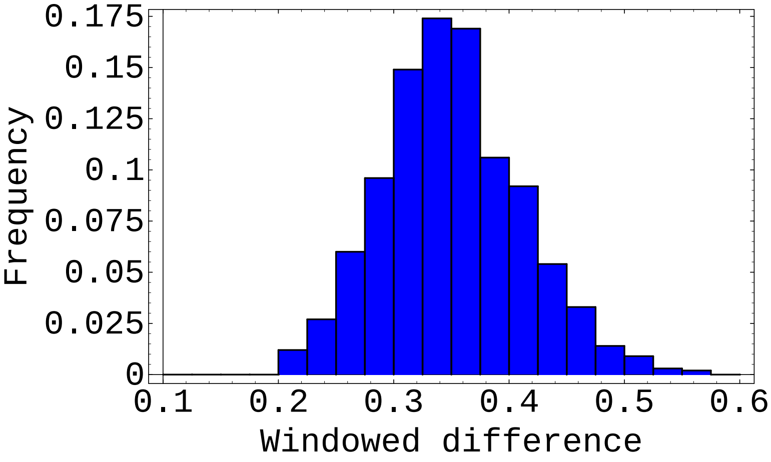


Figure 2.9: Simulated Values of Windowed Difference for Block-Bernoulli Process

compute and report a p-value of the observed value of the parameter. The p-value of an observed value is the probability, under the assumption that the null hypothesis is true, that the parameter will take on a value that is at least as extreme as the observed value.

The phrase "at least as extreme" needs some explanation. Suppose the null hypothesis is of the form "$p = p_0$" for some parameter $p$, and the alternative hypothesis is of the form "$p > p_0$." If a value of $\hat{p}$ is observed for the parameter $p$, then the p-value of this observation is the probability, given the null hypothesis is true, that we would observe a value of $p$ that is at least as large as $\hat{p}$ (the value that was actually observed).

On the other hand, suppose that the alternative hypothesis is of the form "$p \neq p_0$." In this case the p-value of the observed value $\hat{p}$ is the probability, given the null hypothesis is true, that

$$|p - p_0| \geq |\hat{p} - p_0| \ .$$

For a given hypothesis, the smaller the p-value, the less faith we have in the model as an explanation of the observed data. On the other hand, if we have many data sets that are supposedly explained by a certain model, then some of the data sets have small p-values, even if the model is correct. To see why this is true, consider the following simple experiment. We have a coin, and we are testing the hypothesis that it is a fair coin, i.e. that the probability of a head coming up when we flip the coin is .5. Suppose the coin is fair. Suppose we perform 100 tests in each of which the coin is flipped 500 times. Then about 5% of the tests will report a p-value of less than .05. The reason for this is that the observed value $\hat{p}$ has a certain known distribution, because the null hypothesis is true, and we are using this distribution to calculate the p-values of the observations. Since the observations are distributed according to this known distribution, about 5% of them will have p-values less than .05 (just as about 45% of them, say, will have p-values less than .45).

Another way of looking at this situation is as follows: If we have a large number of observations under a single model, and the p-values of these observations are fairly uniformly distributed across the interval $[0, 1]$, then the model is doing a good job of fitting the data. Of course this does not mean that the model is the only one that fits the data well, or that it the 'correct' one. One should consider other factors, such as the simplicity (or lack thereof) of the various models under consideration, before deciding which model one likes best.

### 2.4.1 Selection Bias

If one calculates the value of a parameter for a given observation, and the value is seen to be closer to the expected value of that parameter under one model than under a second model, one is tempted to state that the first model does a better job of explaining the observation than does the second model. However, if one selects only those observations for which the parameter is closer to the expected value in the first model than in the second model, then one is guilty of selection bias.

It is also, in general, the case that if one considers more parameters in creating a model, the model may fit the observed data better than the original, simpler model. For example, since the Markov model subsumes the Bernoulli model, if one includes $p_1$ and $p_2$ as parameters, thereby considering the Markov model, this model will do a better job of fitting a sequence of 0's and 1's than will the Bernoulli model. The trade-off is that the Markov model is more complicated than the Bernoulli model. If the Bernoulli model does a good job of describing the observations, than its simplicity should be an argument against its rejection.

## Exercises

**1** Suppose that we have a sequence of successes and failures of length $n$. Define $\hat{p}$ to be the observed proportion of successes, and define $\hat{p}_1$ to be the observed proportion of successes (excluding the last entry in the sequence, if it is a success) that are followed by another success. Define $\hat{d}$ to be the observed number of consecutive pairs of entries that are both successes (i.e. the number of doubletons). Show that if the last entry in the sequence is a failure, then

$$\hat{d} = n\hat{p}\hat{p}_1 \ ,$$

while if the last entry is a success, then

$$\hat{d} = (n\hat{p} - 1)\hat{p}_1 \ .$$

Thus, in either case, $\hat{d}$ is within one of $n\hat{p}\hat{p}_1$.

**2** Can you describe a win-loss sequence for a baseball team that you would say exhibits streakiness, yet for which the parameter Black is very small?

## 2.5 Data from Various Sports

### 2.5.1 Baseball

In baseball, the sequence of hits and outs recorded by a batter in consecutive at-bats is a candidate for study. A plate appearance that results in a walk (or the batter being hit by a pitch, or a sacrifice bunt) is not counted as an at-bat, because in those cases, it is thought that the batter has not been given a chance to hit a well-pitched ball. We remind the reader that even if one uses a Bernoulli trials approach to model such sequences, one should admit that this is only a first approximation to what is actually occurring on the field; presumably the batter's chance of success in a given at-bat is affected by who is pitching, whether the game is being played during the day or at night, how many runners are on base, and many other factors. However, it is still of interest to test the Bernoulli trials model to see how well it fits the data.

We will now apply the various tests described above to our data set. This data set consists of all players who played in the major leagues between 1978 and 1992, and is taken from the www.retrosheet.org website. We greatly appreciate their labor in creating the files of raw data that we used.

The first test we will apply is the chi-squared test of Albright, described above. For each player and for each season in which that player had at least 150 at-bats, we computed the chi-squared value of his sequence of hits and outs, and then computed the p-value of this chi-squared value. If there is no overall departure from the Bernoulli trials model, we would expect to see the p-values uniformly distributed across the interval $[0, 1]$. The point here is that even though one might find some seasons for some players for which the p-value is less than some prescribed level of significance, say .05, we should not reject the hypothesis that the Bernoulli trials model fits the data well unless we observe many more than 5% of the p-values below .05.

There were 4786 player-seasons in our data set with the at-bat cutoff of 150. Figure 2.10 shows a histogram of the p-values associated with the chi-squared values for these player-seasons. Since the interval $[0, 1]$ has been partitioned into 20 subintervals, uniformity would imply that about 5% of the values occur in each sub-interval. It can be seen that the p-values are essentially uniform across the unit interval. In particular, there is no apparent excess of small p-values that would correspond to large values of the chi-squared parameter (and therefore to either streaky or anti-streaky behavior).

Another pair of parameters to look at is the pair $(\hat{p}_1, \hat{p}_2)$; these are the observed probabilities of a hit after a hit and after an out, respectively. Under the assumption of Bernoulli trials, it can be shown that $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed with a mean and a variance that depend on $n$ and $p$ (we have shown this in the Appendix). In fact, the mean is 0 and the variance is

$$\frac{n + 3np - 1 - 5p}{qn^2} \; .$$

We can test to see if the hits-outs sequences of baseball players fit the Bernoulli

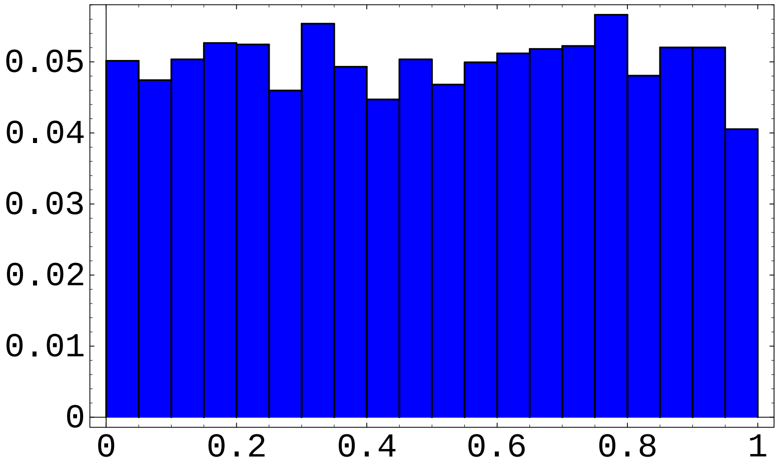


Figure 2.10: p-Values for Albright's chi-squared test in baseball.

model by looking at the values of $\hat{p}_1 - \hat{p}_2$. It should be noted that the $n$ and $p$ values change as we move from player to player, so we will aggregate the values in the following way. We partition the interval $[0, 750]$ into 15 subintervals of length 50, and we partition the interval $[0, .400]$ into 16 subintervals of length .025. Given a pair of subintervals, one of each type, we take all player-seasons in which the number of at-bats is in the first subinterval and the batting average (the value of $p$) is in the second interval. So, for example, the pair $[400, 450]$ and $[.300, .325]$ consists of all player-seasons in which the player had between 400 and 450 at-bats and a batting average of between .300 and .325. To avoid duplications, the subintervals do not contain their left-hand endpoint. The reason that we do this is so that we have enough data for certain values of $n$ and $p$ to be able to say something meaningful.

Of the 240 possible pairs of subintervals, 22 of them have data sets of size at least 100. There were 3188 player-seasons in these data sets (which is more than two-thirds of the data). For each of these data sets, we calculated the average value of $\hat{p}_1 - \hat{p}_2$ (recall that in the Bernoulli model, the mean of $\hat{p}_1 - \hat{p}_2$ is 0). The variance of the average value of $\hat{p}_1 - \hat{p}_2$ is the variance of $\hat{p}_1 - \hat{p}_2$ divided by the square of the size of the data set. Since $\hat{p}_1 - \hat{p}_2$ is approximately normal, so is the average value of $\hat{p}_1 - \hat{p}_2$. So one way to see how well the data fits the Bernoulli model is to compute, for each data set, the $z$-value of the average value of $\hat{p}_1 - \hat{p}_2$; a $z$-value that is less than -2 or larger than 2 is significant at the .05 level. The $z$-value is obtained by dividing the observed value of the average of $\hat{p}_1 - \hat{p}_2$ in the data set by the standard deviation of the average value of $\hat{p}_1 - \hat{p}_2$.

The results are shown in Figure 2.11. Of the 22 $z$-values, one is of absolute value greater than 2 (which is about what one would expect if one were to choose 22 random values from a normal distribution).

We turn next to the length of the longest success run (i.e. the longest run of hits). For each player-season in our data, we can compute the length of the longest success

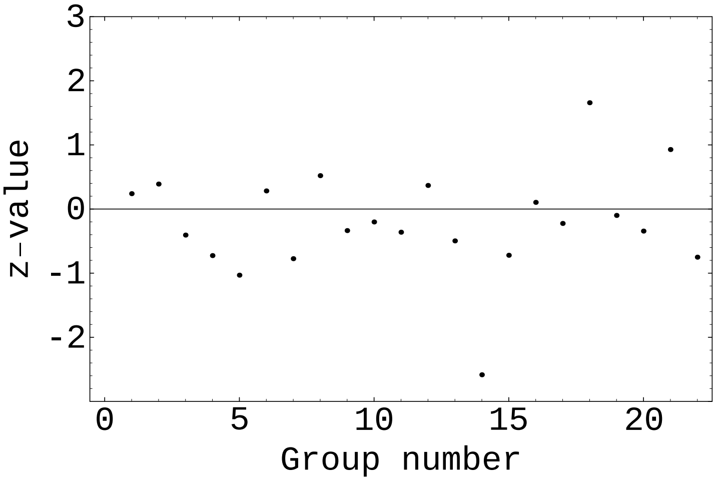


Figure 2.11: z-Values for the Average of $\hat{p}_1 - \hat{p}_2$.

run. However, in the Bernoulli model, the distribution of this length depends upon both $n$, the length of the sequence (in this case, the number of at-bats), and $p$, the probability of a success in an individual trial (in this case, the batting average). The Markov model depends upon $n$, $p$, and $p_1$, the probability of a success following a success. We would like to be able to display the aggregate data against the predictions made by both models.

In the case of the Bernoulli model, we can proceed in several ways. Our first method of comparison will be carried out as follows. For each of the 4786 player-seasons with at least 150 at-bats, we will simulate a season using Bernoulli trials with the same number of at-bats and batting average as the actual season. Then we will observe the longest run of successes in each of these simulated seasons. Finally, we will compare the distribution of the lengths of the longest run of successes in the simulated seasons with the corresponding distribution in the actual data.

When we carry this procedure out, we obtain the results shown in Figure 2.12. The horizontal coordinate represents the length of the longest success run in each player-season, and the vertical coordinate represents the observed frequencies. The fit between the simulated and actual data is quite good. In addition, the reader will recall Figure 2.3, which shows that in the Markov model, if $p_1 > p$, then the distribution of the longest success run is shifted to the right from the corresponding distribution in the Bernoulli model. The aggregate data does not show any such shift.

Another way to compare the Bernoulli model with the data is to proceed as we did above with the distribution of $\hat{p}_1 - \hat{p}_2$. We will group the data using pairs of subintervals, so that the numbers of at-bats and the batting averages are close to equal across the group. Then, for each group, we will compute the distribution of the longest success run and compare it with the theoretical distribution. This comparison will be carried out using a chi-square test. For each of the 22 groups we

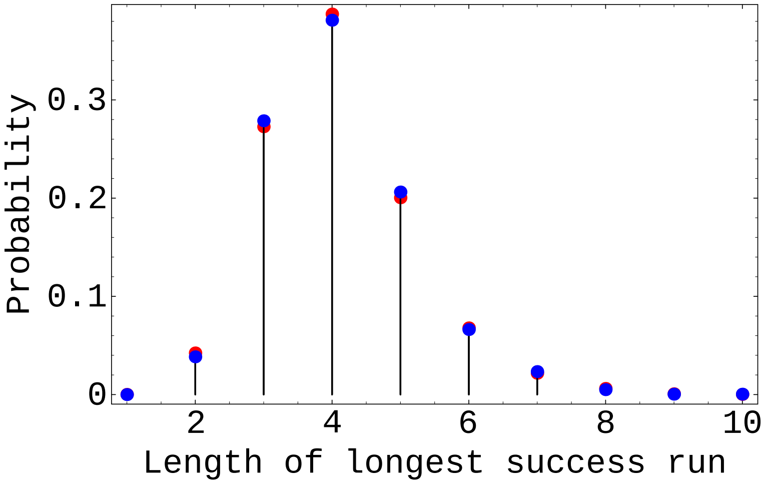


Figure 2.12: Distributions of Simulated and Actual Longest Success Runs.

used above (those containing at least 100 player-seasons) we will report a p-value for the chi-square value.

In Figure 2.13, we show the results of the above calculations. For each of the 22 groups of data containing at least 100 player-seasons, the observed and theoretical average length of the longest success run; the theoretical average is obtained by using, for each group, the average number of at-bats and the average success probability. As can be seen, the fit is very good; the relative error is never greater than 6%, and only twice is it as large as 4%. For each of the 22 data sets, we also compute the chi-squared value of the comparison between the theoretical and observed distributions, and then compute the p-values corresponding to these chi-squared values. Figure 2.14 shows the 22 p-values. These p-values show that the Bernoulli model does a very good job of fitting the data.

It has been suggested that there might be a qualitative difference between success runs and failure runs. The thought was that while success runs are limited in length by the skill of the batter, failure runs might, in some cases, continue longer than predicted because the hitter will spiral downwards psychologically during a slump. We ran the above tests again on the same 22 groups, this time looking at the longest failure runs. The p-values, corresponding to those in Figure 2.14, are shown in Figure 2.15. It can be seen that once again, the fit is very good.

Another aspect of baseball that has been taken to provide evidence in favor of streakiness is the idea of a hitting streak. A hitting streak by a player is a set of consecutive games played by the player in which he gets at least one hit in each game. In order to qualify as a hitting streak, the games must all occur in a single season. The longest hitting streak that has ever occurred in the major leagues lasted 56 games; it happened in 1941, and the player who accomplished this streak was Joe DiMaggio. Much has been written about this streak (see, for example, [3], [6], and [10]). The question that we wish to consider is whether the fact that this

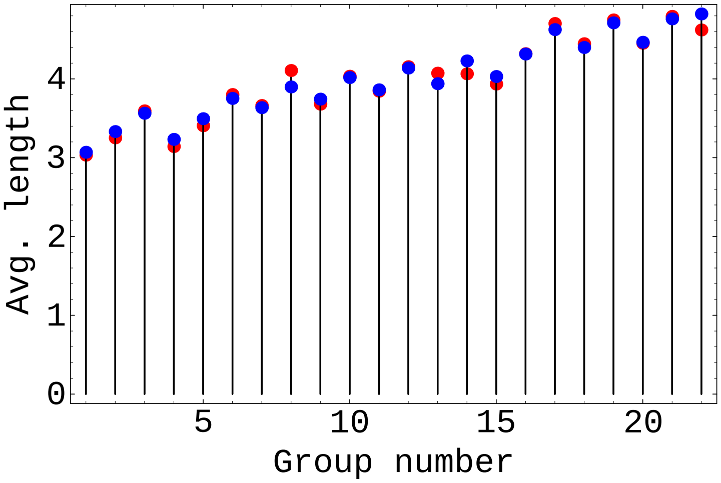


Figure 2.13: Observed vs. Theoretical Longest Success Run Lengths.

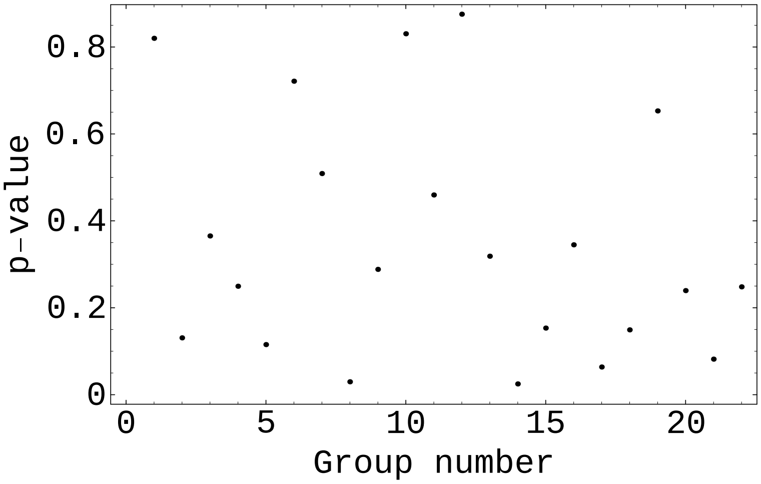


Figure 2.14: p-Values for Chi-Squared Comparison of Longest Success Run Data.

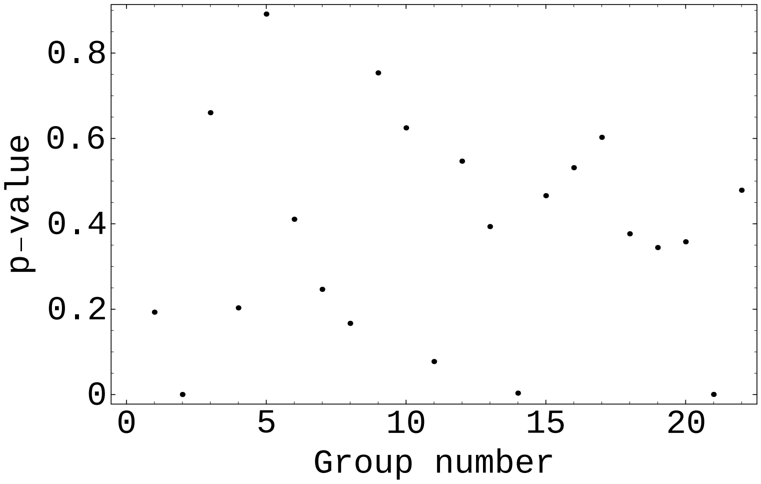


Figure 2.15: p-Values for Chi-Squared Comparison of Longest Failure Run Data.

streak occurred provides evidence against the Bernoulli trials model.

If a hitter has a fixed number of at-bats in each game, then it is fairly easy to calculate the probability that he will hit safely in $n$ consecutive games, assuming the Bernoulli model. For example, if a hitter has a .350 batting average, and has four at-bats in each game, then, under the Bernoulli model, the probability that he gets at least one hit in a given game is

$$1 - (.650)^4 = .8215 .$$

Thus, in a given set of 56 consecutive games, the probability that the hitter gets at least one hit in each game is

$$(.8215)^{56} = .0000165 .$$

This sounds like a very unlikely event, but we're not really asking the right question. A slightly better question might be the following: Given that a hitter plays in 154 games in a season, and has a season batting average of .350, what is the probability that he bats safely for at least 56 games? (We use 154 for the length of a season, since almost all of the player seasons with batting averages of at least .350 occurred before the seasons were lengthened to 162 games.) Once again, under the Bernoulli model, we can answer this question rather easily. Putting this question in terms of ideas that we have already introduced, this question is equivalent to asking what is the probability, in a Bernoulli trials sequence of length 154, and with success probability .8215, that the longest success run is at least of length 56. The question, in a form similar to this one, was considered in [14]. The answer, in this case, is .0003.

Of course, there have been many batters who have had season batting averages of much higher than .350, and it is clear that as the batting average climbs, so does

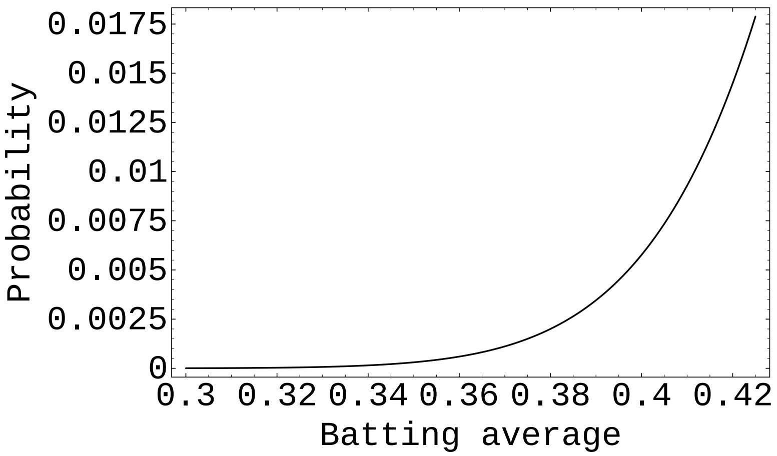


Figure 2.16: Probability of a hitting streak of at least 56 games.

the probability of having a long hitting streak. Figure 2.16 shows the probability of a hitting streak of at least 56 games in a 154-game season as a function of the season batting average. This figure shows that the season batting average has a large effect on the probability of a long hitting streak. Figure 2.17 shows a histogram of player seasons since 1901 for which the batting average was at least .340.

We still are not asking quite the right question. There are two changes that we should make. First, we do not want to know the probability that a given hitter will have a long hitting streak; rather, we want to know the probability that at least one of a set of hitters will have a long hitting streak. We will take this set of hitters to correspond to all hitters since 1901 who batted at least .340 in a season and had at least 300 at-bats (there were 430 such seasons). We are arbitrarily disregarding hitters whose batting averages are below .340.

The second change that should be made concerns the variability of the number of at-bats belonging to an individual player over a set of games. It should be clear that if a batter averages four at-bats per game, but the number of at-bats varies widely over a set of games, then it is less likely that he will have a long hitting streak. As a simple example to help show this, consider a batter with a .350 batting average under the following two scenarios: first, he gets exactly four at-bats in each of 10 consecutive games, and second, he gets, alternatively, two and six at-bats in 10 consecutive games. In the first case, the probability that he gets at least one hit in each game is

$$\Bigl(1 - (.650)^4\Bigr)^{10} = .13997 \ .$$

In the second case, the probability is

$$\Bigl(1 - (.650)^2\Bigr)^5 \Bigl(1 - (.650)^6)^5\Bigr) = .04400 \ .$$

So, here is the question that we want to try to answer: Suppose we restrict our

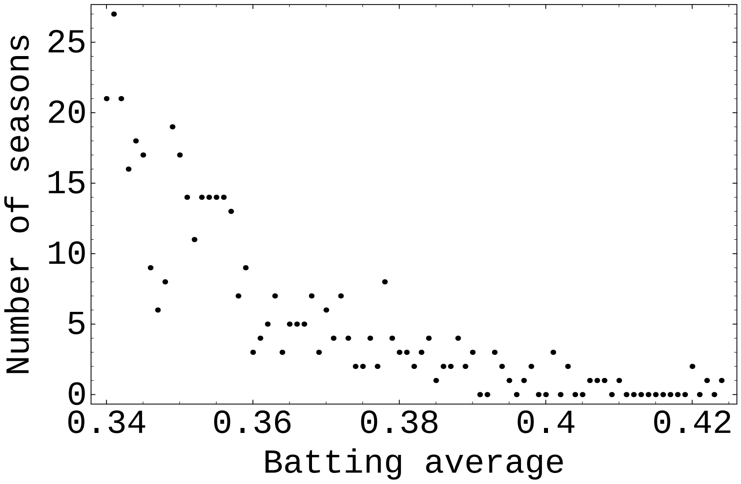


Figure 2.17: Number of player seasons with a given batting average.

attention to those player-seasons in which the player batted at least .340 (we will call this set our good-hitter data set), and suppose we take into account the typical variation between games of the number of at-bats of a player in a game. What is the probability, under the Bernoulli trials model, that at least one of the players would have a hitting streak of at least 56 games?

This is too complicated a question to hope for a theoretical answer, but we can simulate the seasons of these hitters, using the observed batting averages and numbers of games and at-bats. Here is how the simulation is carried out. We begin by finding an estimate for the standard deviation in the sequence of numbers of at-bats in a game, over a typical player season. We can use our original data set from the years 1978 to 1992 to obtain this estimate. We restrict our attention to those players who batted at least .300 for the season, and who had at least 300 at-bats. There were 369 player seasons that satisfied these requirements. For each such season, we found the sequence of numbers of at-bats in all games *started* by the player. The reason for this restriction is that the players in our good-hitter data set started almost every game in which they played. For each of these 369 seasons, we compute the standard deviation of the sequence of at-bats. Then we compute the average of these standard deviations. When we do this, we obtain a value of .8523.

For each player-season in our good-hitter data set, we will produce a sequence of at-bats per game for the number of games in which the player participated that season. The terms of this sequence will be drawn from a normal distribution with mean equal to the average number of at-bats by the player during that season, and with standard deviation equal to .8523. The terms will be rounded to the nearest integer, so that they represent numbers of at-bats (which must be integers). If an integer in the sequence is less than or equal to 0, we throw it out. The reason for this is that a hitting streak is not considered to be interrupted if a player appears

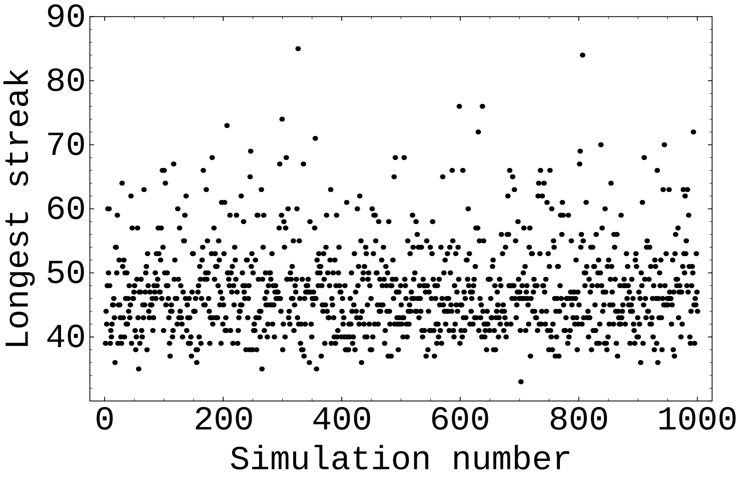


Figure 2.18: Simulated longest hitting streak data.

in a game but does not record any official at-bats.

Since we are operating under the assumption of Bernoulli trials, we use the player's season batting average to simulate his season with the above simulated at-bat sequence. We then record the longest hitting streak in the season. We do this for each player, and record the longest hitting streak among all of the players. To estimate the probability that someone would have a hitting streak of at least 56 games, we carry out the above procedure many times, and calculate the fraction of those trials that lead to at least one such streak.

The results of the simulation are shown in Figure 2.18. The above procedure was carried out 1000 times. In the simulated data, the length of the longest hitting streak ranged from 33 to 85, and 126 times the length was at least 56 games. This simulation shows that the probability, under the Bernoulli trials model, that we would observe a hitting streak of at least 56 games in length is about 1/8 ($\approx$ 126/1000) and thus Joe DiMaggio's streak does not cast doubts on the merits of this model. The simulation also shows that viewed from this angle, DiMaggio's feat is probably not the most amazing feat in all of sports, or even in baseball.

One can consider another question that is usually raised when DiMaggio's streak is discussed; that is the question of whether anyone will ever break his record. Predicting the future is a dangerous business, but one can use the above simulations, together with known data, to say something of interest. Of the 430 player seasons in which the player batted at least .340, about three-quarters of them occurred in the period 1901-1950. Thus, the rate at which .340 seasons occur has changed dramatically over the years. Figure 2.19 shows the moving average of the number of .340-plus player seasons, averaged over ten-year windows. Currently, the average number of such player seasons is about 4 per year. Thus we assume that there will be about 4 such player seasons in each year in the future. We have seen that in 430 such player seasons, the probability is about 1/8 that someone will have a 56-game

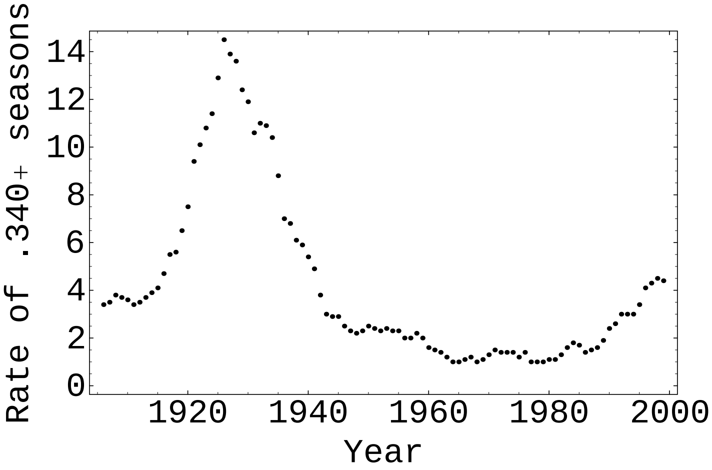


Figure 2.19: Moving average number of .340+ player seasons.

hitting streak. Thus, we should expect to see one in the next 3440 (= 8 × 430) such player seasons, which translates to 860 years. Putting it another way, it seems very unlikely that we will see such a streak in any of our lifetimes. Viewed from this perspective, DiMaggio's streak seems much more impressive than before.

There is another way to estimate how long it might be until the record is broken. It was stated above that the probability is .0003 that a .350 hitter will have a hitting streak of at least 56 games in a season. If there are about 4 such seasons per year in the future, we would expect to see such a streak, assuming the Bernoulli trials model, every $1/(.0003 \times 4) = 833$ years.

The above argument does not take into account the incredible pressure that will surely be felt by any player who approaches the record. Joe DiMaggio must have felt pressure as well during the streak, but once he had broken the existing record of 44 consecutive games, the pressure probably abated quite a bit. Thus, any player who threatens DiMaggio's record will feel pressure for quite a bit longer than DiMaggio did. Of course, those who believe that Bernolli trials are a good model for hitting in baseball might argue that the pressure under which a player finds himself in such a situation is irrelevant.

We now turn to team statistics; we seek to answer the question of whether teams exhibit streaky behavior. One way to check this is to recall that if a process has a positive autocorrelation (meaning that the probability of a success in a given trial increases if the previous trial is a success), then the lengths of the longest streaks will, on average, be longer than in the Bernoulli trials model. This will be true of both winning and losing streaks.

There are 390 team seasons in our data set. For each of these seasons, we computed the lengths of the longest winning and losing streaks. Then we grouped the team seasons by rounding the winning percentages to the nearest multiple of .01 (so, for example, a winning percentage of .564 is rounded to .56). For each

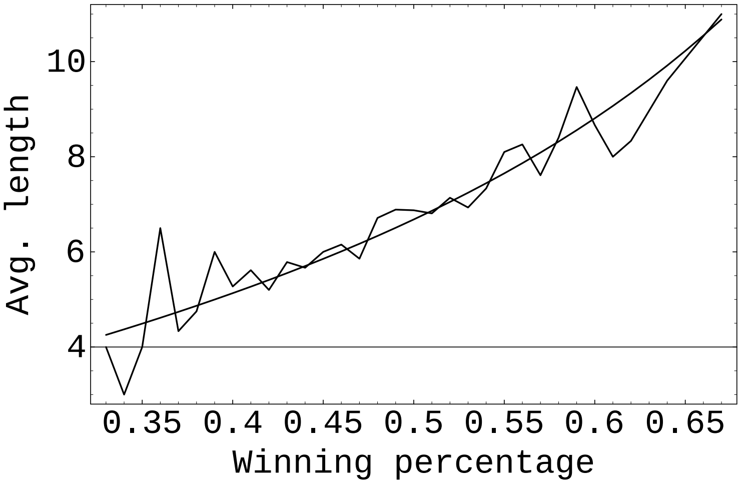


Figure 2.20: Actual and Theoretical Longest Winning Streak Lengths.

group, we calculated the average lengths of the longest winning and losing streaks. Figures 2.20 and 2.21 show these average lengths, as well as the average lengths in the Bernoulli trials model. We note that the fit between the actual and the theoretical longest streak lengths is very good, except perhaps at the ends of the graphs. But even at the ends, there is no bias; the reason for the large deviations is that the group sizes are so small. Of the 390 team seasons in the data set, only 9 have rounded winning percentages that lie outside of the interval $[.36, .64]$.

In the book Curve Ball[1], the authors, Jim Albert and Jay Bennett, study the question of team streakinenss, in the sense of the block-Bernoulli process. Specifically, their model of streaky behavior is as follows: A hypothetical season is split into nine non-overlapping blocks of 18 games each (so in our notation, $a = b = 18$). Three winning percentages are computed, denoted by $p_{\mathrm{C}}$, $p_{\mathrm{av}}$, and $p_{\mathrm{H}}$ (for cold, average, and hot). The percentage $p_{\mathrm{av}}$ is the team's season winning percentage. The percentages $p_{\mathrm{C}}$ and $p_{\mathrm{H}}$ are defined to be .1 less than and more than $p_{\mathrm{av}}$, respectively. So this is somewhat like our $p_{min}$ and $p_{max}$, except that in their model, the block winning percentages are not uniformly chosen in the interval $[p_{\mathrm{C}}, p_{\mathrm{H}}]$; rather, there are only three possible block winning percentages. Finally, each of these three block winning percentages is assigned randomly to three of the nine 18-game blocks. Figure 2.22 shows an example of the block winning percentages for a team whose season percentage is .55.

A team whose win-loss sequence arises from the above block-Bernoulli process will be said to be streaky, while one whose win-loss sequence arises from a Bernoulli process will be called consistent. Of course, we do not expect all teams to behave in one of these two ways; we need a parameter that can be calculated, and that does a good job of distinguishing between these two models. As we said above, our windowed difference parameter does not distinguish very well between the two models we posited.

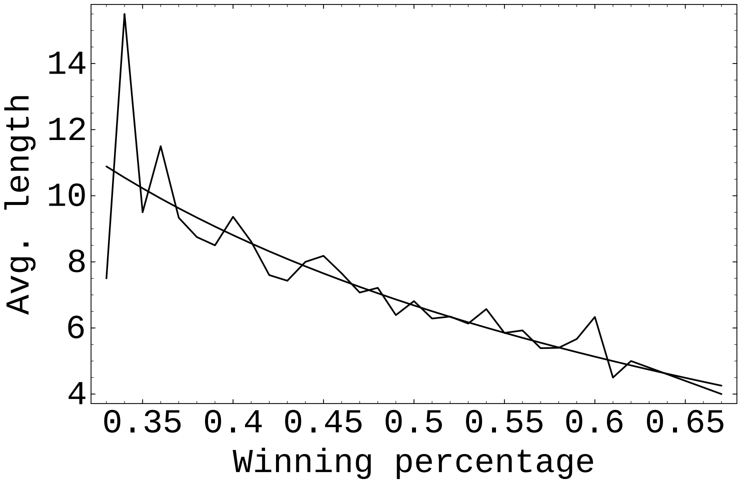


Figure 2.21: Actual and Theoretical Longest Losing Streak Lengths.

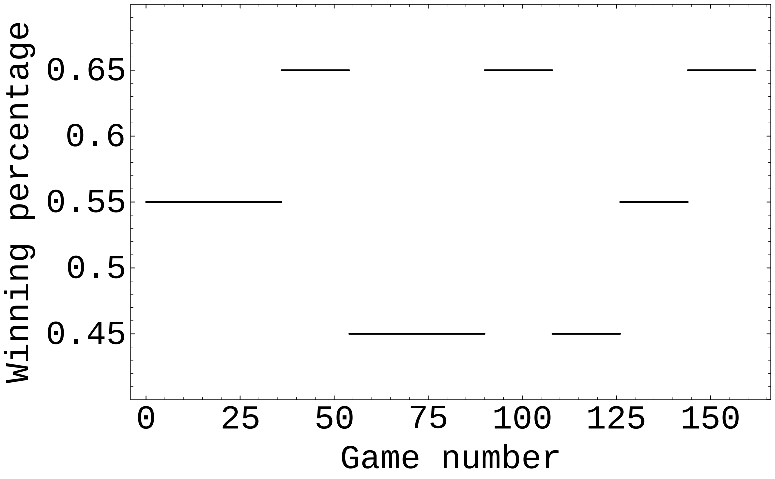


Figure 2.22: Block winning percentages for a streaky team

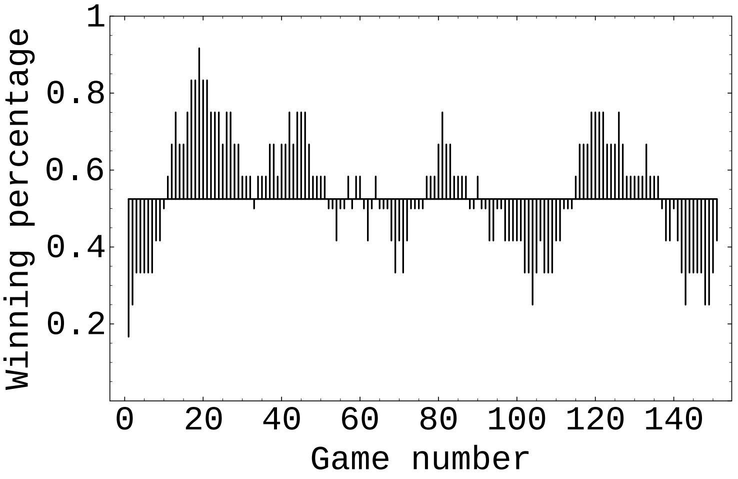


Figure 2.23: Windowed winning percentage for 1984 Baltimore Orioles

Albert and Bennett define their parameter, called Black, as follows. Given a team's win-loss sequence for a season, they compute the windowed winning percentages for all blocks of 12 games (so in a 162-game season, there are 151 such blocks, starting at games 1 to 151). They plot these percentages, along with a horizontal line whose $y$-value is the season winning percentage. An example of this plot is shown in Figure 2.23; the team is the 1984 Baltimore Orioles. The parameter Black is easy to describe in terms of this figure; it is the area of the black region.

It is clear that if a team is consistent, then Black will be very small, while if a team is streaky, Black will probably be large. Thus, this parameter might be able to distinguish between the block-Bernoulli and the Bernoulli models. To decide which model fits a given team better, Albert and Bennett compute the winning percentage, call it $p$, of the team. Then they obtain, by simulation, the distribution of the parameter Black under both models. The actual value of Black for the team is computed, and compared with the two distributions. For each distribution, a p-value is reported. If the first p-value of the observation is larger than the second p-value, then they claim that the first model fits the team's performance better than the second model.

In fact, they go further and use the ratio of the p-values as the odds that the team is consistent (or streaky). For example, if the p-values for a given team, under the consistent and streaky models, are .08 and .30, respectively, then they say that the probability that the team is streaky is .79 ($= .30/(.08 + .30)$). This is a fairly loose use of the word 'probability,' since there is no obvious experiment or sample space on which this probability is based.

We will proceed in a different direction with the parameter Black. We are trying to determine whether the Bernoulli model does a good job of fitting the win-loss sequences in our data set. For each of the 390 teams in this data set, we can calculate a p-value for the observed value of Black, under the null hypothesis of

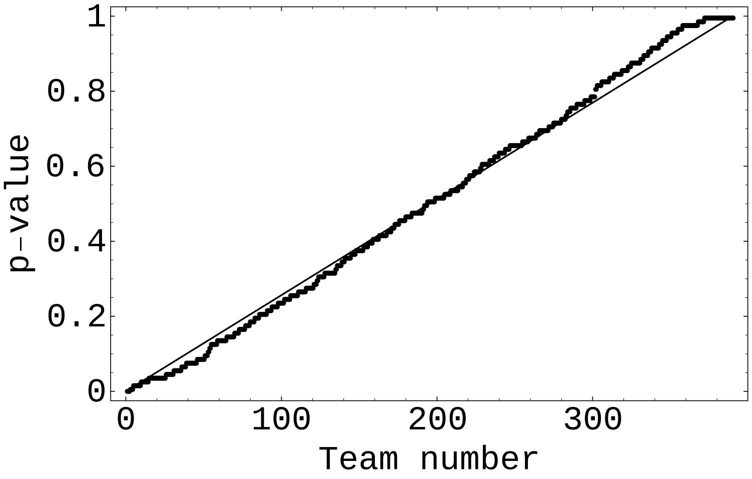


Figure 2.24: p-values for the Black parameter

the Bernoulli model. For each team, we use the winning percentage of that team, rounded to the nearest .01, to compute a simulated distribution for Black. If there are more streaky teams than can be explained well by the Bernoulli model, we should see more small p-values than expected, i.e. the set of p-values should not be uniformly distributed across the interval $[0, 1]$. (The reason that the p-values of streaky teams are small is because the parameter Black is very large for such teams, and large values of Black correspond to small p-values.) Figure 2.24 shows this set of p-values. The set has been sorted. If the set were perfectly uniformly distributed across $[0, 1]$, the graph would be the straight line shown in the figure. One sees that the set of p-values is very close to uniform, and in addition, there are certainly no more small p-values than expected under the Bernoulli model. In fact, we see that there are slightly more teams with large p-values than might be expected. Since large p-values correspond to small values of Black, this says that the number of very consistent teams is slightly more than expected.

### 2.5.2 Basketball

The most widely referenced article on streaks in sports is undoubtedly the paper by Gilovich, Vallone, and Tversky[5]. This article examines whether or not there is auto-correlation in the shooting sequences of basketball players. The authors surveyed fans, collected data from games at both the professional and college level, and carried out a set of experiments with amateur subjects. We will give some of their most intriguing findings here.

In a survey of more than 100 basketball fans, recruited from the student bodies of Cornell and Stanford, 91% believed that a player was more likely to make his next shot if he had made the last two or three shots than if he had missed the last two or three shots. In addition, 68% said the same thing when asked about free throws. The fans were then asked to consider a hypothetical basketball player who has a field goal percentage of 50% (i.e. he makes about one-half of his shots). They were asked to estimate the player's field goal percentage for those shots that occur after he has made a shot, and for those shots that occur after he has missed a shot (these numbers are our $p_1$ and $p_2$). The average value of their estimate for $p_1$ was .61, and for $p_2$ it was .42. In addition, the former estimate was greater than the latter estimate for all of the fans in the survey.

The authors compared these estimates with data obtained from 48 games involving the Philadelphia 76ers in the 1980-81 season. The data consisted of the sequences of hits and misses for all of the field goal attempts for all of the players on the 76ers. Among the more interesting observations is the fact that for eight of the nine major players on the team, the probability that they made a shot was higher after a miss than after a hit. For these players, the weighted mean of their field goal percentages (the average of their probabilities of making a field goal, weighted by the number of attempts) was .52. The weighted mean of their field goal percentages after a miss and a hit were .54 and .51, respectively.

In their paper, the authors performed other statistical tests, including a runs test, a test of stationarity, and a test for stability of shooting percentage across games. None of these test provided any strong evidence for rejecting the Bernoulli model.

They also looked at free throw data from the Boston Celtic teams in the 1980-81 and 1981-82 seasons, and again found no evidence of streakiness. Finally, they recruited members of the Cornell basketball teams (both men and women) to take part in a controlled shooting experiment. Each participant was asked to take 100 shots from a distance at which their accuracy was roughly 50%. They were paid money, where the amount was based both on how accurately they shot and how accurately they predicted their next shot. Once again, there was no evidence suggesting the existence of streakiness.

We now turn to another aspect of basketball in which there is some evidence that fans think there is streakiness. In much of the world, money is bet on almost all aspects of sports. In the United States, for example, one can bet on professional basketball in several ways. One of the most popular ways proceeds as follows. Professional odds-makers (called bookies) set a point spread before the beginning of a game. For example, if the Lakers are playing the Spurs, the bookies may say that the Spurs are favored by 5.5 points. A bettor can bet on either team; if he bets on the Lakers, he is given 5.5 points, meaning that if the Lakers win the game, or if they lose by 5 points or fewer, then the bettor wins. Conversely, if he bets on the Spurs, then the bettor wins only if the Spurs win by at least 6 points.

The bookies' take 10% (called a vigorish, or vig) of the winning bets and all of the losing bets. In many cases, the bookies do not set the point spread at their prediction for the game. The bookies' objective is to have an equal amount of money bet on both teams. To see why they want this, suppose first that \$10,000 is bet on each team. Then the bookies make \$1,000, regardless of which team beats the spread. However, if \$15,000 is bet on the Lakers, and \$5,000 is bet on the Spurs, and the Lakers beat the spread, then the bookies must pay out \$13,500 and only take in \$5,000, so they lose \$8,500.

In a paper that appeared in 1989[4], Colin Camerer showed that the bettors, and hence the bookies, believed in streaky behavior in the NBA (the professional basketball league in the United States). Camerer collected data on all games played in three seasons of the NBA (from 1983 to 1986). His data set consists of the scores of all of the games, and the point spreads set by a popular bookie in Las Vegas.

At the beginning of every game (except for the first game played by a team in a season), each team has a winning streak or a losing streak. Each game is put into a subset depending upon the lengths of these two streaks. For example, if the first team has a three-game winning streak and the second team has a two-game losing streak, then the game is put into the $(+3, -2)$ subset. The longer of the two streaks determines which of the two teams is the 'first' team; if the streaks are of equal length, then a coin flip determines which team is the first team. Thus, each game appears in exactly one subset. For each subset, the fraction of those games in which the first team beat the spread was calculated.

This fraction is very important for the following reason. Camerer's data show, for example, that the subsets of the form $(+4, k)$, with $1 \le k \le 4$, have a combined fraction of .46, this means that 46% of the teams with four-game winning streaks who played teams with shorter streaks managed to beat the spread. This can be taken as evidence that the bettors (and hence the bookies) overvalued the teams with four-game winning streaks. The question of whether such a fraction is significant can be answered using standard statistical methods. We will now show how this is done using the above data.

In order to proceed, we need one more parameter about our data subset, namely its size. There were 159 games in the subsets that we are dealing with. Suppose that we assume the probability that the first team will beat the spread is .5. What is the probability that in 159 trials, the first team will actually beat the spread 46% of the time or less, or equivalently, in at most 73 games? The answer is about .17, which a statistician would not regard as significant.

The data can be pooled by considering all of the subsets $(j, k)$ with $j$ positive (and $j \ge |k|$). There were 1208 games of this type, and the first team beat the spread 47.9% of the time. We can once again ask the question, if a fair coin is flipped 1208 times, what is the probability that we would see no more than 579 $(= .479 \times 1208)$ heads? We can calculate this exactly, using a computer, or we can find an accurate approximation, by recalling that the number of heads $N_H$ is a binomially distributed random variable with mean equal to 604 $(= 1208 \times .5)$ and standard deviation equal to 17.38 $(= \sqrt{1208 \times .5 \times .5})$. Thus, the expression

$$\frac{N_H - 604}{17.38}$$

has, approximately, a standard normal distribution. The value of this expression in the present case is -1.44. The probability that a standard normal distribution takes on a value of -1.44 or less is about .0749, which is thus the p-value of the observation. This observation is therefore significant at the 10% level, but not at the 5% level.

If one instead takes all of the subsets of the form $(j, k)$ with $j \geq 3$ and $j \geq |k|$, there are 698 games, and the first team beat the spread in 318 of these games. This represents 45.6% of the games. What is the p-value of this observation? It turns out to be about .0095, which is a very small p-value.

One can also look at the corresponding pooled data for teams with losing streaks. There were 1140 games in the subsets of the form $(j, k)$, with $j \leq -1$ and $|j| \geq |k|$, and the first team beat the spread in 597 of these games. The p-value of this observation is .055. If we restrict our attention to those subsets for which $j \leq -3$ and $|j| \geq |k|$, there were 643 games, and the first team beat the spread in 340 of these games. The p-value of this observation is .0721.

There are two more collections of subsets that are worth mentioning. The first is the collection of subsets of the form $(j, k)$, with $j$ negative and $k$ positive (and $|j| \geq |k|$). If one does not believe in streaks, then one would think that in this case, bettors would undervalue the first team's chances of beating the spread, since the first team is on a losing streak and the second team is on a winning streak. Thus, the first team should beat the spread more than half the time. There were 670 games in this collection of subsets, and the first team beat the spread in 356 of these games, or 53.1% of the time, leading to a p-value of .0526. Interestingly, in the corresponding collection, in which the first team is on a winning streak and the second team is on a losing streak of equal or smaller length, the first team beat the spread in 358 games, which is 50.1% of the time. The p-value of this observation is, of course, .5.

Although only one of the six p-values reported above is less than .05, five of them are rather small, and thus there is some evidence that the bettors are overvaluing teams with winning streaks and under-valuing those with losing streaks.

If the astute bettor realizes that the average bettor is overvaluing teams with long winning streaks and undervaluing teams with long losing streaks, can he or she make money on this information? In Exercise 1, the reader is asked to show that if there is a 10% vig, then the bettor needs to win 52.4% of the time (assuming he is betting constant amounts) to break even. A few of the pooled subsets mentioned above had winning percentages that were at least 2.4% away from 50% in one direction or the other, meaning that had one made bets, in the correct direction, on all of the games in that pooled subset, one could have made money. Unfortunately, if one tests the hypothesis that the winning percentage in any of the pooled subsets is at least 52.4%, one finds that none of the results are significant at the 5% level.

The results for the 2003-04 season are qualitatively different than those found by Camerer. Through the All-Star break (February 15, 2004), teams having a winning streak, playing teams with equal or shorter streaks, beat the spread in 193 out of 370 games, or 52.2% of the games. Teams having a losing streak, playing teams with equal or shorter streaks, beat the spread in 179 out of 375 games, or 47.7% of the games. Thus, betting on teams with winning streaks, and betting against teams with losing streaks, would have resulted in a winning percentage of 52.2%. Note that this is the opposite of what happened in Camerer's data. Does the reader think that this change will persist, and if so, is it because the bettors have gotten smarter (i.e. have they incorporated the fact that streaks are over-rated into their betting practices)?

## Exercises

**1** Show that if there is a 10% vig, and a bettor makes bets of a constant size, then the bettor needs to win 52.4% of the time to break even.

### 2.5.3 Horseshoes

The game of horseshoes differs in many ways from the games of baseball and basketball. In studying streakiness, some of these differences make it easier to decide whether the results in horseshoes diverge from those that would be predicted under the assumptions of the Bernoulli trials model.

In the game of horseshoes, two contestants face each other in a match. A match consists of an indefinite number of innings. In each inning, one player pitches two shoes, and then the other player pitches two shoes. The shoes are pitched at a stake that is 37 feet from the pitching area. If the shoe encircles the stake, it is called a ringer. A nonringer that is within 6 inches of the stake is called a 'shoe in count.' The former is worth three points and the latter is worth one point. If one player throws $j$ ringers, and the other player throws $k$ ringers, where $j \geq k$ and $j \geq 1$, then the first player gets $3(j-k)$ points, and in this case, shoes in count do not count. If neither player throws a ringer, then the closest shoe in count is worth one point for the player who threw it (if that player threw two shoes in count that are closer than either of his opponent's shoes, that player gets two points). The first player to score 40 points or more is the winner of the match.

Unlike baseball and basketball, the game situation does not affect a player's strategy. In addition, the players typically throw many times in a relatively short time period, and are attempting to do the same thing every time they throw.

Gary Smith has analyzed the data from the 2000 and 2001 Horseshoe World Championships (see [15]). He is particularly interested in whether the data exhibit streakiness. In the cited article, Smith concentrated on doubles (i.e. two ringers thrown by one player in one inning) and non-doubles (i.e. everything else). At the championship level, players throw doubles about half the time. We show some of the data from this paper in Table 2.5.3.

| Group | After 1 Nondouble | After 1 Double |
|---|---|---|
| Men 2000 | .480 | .514 |
| Women 2000 | .501 | .544 |
| Men 2001 | .505 | .587 |
| Women 2001 | .516 | .573 |

Table 2.5.3. Frequency of Doubles Following Non-doubles or Doubles.

Each line of the table represents 16 players. It can be seen that the players were more likely to throw a double following a double than following a non-double. The table gives the average frequencies over sets of players. A breakdown of the data shows that of the 64 players, 25 of the men and 26 of the women were more likely to throw a double following a double than following a non-double (we will refer to this situation as positive auto-correlation, as before). That this is evidence of streakiness can be seen as follows. If players were not affected by their throws in the preceding inning, then about half of them would have positive auto-correlation. Continuing under the assumption of independence between innings, we see that the probability of observing as many as 51 of the 64 players with positive auto-correlation is the same as the probability that if a fair coin is flipped 64 times, it comes up heads at least 51 times. This probability is approximately .0000009.

At the championship level, the players' probabilities of throwing a double in any given inning is so high that using either the length of the longest run of doubles or the number of runs of doubles and non-doubles cannot be used to reject the null hypothesis of Bernoulli trials at the 5% level. For example, one of the players in the 2000 championships pitched 13 doubles in 14 innings. This means that there were either two or three runs of either type. But under the null hypothesis of Bernoulli trials, using the value $p = 13/14$, there is a probability of ,143 of two runs (and a probability of .857 of three runs), so in this case, even if there are two runs, the p-value is much larger than .05.

Smith gets around this problem by calculating, under the null hypothesis, the expected number of runs by a given player in a given game, and then tabulating the number of games in which the actual number of runs was above or below the expected number. Fewer runs than expected mean streakiness. Table 2.2 shows the number of games with fewer or more runs than expected for each championship, together with the p-values for the observations, under the null hypothesis.

| Group | Fewer Runs | More Runs | p-value |
|---|---|---|---|
| Men 2000 | 129 | 107 | 0.0857 |
| Women 2000 | 136 | 103 | 0.0191 |
| Men 2001 | 137 | 98 | 0.0065 |
| Women 2001 | 138 | 99 | 0.0067 |

Table 2.5.3. Number of Games with Fewer or More Runs than Expected.

### 2.5.4 Tennis

The game of tennis is interesting in probability theory because it provides an example of a nested set of Markov chains. The reader will recall that, roughly speaking, a Markov chain is a process in which there is a set of states and a transition matrix whose entries give the probabilities of moving from any state to any other state in one step. The chain can either be started in a specific state, or it can be started with a certain initial distribution among the states. We will describe the various Markov chains that make up a tennis match, and then give some results about tennis that follow from elementary Markov chain theory. We will then look at whether or not tennis is streaky at the professional level.

A tennis match is divided into sets; in most cases, the first person to win two sets is the winner of the match. (There are a few professional tournaments in which the winner is the first person to win three sets.) The set scores in an on-going tennis match can be thought of as labels in a Markov chain. The possible scores, from the point of view of one player are 0-0 (at the beginning of the match), 1-0, 0-1, 1-1, 2-0, 2-1, 1-2, and 0-2. The last four of these states are said to be absorbing states, because once the match enters one of these states, it never leaves the state. The other four states are called transient states, because if the match is in one of these states at a given time, it will leave at the next step. (In a general absorbing Markov chain, a state is called transient if it possible for the chain to go from that state to an absorbing state in a finite number of steps. It is not necessary that the chain always leave the transient state in one step.)

Suppose that player $A$ has probability $p$ of winning a set against player $B$. Then, for example, the probability that the Markov chain will move from state 0-1 to state 1-1 is $p$, while the probability that it will move from state 1-1 to state 1-2 is $1-p$. The reader can check that if the above states are numbered from 1 to 8, and we denote by $p_{ij}$ the probability of moving from state $i$ to state $j$ in one step, then the transition matrix $\mathbf{P} = (p_{ij})$ is given by

$$\mathbf{P} = \begin{array}{c} \\ \text{0-0} \\ \text{1-0} \\ \text{0-1} \\ \text{1-1} \\ \text{2-0} \\ \text{2-1} \\ \text{1-2} \\ \text{0-2} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{0-0} & \text{1-0} & \text{0-1} & \text{1-1} & \text{2-0} & \text{2-1} & \text{1-2} & \text{0-2} \end{array} \\ \begin{pmatrix} 0 & p & 1-p & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-p & p & 0 & 0 & 0 \\ 0 & 0 & 0 & p & 0 & 0 & 0 & 1-p \\ 0 & 0 & 0 & 0 & 0 & p & 1-p & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

It turns out that we will want to split most of the above states into two states in our final model, because it is thought that the serve in tennis has a large effect on who wins a given game; thus, it will be important for us to record who is serving at the beginning of each set. We will discuss this in more detail below.

Each set in tennis consists of a number of games. The first player to win six games, if his or her opponent has won at most four games, wins the set. If the score is 6-5, the set continues until it is either 7-5, in which case the player with seven games wins, or it is 6-6, in which case a tie-breaker is played. Whoever wins the tiebreaker wins the set. Thus a set can be modeled by an absorbing Markov chain.

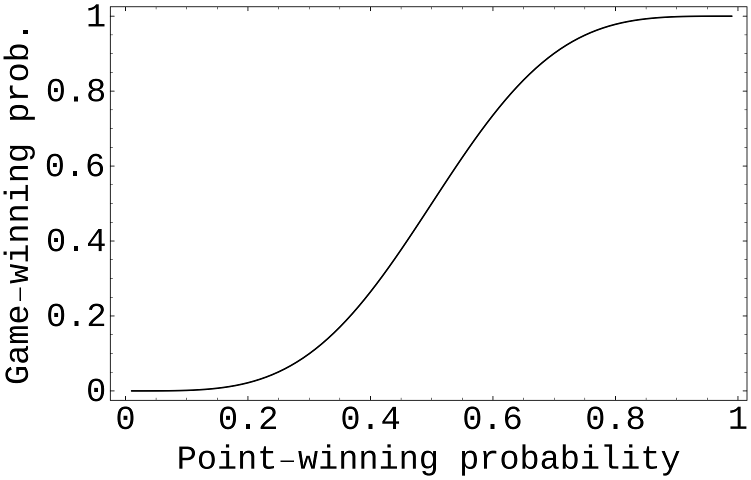


Figure 2.25: Game-winning vs. point-winning probabilities in tennis.

Both a regular game and a tie-breaker are themselves absorbing Markov chains. In a game, the first person to win at least four points and be at least two points ahead wins the game. In a tie-breaker, the first person to win at least seven points and be at least two points ahead wins the tie-breaker.

At the professional level, the person who is serving a point has a much greater than even chance of winning the point. In the authors' opinion, this advantage may not extend down to their level. Since our data comes from professional matches, we need to take account of the serve. In each non-tie-break game, one person serves all of the points. The service rotation in a tie-break is more complicated; one player starts the tie-break by serving one point, and then the players take turns serving two consecutive points.

We show, in Figure 2.25, the probability that a player wins a game that he is serving, if he has a probability $p$ of winning any given point. This graph shows that if $p = .6$, then the player will win the game with probability .74, and if $p = .8$, then the player will win the game with probability .98.

Now suppose that two players are playing a best-of-three set match in which the first player has a probability of $p$ of winning any given point (so in this case, we are assuming that the serve does not affect the outcome of the point). In Figure 2.26, we show the probability of the first player winning the match, as a function of $p$. Note that even if $p = .51$, the probability that the first player wins the match is .748.

How are such probabilities determined? We will briefly describe the calculations that are needed, and the theorems on which these calculations are based. For more examples, and for proofs of the theorems, the reader is referred to [7].

If $\mathbf{P}$ is the transition matrix of an absorbing Markov chain, then we can relabel the states so that the first set of states are the transient ones, and the last set of states are the absorbing ones. If we do so, the matrix $\mathbf{P}$ assumes the following

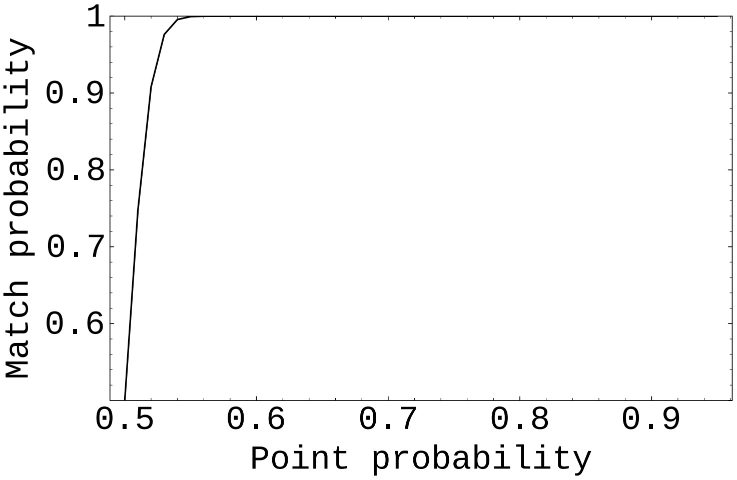


Figure 2.26: Match-winning vs. point-winning probabilities in tennis.

canonical form:

$$\mathbf{P} = \begin{array}{cc} & \begin{array}{cc} \text{TR.} & \text{ABS.} \end{array} \\ \begin{array}{c} \text{TR.} \\ \\ \text{ABS.} \end{array} & \left(\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right) \end{array}$$

The four expressions $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{0}$, and $\mathbf{I}$ are rectangular submatrices of $\mathbf{P}$. If there are $t$ transient states and $r$ absorbing states, then $\mathbf{I}$, for example, is an $r$-by-$r$ matrix. The reason that it is denoted by $\mathbf{I}$ is that it is an identity matrix, since the probability of moving from one absorbing state to a different absorbing state is 0, while the probability of remaining in an absorbing state is 1.

If $\mathbf{P}$ denotes the transition matrix for the set scores in tennis (which was given above), then the matrices $\mathbf{Q}$ and $\mathbf{R}$ are as follows:

$$\mathbf{Q} = \begin{pmatrix} 0 & p & 1-p & 0 \\ 0 & 0 & 0 & 1-p \\ 0 & 0 & 0 & p \\ 0 & 0 & 0 & 0 \end{pmatrix} ,$$

$$\mathbf{R} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ p & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-p \\ 0 & p & 1-p & 0 \end{pmatrix} .$$

Given a transition matrix $\mathbf{P}$, in canonical form, for an absorbing Markov chain, assume that there are $t$ transient states and $r$ absorbing states. The matrix $\mathbf{N}$, defined by the equation

$$\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1} ,$$

is called the fundamental matrix for the chain. The following theorem shows one reason that $\mathbf{N}$ is useful.

**Theorem 2.1** Let $b_{ij}$ be the probability that an absorbing chain will be absorbed in the $j$'th absorbing state if it starts in the $i$'th transient state. Let $\mathbf{B}$ be the matrix with entries $b_{ij}$. Then $\mathbf{B}$ is an $t$-by-$r$ matrix, and

$$\mathbf{B} = \mathbf{NR} ,$$

where $\mathbf{N}$ is the fundamental matrix and $\mathbf{R}$ is as in the canonical form. □

As an example of how this theorem is used, suppose that the first player has probability $p = .6$ of winning a given set against his opponent. Then

$$\mathbf{Q} = \begin{pmatrix} 0 & .6 & .4 & 0 \\ 0 & 0 & 0 & .4 \\ 0 & 0 & 0 & .6 \\ 0 & 0 & 0 & 0 \end{pmatrix} ,$$

so one can calculate that

$$\mathbf{N} = \begin{pmatrix} 1 & .6 & .4 & .48 \\ 0 & 1 & 0 & .4 \\ 0 & 0 & 1 & .6 \\ 0 & 0 & 0 & 1 \end{pmatrix} .$$

Thus, the matrix $\mathbf{B} = \mathbf{NR}$ is given by

$$\mathbf{B} = \begin{pmatrix} .36 & .288 & .192 & .16 \\ .6 & .24 & .16 & 0 \\ 0 & .36 & .24 & .4 \\ 0 & .6 & .4 & 0 \end{pmatrix} .$$

The first row of this matrix is of particular interest, since it contains the probabilities of ending in each of the absorbing states, if the chain starts in the state 0-0 (i.e. the match starts with no score). We see that with the given value of $p$, the probability that the first player wins both of the first two sets is .36, and the probability that he wins the match is $.36 + .288 = .648$.

There are 43 states in the Markov chain representing a set of tennis. There are four absorbing states, with two corresponding to a win by the first player and two to a win by the second player. The reason that we need two winning states for each player is that we need to keep track of who serves first in the subsequent set (if there is one). The transition probabilities are given by the values in Figure 2.25; for example, if the first player is serving, the game score is 2-1, and he has a probability of .6 of winning a given point, then the game score will become 3-1 with probability .74.

Once again, we are interested in the probability that the player who serves first wins the set. This time, there are two parameters, namely the probabilities that each of the players wins a given point when they are serving. We denote these two

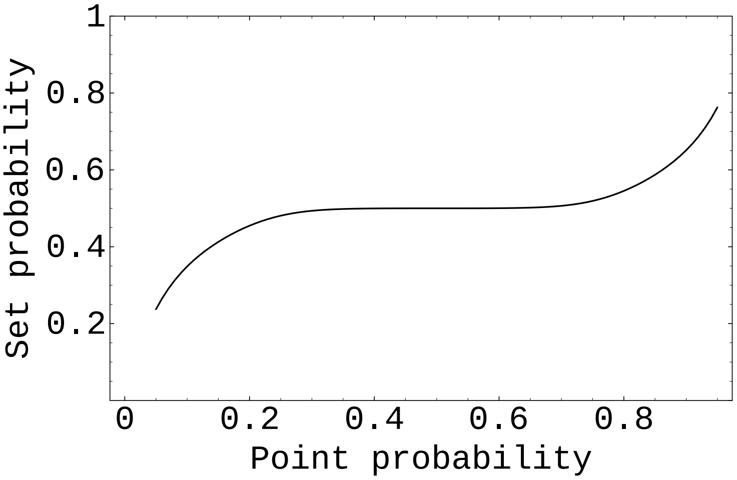


Figure 2.27: Set-winning probabilities for players of equal abilities.

parameters by $p_1$ and $p_2$. If we let $p_1 = p_2$, then Figure 2.27 shows the probability that the first player wins the set as a function of $p_1$.

There is nothing very remarkable about this graph. Suppose instead that $p_1 = p_2 + .1$, i.e. the first player is somewhat better than the second player. In this case, Figure 2.28 shows the probability that the first player wins the set as a function of $p_1$ (given that the first player serves the first game). If, for example, $p_1 = .55$ and $p_2 = .45$, then the probability that the first player wins the set is .65.

Finally, suppose that two players play a best-of-three set match. The Markov chain corresponding to this situation has 11 states, because we must keep track of who begins serving each set. Suppose that the first player has probability $p$ of winning a given point on his serve, and his opponent has probability $p - .1$ of winning a given point on his serve. Figure 2.29 shows the probability that the first player will win the match as a function of $p$. Note that even though there is a dip in the graph, the probability that the first player wins the match is always at least .9, even though the probability that he wins a given point is only slightly greater than his opponent's probability. What kind of streakiness, if any, is evident in professional tennis? The above models show that even with slight differences in the players' abilities, many of the matches are likely to be one-sided. One can turn this around and say that the fact that there are so many close matches among the top professional tennis players means that these players must be very close in ability.

In [8], Jackson and Mosurski describe two models, and some variations on these models, that deal with match scores in tennis. They were interested in the apparent overabundance of 'heavy defeats' in tennis matches, i.e. best-of-three matches in which the loser does not win a set, or best-of-five matches in which the loser wins 0 or 1 sets.

One can model the sequence of sets in a tennis match in several ways. The simplest model is a Bernoulli trials model, in which a given player has the same

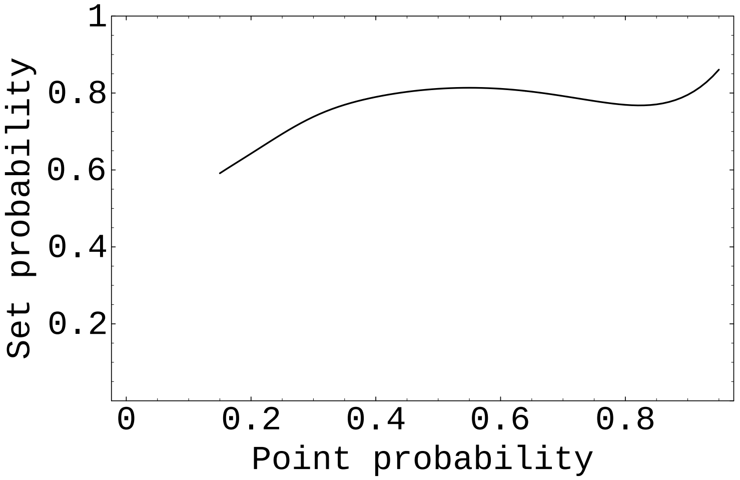


Figure 2.28: Set-winning probabilities for players of unequal abilities.

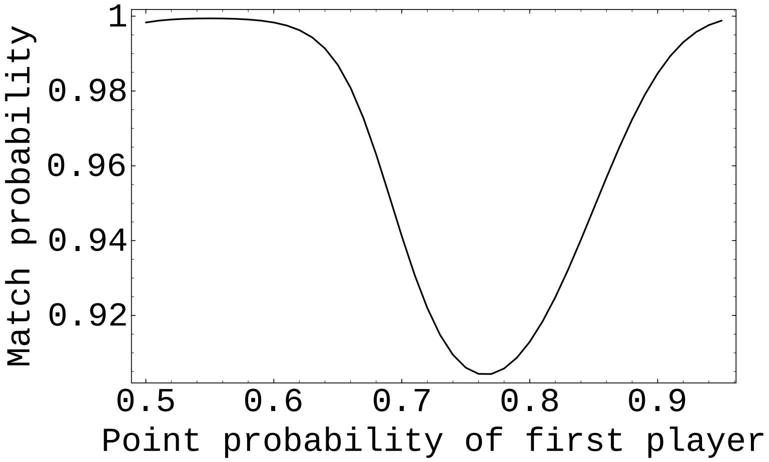


Figure 2.29: Match-winning probabilities for players of unequal abilities.

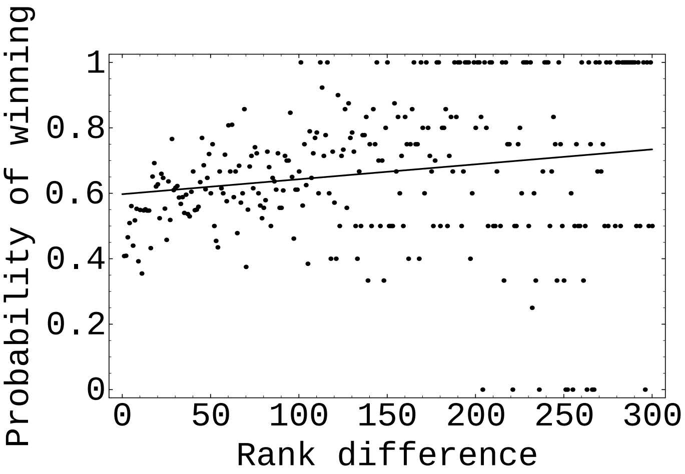


Figure 2.30: Observed winning probabilities vs. rank difference in 2002.

probability $p$ of winning each set. In Exercise 1, we ask the reader to determine the distribution of match scores under this assumption. In order to see if this model does a good job of explaining the data, we need to have some way of estimating $p$, since we certainly do not assume, even among professional tennis players, that $p = .5$.

A reasonable way to proceed is to use the rankings of the players to estimate $p$. Tennis rankings of professional players are determined by the players' performance in the previous twelve months. These rankings are updated each week. They are not perfect because, for example, if a very good player is injured and doesn't play for awhile, his ranking can change by quite a bit, even though once he has recovered, he is presumably as good as he was before the injury.

We have obtained data from the website www.stevegtennis.com. We very much appreciate the work that was done to create the raw data files that we used. Our data set consists of all men's singles matches on the professional tour between 2000 and 2002.

There are at least two ways to use player rankings to estimate $p$. One idea is to fit the observed probabilities to the difference of the two players' ranks. A plot of the observed probabilities for the first sets in mens' singles matches in 2002 versus the difference in the ranks of the two players is shown in Figure 2.30. We only show the results of matches where the rank difference was at most 300, since there are comparatively few matches for each difference larger than this. There are 3925 matches in the resulting data set. This figure also shows a best fit line. The reader can see that the fit is not very good; in fact, the correlation coefficient is .0298.

Jackson and Mosurski adopt a different strategy for estimating $p$. Denote by $r$ and $s$ the two players' ranks, and assume that $r \leq s$. They then define $O(r, s)$ to be the odds (not the probability) that the better player wins the first set. The reason that they define these odds in terms of the first set, rather than the match,

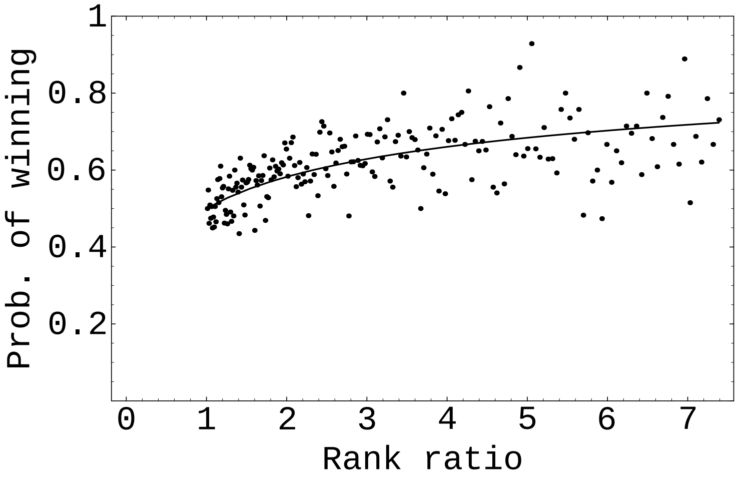


Figure 2.31: Observed winning probabilities vs. rank ratio in 2000-2002.

is that they are concerned about the outcomes of sets affecting the outcomes of later sets in the same match. If we are considering a model in which the sets are considered to be independent events, then we can use $O(r, s)$ as the odds that the better player wins any given set in the match. However, we must be careful with our use of $O(r, s)$ if we are not assuming the sets are independent.

They next state the following assumed relationship among $O(r, s)$, $r$ and $s$:

$$O(r, s) = (\text{ratio of ranks})^\alpha = \left(\frac{s}{r}\right)^\alpha .$$

The parameter $\alpha$ is to be determined by the data. The above relation is equivalent to the relation

$$\log(O(r, s)) = \log(\text{odds of success}) = \alpha \log(s/r) .$$

This is a line through the origin with slope $\alpha$.

We carry out this regression for all matches in the years 2000 through 2002 in which the ratio of the ranks of the players is at most 7.4 (this is a completely arbitrary cut-off; it is about $e^2$). There were 12608 matches in this data set. We obtain a value of $\alpha = .480$, and a correlation coefficient of .363. Figure 2.31 shows the observed probability of winning the match versus the ratio of the ranks, together with the graph of the function $(\text{ratio})^\alpha$.

Now that we have a way to estimate the probability that a given player will win the first set against a certain opponent, we can proceed in several directions. First, we can explain and test the model of Jackson and Mosurski, which they call the odds model. Second, we can test for independence among the sets in professional tennis matches.

In the odds model, it is assumed that the odds the better player will win a given set depend upon the set score of the match. We let $O_{ij}$ denote the odds that the better will player will win the next set if the set score is $i$ (for the better player) to $j$. The discussion above gives us an estimate for $O_{00}$. The model asserts that each time a player wins a set, the odds for the next set change (in one direction or the other) by a factor of $k$, where $k$ is to be determined by the data. So, for example, if the better player wins the first set, then the odds that he wins the second set equal $kO_{00}$. In general, we have

$$O_{ij} = k^{i-j} O_{00} \ .$$

Using our estimate for $O_{00}$, and taking logarithms, we have

$$\log(O_{ij}) = \alpha \log(s/r) + (i-j)\log(k) \ .$$

Note that if we take $k = 1$, then we have the model where the sets are assumed to be independent.

The above equation can be used, along with the data, to estimate $\alpha$ and $k$. The procedure we now describe results in the maximum likelihood estimates for $\alpha$ and $k$. In a nutshell, a maximum likelihood estimate for a set of parameters is the set of values for those parameters that leads to the highest probability for the data set. As an example of this idea, suppose that we flip a coin 20 times and observe 12 heads. We wish to find the maximum likelihood estimate for $p$, the probability of a head on a single toss. We certainly hope that this estimate is 12/20. Let us see if this is the case. For each $p$ between 0 and 1, we compute the probability that if a coin has probability $p$ of coming up heads, it will come up heads 12 times in 20 tosses. This probability is

$$\binom{20}{12} p^{12}(1-p)^8 \ .$$

We wish to maximize this expression over all $p$. If we denote this expression by $f(p)$, then we have

$$f'(p) = \binom{20}{12}\biggl(12p^{11}(1-p)^8 - 8p^{12}(1-p)^7\biggr) \ .$$

Setting this equal to 0 and solving for $p$, we obtain

$$p = \frac{12}{20} \ ,$$

as we hoped. (It is easy to check that this value of $p$ corresponds to a maximum value of $f(p)$.) To reiterate, this means that the probability that we would actually obtain 12 heads in 20 tosses is largest if $p = 12/20$.

In the case of set scores in tennis, we treat each match as an independent event, and use the relationship given above among the odds that the better player wins a given set, $\alpha$, and $k$, to compute the probability that we would obtain the actual data set. Since the matches are assumed to be independent, this probability is the product of the probability that each match in the data set occurs, given values of $\alpha$ and $k$. This calculation is easy to carry out with a computer. The number of completed best-of-three set matches in our data set is 12608. The maximum likelihood estimates for the parameters are

$$\alpha = .37$$

and

$$k = 1.78 \ .$$

Note that this value of $\alpha$ is quite a bit different than the one we obtained using just the first sets of the matches.

Now that we have estimated $\alpha$ and $k$, we can proceed, as in [8], to see how well the model fits the observed distribution of set scores in our data. The way that we do this is to simulate the matches many time on a computer, using the actual pairs of rankings in each match in the data set. When we did this 100 times, the distribution of match results was

$$\{5271.77, 1196.99, 881.01, 941.8, 1317.93, 2998.5\} \ ,$$

where the $i$'th number in the above list is the number of matches whose set outcomes (from the point of view of the better player) matched the $i$'th element in the following list:

$$\{\{1,1\},\{1,0,1\},\{1,0,0\},\{0,1,0\},\{0,1,1\},\{0,0\}\} \ .$$

This simulation gives us a decent approximation to the theoretical distribution. To see how well the model fits the actual distribution, we use a chi-squared test. The actual distribution of the data is

$$\{5453, 1157, 993, 813, 1338, 2854\} \ .$$

If we let $E_i$ denote the $i$'th value in the theoretical distribution, and $O_i$ denote the $i$'th value in the actual distribution, then value of the expression

$$\sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

is approximately chi-squared distributed with 5 degrees of freedom (see [7] for examples of this technique). The value of the above expression is 46.69, which has a p-value of less than .01. So the model does not fit the data very well.

We can also carry out the above calculations under the assumption of independence of sets. To do this, we simply assume that the odds, $O_{00}$, of the better player winning the first set continue to hold for all subsequent sets in the same match. This corresponds to letting $k = 1$. However, we must re-estimate $\alpha$, since our maximum likelihood estimate of $\alpha$ was found simultaneously with our maximum likelihood estimate of $k$. When we do this, we find a new estimate: $\alpha = .41$. Using this value, and simulating the resulting distribution, we obtain the following approximation:

$$\{4459.56, 1742.79, 1235.17, 1241.94, 1746.56, 2181.98\} \ .$$

The chi-squared value in this case is 916.374, which is much larger than in the previous case. So we might conclude that the odds model fits the data better than the independent set model, but neither fits the data well.

Before moving to the best-of-five set data, we digress slightly to suggest another way to estimate $\alpha$ and $k$. This method involves minimizing the value of the chi-squared distribution. More precisely, we try to find the values of $\alpha$ and $k$ so that when the theoretical values of the distribution are compared with the actual values, the chi-squared value is as small as possible. This method is due to Karl Pearson, who was the originator of the chi-squared method of comparing distributions.

The values of $\alpha$ and $k$ that minimize the value of the chi-squared distribution for the best-of-three set matches are .41 and 1.75, which we note are very close to the maximum likelihood estimates for these parameters. For these values of $\alpha$ and $k$, the p-value of the observation is still less than .01, once again signifying that the fit is not good.

When we repeat the maximum likelihood calculations for all of the best-of-five set matches played between 2000 and 2002, the fit of the odds model is much better. There were 1571 completed best-of-five set matches in the data set. The maximum likelihood estimates for $\alpha$ and $k$ are .31 and 1.51. The chi-squared value of the observation is 17.54, with a p-value of .55. We use a chi-squared distribution with 19 degrees of freedom, since there are 20 possible outcomes in a best-of-five set match.

If we attempt to fit the independent set model to the data, we find a maximum likelihood estimate for $\alpha$ of .36, and a chi-squared value of 235.2. The p-value for this observation is less than .01. Thus the odds model fits the data fairly well, and does a much better job than the independent set model.

There is another way to help us decide whether professional tennis matches are streaky. Consider a best-of-three set match. If each player has won one set in the match, then we might think that they are equally likely to win the third, and deciding, set. Of the 12608 completed best-of-three set matches in our data set, 4301 required three sets to complete. Of these 4301 matches, 2331 matches were won by the player who won the second set. Thus, among the matches that took three sets to complete, if a player won the second set, the probability that he won the third set is .542. If the sets were independent, we might think that this probability should be close to .5. Since the actual probability is greater than .5, one could say that the player who won the second set has 'momentum', or is 'on a streak.' Before concluding that this is evidence of streaky behavior, we should consider the relative abilities of the players involved in the matches. It is possible that the reason that the winner of the second set does so well in the third set is because he is the better player.

If we simulate a distribution of set scores, using the value of $\alpha = .480$ (obtained earlier by fitting the odds model to just the first sets in the matches in the data), we obtain the following (the average of 20 simulations, using the same sets of opponents as in the actual data set):

$$\{4676.65, 1740.05, 1187.75, 1180.4, 1746.85, 2076.3\}\ .$$

In this simulated distribution, there are 5855 matches that took three sets to complete. In these matches, there were 2935 (which is almost exactly one-half of 5855) in which the player who won the second set won the third. Thus, we can discount any effect due to the relative ranks of the players.

It is also the case that this percentage does not change very much if we vary $\alpha$. For $\alpha = .3, .31, \ldots, .5$, the percentage stays between .497 and .505. Thus, we may assume that in this model, the player who won the second set has about a 50% chance of winning the third set.

How likely is it that in as many as 2331 matches out of 4301, the player who wins the second set wins the third set as well, given our assumption about independence of sets? This is essentially the same question as asking how likely a fair coin, if tossed 4301 times, will come up heads at least 2331 times. The number of heads in a long sequence of coin flips is approximately normal. If the coin is fair, and the number of tosses is $n$, then the mean is $n/2$ and the standard deviation is $\sqrt{n}/2$. The standardized value corresponding to 2331 is

$$\frac{2331 - n/2}{\sqrt{n}/2} = 5.505 .$$

The probability that a standard normal random variable takes on a value greater than 5.505 is much less than .01. Thus, we can claim that the data exhibits streakiness.

### Exercises

**1** Assume that two players are playing a best-of-three set tennis match, and the first player has probability $p$ of winning each set (i.e. the sets are independent trials). Find the distribution of the four possible match outcomes: 2-0, 2-1, 1-2, and 0-2.

## 2.6 Adaptive Thinking

## 2.7 Runs in the Stock Market

It is accepted that over the long run, most stocks go up in price. Thus, one way to model the price of a stock is to imagine that there is a line, with positive slope, that represents the long-term trend of the stock, and then consider the stock's variation about this line. It is typically the case that instead of using the daily prices of the stock, one uses the logarithms of these prices. In this case, a straight trend line corresponds to the stock price changing by a constant percentage each day. The slope of the trend line can be found by either fitting a straight line to the log price data, or simply by using the starting and ending points of the data as anchor points for the line. In either case, the slope cannot be found unless one knows the data.

As might be imagined, an incredible amount of effort has been directed to the problem of modeling stock prices. In Chapter 4, we will focus on what is known about the distribution of the residual movements (those that remain after the trend line has been subtracted from the data) of stock prices. In this section, we will consider whether the stock market exhibits streaky behavior (or perhaps other forms of non-randomness).

One obvious way in which stock prices could be streaky concerns their daily up-and-down motions. Most stocks have more up days than down days, so one might model them with coins whose success probabilities are greater than .5. Using the data, one can compute the observed success probability, and then count the number of success (or failure) streaks of a given length, or of any length. Then one can compare these observed values with the theoretical values.

Our data set consists of daily prices of 439 of the stocks that make up the S&P 500 list. These prices have been adjusted to take into account dividends and stock splits. For this reason, many of the stock prices are very small near the beginning of the data set, and hence are subject to rounding errors. We typically get around this problem by using only the part of the data set in which a stock's price is above some cutoff value. Also, we throw out all days for a given stock on which the stock price was unchanged.

Suppose that we want to compare the expected and the observed number of pairs of consecutive days in which a given stock's price went up on both days. If we have $n$ data points, then we can define the random variable $X_i$ to equal 1 if the $i$'th and $(i+1)$'st price changes are both positive, and 0 otherwise. Under the assumption that the signs of the daily price changes are independent events, the probability that $X_i = 1$ is just $p^2$, where $p$ is the observed long-range probability of success for that stock. Thus, it is easy to calculate the expected number of up-up pairs. There are $n-1$ daily changes (since there are $n$ data points), so there are $n-2$ random variables $X_i$, each with the same distribution. Thus, the expected number of up-up pairs is just

$$(n-2)p^2\ .$$

Unfortunately, the $X_i$'s are not mutually independent. For example, if $X_6 = 1$ and $X_8 = 1$, then it is not possible for $X_7$ to equal 0. Nonetheless, the $X_i$'s are $m$-independent, for $m = 2$. This concept means that it is possible to partition the sequence $\{X_i\}$ into $m$ subsets such that the random variables in each subset are mutually independent. In this case, we can use the partition $\{X_1, X_3, \ldots\}$ and $\{X_2, X_4, \ldots\}$. If a sequence of random variables is $m$-independent, then it satisfies a modified version of the Central Limit Theorem. (The modification comes in how the variance of the sum of the random variables is calculated.)

Using this modified version of the Central Limit Theorem, we can transform the sum $X_1 + X_2 + \ldots + X_{n-1}$ into a random variable that is approximately standard normal (this is done in the usual way; we subtract the mean and divide by the standard deviation). This gives us, for each stock, a $z$-value, i.e. a value of a standard normal distribution that represents how far above or below the mean the observed number of up-up pairs is. This approach works for any other pattern, such as down-down, or up-down-up, etc.

If stocks are streaky, then the $z$-values for up-up pairs (and for down-down pairs) should be significantly greater than 0. The same thing should be true for up-up-up triples. For each of several different patterns, we calculated the set of $z$-values for all 439 stocks in our data set. We used 1 as our cutoff value, meaning that for each stock, we only used those log prices after the last time the log price failed to exceed 1. The average number of log prices per stock that this gives us is 3861, or almost 15 years' worth of data.

At this point, we have 439 $z$-values. If these were drawn from a standard normal distribution, the distribution of their average would have mean 0 and standard deviation $1/\sqrt{439} \approx .0477$. The table below shows, for various patterns (up-up is denoted by UU, for example), the average $z$-value over our set of stocks. For each pattern, the distance of the average from 0, in units of standard deviation, and the percentage of stocks whose z-value is positive, are given. If the price movements are independent, then one would expect about half of the z-values to be positive.

| Pattern | Average $z$-Value | Standard Deviation Units | Percent Positive |
|---|---|---|---|
| UU | -0.017 | -0.35 | 49.7 |
| DD | -0.016 | -0.34 | 50.1 |
| UD | 0.066 | 1.37 | 50.8 |
| DU | 0.079 | 1.65 | 51.0 |
| UUU | -0.192 | -4.03 | 43.3 |
| UUUU | -0.335 | -7.02 | 37.6 |
| UUUUU | -0.461 | -9.67 | 33.0 |
| DDD | -0.280 | -5.86 | 39.6 |
| DDDD | -0.485 | -10.16 | 35.3 |
| DDDDD | -0.621 | -13.02 | 27.6 |

Table 2.7. $z$-Values for Various Patterns in Daily Price Changes.

We see that these stocks, in general, are anti-streaky. The numbers of UU and DD streaks are not significantly less than what would be expected, but the numbers of streaks of length three and four of both types are significantly smaller than expected.

The next table shows the results of similar calculations involving weekly prices. Specifically, the closing price at the end of the first day of each week of trading was used. Once again, our set of stocks show strong evidence of being anti-streaky.

| Pattern | Average $z$-Value | Standard Deviation Units | Percent Positive |
|---|---|---|---|
| UU | -0.452 | -9.46 | 18.0 |
| DD | -0.520 | -10.90 | 17.5 |
| UD | 1.070 | 22.42 | 81.1 |
| DU | 1.064 | 22.30 | 81.5 |
| UUU | -0.587 | -12.29 | 19.6 |
| UUUU | -0.577 | -12.09 | 21.9 |
| DDD | -0.606 | -12.70 | 21.0 |
| DDDD | -0.579 | -12.14 | 21.9 |

Table 2.7. $z$-Values for Various Patterns in Weekly Price Changes.

The results in these tables should be compared with simulated data from the model in which weekly changes for a given stock are mutually independent events. For each stock in our set, we used the observed probabilities of a positive and a negative weekly change in the stock price to create a simulated set of stock prices. Then, using the same algorithms as were used above, we calculated the average z-values for various patterns over our set of stocks. The results are shown in the table below.

| Pattern | Average $z$-Value | Standard Deviation Units | Percent Positive |
|---|---|---|---|
| UU | -0.010 | -0.21 | 50.1 |
| DD | -0.074 | 1.56 | 51.7 |
| UD | 0.005 | 0.10 | 51.0 |
| DU | 0.002 | 0.04 | 51.0 |
| UUU | -0.035 | -0.74 | 44.9 |
| UUUU | -0.020 | -0.41 | 46.2 |
| DDD | 0.064 | 1.34 | 50.8 |
| DDDD | -0.060 | -1.26 | 43.7 |

Table 2.7. Simulated $z$-Values for Various Patterns in Weekly Price Changes.

The simulated data from this model is much different than the actual data. This supports the observation that both daily and weekly stock prices are anti-streaky.

In their book[11], Andrew Lo and Craig MacKinlay discuss a parameter they call the variance ratio. Suppose that $\{X_i\}_{i=0}^n$ is a sequence of $n$ logarithms of a stock's price. The time increment between successive values might be days, or weeks, or even something as small as minutes. The log price increments are the values $\{X_{i+1} - X_i\}_{i=0}^{n-1}$. A central question that concerns this sequence of log price increments is whether it can be modeled well by a process in which the increments are assumed to be mutually independent.

Lo and MacKinlay begin by fixing a stock and a sequence of log prices $\{X_i\}_{i=0}^n$ for that stock. All prices have been adjusted for splits and dividends, and we assume in what follows that the time increment is a week. Next, they calculate the estimator

$$\hat{\mu} = \frac{1}{n}\sum_{i=0}^{n-1}\bigl(X_{i+1} - X_i\bigr) ,$$

which is an estimate of the average change per week in the logarithm of the stock price. This expression can be simplified to

$$\hat{\mu} = \frac{1}{n}\bigl(X_n - X_0\bigr) .$$

Next, they calculate the estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=0}^{n-1}\bigl(X_{i+1} - X_i - \hat{\mu}\bigr)^2 ,$$

which is an estimator of the variance of the weekly change in the logarithm of the stock price. If we assume for the moment that $n$ is even, we could instead look at the estimator

$$\frac{1}{n/2}\sum_{i=0}^{n-1}\bigl(X_{2i+2}-X_{2i}-2\hat{\mu}\bigr)^2\ ,$$

which is an estimator of the variance of the increments in even-numbered observations. Under the assumption that the increments are independent, the variance of the differences

$$\{X_{2i+2}-X_{2i}\}$$

is twice the variance $\sigma^2$ of the differences

$$\{X_{i+1}-X_i\}\ .$$

This provides a way to test whether the assumption of independent increments in stock prices is a reasonable one. One can compute both estimators for a given stock, and see how close the ratio is to two.

Lo and MacKinlay slightly change the second variance estimator above, by dividing by two. So let us define

$$\hat{\sigma_2}^2=\frac{1}{n}\sum_{i=0}^{(n-1)/2}\bigl(X_{2i+2}-X_{2i}-2\hat{\mu}\bigr)^2\ .$$

Under the assumption of independent increments, the theoretical value of $\sigma_2^2$ (for which $\hat{\sigma_2}^2$ is an estimator) equals the value of $\sigma^2$. Thus, still under this assumption, the ratio of the estimators should be close to 1. Of course, one can, for any integer $q>1$, define the estimator $\hat{\sigma_q}^2$ in the same way. Once again, under the assumption of independent increments, $\sigma_q^2=\sigma^2$.

Lo and MacKinlay modify the above set-up in one additional way. Instead of using non-overlapping time increments in the definition of $\hat{\sigma_q}^2$, they use all of the time increments of length $q$, obtaining the following definition (note that we no longer need to assume that $q$ divides $n$):

$$\hat{\sigma_q}^2=\frac{1}{q(n-q+1)}\sum_{i=0}^{(n-q+1)}\bigl(X_{i+q-1}-X_i-q\hat{\mu}\bigr)^2\ .$$

The test statistic for the variance ratio $\hat{\sigma_q}^2/\hat{\sigma}^2$ is denoted by $M_r(q)$ (the subscript refers to the fact that we are dealing with a ratio), and is defined by

$$M_r(q)=\frac{\hat{\sigma_q}^2}{\hat{\sigma}^2}-1\ .$$

We have seen that under the assumption of independent increments, this statistic should typically be close to 0. In order for it to be a useful statistic in testing this assumption, we need to know something about how this statistic is distributed. Lo and MacKinlay show that for large $n$, the statistic $\sqrt{n}M_r(q)$ is approximately normally distributed with mean 0 and variance

$$\frac{2(2q-1)(q-1)}{3q}\ .$$

(In fact, to be strictly accurate, they make one further modification, that we will ignore, in order that the estimators are unbiased. The above asymptotic statement does not change when this modification is made.)

Lo and MacKinlay calculated the variance ratio for both individual stocks and for various sets of stocks on the New York and American stock exchanges. Their data consisted of weekly prices for 625 stocks from 1962 to 1985 (1216 weeks). For the values $q = 2, 4, 8$, and 16, their observed values of $M_r(q)$ for the set of all stocks in their data set were $.30, .64, .94$, and 1.05 respectively. These were all statistically different from 0 at the 5% level of significance.

They also computed the average variance ratio for the individual stocks. For the above values of $q$, these average variance ratios were $-.03, -.06, -.08$, and $-.11$. These observed values were not statistically different from 0 at the 5% level of significance. However, it is interesting that these observed values are all of opposite sign from those corresponding to the set of all stocks.

It is possible to describe a situation in which the variance ratio of a stock (for $q = 2$, say) would be negative. Suppose that over blocks of two consecutive weeks, there were more up-down and down-up pairs than expected. This might mean that the average net change in the stock's price over two-week periods might be somewhat less than twice the average net change over one-week periods, and the same might be true of the average variance.

The above is admittedly only speculation. We carried out two sets of calculations with our stocks to see if any of this speculation is correct. First, we calculated the variance ratios for our stocks, for the values of $q$ listed above. The average variance ratios were -.060, -.103, -.135, and -.159.

Our second set of calculations consisted of simulating our stocks by using a two-state Markov chain, in which the transition matrix entries are obtained from the observed probabilities for the four possible pairs up-up, up-down, down-up, and down-down (these were discussed above). For each of the above $q$, we simulated each of our stocks ten times, and computed the average variance ratio. Then we computed the average, over all of the stocks, of these averages. The values obtained were -.088, -.125, -.142, and -.153. Furthermore, of the 439 stocks, the numbers whose average simulated variance ratio had the same sign as the actual variance ratio, for $q = 2, 4, 8$, and 16, were 386, 384, 373, and 358. So the Markov chain model behaves similarly, in terms of the variance ratio, to the actual stock prices.

## 2.8 Conclusions

## 2.9 Appendix

In this section, we will show how one can use generating functions to derive the theoretical distributions of several parameters of interest in this chapter. The distribution of the number of runs in a sequence of $n$ Bernoulli trials was calculated by Mood [12]. An example of this distribution, with $n = 50$ and $p = .2$, is shown in Figure 2.1.

One sees two roughly normal-shaped distributions. The reason for this is that the probability that the number of runs is even does not equal the probability that the number of runs is odd (except if $p = 1/2$). In fact, the number of runs is even if and only if the first trial and the last trial disagree, which happens with probability $2p(1-p)$. Thus, the sum of the distribution values corresponding to even-valued outcomes equals this value, and the odd distribution values sum to 1 minus this value. (Note to us: This picture means that it might be better to work with the number of success runs.) We will give a derivation of this distribution that is different than the one found in [12], using a method that will generalize to the distribution of the number of success runs (i.e. the number of runs in the first state) of a Markov chain with two states. The distribution was first derived by Zaharov and Sarmanov [16].

If we denote by $r_{n,k}$ the probability of $k$ runs in $n$ trials (this number also depends upon $p$), then we will show that the generating function

$$r(x, y, p) = \sum_{n=1}^{\infty} \sum_{k=1}^{n} r_{n,k} x^n y^k$$

equals

$$\frac{xy(1 + 2px(-1+y) - 2p^2x(-1+y))}{1 - x + (-1+p)px^2(-1+y^2)} \ .$$

To derive this expression, we start by defining $f(n, p, k, S)$ to be the probability that in a sequence of $n$ Bernoulli trials with success probability $p$, there are exactly $k$ runs and the sequence ends in a success. The quantity $f(n, p, k, F)$ is defined similarly. Next, we define

$$G(x, y) = \sum_{i=1}^{\infty} \sum_{j=1}^{i} f(i, p, j, S) x^i y^j$$

and

$$H(x, y) = \sum_{i=1}^{\infty} \sum_{j=1}^{i} f(i, p, j, F) x^i y^j \ .$$

How are the coefficients related to one another? If a sequence of length at least two ends in a success, and it has $k$ runs, and we chop off the last term in the sequence, we obtain a sequence of length $n-1$ that still has $k$ runs and ends in a success, or it has $k-1$ runs and ends in a failure. Thus, for $n \ge 2$, we have

$$f(n, p, k, S) = f(n-1, p, k, S)p + f(n-1, p, k-1, F)q \ .$$

Similarly, we have, for $n \ge 2$,

$$f(n, p, k, F) = f(n-1, p, k, F)q + f(n-1, p, k-1, S)p \ .$$

These recursions do not hold when $n = 1$, because $f$ is not defined if the first parameter is 0. When $n = 1$, we have

$$f(1, p, 1, S) = p$$

and

$$f(1, p, 1, F) = q \ .$$

Using the recursions and initial conditions given above, we can write

$$G(x, y) = pxyH(x, y) + pxG(x, y) + pxy \ ,$$

and

$$H(x, y) = qxyG(x, y) + qxH(x, y) + qxy \ .$$

One can solve these equations for $G(x, y)$ and $H(x, y)$, obtaining

$$G(x, y) = \frac{pxy(1 + x(-1 + p + y - py))}{1 - x + (pqx^2(1 - y^2)}$$

and

$$H(x, y) = \frac{(-1 + p)xy(1 + px(-1 + y))}{1 - x + (pqx^2(1 - y^2)} \ .$$

Finally, note that

$$r(n, k) = f(n, p, k, S) + f(n, p, k, F) \ ,$$

so

$$r(x, y, p) = G(x, y) + H(x, y) \ ,$$

giving us the required expression for $r(x, y, p)$.

The above expression for the generating function $r(x, y, p)$ allows one to write an exact expression for $r_{n,k}$. The expression for $r(x, y, p)$ is of the form

$$xy\frac{A + Bx}{C + Dx + Ex^2}$$

where the capital letters represent expressions that do not involve $x$. Using the algebraic method of partial fractions (usually first learned in calculus to help integrate rational functions), one can write this expression as

$$xy\left(\frac{F}{1 - Gx} + \frac{H}{1 - Ix}\right) \ .$$

The two summands can be rewritten, using ideas about geometric series, to obtain

$$xy(F + FGx + FG^2x^2 + \ldots + H + HIx + HI^2x^2 + \ldots) \ .$$

We want the coefficient of $x^ny^k$ in this expression. The coefficient of $x^n$ is

$$y(FG^{n-1} + HI^{n-1}) \ .$$

This is a series involving $y$ but not $x$. We want the coefficient of $y^k$ in this series. The answer, which is obtained after some gruesome algebra (best performed by a computer algebra package) is as follows. Let $q = 1 - p$. If $k$ is odd, then

$$\begin{aligned} r_{n,k} \quad = \quad & \frac{1}{2^{n-1}}\Bigg[\sum_{j=\frac{(k-1)}{2}}^{\lfloor\frac{n}{2}\rfloor}\binom{n}{2j+1}\sum_{i=\frac{(k-1)}{2}}^{j}\binom{j}{i}(4pq)^i\binom{i}{(k-1)/2}(-1)^{i-(k-1)/2} \\ & -4pq\sum_{j=\frac{(k-2)}{2}}^{\lfloor\frac{n-1}{2}\rfloor}\binom{n-1}{2j+1}\sum_{i=\frac{(k-1)}{2}}^{j}\binom{j}{i}(4pq)^i\binom{i}{(k-1)/2}(-1)^{i-(k-1)/2}\Bigg] \ , \end{aligned}$$

while if $k$ is even, then

$$r_{n,k} = \frac{4pq}{2^{n-1}} \sum_{j=\frac{k}{2}}^{\lfloor\frac{n-1}{2}\rfloor} \binom{n-1}{2j+1} \sum_{i=\frac{(k-2)}{2}}^{j} \binom{j}{i} (4pq)^i \binom{i}{(k-2)/2} (-1)^{i-(k-2)/2} \ .$$

These expressions were used to generate Figure 2.1.

We can use the expression for $r(x, y, p)$ to calculate the mean (and variance) of the distribution. We recall that for fixed $n$, the mean of the distribution $\{r_{n,k}\}$ equals

$$\sum_{k=1}^{n} k r_{n,k} \ .$$

The value of this sum can be obtained from the generating function $r(x, y, p)$ by using calculus. If we compute the partial of $r(x, y, p)$ with respect to $y$, and then set $y = 1$, we obtain the expression

$$\sum_{n=1}^{\infty}\sum_{k=1}^{n} k r_{n,k} x^n \ ,$$

which can be written as

$$\sum_{n=1}^{\infty} x^n \sum_{k=1}^{n} k r_{n,k} \ .$$

Thus the mean of the distribution for sequences of length $n$ is just the coefficient of $x^n$ in the above expression. The point of this is that we do not need to use the formulas for $r_{n,k}$ to calculate the mean. Rather, we use the closed-form expression for $r(x, y, p)$, and apply the ideas above to this expression.

If we perform these calculations, we obtain the expression

$$\frac{x}{1-x} + \frac{2pqx^3}{(1-x)^2} + \frac{2pqx^2}{1-x} \ .$$

Using the facts that

$$\frac{1}{1-x} = 1 + x + x^2 + \dots$$

and

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + \dots \ ,$$

in a suitable interval containing the origin, we can expand each of the three summands above as series; they are, respectively,

$$x + x^2 + x^3 + \dots \ ,$$

$$2pq(x^3 + 2x^4 + 3x^5 + \dots) \ ,$$

and

$$2pq(x^2 + x^3 + x^4 + \dots) \ .$$

Now we can easily write down the coefficient of $x^n$; it is

$$1 + 2pq(n-2) + 2pq = 1 + 2pq(n-1) \ ,$$

if $n \geq 2$.

There is an easy way to check this. In fact, the calculation below is an easier way to *find* the mean in this case, but the above method can be used to find other moments (including the variance) and the calculation below does not generalize. Let $X_i$, for $1 \leq i \leq n-1$, denote the random variable that is 1 if the $i$'th and $(i+1)$'st outcomes disagree. Then the average number of runs is just

$$\sum_{i=1}^{n-1} X_i \ ,$$

so

$$\mu = \sum_{i=1}^{n-1} E(X_i) \ .$$

But for each $i$, the probability that $X_i = 1$ is just $2pq$, so for each $i$,

$$E(X_i) = 2pq \ ,$$

so the average number of runs is just

$$= 1 + 2p(1-p)(n-1) \ .$$

In the Markov model, the situation is more complicated(!). We will not go into the details here, but rather give an outline of how to proceed. We write

$$f_{S,S}(n,k)$$

for the probability that a sequence of $n$ trials begins and ends with a success and has $k$ runs. The quantities $f_{S,F}$, $f_{F,S}$ and $f_{F,F}$ are defined similarly. We define the generating function

$$G_{S,S}(x,y) = \sum_{n=1}^{\infty}\sum_{k=1}^{\infty} f_{S,S}(n,k)x^n y^k \ ;$$

the functions $G_{S,F}$, $G_{F,S}$, and $G_{F,F}$ are defined in a similar manner. One can show that

$$\begin{aligned}
G_{S,S}(x,y) &= p_2xyG_{S,F}(x,y) + p_1xG_{S,S}(x,y) + xy \ , \\
G_{S,F}(x,y) &= (1-p_1)xtG_{S,S}(x,y) + (1-p_2)xG_{S,F}(x,y) \ , \\
G_{F,S}(x,y) &= p_2xyG_{F,F}(x,y) + p_1xG_{S,F}(x,y) \ , \\
G_{F,F}(x,y) &= (1-p_1)xyG_{S,F}(x,y) + (1-p_2)xG_{F,F}(x,y) + xy \ .
\end{aligned}$$

Note that two of these equations are homogeneous (there are no summands on the right-hand sides that do not involve the functions $G_{*,*}$) which makes it easy to find $G_{S,F}$, say, once we have found $G_{S,S}$. In addition, by switching the roles of success and failure, one sees that it is easy to find $G_{F,F}$ once we know $G_{S,S}$ (and $G_{F,S}$ once we know $G_{S,F}$).

Thus, we only need to find the coefficients $f_{S,S}(n,k)$, in a manner similar to the one used to obtain the results for Bernoulli trials. One can show that

$$G_{S,S}(x,y) = \frac{xy}{1 - (p_2(1-p_1)x^2y^2)/(1-(1-p_2)x) - p_1x} \ .$$

Note that since the sequences corresponding to $G_{S,S}(x,y)$ begin and end with a success, the only positive probabilities are those corresponding to odd $k$. If $k$ is odd, then

$$f_{S,S}(n,k) = \frac{1}{2^{n-1}} \Bigg[ (-(1-p_1-p_2) \sum_{j=0}^{\lfloor \frac{(n-2)}{2} \rfloor} \binom{n-1}{2j+1} (1+p_1-p_2)^{n-1-(2j+1)} \Bigg($$

$$\sum_{i=\frac{k-1}{2}}^{j} \binom{j}{i} (1+p_1-p_2)^{2(j-i)} (-4)^i (-((1-p_1)p_2))^{(k-1)/2} (p_1(1-p_2))^{i-(k-1)/2} \binom{i}{(k-1)/2} \Bigg) \Bigg)$$

$$+ \sum_{j=0}^{\lfloor \frac{n-1}{2} \rfloor} \binom{n-1}{2j} (1+p_1-p_2)^{n-1-2j} \Bigg($$

$$\sum_{i=\frac{k-1}{2}}^{j} \binom{j}{i} (1+p_1-p_2)^{2(j-i)} (-4)^i (-((1-p_1)p_2))^{(k-1)/2} (p_1(1-p_2))^{i-(k-1)/2} \binom{i}{(k-1)/2} \Bigg) \Bigg) \Bigg] \ .$$

One can use the recursions relating $G_{S,S}(x,y)$ and $G_{S,F}(x,y)$ given above to obtain the following formula for $f_{S,F}(n,k)$:

$$f_{S,F}(n,k) = (1-p_1) \sum_{l=0}^{n-1} (1-p_2)^l f_{S,S}(n-l-1, k-1) \ .$$

These formulas are useful in plotting the distributions (depending upon whether the first trial is a success or a failure) of the number of runs. In Figure 2.2 a plot of the distribution for $n = 50$, $p_1 = .3$, and $p_2 = .1$, where the first trial is a success.

We note that the fixed probability vector of the transition matrix in the Markov chain model equals

$$\left( \frac{p_2}{1-p_1+p_2}, \frac{1-p_1}{1-p_1+p_2} \right) \ .$$

This means that the long-term proportion of successes equals

$$\frac{p_2}{1-p_1+p_2} \ .$$

Note that in this model, we have three quantities that can be estimated from the data, namely $p_1$, $p_2$, and the long-term proportion of successes (which we will call $p$), but there is a relation among them. I have no idea how a statistician would deal with this, since it is unlikely that the observed values satisfy the relation.

In order to create a distribution for the Markov chain model, we will weight the distributions corresponding to starting with a success or a failure by $p$ and $(1-p)$. The resulting generating function is

$$p\Big(G_{S,S}(x,y)+G_{S,F}(x,y)\Big)+(1-p)\Big(G_{F,S}(x,y)+G_{F,F}(x,y)\Big)\ .$$

If we do this, the expected number of runs equals

$$n\Omega_1+p\Omega_2+(1-p)\Omega_4+(p\Omega_3+(1-p)\Omega_5)(p_1-p_2)^{n-1}\ ,$$

where

$$\begin{aligned}
\Omega_1 &= \frac{2p_2(1-p_1)}{1-p_1+p_2}\ ,\\
\Omega_2 &= \frac{2-4p_1+2p_1^2-p_2+3p_1p_2-2p_1^2p_2-p_2^2+2p_1p_2^2}{(1-p_1+p_2)^2}\ ,\\
\Omega_3 &= \frac{-1+2p_1-p_1^2+p_2-p_1p_2}{(1-p_1+p_2)^2}\ ,\\
\Omega_4 &= \frac{1-2p_1+p_1^2-p_2+3p_1p_2-2p_1^2p_2+2p_1p_2^2}{(1-p_1+p_2)^2}\ ,\\
\Omega_5 &= \frac{p_2(1-p_1-p_2)}{(1-p_1+p_2)^2}\ .
\end{aligned}$$

Thus, the expected number of runs is asymptotic to

$$n\Omega_1+\Omega_4\ .$$

If one lets $p_1=p_2$, so that the Bernoulli model is obtained, one can check that the result agrees with the one already obtained.

The two graphs show what one might expect, namely that the expected number of runs in the Markov model is less than in the Bernoulli trials model. It is easy to show this; one need only show that if $p_1>p_2$, then

$$\Omega_1<2p(1-p)\ .$$

Next, it might be nice to show that the runs distribution corresponding to $G_{S,S}(x,y)$ is asymptotically normal. One might need to do this to discuss the power of the test that compares the two models.

We also might try writing approximation algorithms for the calculation of these probabilities, since the probabilities fall off rapidly and it is time-consuming to calculate the sums using the exact expressions.

The distribution of the length of the longest success run in the Markov chain model can be calculated using recursions. Define $A(n,x,k,p_1,p_2)$ to be the probability that a Markov sequence, beginning with a success, has no success run exceeding $x$ in length, and ends with $k$ successes, for $0\le k\le x$. This function satisfies the following equations. First, if $n=1$, then the function is 1 if $k=1$ and 0 otherwise. If $n>1$, then if $k>1$,

$$A(n,x,k,p_1,p_2)=A(n-1,x,k-1,p_1,p_2)p_1\ ,$$

since a sequence of length $n$ that ends in $k$ successes is obtained from one of length $n-1$ that ends in $k-1$ successes by adding one success to the end, and this happens with probability $p_1$, since $k > 1$. If $n > 1$ and $k = 1$, then

$$A(n, x, 1, p_1, p_2) = A(n-1, x, 0, p_1, p_2)p_2 \ ,$$

since in this case we are adding a success to the end of a sequence whose last state was the failure state. Finally, if $n > 1$ and $k = 0$, then

$$A(n, x, 0, p_1, p_2) = \left( \sum_{j=1}^{x} A(n-1, x, j, p_1, p_2)(1-p_1) \right) + A(n-1, x, 0, p_1, p_2)(1-p_2) \ ,$$

since in this case either there were $j$ successes immediately preceding the last trial, which was a failure, for some $j$ between 1 and $x$, or else the penultimate trial was also a failure.

These equations allow one to compute the values of the function $A$. To obtain the desired distribution, namely the set of probabilities that a Markov sequence, beginning with a success, has longest success run exactly equal to $x$, we compute the quantity

$$\sum_{k=0}^{x} A(n, x, k, p_1, p_2) - \sum_{k=0}^{x-1} A(n, x-1, k, p_1, p_2) \ .$$

This adds the weights of all of the sequences with longest success run at most $x$, and subtracts from this the weights of all of the sequences with longest success run at most $x - 1$.

### 2.9.1 Doubletons

We now turn to the question of the distribution of $d$, the number of doubletons (i.e. consecutive pairs of successes in a sequence of trials). The reason for our interest in this quantity is because, as was stated in Exercise 1, the number $d$ is closely related to $p_1$, the probability of a success following a success, and $p_1$ is of interest when studying autocorrelation. We will first consider asymptotic behavior of the distribution of the number of doubletons, in both the Bernoulli trials model and the Markov model, and then we will derive a recursion for the exact distribution in the Bernoulli case.

Since the Bernoulli trials model is a special case of the Markov model, we will work with the Markov model. In this model, the probabilities that a success follows a success or a failure are defined to be, respectively, $p_1$ and $p_2$.

We are interested in the distribution of $d$ for large values of $n$, the length of the sequence. It turns out that if $n$ is large, it does not matter very much in which state the Markov chain started (i.e. whether the first trial resulted in a success or a failure). We have stated above that the long-term fraction of the time that the chain is in state 1 is equal to

$$\frac{p_2}{1 - p_1 + p_2} \ .$$

Note that if $p_1 = p_2$ (i.e. we are in the Bernoulli model) then this expression reduces to $p_2$ $(= p_1)$.

We can find the distribution of the number of doubletons in a sequence of length $n$ generated by the Markov process by considering the related Markov chain that consists of four states: SS, SF, FS, and FF. The first state, SS, means that the first two trials in the original sequence are both successes. It is straightforward to determine the transition probabilities for this new Markov chain. For example, if the chain is in state SS, it stays in state SS or moves to state SF depending upon whether the next trial in the original chain is a success or a failure. Thus,these two transitions occur with probability $p_1$ and moves to state SF with probability $1-p_1$.

In the new Markov chain, we wish to know the distribution of the number of times $Y_{SS}^{(n)}$ that the chain is in state SS in the first $n$ trials. Of course, this distribution depends upon the starting state (or starting distribution), but it turns out that the limiting distribution is independent of the starting state. The Central Limit Theorem for Markov Chains (see [9], pg. 89) states that $Y_{SS}^{(n)}$ is asymptotically normally distributed, with a mean and standard deviation that are straightforward to calculate. Examples of how the mean and standard deviation are calculated are given in [9]. In the present case, the asymptotic value of the mean and standard deviation are

$$\frac{np_1p_2}{1-p_1+p_2}$$

and

$$\frac{np_1p_2(p_1^2(-1+p_2)+(1+p_2)^2-p_1p_2(3+p_2))}{(1-p_1+p_2)^3} \ .$$

Now suppose that we have a process that presents us with a sequence of successes and failures, and suppose that the observed probability of a success is .3. We assume that $p = .3$, and we wish to test the hypothesis that $p_1 = .3$ against the alternative hypothesis that $p_1 > .3$. To carry out this test, using the number of doubletons as a parameter, we first choose an acceptance region around the value $np^2$ in which 95% of the values will fall (remember, in the case that the null hypothesis is true, then $p_1 = p_2 = p$, and we are dealing with Bernolli trials). Since $Y_{SS}^{(n)}$ is asymptotically normal, it is easy to pick this acceptance region. It is an interval of the form $[0, c]$, because the form of the alternative hypothesis is a one-way inequality. The number $c$ is $np^2$ plus 1.65 times the standard deviation of $Y_{SS}^{(n)}$. We obtain the value of

$$c = n(.3)^2 + 1.65\sqrt{.1197n} \ .$$

This leads us to the distributions shown in Figures 2.4 to 2.5.

We now give an outline of the method used to find the exact distribution of the number of doubletons in the Bernoulli model, with parameters $n$ and $p$. We define $r(n,k,p)$ and $s(n,k,p)$ to be the probabilities of exactly $k$ doubleton successes in $n$ trials, with success probability $p$, and with the sequence ending in a failure or a success, respectively. Then the following recursions hold:

$$r(n,k,p) = q * r(n-1,k,p) + q * s(n-1,k,p) \ ,$$

$$s(n,k,p) = p * r(n-1,k,p) + p * s(n-1,k-1,p) \ .$$

The first of these equations says that sequences of length $n$ with exactly $k$ doubletons and which end in a failure arise from sequences of length $n-1$ with exactly $k$

doubletons by adding a failure to the end. The second equation says that sequences of length $n$ with exactly $k$ doubletons and which end in a success arise by adding a success to the end of either a sequence of length $n-1$ with exactly $k$ doubletons, ending in a failure, or to the end of a sequence of length $n-1$ with exactly $k-1$ doubletons, ending in a success.

Next, define

$$R(p,x,y) = \sum_{n=1}^{\infty}\sum_{k=0}^{n-1} r(n,k,p)x^n y^k$$

and

$$S(p,x,y) = \sum_{n=1}^{\infty}\sum_{k=0}^{n-1} s(n,k,p)x^n y^k \ .$$

The distribution of the number of doubletons in sequences of length $n$ is given by the set of values

$$\{r(n,k,p) + s(n,k,p)\} \ .$$

The above recursions, together with attention to initial conditions, give rise to the following functional equations:

$$R(p,x,y) = qx\Bigl(R(p,x,y) + S(p,x,y)\Bigr) + qx$$

and

$$S(p,x,y) = px\Bigl(R(p,x,y) + S(p,x,y)\Bigr) + px \ .$$

These equations can be solved for $R(p,x,y)$ and $S(p,x,y)$:

$$R(p,x,y) = -1 + \frac{1-pxy}{1-qx-pqx^2-pxy+pqx^2y}$$

and

$$S(p,x,y) = \frac{px}{1-qx-pqx^2-pxy+pqx^2y} \ ,$$

where $q = 1-p$.

We can write

$$1 + R(p,x,y) = \frac{A}{1-\gamma_1 x} + \frac{B}{1-\gamma_2 x} \ ,$$

for suitable choices of constants $A$, $B$, $\gamma_1$, and $\gamma_2$. The right-hand side can be expanded to yield

$$(A+B) + (A\gamma_1 + B\gamma_2)x + (A\gamma_1^2 + B\gamma_2^2)x^2 + \dots \ ,$$

allowing us to determine the coefficient of $x^n$. A similar method allows us to deal with $S(p,x,y)$.

If we let

$$\Delta = q^2 - 2pq(-2+y) + p^2y^2 \ ,$$

then one can show, after some work, that the coefficient of $x^n$ in the power series for $R(p,x,y)$ equals

$$\frac{1}{2^n}\left((q-py)\sum_{j=0}^{\lfloor n/2\rfloor}\binom{n}{2j+1}(q+py)^{n-2j+1}\Delta^j+\sum_{t=0}^{\lfloor n/2\rfloor}(q+py)^{n-2t}\Delta^t\right)\ ,$$

and the coefficient of $x^n$ in the power series for $S(p,x,y)$ equals

$$\frac{p}{2^{n-1}}\sum_{j=0}^{\lfloor n/2\rfloor}\binom{n}{2j+1}(q+py)^{n-2j-1}\Delta^j\ .$$

To obtain $r(n,k,p)$ and $s(n,k,p)$ from these expressions, we need to find the coefficients of $y^k$ in the above expressions. Writing $\Delta_l^j$ for the coefficient of $y^l$ in $\Delta^j$, one can show that

$$\Delta_l^j=\sum_{h=0}^{l}\binom{j}{h}p^l(-1)^l(q-2\sqrt{-pq})^{j-h}(q+2\sqrt{-pq})^{j-l+h}\ .$$

Using this abbreviation, one can then show that

$$\begin{aligned}r(n,k,p)\quad &=\quad \frac{1}{2^n}\Bigg(q\sum_{j=0}^{\lfloor n/2\rfloor}\binom{n}{2j+1}\sum_{l=0}^{k}\Delta_l^j\binom{n-2j-1}{k-l}p^{k-l}q^{n-2j-1-(k-l)}\\ &\qquad -p\sum_{j=0}^{\lfloor n/2\rfloor}\binom{n}{2j+1}\sum_{l=0}^{k-1}\Delta_l^j\binom{n-2j-1}{k-1-l}p^{k-1-l}q^{n-2j-1-(k-1-l)}\\ &\qquad +\sum_{j=0}^{\lfloor n/2\rfloor}\binom{n}{2j}\sum_{l=0}^{k}\Delta_l^j\binom{n-2j}{k-l}p^{k-l}q^{n-2j-(k-l)}\ ,\Bigg)\end{aligned}$$

and

$$s(n,k,p)=\frac{p}{2^{n-1}}\sum_{j=0}^{\lfloor n/2\rfloor}\binom{n}{2j+1}\sum_{l=0}^{k}\Delta_l^j\binom{n-2j-1}{k-l}p^{k-l}q^{n-2j-1-(k-l)}\ .$$

As before, we can use the closed-form expressions for $R(p,x,y)$ and $S(p,x,y)$ to find the mean and variance of the distribution of the number of doubletons (and if we compare these expressions with the asymptotic ones obtained above for the Markov case, we can partially check the accuracy of our calculations). If we write

$$F(p,x,y)=R(p,x,y)+S(p,x,y)\ ,$$

then the mean of the distribution of doubletons is obtained by differentiating $F(p,x,y)$ with respect to $y$, setting $y=1$, and asking for the coefficient of $x^n$. To see why this works, note first that since

$$F(p,x,y)=\sum_{n=1}^{\infty}\sum_{k=0}^{n}(r(n,k,p)+s(n,k,p))x^ny^k\ ,$$

if we differentiate $F(p,x,y)$ with respect to $y$, we obtain

$$\frac{\partial}{\partial y}F(p,x,y)=\sum_{n=1}^{\infty}\sum_{k=0}^{n}k(r(n,k,p)+s(n,k,p))x^ny^{k-1}\ .$$

If we set $y=1$ and consider the coefficient of $x^n$, we find that it equals

$$\sum_{k=1}^{n}k(r(n,k,p)+s(n,k,p))\ ,$$

which is clearly the mean of the distribution. In the present case, we find the value of the mean to equal $(n-1)p^2$, which can be checked as being asymptotically equal to the expression we obtained for the asymptotic value of the mean in the Markov case (with $p_1=p_2=p$).

A similar, but more complicated, calculation leads to the variance of the distribution of doubletons; we obtain the expression

$$p^2q(n+3np-1-5p)\ .$$

One can check that this is asymptotic to the expression obtained for the asymptotic value of the variance in the Markov case.

It would be nice to be able to show that the random variable that counts the number of doubletons in the Bernoulli trials case is asymptotically normal. If this were true, then we could use the above values for the mean and standard deviation to give a precise asymptotic description of the distribution of the number of doubletons. It is typically the case that one tries to use the Central Limit Theorem to show that a given sequence of distributions is asymptotically normal. In order to use the Central Limit Theorem, one must write the terms of the sequence as sums of mutually independent random variables.

In the present situation, if we consider sequences of $n$ Bernoulli trials, and we let $X_i$ denote the random variable that is 1 if the $i$'th and $i+1$'st trials are successes, then the number of doubletons in the first $n$ trials is

$$X_1+X_2+\ldots+X_{n-1}\ .$$

Unfortunately, the $X_i$'s are not mutually independent (for example, if $X_{i-1}=X_{i+1}=1$, then $X_i$ must be 1 as well). Nevertheless, it is possible to salvage the situation, because the $X_i$'s are 'independent enough.' More precisely, the sequence $\{X_i\}$ is $m$-independent, i.e. it is possible to find an $m$ (in this case, $m=2$) such that the sequence can be partitioned into $m$ subsets such that the random variables in each subset are mutually independent. In this case, we can take the sets $\{X_1,X_3,X_5,\ldots\}$ and $\{X_2,X_4,X_6,\ldots\}$. If some other conditions (which we will not state here) are satisfied, then the sequence satisfies the Central Limit Theorem, i.e. the sum $S_n=X_1+X_2+\ldots+X_n$ is asymptotically normal. In the present case, all of the necessary conditions are satisfied, so the distribution of the number of doubletons is asymptotically normal.

## Exercises

**1** Let $h_S(n,k)$ denote the probability that in the Markov model, a sequence of $n$ trials begins with a success and has exactly $k$ success runs. Explain how one can write $h_S(n,k)$ in terms of $f_{S,S}(n,i)$ and $f_{S,F}(n,j)$ for appropriate choices of $i$ and $j$.

**2** Define the function $F(r_1, r_2)$ by

$$F(r_1,r_2) = \begin{cases} 2 & \text{if } r_1 = r_2 \\ 1 & \text{if } |r_1 - r_2| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that we have a sequence of $n_1$ successes and $n - n_1$ failures in a Bernoulli trials process with parameters $n$ and $p$. Suppose, in addition, that there are $r_1$ success runs and $r_2$ failure runs.

(a) Show that if $r_1 \geq 1$, then there are $\binom{n_1-1}{r_1-1}$ ways to split the $n_1$ successes into runs.

(b) Using the fact that runs of successes and failures must alternate in a sequence, show that if $r_1 \geq 1$ and $r_2 \geq 1$, then the number of sequences of the given type equals

$$\binom{n_1 - 1}{r_1 - 1}\binom{n - n_1 - 1}{r_2 - 1}F(r_1, r_2)\ .$$

(c) Show that any two sequences of length $n$ with $n_1$ successes are equally likely.

(d) Show that the probability that a Bernoulli trials sequence of length $n$ has exactly $r_1 \geq 1$ success runs and $r_2 \geq 1$ failure runs, given that it has $n_1$ successes and $n - n_1$ failures, equals

$$P_{n_1}(r_1, r_2) = \frac{\binom{n_1-1}{r_1-1}\binom{n-n_1-1}{r_2-1}F(r_1,r_2)}{\binom{n}{n_1}}\ .$$

(e) Determine the corresponding formula if either $r_1 = 0$ or $r_2 = 0$.

(f) If $X$ and $Y$ are events in a sample space, then

$$P(X \wedge Y) = P(X\,|\,Y)P(Y)\ .$$

In the space of all Bernoulli sequences of length $n$, let $X$ denote the event that the number of success runs equals $r_1$ and the number of failure runs equals $r_2$, and let $Y$ denote the event that the number of successes equals $n_1$ and the number of failures equals $n - n_1$. Parts d) and e) calculate $P(X\,|\,Y)$. Find $P(Y)$ and use this to find a formula for $P(X \wedge Y)$.

(g) By summing over the appropriate set of $n_1$'s, find a summation that gives the probability that a Bernoulli sequence of length $n$ has exactly $r_1$ success runs. (This is the form of the expression for this distribution in [12].)

(h) Write a summation that gives the probability that a Bernoulli sequence of length $n$ has exactly $r$ runs of both types (i.e. the total number of runs is $r$).

# Chapter 3

# Lotteries

## 3.1 Introduction

Lotteries are discussed frequently in the news, and they have a huge impact directly and indirectly on our lives. They are the most popular form of gambling and an increasingly important way that states obtain revenue. In this chapter, we will use the Powerball lottery to illustrate some of the statistical ideas associated with lotteries.

## 3.2 The Powerball Lottery

### 3.2.1 The Rules

The Powerball Lottery is a multi-state lottery, a format which is gaining popularity because of the potential for large prizes. It is currently available in 20 states and Washington D.C. It is run by the Multi-State Lottery Association, and we shall use information from their web homepage, http://www.musl.com. We found their "Frequently Asked Questions," (hereafter abbreviated FAQ) to be particularly useful. These are compiled by Charles Strutt, the executive director of the Association.

A Powerball lottery ticket costs $1. For each ticket you are asked mark your choice of numbers in two boxes displayed as shown in Figure 3.1.

You are asked to select five numbers from the top box and one from the bottom box. The latter number is called the "Powerball". If you check EP (Easy Pick) at the top of either box, the computer will make the selections for you. You also must select "cash" or "annuity" to determine how the jackpot will be paid should you win. Finally, there is another option called the "Power Play." In what follows, we will refer to a particular selection of five plus one numbers as a "pick."

The Powerball Lottery was started in 1992. In the history of this lottery, there have been three versions, which we will refer to as old, middle, and new. Before November 2, 1997, there were 45 numbers in the top box and 45 in the bottom box. On that date, these numbers were changed to 49 and 42, respectively. They were changed again on October 9, 2002 to 53 and 42, respectively. Unless stated

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| 29 | 30 | 31 | 32 | 33 | 34 | 35 |
| 36 | 37 | 38 | 39 | 40 | 41 | 42 |
| 43 | 44 | 45 | 46 | 47 | 48 | 49 |
| 50 | 51 | 52 | 53 | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| 29 | 30 | 31 | 32 | 33 | 34 | 35 |
| 36 | 37 | 38 | 39 | 40 | 41 | 42 |

Figure 3.1: Picking your numbers.

otherwise, the calculations below will refer to the new version.

Every Wednesday and Saturday night at 10:59 p.m. Eastern Time, lottery officials draw five white balls out of a drum with 53 balls and one red ball from a drum with 42 red balls. Players win prizes when the numbers on their ticket match some or all of the numbers drawn (the order in which the numbers are drawn does not matter). There are 9 ways to win. Here are the possible prizes and their probabilities:

| You Match | You Win | Probability of Winning |
|---|---|---|
| 5 white balls and the red ball | JACKPOT | 1/120,526,770 |
| 5 white balls but not the red ball | $100,000 | 1/2,939,677 |
| 4 white balls and the red ball | $5,000 | 1/502,195 |
| 4 white balls but not the red ball | $100 | 1/12,249 |
| 3 white balls and the red ball | $100 | 1/10,685 |
| 3 white balls but not the red ball | $7 | 1/260.61 |
| 2 white balls and the red ball | $7 | 1/696.85 |
| 1 white ball and the red ball | $4 | 1/123.88 |
| 0 white balls and the red ball | $3 | 1/70.389 |

Table 3.2.1. The chance of winning.

If a player wins any prize except the jackpot and if she has not selected the Power Play option, she wins the amount of that prize. If she has selected the Power Play option (which costs $1 dollar per ticket), then all prize amounts except the jackpot are multiplied by either 2, 3, 4, or 5. This multiplier is the same for all players in a given lottery, and is determined by the drawing of a Power Play ball at the time the other balls are drawn. There are five Power Play balls; two of them have a 5 on them, and the numbers 2, 3, and 4 each appear on one of the other three balls.

If the player wins the jackpot, she must share it equally with all other players (if there are any) who have also won the jackpot. A few other comments concerning the jackpot are in order. First, the winning players have 60 days after they have won to declare whether they want a lump sum payment or a series of 30 equal payments over the next 29 years (the first payment is made immediately and the others are made at the end of each subsequent year). This latter type of award is called an annuity. Second, the lump sum is typically only about 60% of the announced value of the jackpot. The present value of the annuity is equal to the lump sum amount. The announced value of the jackpot is the undiscounted total of all the annuity payments. An example will help clear this up.

Suppose that the jackpot is announced to be $30 million. If there is only one winner, and he chooses the annuity option, he will receive 30 equal payments of $1 million each. Thus, in some sense he has won $30 million. However, when considering future payments, one usually discounts those future payments. The point here is that having $100 today is in a sense equal to having $110 one year from today, if one can earn 10% interest on cash. In a symmetric fashion, being awarded a payment of $100 that will be paid one year from today is equal to having 10/11 of $100 (or $90.91) today. The present value of the 30 equal payments of $1 million is determined by discounting each of the last 29 payments by an amount determined by the date on which the payment will be made.

### 3.2.2 Calculating the Probabilities

The first question we ask is: how are these probabilities determined? In what follows, we will calculate various probabilities assuming the player has bought one ticket. This is a counting problem that requires that you understand one simple counting rule: if you can do one task in $n$ ways and, for each of these, another task in $m$ ways, the number of ways the two tasks can be done is $mn$. A simple tree diagram makes this principle very clear.

When you watch the numbers being drawn on television, you see that, as the five winning white balls come out of the drum, they are lined up in a row. The first ball could be any one of 53. For each of these possibilities the next ball could be any of 52, etc. Hence the number of possibilities for the way the five white balls can come out in the order drawn is

$$53 \cdot 52 \cdot 51 \cdot 50 \cdot 49 = 344,362,200 .$$

But to win a prize, the order of these 5 white balls does not count. Thus, for a particular set of 5 balls all possible orders are considered the same. Again by our counting principle, there are $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ possible orders. Thus, the number of possible sets of 5 white balls not counting order is $344,262,200/120 = 2,869,685$. This is the familiar problem of choosing a set of 5 objects out of 53, and we denote this by $C(53,5)$. Such numbers are called binomial coefficients. We can express our result as

$$C(53,5) = \binom{53}{5} = \frac{53!}{5!48!} = 2,869,685 \ .$$

Now for each pick of five white numbers there are 42 possibilities for the red Powerball, so the total number of ways the winning six numbers can be chosen is

$$42 \cdot C(53,5) = 120,526,770 \ .$$

We will need this number often and denote it by $b$ (for big).

The lottery officials go to great pains to make sure that all $b$ possibilities are equally likely. So, a player has one chance in 120,526,770 of winning the jackpot. Of course, the player may have to share this prize.

We note that on the Powerball website, the column corresponding to the last column in Table 3.1 is labeled 'odds.' The numbers in the column are in fact probabilities, not odds. Because the probabilities are small, there is not much difference between odds and probabilities. However, this is a good excuse to get the difference between the two concepts straightened out. The media prefers to use odds, and textbooks prefer to use probability or chance. Here the chance of winning the jackpot is 1 in 120,526,770, whereas the odds are 1 to 120,526,769 in favor (or 120,526,769 to 1 against).

To win the $100,000 second prize, the player must get the 5 white numbers correct but miss the Powerball number. How many ways can this be accomplished? There is only one way to get the set of five white numbers, but the player's Powerball pick can be any of the 41 numbers different from the red number that was drawn. Thus, the chance of winning second prize is 41 in 120,526,770; rounded to the nearest integer this is 1 in 2,939,677.

One can find the probability of winning the second prize by pointing out the probability that you get the 5 white numbers correct is 1/C(53,5). The chance of not getting the red ball correct is 41/42. Since these events are independent, the chance that they both happen is the product of their individual probabilities.

When there were 45 white balls and 45 red balls, the ticket listed the chances of getting only the red ball as 1 in 84. This often seemed wrong to players who have had elementary probability, as the following exchange from the Powerball FAQ[1] illustrates:

COULD YOUR ODDS BE WRONG?

> I have a simple question. You list the odds of matching only the powerball as one in 84 on the powerball "ways to win" page. From my understanding of statistics (I could be wrong, but I got an A), the odds of selecting one number out of a group is simply one over the number

[1] From the Multi-State Lottery Association web site at http://www.musl.com/

> of choices. Since there are not 84 choices for the powerball, may I assume the balls are somehow ”fixed” so that some are more common than others? Otherwise, the listed odds are somehow defying the laws of statistics. I am really very eager to hear your explanation, so please return my message. Thank you.
>
> Susan G., via the Internet.
>
> This is one of the most common questions we get about the statistics of the game. If you could play only the red Powerball, then your odds of matching it would indeed be 1 in 45. But to win the \$1 prize for matching the red Powerball alone, you must do just that; match the red Powerball ALONE. When you bet a dollar and play the game, you might match one white ball and the red Powerball. You might match three white balls and the red Powerball. To determine the probability of matching the red Powerball alone, you have to factor in the chances of matching one or more of the white balls too.
>
> C. S.

To win this last prize you must choose your six numbers so that only the Powerball number is correct. In the old version of the Powerball lottery this would be done as follows: there are $45 \cdot C(45, 5) = 54,979,155$ ways to choose your six numbers. But here your first 5 numbers must come from the 40 numbers not drawn by the lottery. This can happen in $C(40, 5) = 658,008$ ways. Now there is only one way to match the Powerball number, so overall you have 658,008 chances out of 54,979,155 to win this prize. This reduces to 1 chance in 83.55, or about 1 chance in 84, in agreement with the official lottery pronouncement.

The same kind of reasoning of course carries over to the present version of the game. To find the chance of winning any one of the prizes we need only count the number of ways to win the prize and divide this by the total number of possible picks $b$. Let $n(i)$ be the number of ways to win the $i$th prize. Then the values of $n(i)$ are shown in Table 3.2 below.

| Number of Ways | Match |
|---|---|
| $n(1) = 1$ | all six balls |
| $n(2) = 41$ | 5 white balls but not the red ball |
| $n(3) = C(5, 4) \cdot C(48, 1)$ | 4 white balls and the red ball |
| $n(4) = n(3) \cdot 41$ | 4 white balls but not the red ball |
| $n(5) = C(5, 3) \cdot C(48, 2)$ | 3 white balls and the red ball |
| $n(6) = n(5) \cdot 41$ | 3 white balls but not the red ball |
| $n(7) = C(5, 2) \cdot C(48, 3)$ | 2 white balls and the red ball |
| $n(8) = C(5, 1) \cdot C(48, 4)$ | 1 white ball and the red ball |
| $n(9) = C(48, 5)$ | only the red ball |

Table 3.2.2. How many ways can you win a particular prize?

Dividing these numbers by $b$, we obtain the chance of winning the corresponding prizes given in Table 3.2.1. Adding all the of $n(i)$ values gives a total of 3,342,046 ways to win something. Thus we get an overall chance of winning of $3,342,046/b = 0.02773$, which is about 1 in 36.

In a textbook, we would be apt to give the results of Table 3.1 as:

| You Match | You Win | Probability of Winning |
|---|---|---|
| 5 white balls and the red ball | JACKPOT | 0.0000000083 |
| 5 white balls but not the red ball | $100,000 | 0.0000003402 |
| 4 white balls and the red ball | $5,000 | 0.0000019913 |
| 4 white balls but not the red ball | $100 | 0.0000816416 |
| 3 white balls and the red ball | $100 | 0.0009358917 |
| 3 white balls but not the red ball | $7 | 0.0038371558 |
| 2 white balls and the red ball | $7 | 0.0014350339 |
| 1 white ball and the red ball | $4 | 0.0080720656 |
| 0 white balls and the red ball | $3 | 0.0142068355 |

Table 3.2.2. The probabilities of winning.

As noted earlier, rounding the reciprocals of these probabilities to the nearest integer gives the numbers reported as ”odds” on the lottery ticket.

Discussion Question: Which of the two methods for presenting the chances of winning, Table 3.1 or Table 3.3, do you think is best understood by the general public? Which do you prefer?

### 3.2.3 What is your expected winning for a $1 ticket?

The value of a gambling game is usually expressed in terms of the player's expected winning. If there are $n$ prizes and $p(i)$ is the probability of winning the $i$th prize $w(i)$, then your expected winning is:

$$E = w(1) \cdot p(1) + w(2) \cdot p(2) + ... + w(n) \cdot (n) \ .$$

We will first discuss the case when the Power Play option is not chosen. Later, we will summarize the corresponding case when this option is chosen. For all prizes, except the jackpot, we can assume we know the value of the prize. However, since the size of the jackpot differs significantly from drawing to drawing, we will want to find the expected winning for different jackpot sizes. In the 508 drawings from the beginning of the lottery on April 22, 1992 through March 1, 1997 the jackpot was won 75 times. It was shared with one other winner 11 times. During this period the jackpot prize varied from $2 million to $314,900,000.

If $x$ is the amount of the jackpot and $p(i)$ the probability of winning the $i$th prize, the expected winning is:

$$E = x \cdot p(1) + 100,000 \cdot p(2) + 5000 \cdot p(3) + 100 \cdot p(4) + 100 \cdot p(5)$$

$$
\begin{aligned}
&\quad +7 \cdot p(6) + 7 \cdot p(7) + 4 \cdot p(8) + 3 \cdot p(9) \\
&= \frac{x}{b} + 0.173 \ ,
\end{aligned}
$$

where $b = 120,526,770$. The last expression above says that the expected value of a ticket has two components, the amount attributable to the jackpot, and the amount attributable to all of the other prizes. As we will soon see, the amount won by a jackpot winner is affected by many things. The second summand is not changed by any of these things.

Using this, we can find the expected winning for various values of the jackpot:

| $x$ = Jackpot (in millions of dollars) | $E$ = Expected winnings (in dollars) |
|---|---|
| 20 | 0.339 |
| 40 | 0.505 |
| 60 | 0.671 |
| 80 | 0.837 |
| 100 | 1.003 |
| 120 | 1.169 |
| 140 | 1.335 |
| 160 | 1.501 |
| 180 | 1.667 |
| 200 | 1.833 |
| 220 | 1.999 |
| 240 | 2.165 |
| 260 | 2.331 |
| 280 | 2.496 |
| 300 | 2.662 |
| 320 | 2.828 |

Table 3.2.3. Expected winnings for different size jackpots.

A game is said to be *favorable* if the expected winning is greater than the cost of playing. Here we compare with the $1 cost of buying a ticket. Looking at Table 3.4, we see that the lottery appears to be a favorable game as soon as $x$ gets up to $100 million.

The jackpot for the Powerball lottery for December 25, 2002 built up to some $314.9 million, as hordes of players lined up at ticket outlets for a shot at what had become the largest prize for any lottery in history. At first glance, it certainly looks like this was a favorable bet!

However, the reader will recall that the winner must choose, within 60 days of winning the jackpot, whether to take a lump sum payment of cash or to take an annuity. In fact, the Powerball website regularly updates the value of the jackpot for each choice. At this writing (17 January, 2004), we find the following report.

Next Powerball Jackpot Estimate

Saturday, January 17, 2004 $37 Million ($20.3 Million - cash option)

> Select the cash option and receive the full cash amount in the prize pool. Select the annuity option and we will invest the money and pay the annuity amount to you over 30 annual payments.

The $37 million here is analogous to the $314.9 million from December 25, 2002, and is the number that the media likes to hype. But note that this corresponds to the annuity amount to be paid out over time, not the immediate cash value. Not only are you not going to get this money tomorrow–the lottery doesn't even have it on hand! This is explained further in an earlier excerpt from the FAQ:

> When we advertise a prize of $25 million paid over 29 years (30 payments), we actually have about $13 million in cash. When someone wins the jackpot, we take bids to purchase government securities to fund the prize payout. We take the $13 million in cash and buy U.S. government-backed securities to fund these payments. We buy bonds that will mature in one year at $1 million, then bonds which will mature in two years at $1 million, etc. Generally, the longer the time to maturity, the cheaper the bonds.

The cash option on the $314.9 million from December 25 was about $170 million. From Table 3.4, we see that this still looks like a favorable bet, with expected value of about 2. We have been assuming that the player has elected the lump sum cash payment, and treating the annuity as equivalent in present value terms. You may want to think harder about this. An article in the Star Tribune by Julie Trip (June 7, 1998, Metro section p. 1D) discusses the question of lump sum or annuity. The article is based on an interview with Linda Crouse, a financial planner and certified public accountant in Portland, Oregon. (Note that this article was written in the middle regime of the lottery.) From this article we read:

> Crouse ran numbers to help determine whether it's better to take a windfall in payments over time - an annuity - or in a lump sum.
>
> Crouse used the Powerball jackpot as an example to determine which pays off in the long run: the ticket that pays the winner $104.3 million now or pays $7.7 million annually for 25 years. (Both are before taxes.)
>
> The annuity represents a 5.4 percent return. That sounds easy to beat if you take the lump sum and invest it - until you consider the huge negative effect of paying all the taxes up front instead of over 25 years. Figure 45 percent of the payout - $46.9 million - goes to state and federal taxes right off the bat. If you invest the remaining $57.4 million and receive an average return of 8 percent, you still can't beat the annuity. After all taxes are paid, you receive $4,235,000 annually for the annuity vs. $3,991,000 for the lump sum you invested at 8 percent.
>
> Beyond about a 9 percent return, you start to beat the annuity.

Of course, one should consider the fact that the annuity is a guaranteed payment while your investments are subject to the volatility of the way you invest your money.

Well, at least with the lump sum above, we convinced ourselves that we had a favorable game. Alas, there is another rub. We have been implicitly assuming that if we hold the lucky numbers, we will get the whole prize! But if other ticket holders have selected the same numbers, the jackpot will be split. This will be a particularly important factor when large number of tickets are sold. As the jackpot grows, an increasing number of tickets are sold. For example, for the December 25, 2002 Powerball lottery, it was estimated that 113 million tickets were sold.

The chance of having to share the jackpot depends upon several factors. First, it depends upon the number of tickets sold. Second, it depends upon whether the numbers are all roughly equally likely to be chosen. In the Easy Pick method, the numbers are chosen by the computer, and we can therefore assume that they are all equally likely. We are told that about 70% of tickets sold in a typical lottery are chosen by the Easy Pick method. Probably this percentage is even larger when the jackpot is large since people tend to buy a number of tickets and would be more likely to use the Easy Pick method when they do this.

The remaining tickets have numbers that are chosen by the ticket buyers. Figure 3.7 (displayed later in the chapter) shows that the numbers are not chosen with equal probabilities by the buyers. In a given winning set of numbers, some of them will be more likely than others to have been chosen by these buyers. We expect that this will have little effect on the number of jackpot winners. *****Simulate this.

Because the effect of the non-Easy Pick tickets is small, we will use $n = 113,000,000$ as the number of tickets sold, and we will assume that they were all chosen by the Easy Pick method. The probability that a particular ticket is the winning ticket is $1/120,526,770$. The probability of $k$ winners can be obtained from a binomial distribution with $p = 1/120,526,770$ and $n = 113,000,000$. The expected number of winning tickets is

$$np = \frac{113,000,000}{120,526,770} = 0.94 \ .$$

Since $p$ is small and $n$ is large we can use the Poisson approximation:

$$p(k) = e^{-m}\frac{m^k}{k!} \ ,$$

where $m = 0.94$. Carrying out this calculation gives:

| Number of Winners | Probability |
|---|---|
| 0 | .3906 |
| 1 | .3672 |
| 2 | .1726 |
| 3 | .0541 |
| 4 | .0127 |
| 5 | .0024 |
| 6 | .0004 |
| 7 | .0001 |

Table 3.2.3. Poisson probabilities for numbers of jackpot winners.

From this table we find that the probability, given that there is at least one winner, that the winner is the only winner is .603. Thus the probability that the winner has to share the prize is .397.

Recall that the cash value of the December 25 jackpot was about $170,000,000. Using the probability $p(k)$ that we will have to share this with $(k-1)$ others, we can find the expected amount that a winning player will end up with. We need only sum the values $(170,000,000/k) * p(k)$ for $k = 1$ to 12. Carrying out this calculation we find that the expected cash value of a jackpot is $80,789,100. Using this number for $x$ in the expression for the expected value of a ticket (the expression that precedes Table 3.4), we get an expected value of $.843.

Unfortunately, the government isn't about to let a lucky winner just walk off with a lump sum without paying taxes. One FAQ inquirer protested the fact that 28% of that payment is withheld in federal tax. In fact, the situation is worse, since some states will take out additional money to cover state tax as well! Here in New Hampshire (at this writing at least!), there is no state income tax. Thus, if we win the jackpot, we finally have to take 72% of our expected winning of $80,789,100 obtaining $58,168,152. In general, we will have to pay taxes on any money we win, so the expected value of a ticket, after taxes, is $.72 \cdot \$.843 = \$.607$.

What happens if the player chooses the Power Play option (this option was described in Section 3.2)? In this case, the expression for the expected value of the ticket changes to

$$\begin{aligned} E &= x \cdot p(1) + 380,000 \cdot p(2) + 19000 \cdot p(3) + 380 \cdot p(4) + 380 \cdot p(5) \\ &\quad +26.6 \cdot p(6) + 26.6 \cdot p(7) + 15.2 \cdot p(8) + 11.4 \cdot p(9) \\ &= \frac{x}{b} + 0.173 \ , \end{aligned}$$

where $x$ is still $58,168,152, since the Power Play option does not affect the jackpot. The other award values have been changed by multiplying by the expected value of the multiplier, which is

$$2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} + 5 \cdot \frac{2}{5} = \frac{19}{5} \ .$$

The expected value is now $1.1412, which sounds like it is now favorable, but we must remember that the player paid $1 for this option, so the expected return per

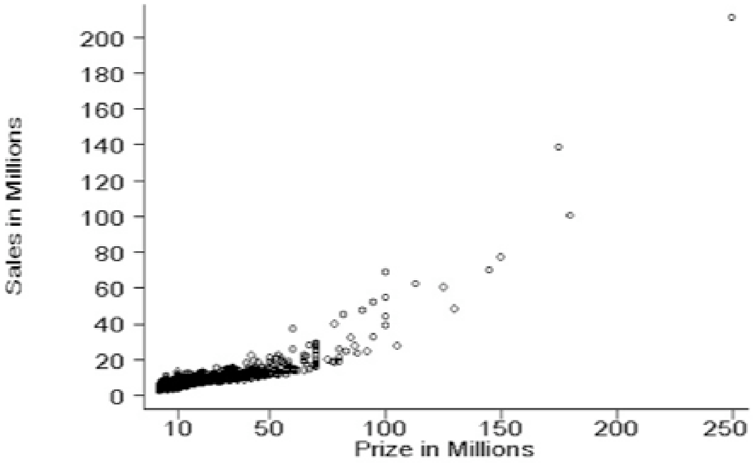


Figure 3.2: Powerball Sales vs. Jackpot Size.

dollar is \$.057, which is slightly worse than it was before. par Perhaps we have now explained the famous quote:

> "The lottery: A tax on people who flunked math."
>
> – Monique Lloyd

Discussion Question: Suppose that someone buys two lottery tickets, and fills them out with the same six numbers. How does this affect his expected value?

### 3.2.4 Is the expected value of a ticket every greater than \$1?

We saw above that even in the record jackpot drawing of December 25, 2002, the expected value of a ticket was significantly less than \$1. We now consider whether, for certain jackpot sizes, the expected value of a ticket ever exceeds \$1. To do this, we need to be able to estimate the number of tickets sold as a function of the jackpot size. This has been done for the Powerball lottery, by Emily Oster, in her senior thesis at Harvard.[2] Figure 3.2 shows the data for the Powerball lottery from 1992 to 2000.

Oster used a log-linear fit to arrive at the following equation. In this equation, $s_i$ denotes the number of tickets sold in the $i$'th lottery, in millions, and $p_i$ denotes the announced size of the $i$'th jackpot, in millions:

$$\log(s_i) = 15.558 + .016p_i \ ,$$

or equivalently,

$$s_i = (5,712,000)(1.0161)^{p_i} \ .$$

[2] Emily Oster, Dreaming Big: Why Do People Play the Powerball?, Harvard University, March 2002.

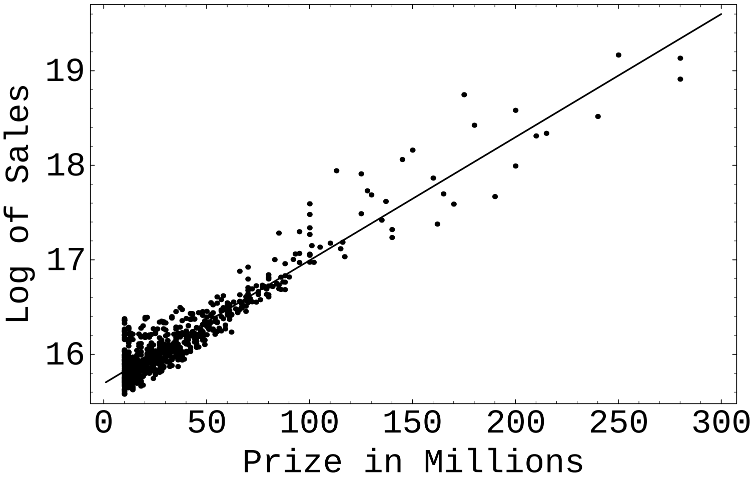


Figure 3.3: Log-Linear Fit of Powerball Sales vs. Jackpot Size.

Thus, for example, if the jackpot is \$100,000,000, we let $p_i$ equal 100, and find that $s_i = 28,210,000$. Figure 3.3 shows this log-linear fit.

It is interesting to compare Figure 3.2 with the corresponding data from the United Kingdom lottery, shown in Figure 3.4. In the latter figure, there seems to be almost no correlation between the size of the jackpot and the number of tickets sold.

There is no reason to think that the above equation will be accurate for announced jackpot sizes larger than about $\$300,000,000$, since there are no data for jackpots larger than this size. In fact, if the announced jackpot size were $\$500,000,000$, say, then the above formula predicts that the number of tickets sold would be more than 4 billion, or more than 10 tickets for each person in the United States. This is clearly unreasonable. Thus, in what follows, we will initially assume that the announced jackpot size is no larger than $\$300,000,000$.

We can now proceed as we did in the preceding subsection. Given an announced jackpot size $j$, we can estimate, using Oster's equation above, the number of tickets that will be sold. Then, using the Poisson distribution, we can calculate the expected pre-tax value of a jackpot-winning ticket. Next, we add 17.3 cents to this value; this represents the contribution to the ticket's expected value by the prizes other than the jackpot. Finally, we multiply the result by .72, obtaining the after-tax expected value of the ticket. Figure 3.5 shows the after-tax expected value of a ticket, as a function of the announced jackpot size. Note that the expected after-tax value never exceeds 65 cents.

As we said above, we cannot extrapolate our estimate of the number of tickets sold past $\$300,000,000$ with any assurance of accuracy. Suppose that we assume the number of tickets sold has a certain maximum value $t$, no matter how large the announced jackpot size is. Given $t$, how large must the announced jackpot size be so that the expected after-tax value of a ticket will exceed \$1?

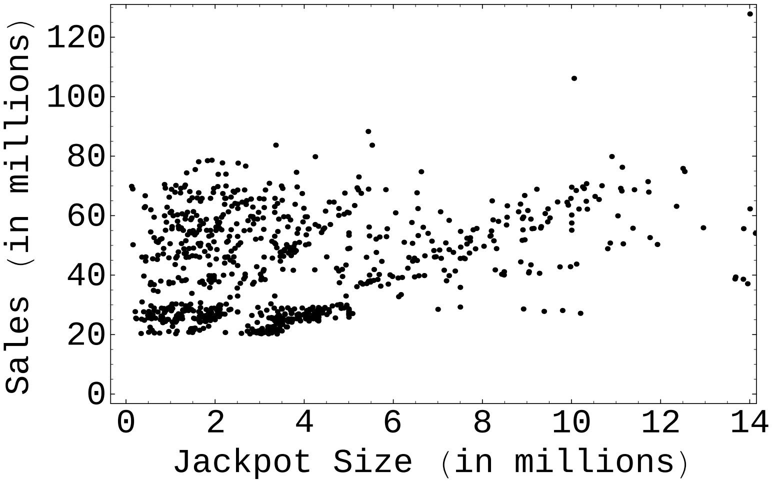


Figure 3.4: United Kingdom Sales vs. Jackpot Size.

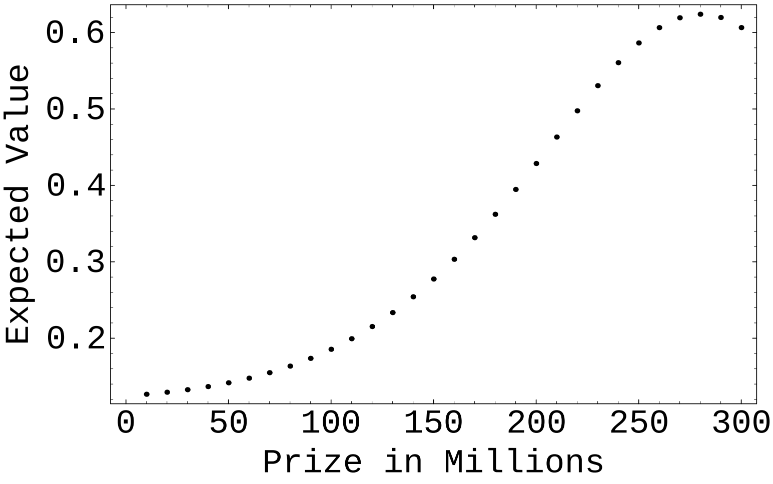


Figure 3.5: Expected after-tax value of a ticket vs. announced jackpot size.

This question is easy to answer, using the same methods as we used above. Note that for fixed $t$, the expected number of jackpot winners is constant, so the expected after-tax value increases linearly in the announced jackpot size. If $t = 200,000,000$, then the expected after-tax value of a ticket equals \$1 if the announced jackpot size is $\$475,000,000$; if $t = 500,000,000$, the corresponding announced jackpot size is $\$783,000,000$. These examples have jackpot sizes that far exceed any Powerball jackpot that has ever occurred. Thus it is safe to say that under almost no conceivable circumstance would the after-tax value of a ticket exceed \$1.

**************Add story about Celestino Mendez?

### 3.2.5 What kind of numbers do lottery buyers choose?

We have suggested that we might at least be able to avoid sharing the jackpot with people who choose their own numbers if we choose our own cleverly. It is well known that people who choose their own numbers do not choose randomly. They use their child's birthday, their "lucky" numbers, arithmetic progressions such as 2-4-6-8-10-12, numbers from sports, etc.

To see what players actually do, we obtained the numbers chosen by players in the Powerball Lottery in one state on May 3, 1996. Recall that at this time the game was played by selecting five of 45 white balls and one of 45 red balls. On this day, 17,001 of the picks were chosen by the buyers, and 56,496 were chosen by the computer (Easy Pick). Thus only about 23% of the picks were chosen by the buyers.

We first compare the distribution of the individual numbers from the picks chosen by the computer and those chosen by the buyers. To make the two sets the same size, we use only the first 17,001 picks produced by the Easy Pick method. Each pick contributed 6 numbers, so in both cases we have 102,006 numbers between 1 and 45. Figure 3.6 is a plot of the number of times each of the numbers from 1 to 45 occurred for the picks chosen by the computer. There does not seem to be very much variation, but it is worth checking how much variation we would expect if the numbers were, in fact, randomly chosen. If they were randomly chosen, the numbers of occurrences of a particular number, say 17, would have a binomial distribution with $n = 102,006$ and $p = 1/45$. Such a distribution has mean $np$ and standard deviation $\sqrt{npq}$, where $q = 1 - p$. This gives values for the mean and standard deviation of 2267 and 47.

It is hard to tell the actual differences from the graph, so we looked at the actual data. We found that, for all but two numbers, the results were within two standard deviations of the mean. For the other 2 numbers the results were within 3 standard deviations of the mean. Thus the picks chosen by the computer do not appear to be inconsistent with the random model. A chi-square test would give a way to proceed with a formal test of this hypothesis.

We look next at the picks chosen by the players. Recall that we have the same number 17,001 of picks so we again have 85,005 individual numbers. Here the numbers are in increasing order of frequency of occurrence:

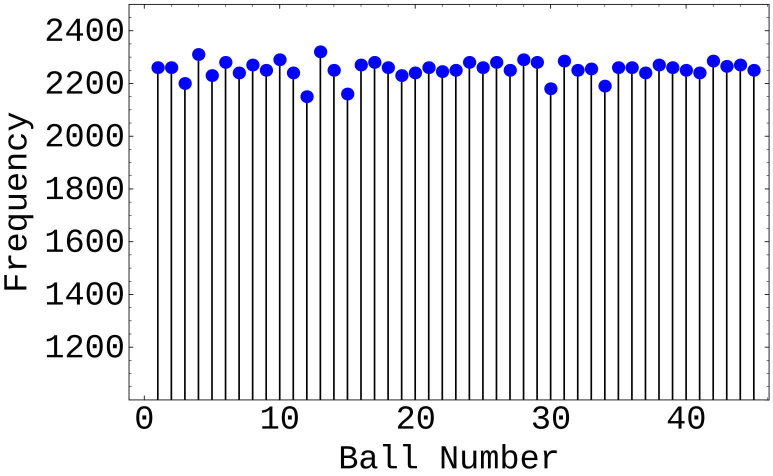


Figure 3.6: Frequencies of numbers chosen by computer.

| Number | Probability | Number | Probability | Number | Probability |
|---|---|---|---|---|---|
| 37 | 0.010 | 29 | 0.020 | 1 | 0.026 |
| 38 | 0.011 | 28 | 0.020 | 22 | 0.026 |
| 43 | 0.012 | 31 | 0.020 | 13 | 0.026 |
| 45 | 0.012 | 18 | 0.022 | 23 | 0.027 |
| 39 | 0.012 | 30 | 0.023 | 6 | 0.028 |
| 44 | 0.012 | 19 | 0.023 | 2 | 0.029 |
| 41 | 0.013 | 27 | 0.023 | 10 | 0.029 |
| 36 | 0.013 | 24 | 0.024 | 4 | 0.029 |
| 42 | 0.014 | 14 | 0.024 | 8 | 0.030 |
| 34 | 0.014 | 26 | 0.024 | 12 | 0.030 |
| 40 | 0.015 | 16 | 0.024 | 11 | 0.030 |
| 32 | 0.015 | 17 | 0.024 | 3 | 0.033 |
| 35 | 0.016 | 21 | 0.025 | 5 | 0.033 |
| 33 | 0.018 | 15 | 0.025 | 9 | 0.033 |
| 20 | 0.019 | 25 | 0.026 | 7 | 0.036 |

Table 3.2.5. Observed probabilities for numbers chosen by players.

These frequencies are plotted in Figure 3.7. You don't have to do any fancy tests to see that these are not randomly chosen numbers. The most popular number 7 was chosen 3,176 times, which would be 19 standard deviations above the mean if the numbers were randomly chosen!

It is often observed that people use birthdays to choose their numbers. If they did, we would expect numbers from 1 to 12 to be most likely to be picked since such numbers can occur both in the month and the day. The next most likely numbers to be picked would be those from 13 to 31 where the remaining days of the months could occur. The least likely numbers would then be those from 32 to 45 where the

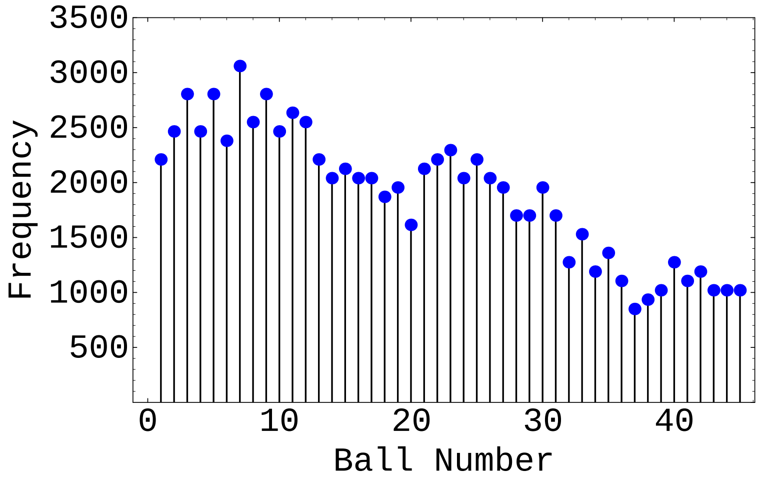


Figure 3.7: Frequencies of numbers chosen by players.

year of the birthday could occur but only for those at least 52 years old. Note that this is indeed what appears to be happening.

Discussion Questions:

1. We have seen that the numbers picked by the players fall into three sets, with the property that numbers in the same set are approximately equally likely to be chosen, but numbers in different sets are not equally likely to be chosen. Let us denote the three sets by

$$S_1 = \{1, 2, \ldots, 12\} ,$$
$$S_2 = \{13, 14, \ldots, 31\} ,$$
$$S_3 = \{32, 33, \ldots, 45\} .$$

The collection of all sets of five unequal numbers, between 1 and 45, written in increasing order, serves as the sample space of all possible choices by the players. The numbers in $S_1$, $S_2$, and $S_3$ occur with average frequencies .0306, .0234, and .0134, respectively. Using these frequencies, show that if a chosen set of numbers has all of its numbers in $S_1$, then it occurs with probability $2.7 \times 10^{-8}$, while if all of its numbers are in $S_3$, then it occurs with probability $4.3 \times 10^{-10}$. How does this affect the expected value of these two tickets, i.e. is there a difference between the expected values of tickets of the above two types?

2. Using the above data from the December 25, 2002 lottery, we saw that the expected value of one lottery ticket, after taxes, is \$.607. Suppose someone buys two tickets and puts the same numbers on both tickets. What is the expected value of each ticket?

Finally, we look at the winning numbers to see if they could be considered to be randomly chosen. Recall that the lottery officials try to ensure that they are.

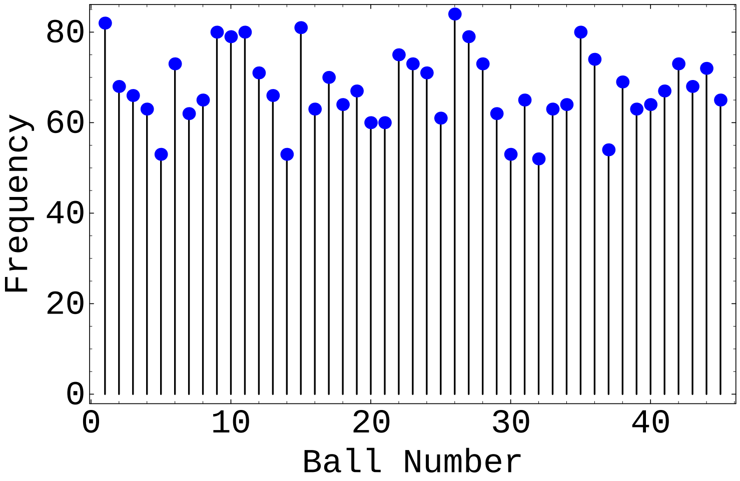


Figure 3.8: Winning number frequencies.

Here we have many fewer numbers so we expect more variation even if randomly chosen. Since there were 508 drawings in the period we are considering, we have $6 \cdot 508 = 3048$ numbers. Now, if the numbers are randomly chosen, the number of times a particular numbers occurs has a binomial distribution with $n = 3048$ and $p = 1/45$. Such a distribution has a mean of 67.7 and standard deviation 8.14. The biggest deviations from the mean are about 2 standard deviations so this looks consistent with the hypothesis that the numbers were randomly chosen. Again, we could check this with a chi-square test.

### 3.2.6 Finding patterns

Recall that players choose their first five numbers to match the white balls separately from their choice of the Powerball number to match the red ball. Thus, if we are looking for patterns in the way people choose their numbers, it is best to consider the first five numbers by themselves. We recall that our data set consists of two sets of picks of size 17,001; the first set contains picks chosen by the Easy Pick method, and the second set contains picks chosen by the players. For the Easy Picks, we found that 136 of these were represented twice and 2 were represented 3 times.

To see if we should have expected to find 3 picks the same, we use the solution of the birthday problem, which is a well-known problem in probability. The most basic form of this problem asks for the probability that at least two people among a set of $k$ people will share a birthday. Part of the reason that this problem is interesting stems from the fact that if one asks how many people are needed in a room in order to have a fair bet that at least two of these people have the same birthday, then the surprising answer is that only 23 people are needed.

In our case, we are asking a more difficult question: Given 17,001 choices from a set of size $C(45, 5) = 1,221,759$ (this is the number of possible choices of five numbers from 45 numbers), what is the probability that at least three of the choices are equal? So, instead of 366 possible birthdays, there are now 1,221,759 possible birthdays. It can be calculated that in this case, the probability of finding three or more choices the same is about .42. To see how this calculation can be done, see[3]. One can also calculate that there is only a probability of .002 of finding 4 or more the same birthday. Thus we should not be surprised at finding 3 picks the same and should not expect to find 4 the same. Again, the computer picks seem to conform to random choices.

We look next at the 17,001 picks of 5 numbers chosen by the lottery players. We found 966 sets of numbers that were represented more than once (compared to 138 for the Easy Pick numbers). The largest number of times a particular set of numbers was chosen was 24. This occurred for the pick 02-14-18-21-39. Looking at the order in which the picks were given to us, we noticed that these occurred consecutively in blocks of 5, with the blocks themselves close together. The ticket on which you mark your numbers allows room for 5 sets of numbers. We concluded that one player had made 24 picks all with the same five numbers. He at least chose different Powerball numbers. The same explanation applied to the next most popular pick 08-12-24-25-27, which occurred 16 times.

The third most popular set 03-13-23-33-43 was picked by 13 people and was more typical of the patterns that people chose. In this version of Powerball, the numbers were arranged on the ticket as shown below:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Pick 5 | | | | | | | EP___ | |
| 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 |
| 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 |

Note that the pick 03-13-23-33-43 is an arithmetic sequence obtained by going down a diagonal starting with 03. Similarly, the set of numbers 01-11-21-31-41, which was chosen 10 times, corresponds to going down a diagonal starting with 01 and the set 06-15-24-33-42, chosen 9 times, corresponds to going down a column starting with 06. The most interesting pattern noticed was 01-09-23-37-45, occurring 8 times, which results from choosing the corner points and the middle point. Since we do not expect repetitions of 4 or more to occur by chance, we looked at all those that occurred 4 or more times. We could explain all but three such sets of 5 numbers. These were:

01-03-09-30-34 (occurred 5 times, always with Powerball number 40)
05-06-16-18-23 (occurred 4 times, always with Powerball number 31)
02-05-20-26-43 (occurred 4 times with different Powerball numbers) .

Here are two letters that appeared in *The Times* (London) related to the problem of people choosing popular numbers. The letters followed an article in *The Times*

[3] MathWorld, "Birthday Problem," http://mathworld.wolfram.com/BirthdayProblem.html

stating that the inaugural drawing of the new British Lottery had five times the number of winners expected, including seven people who had to share the jackpot. They blamed this on the fact that the six winning numbers 03-05-14-22-30-44 had five numbers under 31 and most people chose low numbers. In this lottery, you choose 6 numbers between 1 and 49 and have to get them all correct to win the jackpot. If you get three numbers correct you win £10. The amount you win for any other prize depends on the number of other people who win this prize.

> The Times, 24 November 1994, letters to the editor.
>
> Slim pickings in National Lottery
>
> From Mr. George Coggan
>
> Sir, With random choices, the odds against there being seven or more jackpot winners in the National Lottery when only 44 million tickets have been sold are 23-1. This suggests that those who predicted that low numbers would be popular were right as the slightly disproportionate number of single digits (3 and 5 came up) would combine to produce more winners than would be produced by entirely random selections.
>
> Mildly interesting, one might think, but then one suddenly realizes that there is a lurking danger that the rules create the possibility that when (as will happen sooner or later) three single digit numbers come up the prize fund may not be enough to cover the Pounds 10 guaranteed minimum prize, never mind a jackpot. I estimate that if the number 7 had come up instead of say 44 the prize fund in this first lottery would have been about Pounds 5 million short of the guarantee. What then panic?
>
> Yours sincerely,
>
> GEORGE COGGAN,
>
> 14 Cavendish Crescent North,
>
> The Park, Nottingham.
>
> November 21. 1994 Tuesday

> The Times, 29 November 1994, letter to the editor.
>
> No need to fear a lottery shortfall
>
> From the Director General of the National Lottery
>
> Sir, Mr. George Coggan (letter, November 24) raises concerns about the National Lottery Prize Fund's ability to pay winners when ”popular” numbers are selected in the weekly draw.
>
> We are aware that players do not choose numbers randomly but use birthdays, sequences or other lucky numbers. This causes the number

> of winners to deviate each week from the number predicted by statistical theory. Experience from other lotteries shows that the number of winners of the lower prizes can vary by up to 30 per cent from the theoretical expectation.
>
> In the first National Lottery game there were many more Pounds 10 prize-winners than theory predicted. It is just as likely that future draws will produce fewer than expected winners and, because each higher prize pool is shared between the winners, prize values will rise accordingly.
>
> Mr. Coggan suggests a pessimistic scenario in which the cost of paying the fixed Pounds 10 prizes to those who choose three correct numbers exceeds the prize fund. Best advice, and observations from other lotteries around the world, is that, even after allowing for the concentration on "popular" numbers, the chances of this happening are extremely remote.
>
> Your readers will be reassured to know, however, that I have not relied totally upon statistics or evidence from other lotteries. Camelot's license to operate the National Lottery also requires them to provide substantial additional funds by way of deposit in trust and by guarantee to protect the interests of the prize-winners in unexpected circumstances.
>
> Yours faithfully,
>
> PETER A. DAVIS
>
> Director General, National Lottery
>
> PO Box 4465
>
> London SW1Y 5XL
>
> November 25.

Of course, it is interesting to look at this problem for the Powerball lottery. We noted that, in our sample of 17,001 numbers where players picked their own numbers, there were particular sets of five numbers for the white balls that were chosen as many as 10 times. For example the set of numbers 01-11-21-31-41 obtained by going down a diagonal starting at 1 in the box where you mark your numbers was chosen 10 times in our sample of 17,001.

The July 29, 1998 lottery set a new record for the jackpot size (this record was broken on December 25, 2002, as noted earlier). For the July 29, 1998 drawing there were 210,800,000 tickets sold. If we assume that about 30% of the players pick their own numbers, then using the above example, it is possible that the there exists a set of white numbers that was picked by

$$\frac{.3(210800000)}{1700} = 37200$$

players. It will be recalled that a player who picks all five white numbers gets \$100,000, so if the lottery officials had the bad luck to also choose this set of five

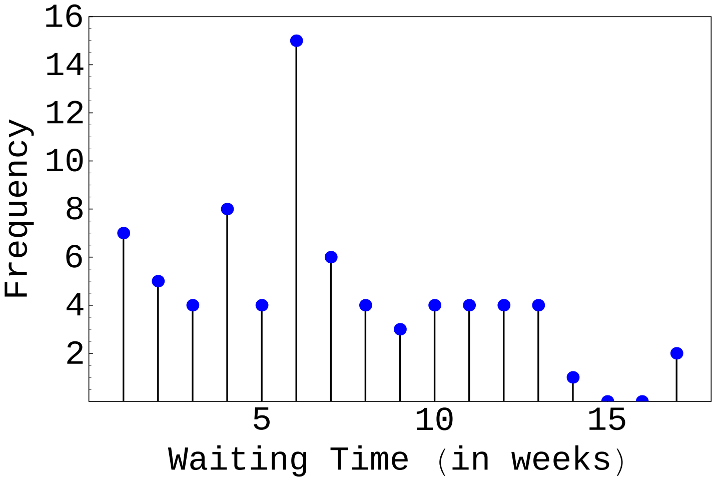


Figure 3.9: Waiting times between jackpots.

numbers this would cost them 3.72 billion dollars! The new boxes are not as symmetric as the old ones to which our data applied. This may help them with this potential problem. It is also the case that the lottery computer is very unlikely to choose a popular set of numbers. However, the lottery is protected in still another way. In the fine print that describes the lottery, it is stated that if there are too many awards won, the officials reserve the right to divide a certain amount of money among the prize winners, just like what is done with the jackpot. This last event has never happened in the history of the Powerball lottery.

### 3.2.7 How often is the jackpot won?

The size of the jackpot changes from one drawing to the next. If, on a given drawing, no one chooses the winning numbers, the jackpot is increased several million dollars for the next drawing. When there is a winner, the jackpot goes back to the minimum amount, which currently is 10 million dollars. The size of the increase when there is no winner depends upon the number of tickets sold for the previous drawing. We investigate the size of the jackpots through the years of the original rules.

We find, on the Powerball homepage, the amounts of the jackpot in all the drawings under the original rules. The jackpots, from the beginning of the Powerball lottery on April 22, 1992 until March 1, 1997, range from 2 million to 111,240,463. The jackpot was won in 75 of these 805 drawings. 11 of these times there were two winners and never more than two winners. The total of all these jackpots was $2,206,159,204 with an average of $29,415,456. The average number of drawings between jackpots being won was 6.72 or, since there are two drawing a week, about 3 weeks. The distribution of times between jackpots is shown in Figure 3.9.

*****************Add the data for the waiting times between 1997 and the present.

DISCUSSION QUESTION: The mode 6 seems rather over-represented. Can you think of any explanation for this, or is it just chance?

It is interesting to ask what kind of a probability model would be appropriate to describe the time between winning the jackpot. The probability of the jackpot being won depends upon the number of tickets sold. (Actually, it depends upon the number of different picks chosen.) If the same number of tickets were sold on each drawing then the appropriate model would be: toss a coin with probability $p$ for heads until the first time heads turns up where $1/p$ is the average time between heads. Unfortunately, it is not at all reasonable to assume the same number of tickets will be sold. Here is what the Powerball FAQ says about this.

> For a \$10 million jackpot draw we sell about \$11 million. For a \$20 million jackpot we sell about \$13 million. With a \$100 million jackpot we sell \$50 to \$70 million for the draw (depending on time of year and other factors).

Let's assume that, for a particular jackpot, $n$ tickets are sold. Then the probability that a particular person does not win the jackpot is $(a-1)/a$, where, for the old version of the game,

$$a = 45 \cdot C(45,5) = 54,979,155 \ .$$

The probability that none of the tickets sold wins the jackpot is

$$\left(\frac{a-1}{a}\right)^n \ .$$

Here is a table of these probabilities for some different values of the number of tickets sold.

| Millions of tickets sold | Probability no one wins |
|---|---|
| 10 | .834 |
| 20 | .695 |
| 30 | .579 |
| 40 | .483 |
| 50 | .403 |
| 60 | .336 |
| 70 | .280 |

Table 3.2.7. Probability no one wins the jackpot.

### 3.2.8 Other lotteries pose new questions

There are many other interesting questions that can be explored about lotteries. The questions that one asks depend, to some extent, on the nature of the lottery. For example, in September 1996 the Multi-State lottery introduced a new lottery called Daily Millions where the amount of the jackpot is always \$1 million and, if you win, you don't have to share it with another person who also has the same winning pick (actually, if there are more than 10 such winners they share a $10 million prize.) An article appeared in the Star Tribune shortly after this lottery was introduced[4]. The article began as follows:

> The lottery wizards said it wasn't supposed to work this way.
>
> Nearly five months after the Daily Millions lottery began, none of the 34 million tickets sold has won the $1 million jackpot. The probability of such a long losing streak is 1 in 38.
>
> The drought has lasted so long, "We even have solid believers in statistics questioning the wisdom of numbers," quipped Charles Strutt, executive director of the Multi-State Lottery Association, which runs Daily Millions and Powerball.
>
> They're not the only ones.
>
> At the East Grand Forks, Minn., Holiday store, "People are starting to get a little disgusted with it," said cashier Steve Nelson. The store has been among the top sellers of Powerball tickets, but sales of Daily Millions tickets have declined.

The article states that the ticket sales in the first week of the lottery were $2.75 million, but five months later, they had declined to $1.23 million.

The day after the above article appeared, the Daily Millions lottery had its first winner.

### 3.2.9 Using lottery stories to discuss coincidences

James Hanley[5] has discussed how stories about lottery winners provide good examples to discuss the meaning of apparent coincidences. Here is his first example[6].

> Lottery officials say that there is 1 chance in 100 million that the same four-digit lottery numbers would be drawn in Massachusetts and New Hampshire on the same night. That's just what happened Tuesday.
>
> The number 8092 came up, paying $5,482 in Massachusetts and $4,500 in New Hampshire. "There is a 1-in-10,000 chance of any four digit number being drawn at any given time," Massachusetts Lottery Commission official David Ellis said. "But the odds of it happening with two states at any one time are just fantastic," he said.

What is the probability that the same four-digit lottery number would be drawn in Massachusetts and New Hampshire on the same night? What is the probability

[4] Doyle, Pat, "Daily Millions Beats Odds: No One Wins – 5-Month Losing Streak Puzzles Even Statisticians," Star Tribune, Minneapolis, 7 Feb. 1997, pg. 1B.

[5] Hanley, James A. "Jumping to Coincidences, " The American Statistician, Vol 46, No. 3, pp 197-201.

[6] "Same Number 2-State Winner," Montreal Gazette, September 10, 1981.

that some two such lotteries have the same two numbers during a given period of time? Is this different from a reporter noticing that the number that turned up in the lottery in New Hampshire on Wednesday happened also to occur in the Massachusetts lottery on Saturday?

Here is another of Hanley's examples[7].

> Defying odds in the realm of the preposterous–1 in 17 million–a women who won $3.9 million in the New Jersey state lottery last October has hit the jackpot again and yesterday laid claim to an addition $1.5 million prize...
>
> She was the first two time million-dollar winner in the history of New Jersey's lottery, state officials said. They added that they had never before heard of a person winning two million-dollar prizes in any of the nation's 22 state lotteries.
>
> For aficionados of miraculous odds, the numbers were mind boggling: In winning her first prize last Oct. 24, Mrs. Adams was up against odds of 1 in 3.2 million. The odds of winning last Monday, when numbers were drawn in a somewhat modified game, were 1 in 5.2 million.
>
> And after due consultation with a professor of statistics at Rutgers University, lottery officials concluded that the odds of winning the top lottery prize twice in a lifetime were 1 in about 17.3 trillion–that is, 17,300,000,000,000.

Does it matter that she played the lottery many times, often buying more than one ticket? Again, are we talking about this happening somewhere, sometime? Should we ever believe that something with these odds has happened?

### 3.2.10 Lottery systems

Richard Paulson[8] observes that claims made about systems for improving your chances at lotteries illustrate important statistical concepts. For example, people claim that, by analyzing the historical data of winning numbers, it is possible to predict future winners. Indeed, lottery sites encourage this by making this historical data available. Sometimes the argument is simply that, when a particular number has not turned up as often as would be expected, then this number is more likely to come up in the future. This is often called the "gambler's fallacy" and all too many people believe it. The fact that it is not true is the basis for many beautiful applications of chance processes called martingales.

Paulson remarks that he particularly enjoys discussing the following system. Consider, the six winning numbers in the Powerball Lottery. If they occur randomly their sum should be approximately normally distributed with mean $6(1+45)/2 =$

[7] "Odds-Defying Jersey Woman Hits Lottery Jackpot 2nd Time," New York Times, February 14, 1986.

[8] Paulson, Richard A. "Using Lottery Games to Illustrate Statistical Concepts and Abuses", The American Statistician , Vol. 45, No. 3, pp. 202-204.

138 and standard deviation approximately 32. Thus, sets of six numbers whose sum is more than 64 away from the mean 138, are not likely to occur as winning sets and should be avoided. It is better to pick six numbers whose sum is near 138. We leave the reader to ponder this last system.

One of our teachers, the well-known probabilist Joseph Doob, was often asked for a system to use when playing roulette. His advice ran as follows. Play red once. Then wait until there have been two blacks and play red again. Then wait until there have been three blacks and play red again. Continue in this manner. You will enjoy playing and not lose too much.

# Chapter 4

# Scaling in Human Endeavor

## 4.1 Zipf's Law

In 1916, the Frence stenographer J. B. Estoup[1] noted that if one listed words in a text according to the frequency of their occurrence, the product of frequency ($F$) and rank ($R$) of the words is approximately constant. Pereto[2] observed the same phenomena relating to corporate wealth. In his book[3], George Zipf showed that this constant product law applied to the ranking of populations of cities and to many other data sets. This became known as Zipf's law.

One can check this law on large sets of text, using computers. The British National Corpus (BNC) is a 100-million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. If the product of rank and frequency is a constant, then a log-log plot of frequency versus rank should be a straight line with slope -1. (To see why this should be so, assume that

$$F = \frac{c}{R} \ ,$$

where $c$ is a constant. Then

$$\log F = \log c - \log R \ ,$$

which is a straight line with slope -1.)

The log-log plot of the data is shown in Figure **??**. The line with slope -1 fits the higher part of the graph corresponding to the most common words. But to fit the lower part of the graph, corresponding to words used infrequently, we had to choose a line with larger negative slope -1.64. The need for different lines forthe frequent and the infrequent words was suggested by Ferrer-i-Cancho and Sole in their paper

[1] J. B. Estoup, Gammes Stenographiques, Intstitut Stenographicque de France, Paris, 1916.

[2] Cours d'economie politique, Drox, Geneva, Switzerland, 1896 and Rouge, Lausanne et Paris, 1897

[3] G. K. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley, 1949

"Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited."[4]

An interesting application of Zipf's law was carried out by Gabriel Landini. The Voynich manuscript is a famous manuscript shrouded in mystery. this mystery is described by Victory Segal in her brief account of a BBC4 documentary in the Sunday Times.[5] She writes:

> While it is easy to mock conspiracy theorists as loners in X Files T-shirts, this documentary explores a puzzle whose appeal stretches way beyond the glow of the internet chat room. The Voynich manuscript was discovered in an Italian villa by an antiquaian book dealer, Wilfred Voynich, early last century. Decorated with pictures of plants, stars, and "naked nymphs," its 238 pages are written in an unreadable language, a peculiar grid of symbols that have resisted all attempts at translation.
>
> Some think it is a code, others a language in its own right. It might come from the 13th century or it might be a modern forgery, perpetrated by Voynich himself. This is a perfect introduction to the manuscript and its lore, veering between theories at a dizzying rate. Roger Bacon, John Dee, and Emperor Rudolph II all play their part, as do National Security Council cryptographers, Yale maths professors and an array of eccentric amateurs, from computer games programmers to herbalists. And, of course, there is a man in South Carolina who thinks extraterrestrials hold the key.

One can find more details of the history and current research on the web. We enjoyed the discussion here which has several nice pictures from the manuscript including the one in Figure **??**.

Several electronic versions of the Voynich manuscript which use arbitrary alphabets to transcribe the symbols are available. These have been used to study statistical properties of the manuscript. In particular, Landini used two of these to see if Zipf's law applied to the frequencies of the Voynich text. His results are described here. Figurer̃effig shows a plot that he gives for the rank-frequency relation using a translation called voynich.now. For comparison, Landini gives similar plots (shown in Figure **??**) for several well-known text sources including Alice in Wonderland.

Of course, it would be nice if one could conclude that the Voynich manuscript is not a hoax from the fact that the frequencies of words in the Voynich manuscript satisfy Zipf's law. However, it may not be correct to conclude this, since it has been verified by a number of studies that words generated randomly also appear to satisfy Zipf's law. ********Put in description of power laws and Polya models.

Zipf did not give a mathematical formulation but, as the title of his book suggests, he had the idea that it was related to a principle of least effort. For the case of development of a language Zipf wrote:

[4] Ferrer-i-Cancho, R. and Sole, R. V., Journal of Quantitative Linguistics, 8 (3), pgs. 165-173.

[5] Segal, Victory, London Times, 8 Dec. 2002

> Whenever a person uses words to convey meanings he will automatically try to get his ideas across most efficiently by seeking a balance between the economy of a small wieldy vocabulary of more general reference on the one hand, and the economy of a larger one of more precise reference on the other hand, with the result that the vocabulary of different words in his resulting flow of speech will represent a vocabulary balance between our theoretical Forces of Unification and Diversification.

## 4.2 Lotka Distributions

## 4.3 Appendix

In this section, we show how to find the limiting distribution of one of the Polya models. We begin with a description of the model.

Description of Model: Start with one bin with one ball at time $t = 1$. We denote this first ball as $b_1$, and set $b_1 = 1$. At every integer step $t$, starting with $t = 2$, add one ball $b_t$. The value of $b_t$ will be the number of the bin into which the ball is placed. With probability $p$, the ball $b_t$ goes into a new bin, so in this case,

$$b_t = 1 + \max_{j<t} b_j \ .$$

With probability $(1 - p)$, we pick a ball $b_j$ uniformly over $j < t$, and add $b_t$ to the bin that contains $b_j$; in other words, in this case we set $b_t = b_j$.

For each $i \geq 1$, define $f_{i,t}$ to be the random variable whose value is the fraction of bins at time $t$ that contain exactly $i$ balls. We want to show that for each $i \geq 1$,

$$\lim_{t \to \infty} f_{i,t}$$

exists with probability 1.

Define $n_{i,t}$ to be the random variable whose value is the number of bins at time $t$ with exactly $i$ balls, and define $n_t$ to be the random variable representing the number of bins at time $t$ with at least one ball. Clearly,

$$f_{i,t} = \frac{n_{i,t}}{n_t} \ .$$

Consider first the case $i = 1$. At time $t + 1$, with probability $p$ we add $b_{t+1}$ to a new bin, which means that $n_{1,t+1} = n_{1,t} + 1$. Otherwise, there are $n_{1,t}$ balls in bins with one ball, and $t$ balls already placed, so with probability $(1 - p)(n_{1,t}/t)$, we put $b_{t+1}$ in a bin that contains one ball; in this case, $n_{1,t+1} = n_{1,t} - 1$. The final possibility is that we put $b_{t+1}$ in a bin that contains more than one ball; in this case, $n_{1,t+1} = n_{1,t}$. Thus, we have

$$E(n_{1,t+1}) \quad = \quad E\Big(p(n_{1,t} + 1) + (1 - p)\big(\frac{n_{1,t}}{t}(n_{1,t} - 1) + \big(1 - \frac{n_{1,t}}{t}\big)n_{1,t}\big)\Big)$$

$$= E\biggl(n_{1,t} + p - (1-p)\frac{n_{1,t}}{t}\biggr) .$$

Using standard generating function techniques, one can show that for $t \geq 1$,

$$\begin{aligned} E(n_{1,t}) &= \frac{p}{2-p}t + \frac{2t}{p-2}(-1)^t\frac{1-p}{t} \\ &= \frac{p}{2-p}t + o(t) . \end{aligned}$$

Since $n_t$ is counting the number of new bins up to and including time $t$, together with the original bin, we see that

$$E(n_t) = 1 + p(t-1) \sim pt .$$

Of course, one cannot conclude from these statements about expected value that

$$E(f_{1,t}) \sim \frac{1}{2-p} ,$$

although one could conjecture that this is true. In order to prove that such a statement is true, it suffices to show that both of the sets of random variables $\{n_{i,t}\}$ and $\{n_t\}$ are usually very close to their mean. This would show that with high probability, the ratio $n_{i,t}/n_t$ is near the ratio of the means. The sequence $\{n_t\}$ is essentially a sequence of binomially distributed random variables, so it is easy to deal with. The sequence $\{n_{i,t}\}$ is a little more difficult. We will consider how to do this once we have determined what the limit of $f_{i,t}$ should be.

If $i \geq 2$, we have

$$\begin{aligned} E(n_{i,t+1}) &= E\biggl(pn_{i,t} + (1-p)\biggl(\frac{in_{i,t}}{t}(n_{i,t}-1) + \frac{(i-1)n_{i-1,t}}{t}(n_{i,t}+1) + \\ &\quad \biggl(1 - \frac{(i-1)n_{i-1,t} + in_{i,t}}{t}\biggr)n_{i,t}\biggr)\biggr) \\ &= \biggl(1 - \frac{i(1-p)}{t}\biggr)E(n_{i,t}) + \frac{(i-1)(1-p)}{t}E(n_{i-1,t}) . \end{aligned}$$

Suppose that we have shown that

$$E(n_{i-1,t}) \sim K_{i-1}t$$

for some constant $K_{i-1}$. (We have shown that $K_1 = p/(2-p)$.) If there is a constant $K_i$ such that

$$E(n_{i,t}) \sim K_i t ,$$

what would the value of $K_i$ be? One can show that

$$K_i \sim i(1-p)(K_{i-1} - K_i) - (1-p)K_{i-1} ,$$

which implies that for $i \geq 2$,

$$K_i = \frac{(i-1)(1-p)}{1+i(1-p)} K_{i-1} \ .$$

Since we don't know that $K_i$ exists for $i \geq 2$, let us define $K_i^*$ by the equations

$$K_1^* = \frac{1}{2-p}$$

and for $i \geq 2$,

$$K_i^* = \frac{(i-1)(1-p)}{1+i(1-p)} K_{i-1}^* \ .$$

(We would like to show that $K_i$ exists and equals $K_i^*$.) If any of this is right, then it must be true that

$$\sum_{i=1}^{\infty} K_i^* = p \ ,$$

since the left-hand side, multiplied by $t$, is asymptotically the number of bins filled with balls at time $t$, which we know is close to $pt$ with high probability.

Here are the first few values of $K_i^*$:

$$\begin{aligned} K_1^* &= \left(\frac{p}{2-p}\right) \\ K_2^* &= \left(\frac{1-p}{3-2p}\right)\left(\frac{p}{2-p}\right) \\ K_3^* &= \left(\frac{2-2p}{4-3p}\right)\left(\frac{1-p}{3-2p}\right)\left(\frac{p}{2-p}\right) \ . \end{aligned}$$

The sum of this series can be calculated using hypergeometric functions. Specifically, the hypergeometric function $F(\alpha, \beta, \gamma; z)$ is defined as follows:

$$\begin{aligned} F(\alpha, \beta, \gamma; z) &= 1 + \frac{\alpha\beta}{\gamma 1!} z + \frac{(\alpha)(\alpha+1)(\beta)(\beta+1)}{(\gamma)(\gamma+1)2!} z^2 + \\ &\quad \frac{(\alpha)(\alpha+1)(\alpha+2)(\beta)(\beta+1)(\beta+2)}{(\gamma)(\gamma+1)(\gamma+2)3!} z^3 + \dots \ . \end{aligned}$$

It can be shown that

$$F(\alpha, \beta, \gamma; z) = \frac{1}{B(\beta, \gamma-\beta)} \int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-tz)^{-\alpha} \, dt \ ,$$

if $Re(\gamma) > Re(\beta) > 0$. In the above equation, $B(x, y)$ is the beta function, and is defined by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} \, dt \ .$$

It is relatively straightforward to show that

$$\sum_{i=1}^{\infty} K_i^* = \frac{p}{2-p} F(1, 1, 2 + 1/(1-p); 1) \ ,$$

and it is easy to calculate that the right-hand side of the above equation equals $p$.

We can now proceed to the general case. Let

$$a_{i,t} = E(n_{i,t}) \ ,$$

and let

$$b_{i,t} = \frac{a_{i,t}}{t} \ .$$

Define

$$g_i(x) = b_{i,1}x + b_{i,2}x^2 + \dots \ .$$

We have already shown that

$$g_1(x) = \frac{p}{2-p}(1-x)^{-1} + 1 - \frac{2}{2-p}(1-x)^{1-p} \ .$$

(This is the equation that leads to our expression for $E(n_{1,t})$.) In general, it is the case that

$$g_i(x) = A_i(1-x)^{-1} + B_i + \sum_{j=1}^{i} C_{i,j}(1-x)^{j(1-p)} \ ,$$

where the coefficients on the right-hand side are defined by

$$A_i = \frac{(i-1)(1-p)}{1+i-ip}A_{i-1} \ ,$$

if $i > 1$, and $A_1 = p/(2-p)$;

$$B_i = \frac{i-1}{i}B_{i-1} \ ,$$

if $i > 1$, and $B_1 = 1$;

$$C_{i,j} = (i-1)\frac{C_{i-1,j}}{i-j} \ ,$$

if $i > 1$ and $j < i$;

$$C_{i,i} = -\frac{(i-1)(1-p)}{(1+i-ip)}A_{i-1} - \frac{(i-1)(1-p)}{(i-ip)}B_{i-1} - (i-1)\sum_{j=1}^{i-1}\frac{C_{i-1,j}}{i-j} \ ,$$

if $i > 1$, and $C_{1,1} = -2/(2-p)$.

Where does this leave us? The above can be used to show that for all $i \geq 1$,

$$E(n_{i,t}) \sim A_i t = K_i^* t \ ,$$

which is what we conjectured is true.

Now, following Bollobas et al., we fix $i \geq 1$ and $n \geq 1$, and define a sequence of random variables $\{X_t\}_{t=1}^n$ as follows:

$$X_t = E(n_{i,n}|b_1, b_2, \dots, b_{t-1}) \ .$$

This sequence is a martingale. We want to show that there is a $c$ such that

$$|X_{t+1} - X_t| \leq c$$

for $1 \leq t < n$. The reason that this is useful is the following theorem, due to Azuma[6] and Hoeffding[7].

**Theorem 4.1** (Azuma-Hoeffding inequality) Let $\{X_t\}_{t=1}^n$ be a martingale with $|X_{t+1} - X_t| \leq c$ for $t = 1, \ldots, n$. Then

$$P\big(|X_n - X_1| \geq x\big) \leq \exp\left(-\frac{x^2}{2c^2 n}\right) .$$

□

Here is an argument (modeled on the one in Bollobas et al.[8]) that shows that $c$ can be taken to be 2 in our model. At time $t+1$, we choose, with probability $p$, a new bin numbered $n_t + 1$ (in other words, we set $b_{t+1}$ equal to $n_t + 1$) or else we choose, with probability $(1-p)/t$, a $j$ with $1 \leq j \leq t$, and set $b_{t+1} = b_j$.

Now consider two possible choices that were made at time $t+1$ in the manner described above. We want to understand how much these two choices can affect the number $n_{i,n}$. So, for example, let's assume that one choice was a new bin $b^*$, and the other choice was a $b_j$, with $j \leq t$. Any subsequent choice in either case involving bins other than $b^*$ or $b_j$ will proceed with the same probability in either case. Thus, the only two bins which might have a different number of balls in the two cases are bins $b^*$ and $b_j$. This means that in the two cases, the values of $n_{i,n}$ differ by at most 2.

We can now apply the Azuma-Hoeffding inequality to the sequence $\{n_{i,t}\}$. We see that $X_1 = E(n_{i,n})$ and $X_n = n_{i,n}$, so the inequality implies that

$$P\big(|n_{i,n} - E(n_{i,n})| \geq x\big) \leq \exp\left(-\frac{x^2}{8n}\right) .$$

Let $\epsilon > 0$. The law of large numbers implies that

$$P\left(\left|\frac{n_n}{n} - p\right| < \epsilon\right) > 1 - \epsilon$$

for sufficiently large $n$ (say $n \geq n_0$). Thus, for all $n \geq n_0$, with probability greater than $1 - \epsilon$,

$$\frac{n_{i,n}}{(p+\epsilon)n} < \frac{n_{i,n}}{n_n} < \frac{n_{i,n}}{(p-\epsilon)n} .$$

Now choose $x$ so that

$$\frac{K_i + x}{p - \epsilon} < \frac{K_i}{p} + 2\epsilon$$

and

$$\frac{K_i - x}{p + \epsilon} > \frac{K_i}{p} - 2\epsilon .$$

[6] K. Azuma, Weighted sums of certain dependent variables, Tohoku Math J. 3 (1967), pgs. 357-367.

[7] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Stat. Assoc. 58 (1963), pgs. 13-30

[8] Bollobas, B., O. Riordan, J. Spencer, and G. Tusnady, The degree sequence of a scale-free random graph process,

If $n$ is sufficiently large, then

$$\exp\left(-\frac{x^2}{8n}\right) < \epsilon \ .$$

Thus, for all large $n$, with probability greater than $1 - 2\epsilon$,

$$\frac{K_i}{p} - 2\epsilon < \frac{n_{i,n}}{(p+\epsilon)n} < \frac{n_{i,n}}{n_n} < \frac{n_{i,n}}{(p-\epsilon)n} < \frac{K_i}{p} + 2\epsilon \ .$$

This implies that

$$\lim_{n\to\infty} f_{i,t} = K_i \ .$$

# Bibliography

[1] Jim Albert and Jay Bennett. *Curve Ball.* Copernicus Books, 2001.

[2] S. Christian Albright. A statistical analysis of hitting streaks in baseball. *Journal of the American Statistical Association*, 88(424):1175–1183, December 1993.

[3] Scott Berry. The summer of '41: A probabilistic analysis of DiMaggio's "streak" and Williams's average of .406". *Chance*, 4(4):8–11, 1991.

[4] Colin Camerer. Does the basketball market believe in the 'hot hand'? *The American Economic Review*, 79(5):1257–1261, 1989.

[5] Thomas Gilovich, Robert Vallone, and Amos Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17:295–314, 1985.

[6] Stephen Jay Gould. The streak of streaks. *Chance*, 2(2):10–16, 1989.

[7] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability.* American Mathematical Society, 1997.

[8] D. A. Jackson and K. Mosurski. Heavy defeats in tennis: Psychological momentum or random effect? *Chance*, 10(2):27–34, 1997.

[9] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains.* D. Van Nostrand and Company, 1960.

[10] James R. Lackritz. Two of baseball's great marks: Can they ever be broken? *Chance*, 9(4):12–18, 1996.

[11] Andrew Lo and Craig MacKinlay. *A Non-Random Walk Down Wall Street.* Princeton University Press, 1999.

[12] A. M. Mood. The distribution theory of runs. *The Annals of Mathematical Statistics*, 11(4):367–392, December 1940.

[13] Mark F. Schilling. The longest run of heads. *The College Mathematics Journal*, 21(3):196–207, May 1990.

[14] Tom Short and Larry Wasserman. Should we be surprised by the streak of streaks? *Chance*, 2(2):13, 1989.

[15] Gary Smith. Horseshoe pitchers' hot hands.

[16] V. K. Zaharov and O. V. Sarmanov. Distribution law for the number of series in a homogeneous Markov chain. *Soviet Math. Dokl.*, 9(2):399–402, 1968.