Notes on Cinemetric Data Analysis

Mike Baxter

September 2014
Contents

Preface viii

1 Cinemetrics and R 1
  1.1 Introduction 1
  1.2 Cinemetrics 2
    1.2.1 The idea 2
    1.2.2 Pattern recognition in cinemetrics 3
    1.2.3 Lies, damned lies and statistics (and fear) 3

2 Examples 5
  2.1 Preamble 5
  2.2 Examples 5
    2.2.1 ASLs over time 5
    2.2.2 Comparison of ASL distributions 6
    2.2.3 Pattern and SL distributions 8
    2.2.4 Internal pattern - individual films 10
    2.2.5 Internal pattern - an example of global analysis 11
    2.2.6 Shot-scale analysis 13

3 Getting R, getting started 21
  3.1 Finding R 21
  3.2 Data entry 21
    3.2.1 General 21
    3.2.2 Using the Cinemetrics database 22
  3.3 Packages 22
  3.4 Reading 23

4 Descriptive statistics 24
  4.1 Introduction 24
  4.2 Basics 24
  4.3 Functions 25
  4.4 Data manipulation 26
  4.5 Illustrative graphical analyses 27
  4.6 Definitions and comments on some simple statistics 31

5 Graphical analysis – basics 35
  5.1 Histograms 35
    5.1.1 Basics – an example 35
    5.1.2 Technicalities 36
    5.1.3 Example continued - log-transformation 37
  5.2 Kernel density estimates 38
  5.3 Boxplots 41
9.5.2 Analysis of individual shot-scale types

10 Large data sets
10.1 Introduction
10.2 Data entry and basic analysis
  10.2.1 Data entry
  10.2.2 Data analysis
  10.2.3 Dealing with several films at once
10.3 Further basic code
10.4 Hypothesis tests

11 Median or mean? (and more)
11.1 Introduction
11.2 Some theory
11.3 Outliers and robustness
  11.3.1 General remarks
  11.3.2 Outliers and Brief Encounter
11.4 Case studies
  11.4.1 ASL and MSL trends over time
  11.4.2 ASLs, MSLs and the silent-sound transition
  11.4.3 ASLs and MSLs in D.W. Griffith’s one-reelers

12 Lognormality of SLs – a case study
12.1 The lognormality issue
12.2 Exploratory analyses
  12.2.1 Graphical inspection
  12.2.2 Comparing estimates of the median
  12.2.3 Skewness and kurtosis I
  12.2.4 Skewness and kurtosis II
12.3 Omnibus tests of (log)normality
12.4 On distributional regularity – an alternative approach
12.5 R code

13 Multivariate methods – basics
13.1 Introduction
13.2 Data processing
  13.2.1 Brisson’s partitioning method
13.3 Smoothing
  13.3.1 General issues
  13.3.2 Scaling
  13.3.3 Degrees of smoothing
13.4 Distance measures
13.5 Principal component (PCA) and cluster analysis
  13.5.1 Principal component analysis
  13.5.2 Cluster analysis
13.6 R code for Brisson’s partition method

14 Structure in Griffith’s Biographs
14.1 Introduction
14.2 Initial cluster analyses
14.3 Principal component analysis (PCA)
14.4 Varying the level of smoothing
14.5 Discussion
List of Figures

2.1 ASLs for American films (1955-2005). 6
2.2 Histograms for the ASLs of American and European silent films 1918-1923. 7
2.3 Using KDEs and CFDs to compare ASL distributions. 8
2.4 Histograms and KDEs for the SL distributions of Brief Encounter and Harvey. 9
2.5 Smoothed SL data for The Blue Angel. 10
2.6 Time-series analyses of SL data from King Kong (2005). 12
2.7 Bar charts of shot-scale data for 24 films of Fritz Lang. 14
2.8 A correspondence analysis for the shot-scale data for films of Fritz Lang. 15
2.9 Correspondence analyses of shot-scale data for films of Lang, Hitchcock and Ophuls. 17
2.10 Bar charts of shot-scale data for films of Hitchcock’. 18
2.11 Bar charts of shot-scale data for films of Hitchcock, Lang and Ophuls. 19

4.1 Scatterplot matrices for variables descriptive of SL distributions. 27
4.2 A ‘magnified’ version of part of Figure 4.1. 29
4.3 A scatterplot matrix of robust measures of dispersion for 134 Hollywood films. 30

5.1 Histograms for the SLs of A Night at the Opera. 36
5.2 Histograms of logged SLs for A Night at the Opera. 38
5.3 The default KDE for the SLs of A Night at the Opera. 39
5.4 The KDE for A Night at the Opera, using SLs less than 40. 40
5.5 Different KDEs for the logged SLs of A Night at the Opera. 40
5.6 A boxplot and violin plot for A Night at the Opera. 41
5.7 A variety of plots for logged and unlogged SLs for Harvey. 42
5.8 A histogram for the SLs of The Scarlet Empress and a variety of boxplots. 44

6.1 A comparison of SL histograms for Easy Virtue and The Skin Game. 47
6.2 A comparison of KDEs for the SLs of Easy Virtue and The Skin Game. 48
6.3 A comparison of KDEs for the logged SLs of Easy Virtue and The Skin Game. 48
6.4 Boxplots and violin plots, log-scaled, for Easy Virtue and The Skin Game. 49
6.5 A comparison of CFDs for the SLs of Easy Virtue and The Skin Game. 50
6.6 A probability plot for logged SL data for a A Night at the Opera. 53
6.7 Different comparisons of SLs for Exodus and Aristocats using KDEs. 53
6.8 A histogram, KDE and probability plot for logged SLs of Walk the Line. 54
6.9 KDE and probability plots for A Night at the Opera and Walk the Line. 55
6.10 Comparisons of KDEs of films for logged SLs compared to lognormality. 56

7.1 Moving average of the SLs for Ride Lonesome from Salt (2010). 59
7.2 Fifth and first order polynomial fits to the SL data for Intolerance. 60
7.3 Linear and polynomial fits for the SLs of the different stories within Intolerance. 61
7.4 Moving averages for Ride Lonesome. 62
7.5 Moving average and loess smooths for Ride Lonesome. 63
7.6 Plots of SL against cut-point for Top Hat with a moving average and loess smooths. 65
7.7 Different loess smooths for the SLs of Top Hat. 65
7.8 Different loess smooths for the logged SLs of Top Hat.
7.9 Moving average, moving median and loess smooths for Top Hat.
7.10 A time-series analysis of ranked SL data for Top Hat from Redfern.
7.11 A time-series analysis of ranked SL data for Top Hat.
7.12 An order structure matrix for the SL structure of Top Hat, after Redfern.
7.13 A KDE of the time-series structure of Top Hat, from Redfern.
7.14 KDE and loess smooths of the time-series for Top Hat.

8.1 Loess smooths for Lights of New York.
8.2 ‘Wallpaper’ plots and loess smooths for Lights of New York.
8.3 An order structure matrix plot for Lights of New York and a variant of it.
8.4 Heat color representations of SL data for Halloween with loess smooths.
8.5 Heat color representations of SL data for Halloween with KDEs.
8.6 Comparative heat color representations of Halloween and Lights of New York.
8.7 Color-coded plots of KDEs for The House on Sorority Row and Pinocchio.
8.8 Gray-scale plots for the ranked data of Pinocchio and The House on Sorority Row.
8.9 Plots showing the effect of bandwidth choice on KDE comparisons.

9.1 Bar charts of shot-scale data for Hitchcock’s silent films.
9.2 A basic correspondence analysis of the shot-scale data from Table 9.1.
9.3 An enhanced correspondence analysis of the shot-scale data from Table 9.1.
9.4 Separate row and column plots for a CA of Table 9.1.
9.5 The row plot from a CA of Table 9.1 with enhanced labeling.
9.6 A CA for shot-scale data illustrating the use of supplementary points.
9.7 Mean relative frequency plotted against shot scale ranks for three categories of Hitchcock films.
9.8 Shots-scale variation over time for the films of Hitchcock.

10.2 A scatterplot matrix for date, sample size, ASL and MSL for 134 films.

11.1 Idealized lognormal SL distributions.
11.2 Histogram and boxplot displays of SLs for Brief Encounter.
11.3 Plots of MSL and ASL against date for 134 Hollywood films.
11.4 A plot of ASL against MSL for 160 films from the period 1920–1933.
11.5 KDEs of the MSL/ASL ratio for 54 silent films 106 early sound films.
11.6 Boxplots of the MSL/ASL ratio for silent films and sound films from 1929-1933.
11.7 KDEs for logged SLs of Rain (1932) contrasting action and dialog shots.
11.8 A plot of ASL against MSL for 62 of D.W. Griffith’s Biograph one-reelers.
11.9 A magnified plot of Figure 11.8 for films with n = 55 or greater.

12.1 Histograms and probability plots for Harvey (1950).
12.2 Histograms and probability plots for Brief Encounter (1945).
12.3 Plots based on different estimates of the median of SL distributions.
12.4 A plot of skewness against kurtosis for 134 films labelled by date.
12.5 An enhanced plot of skewness against kurtosis for 134 Hollywood films.
12.6 The Shapiro-Wilk statistic for 134 Hollywood films.
12.7 A plot of skewness against kurtosis for 134 films after double-transformation.

13.1 The SL distribution of A Corner in Wheat using the Brisson method.
13.2 Smoothed SL distributions for some of D.W. Griffith’s Biograph films.
13.3 Smoothed SL distributions using different degrees of smoothing.

14.1 Average-linkage cluster analysis based on 25 D.W Griffith’s Biograph films from 1911–12.
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.2</td>
<td>An enhanced version of Figure 14.1</td>
<td>153</td>
</tr>
<tr>
<td>14.3</td>
<td>Single-linkage cluster analysis based on 25 D.W Griffith’s Biograph films from 1911–12.</td>
<td>154</td>
</tr>
<tr>
<td>14.4</td>
<td>PCA plot based on a sample of D.W.Griffith’s films from 1911–1912</td>
<td>155</td>
</tr>
<tr>
<td>14.5</td>
<td>A scatterplot matrix for the first three PCs from a PCA of Griffith films.</td>
<td>156</td>
</tr>
<tr>
<td>14.6</td>
<td>Smoothed SL distributions for three Griffith Biographs that cluster together.</td>
<td>157</td>
</tr>
<tr>
<td>14.7</td>
<td>Smoothed distributions for films in the ‘red’ and ‘blue’ clusters.</td>
<td>157</td>
</tr>
<tr>
<td>14.8</td>
<td>Average-linkage cluster analysis based on 25 D.W Griffith’s Biograph films from 1911–12.</td>
<td>158</td>
</tr>
<tr>
<td>14.9</td>
<td>Principal components analysis based on a sample of D.W.Griffith’s films from 1911–1912</td>
<td>159</td>
</tr>
<tr>
<td>14.10</td>
<td>Smoothed SL distributions for three Griffith Biographs.</td>
<td>160</td>
</tr>
<tr>
<td>14.11</td>
<td>Smoothed SL distributions for the ‘red’ cluster using two different spans.</td>
<td>160</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Shot scale data (%) for films of Fritz Lang  
9.1 Shot scale data (%) for films of Alfred Hitchcock.  
9.2 Shot scale data (%) for films of Max Ophuls. 
14.1 D.W. Griffith Biograph films from the years 1912–1913.
Preface

These notes are a revision of ones I wrote after about nine months exposure to the idea of cinemetrics. Retrospectively it seems somewhat presumptuous to have attempted this, given that I have no background in film studies, but I became interested. This is essentially a book that attempts to both summarize work that has taken place in cinemetrics and has case studies on topics that have attracted my interest. There is some emphasis on computational aspects.

The interest started with an enforced period of convalescence. Anticipating a period of idleness I acquired a DVD player – not something I’d previously owned. It was bought to while away time, and prompted by an irrational urge to see Quentin Tarantino’s *Kill Bill* films, previously unseen. Stocking up on these and other films to watch I took advantage of cheap offers on *The Cabinet of Dr Caligari*, Murnau’s *Nosferatu* etc. that rekindled a dormant interest in German expressionist and other early films of the kind screened by the better university film societies when I was a student (Edinburgh University). Things got out of hand and I ended up with a quite large collection of DVDs of silent film.

As academics do, it occurred to me that I ought to know more about the subject beyond just watching the films. There is copious (mis)information available on the web but something more seemed to be needed. This ‘something’ is called a ‘book’, and you can’t stop at one; they’re addictive. My tastes are eclectic, and everything from pulp biography, through serious (and in some cases seriously long) historical studies, film theory, feminist readers on silent film etc. became grist to the mill. At some point I decided that my understanding of film technology, and the way it might have conditioned the appearance of films I was watching and why I appreciated them, was one of many major gaps in my knowledge. Having seen several intriguing references to Barry Salt’s *Film Style & Technology: History & Analysis* it found its way onto my shelves. It wasn’t quite what I was expecting.

Apart from the bracing views on film theory and theorists expressed in the opening chapters (and I’ve since read several academic critiques of this) it was the several chapters on statistical style analysis of motion pictures that intrigued me. This was an unexpected intersection between leisure and professional interests. This is rare, though not unknown; much of my research activity as an academic statistician has been in the realm of archaeology, which started as a leisure pursuit without any thought given to statistical applications in the area.

My initial interest in Salt’s book was piqued and furthered by the discovery that what he originally called *The Statistical Style Analysis of Motion Pictures* (Salt, 1974) – what I view as the statistical analysis of ‘filmic’ data – had a name, *Cinemetrics*, and an eponymous web-site. It was also evident that some practitioners got quite excited (or is it excitable) about how to ‘correctly’ apply statistical methods to cinemetric data. That the way statistics are used can inspire passion wasn’t news. Contributions to discussed papers of the *Royal Statistical Society*, as recorded in its journals, and particularly about the mid-twentieth century when foundational issues were a major concern, can be vitriolic. Similarly, debate (polemic) about the merits of the ‘quantitative revolution’ in archaeology in the 1960-70s was not always characterized by rational argument that showed any attempt to understand an opposing ‘philosophical’ viewpoint.

I always rather enjoyed reading about this kind of thing. I lead an unexciting life and one of my ideas of fun is doing data analysis (for interest, but not of the monetary kind). Cinemetrics looked like fun. I began to play around with some of the data available, looking at issues that had been raised in the cinemetric literature and emulating analyses that had been done – not always
reaching the same conclusions. These notes, undreamed of in the not too distant past, and never mind a revision, are one consequence.

One purpose was to gather together ideas that have been put forward for statistical data analysis in cinematics. This seemed useful, since the material is scattered and much of it is web-based. Since then a series of discussions on the Cinemetrics website, initiated by Yuri Tsvian and to which I was invited to contribute, have helped crystallize differing approaches to cinemetric data analysis. These are referred to as the On Statistics discussion in the notes. It seemed useful to update the notes to take these discussions into account, as well as recently published work of interest.

There is an emphasis on how to implement ideas using the statistical software, R, an open-source, state-of-the-art, highly flexible package. That a lot might be done in cinemetric analysis using modern statistical software, but largely wasn’t, was one of my initial impressions – hence the emphasis. It also struck me that some standard approaches to data analysis and presentation could be done differently and possibly more efficiently. All that’s really being said here is that modern statistical software, and some methodology, can usefully add to what is already done, sometimes routinely, and that there is nothing terribly demanding about either using the software or implementing and understanding the methodologies.

My intellectual debt to, and the inspiration I have drawn from, the work of Barry Salt has already been acknowledged. These notes would not have been possible without the contributions of many people to the Cinemetrics website, both to the database that I have made extensive use of and the more reflective discussions about how the data can be analyzed. I am particularly grateful to Yuri Tsvian for his invitation to contribute to these discussions and for his encouragement.

I’d also wish to acknowledge the support of the Neubauer Collegium at the University of Chicago for funding a visiting fellowship that enabled me to meet Yuri and colleagues and discuss issues that pertain to these notes. I’d particularly like to thank Josh Beck, Jamie Bender and Sarah Davies Breen for their hospitality during my visit. One outcome was a conference and recordings of the presentations can be viewed on the Neubauer Collegium website at http://neubauercollegium.uchicago.edu/events/uc/Cinemetrics-Conference/.
Chapter 1

Cinemetrics and \texttt{R}

1.1 Introduction

As indicated in the preface, I am a relative newcomer to ‘cinemetrics’ but have a background as an applied statistician with a long-standing interest in applications outside the ‘scientific mainstream’, particularly, but not exclusively, archaeology. That I have no prior background film scholarship has obvious disadvantages, though a possible compensation is that one can look at the way statistics has been used in cinemetrics with a dispassionate eye. To that eye an important body of technique looks a little ‘old-fashioned’, not in the use of long-established methodology but more in its implementation, which does not exploit the enormous and accessible computational power now available. Spreadsheet packages like \texttt{Excel} have their uses, are readily available, and people are comfortable with them, but they are limited as far as statistical analysis goes compared with the open-source package \texttt{R} used for all the analyses in what follows.

The aim in the notes is to provide a fairly thorough coverage of how statistical methods have been used in cinemetrics (from a statistician’s perspective) \textit{and} to show how these methods can be implemented in a fairly painless way. My understanding of what cinemetrics is about is outlined in Section 1.2 and Chapter 2 provides a foretaste of applications discussed in later chapters, without the distraction of computational details. The software used is introduced in Chapter 3 with some simple illustrations of its use for the calculation and display of familiar descriptive statistics, such as the average shot-length (ASL), provided in Chapter 4. The focus in Chapters 5 to 8 is on graphical data analysis. Some of the methods should be familiar, others less so. While the aim has been to collect together and organize material, a critical appraisal comparing the merits and limitations of different methods is provided where appropriate.

Graphical analysis is also the subject of Chapter 9, but there the focus is on the analysis of shot-scale data rather than shot-length data, which dominate the earlier chapters. Some of what is covered will be familiar to anyone acquainted with Barry Salt’s book \textit{Film Style \& Technology: History \& Analysis} (various editions), but the idea of applying the multivariate method of correspondence analysis to such data is fairly new. Chapter 10 explains how to construct large data sets in \texttt{Excel} suitable for import and manipulation in \texttt{R}. Subsequent chapters mainly consist of ‘case studies’ designed both to illustrate methodology and focus on particular topics that have attracted my interest or which have been the occasion for debate in the literature (Chapter 9 might also be viewed in this light).

Chapter 11 reviews some of the issues involved in choosing between the average and median SL, and other descriptive statistics, for summarizing SL distributions. This has been a surprisingly contentious topic and was the subject of the first discussion topic, \textit{Median or Mean?}, in the \textit{On Statistics} debates on \textit{Cinemetrics}. The opportunity is taken to present some small case studies, not previously published, to illustrate applications of the more general ideas addressed. Chapter 12 likewise addresses an issue that has proved contentious – the lognormality or otherwise of SL distributions – and, in particular, attempts to summarize some of the debate that has occurred.
in papers published in 2012–2013 in the journal *Literary and Linguistic Computing*. Multivariate statistical methods, as often applied, are concerned with exploring patterns in large tables of data. Some other commonly used methods, with applications to cinemetric data (apart from correspondence analysis discussed in Chapter 9), are discussed and illustrated in Chapters 13 and 14. This is fairly ‘cutting-edge’ as far as cinematics goes, though some of the ideas were used in the third of the *On Statistics* discussions, *Looking for Lookalikes*.

Notation and mathematical formulae are unavoidable in some instances, particularly when discussing the work of others, but as it’s claimed that a user can effectively apply statistics without needing a deep understanding of the more complex mathematics involved I’ve tried to keep it to a minimum. The main exception is the Appendix, where the lognormal distribution, and the idea of logarithmic and other data transformations are discussed.

1.2 Cinemetrics

1.2.1 The idea

What I see as some of the highlights of the development of cinemetrics is outlined below. Salt’s (1974) seminal paper is important in several respects. From this, and his other writings, and in the context of these notes, I want to highlight what I see as his emphasis on comparison, involving pattern recognition, and the way this can be approached using quantified data. Section 1.2.2 elaborates on this (with some added polemic in Section 1.2.3).

‘Cinemetrics’ is understood to mean the statistical analysis of quantitative data, descriptive of the structure and content of films that might be viewed as aspects of ‘style’. For practical purposes this usually means the analysis of data that describe shots in some way, for example their length or a characterization of their type. Shot lengths (SLs) have attracted most attention in the literature; examples of type include classification by shot scale, content (e.g., action/dialogue), camera movement and so on.

Everything must have a beginning, and Barry Salt’s article *The Statistical Style Analysis of Motion Pictures* published in *Film Quarterly* (Salt, 1974) is a good place to start. The paper is usefully discussed in Buckland’s (2008) review of Salt’s book *Moving Into Pictures* (2006). Salt’s earlier book *Film Style & Technology: History & Analysis*, first published in 1983 and now in its third edition (Salt, 2009), is important. It was fairly recent exposure to this that alerted me to the existence of cinemetrics and aroused my interest.

It cannot be claimed that the subject took the world by storm. Finding relevant publications prior to the last few years is not easy. One obvious reason for this is that the data collection needed to make cinematic analysis possible is both time-consuming and (I assume – I’m an ‘end-user’ and have not done this myself) tedious. In easing this problem the development of the Cinemetrics website (http://www.cinemetrics.lv/), headlined as a ‘Movie Measurement and Study Tool Database’ on its front page, seems to me to be a major landmark. It simultaneously facilitates data collection and acts as a repository for data collected and articles published on cinemetric data analysis. Quantified information is now available for thousands of films, so the absence of data is no longer a reason for not engaging in cinemetic studies.

Two other bodies of work merit attention. One is the diverse set of postings related to cinemetrics and statistics on Nick Redfern’s research blog (http://nickredfern.wordpress.com/). Where possible I’ve preferred to reference his more structured articles – in a format associated with conventional journal publication – available as pdf files on the web. The blog, though, is the only source I’ve seen for the application of some statistical ideas (both old and relatively new) to cinemetric data analysis, and some of this deserves both notice and evaluation.

The other body of work I have in mind is that of James Cutting and his colleagues, mostly Jordan DeLong and Kaitlin Brunick, published in, mostly, journal articles in the period 2010-2013. I have seen some of this work labeled as ‘cognitive film theory’. From a statistical point of view it is technically more demanding than most of what is covered in these notes. The applications involve fairly complicated approaches to pattern recognition in cinemetic data. The idea, as I
interpret it, of pattern recognition facilitated by the quantification of ‘filmic’ entities is one of the major innovations in Salt’s 1974 paper and later work.

1.2.2 Pattern recognition in cinemetrics

It is convenient to begin by quoting, at a little length and with comment, from Buckland (2008), a sympathetic reviewer of Salt’s work. References are from pages 22–24 of his paper. Buckland suggests that for Salt, style designates a set of measurable patterns that significantly deviate from contextual norms. He, therefore, has to establish those norms before he can define how a director deviates from them.

The purpose of tables and charts is not only to display the data itself, but also to show patterns and exceptions in the data. ... Whereas tables report actual figures, charts are designed to leave a general impression, rather than communicate exact numbers. ... The representation of data in charts is more visual, for it immediately reveals patterns and exceptions in the data.

In context, Buckland is referring to Salt’s (1974) focus on directorial style, relating it to auteur criticism. The important idea, of more general application, is that of pattern recognition and comparison. The idea of objective comparison is central to Salt’s thinking and quantification is central to this. Buckland recognizes that the data [Salt] collects on individual directors can be recontextualized into broader frameworks to characterize the style of films genres, national cinemas, or historical periods, which Salt carried out in *Film Style and Technology*.

Salt’s (1974, p.14) original thoughts are worth remembering. To establish the existence of an individual formal style in the work of a director, it is necessary to compare not only a sufficient number of his films with each other, but also – which is always forgotten – to compare his films with films of similar genre made by other directors at the same time. ... An even more absolute norm for any period is really needed as well, to give a standard of comparison that reflects the technical and other constraints on the work of filmmakers at that time and place ... This was written with reference to directorial style but, in Buckland’s word, can be ‘recontextualised’. The strong emphasis on comparison is there, as is the need to establish norms, which I’m interpreting as ‘pattern’. Comparison can operate at various levels, most simply in the comparison of some aspect of a small body of films, such as their shot-length distributions. A more challenging problem is to identify patterns across a large body of films (Chapters 9 to 14).

While I find it convenient to think that a lot of cinemetric data analysis can be viewed as pattern recognition, it is difficult to define precisely what I mean by ‘pattern’. This is dealt with, at least partially, in Chapter 2 which provides a collection of examples, inspired by what has been done in the literature, where what I’d regard as ‘pattern recognition’ is going on.

1.2.3 Lies, damned lies and statistics (and fear)

The first six chapters (and the preface) of Salt (2009), are largely occupied with an entertaining if possibly over-the-top polemic against the iniquities of ‘film theory’, where it militates against (I think this is the argument) the pursuit of objective ‘scientific’ study and analysis of film. It reminded me, in some ways, of arguments that raged in archaeology between the 1960s and 1990s.

About the time that Salt (1974) was published, archaeology was in the throes of a ‘quantitative revolution’ that had begun in earnest in the 1960s. The ideas of quantification in archaeology were initially labelled the ‘New Archaeology’ and came to be known as processualism. The more philosophically minded protagonists explicitly linked their ideas to (logico-deductive) positivism as expounded, for example, in the work of Hempel (1965), attempting to explain variation in the archaeological record in terms of law-like relationships. Some protagonists were rather evangelical in their promotion of ‘scientific’ archaeology and made seriously inflated claims about what quantification could achieve. As Brandon, in the preface to Westcott and Brandon (2000), noted...

---

1 A review of some of the statistical highlights is provided in Chapter 1 of Baxter (2003).
‘during its heyday, statistics had been waved above archaeologists’ heads as an "answer" to dealing with a multitude of archaeological problems’.

As might be expected there was a theoretical reaction, from the 1980s on, often labelled post-processualism, a generic term for a diverse and incompatible set of ‘philosophies’. This embraces different forms of interpretive archaeology and some of the same -isms (structuralism, post-structuralism, Marxism etc.) and thinkers that Salt (2009) castigates. On both sides of the argument as much heat as light, if not more, was generated, but things settled down. As Brandon concluded, ‘after much yelling and arm-waving, most agreed that statistics were not an answer in themselves but [were] an extremely important tool available for archaeological use’.

There are still those who regard the application of scientific methodology, and more specifically statistical analysis, to archaeological understanding as de-humanising, demeaning, and inappropriate for ‘humanistic’ study. This is an extreme position but not a caricature. That this view can be deeply felt is not in doubt; the suspicion exists that ‘ideology’ can be worn as a ‘principled’ philosophical cloak behind which fear and ignorance of matters mathematical and statistical is concealed.

There may or may not be analogies here with the place of cinemetrics in film scholarship. You can’t do much about ideology; fear, if acknowledged, can be confronted and dealt with. My general attitude to the use of statistics by non-specialists is that you should dip your toe in the water and at some point jump in. Most will learn to swim well enough for their purposes (though it has to be acknowledged that ‘collateral damage’ in the form of drowning, can occur). You need the motivation to learn to swim; not everyone has it and they are usefully employed doing other things. This, so long as swimming is tolerated, is fine.

From the perspective of the 2010s I don’t share the pessimistic view that Salt (2012) had about what his intended audience in 1974 could reasonably be expected to cope with. The mathematics needed for some statistical methodology can be complex but, conceptually, not the idea that often underpins an applied statistical technique. Computational ease is the key to the successful implementation of (possibly) mathematically complex, conceptually simple ideas and this is not the problem it once was. These notes are based on the belief that if you understand an idea, and it is easy to apply it, then it is straightforward. You don’t need to worry about the mathematical detail, and any complex computer programming necessary has been done for you.
Chapter 2

Examples

2.1 Preamble

The intention here is to illustrate the kind of thing I understand by ‘pattern’ by way of examples. These serve as an introduction to later chapters where more technical discussion, including computational aspects, is provided.

If forced to define ‘pattern’ it would be something rather vague, like an organisation of the elements of analysis that departs in a recognisable and describable fashion from what would be expected from ‘random’ organisation (‘random’ not always being easy to define); or a marked departure from the pattern exhibited by some hypothetical ‘base’ model, not itself of intrinsic interest.

Some patterns can be described qualitatively; for example, many SL distributions have a characteristic shape, singly peaked and skew, that differs from a hypothetical and uninteresting ‘flat’ distribution. A subset of these exhibit a stronger pattern, that they have in common and which potentially establishes a ‘norm’, in the sense that they can be described mathematically using the lognormal distribution (Section 2.2.3). This might be thought of as mathematically describable regularity. ‘Describable regularity’ is manifest in the general decline in ASLs of films over the last 60 years or more; it is less easy to characterize using a precise mathematical formula, but readily displayed using graphical statistical methods (Section 2.2.1).

Other forms of pattern require the use of external knowledge to identify them. For example, Figure 2.8 reduces shot-scale data for 24 of Fritz Lang’s films to 24 points on a map. Unannotated, their disposition looks fairly random; add labeling to differentiate between the early German films (mainly silent) and later American ones and a pattern in the form of separation between the two is apparent. In this sort of application, if you are lucky, you may get clear white space between distinctive clusters of points. This is indicative of a form of pattern that can be identified without reference to external knowledge, but may be interpreted in the light of such.

This brief discussion is intended to be indicative rather than exhaustive. Issues raised in the identification, display, and comparison of the pattern of SLs within films, are illustrated by example in Sections 2.2.4 and 2.2.5. The point to reiterate is that much cinemetric data analysis is motivated by the idea of comparison which, in turn, is facilitated by the ability to recognize and contrast patterns within bodies of data.

2.2 Examples

2.2.1 ASLs over time

One of the more obvious and commented on examples of a ‘pattern’ in cinemetric data is the decline in average shot-lengths (ASLs) over a period since about the early 1950s. This is illustrated, for example, in Figure 1 of Cutting et al. (2011a) using their sample of 150 films and data collected
by Barry Salt, and Salt’s (2009, p.378) own similar demonstration for 7,792 US feature films. This sort of pattern is illustrated for US films (1955-2005) in the left-hand plot of Figure 2.1. ASLs are shown for individual films, the scale being truncated at an ASL of 20 for display purposes, with a line joining the mean ASLs for each year.

Figure 2.1: The left-hand plot is of ASLs for American films (1955-2005), the line joining the mean ASLs for each year. This is repeated in the right-hand plot along with a smoothed version; the overall ASL for all films is also shown.

The general decline in ASLs is clear enough. A magnified plot, omitting individual ASLs is shown to the right, where the decline in mean ASLs of about 7 seconds over a period of 50 years is more apparent. The general pattern is usefully summarized by fitting a smooth curve to the yearly mean ASLs, which better allows general trends to be discussed. Thus, there is a fairly sharp and steady decline from the mid-1950s to late 1960s, a gentler decline with some variation to the early 1990s, then a pick-up in the rate of decline. As ASLs can’t go below zero, and there is a practical limit, it is obvious that the pattern will eventually ‘bottom-out’ at some limit above zero, unless the trend is reversed.

The smoothing method used is discussed in Section 7.5. The right-hand plot shows the mean ASL for all the films used in the period concerned. Pattern here is manifest as a regular and describable deviation from a model — that there is no variation in mean ASLs over time — that can be viewed as a baseline for comparison, which may or may not be of intrinsic interest in its own right.

2.2.2 Comparison of ASL distributions

The comparison effected here using histograms, in Figures 2.2 and 2.3, is between samples of American and European silent films for the period 1918-1923. The first figure emulates a comparison from Salt (2009, p.192); the data used are from Salt’s database and are not identical to that he used, but the figures are close enough to his for the illustrative purposes intended.

Figure 2.2 compares histograms of ASLs for the two bodies of films. This is a fairly common approach to this kind of comparison; O’Brien’s (2005, p.92) study of American and French films during the period of transition to sound provides another example. Pattern can be thought of in more than one way. At the level of qualitative description both sets of films are similar in that they have a single main peak, but the distribution is more spread out for European films.

1The data were extracted from Barry Salt’s database on the Cinemetrics website.
Here, and differently from the previous example, the focus is on comparison between two bodies of data. A quantitative comparison of differences in the qualitatively similar ASL distributions can be effected by calculating statistics that summarize properties of the distribution. One such is the mean of the ASLs, indicated on the histograms for the two sets of films. That for the European films is larger than that for the American ones, suggesting its pattern is shifted to the right by comparison. The difference can be measured, giving a quantitative ‘edge’ to the qualitative analysis. The use of the ASL, alternatives to it, and measures of spread is discussed in Chapter 11.

Histograms, discussed further in Section 5.1, are quite a clumsy way of effecting comparisons. An alternative is the use of kernel density estimates, KDEs (Section 5.2), which can be thought of, for present purposes, as a smoothed form of histogram. They are much more readily overlaid, for comparative purposes, as the left-hand panel of Figure 2.3 shows. It is immediately evident that the ASLs for the American films are much more concentrated; that the body of the European films is shifted to the right by comparison (i.e. ASLs tend to be larger); and that they spread well beyond the upper limit of ASLs for American films. Another feature of possible interest, that KDEs reveal better than histograms, is that there are two apparent peaks in the distribution for American films. This may turn out not to mean anything, but one is being alerted by the display to a less obvious aspect of pattern that might merit exploration.

The right-hand panel is another way of comparing the same data, using cumulative frequency diagrams (Section 6.3). The vertical axis shows proportions, so if you read across from some value then drop down when you hit a curve you get the ASL which has that proportion of films with smaller ASLs. Such diagrams are common enough in the statistical literature, though not, I think, as easily ‘read’ as histograms or KDEs. The present example shows, fairly starkly, that a greater proportion of the European films exceed almost any given ASL than the American films.

We have here three different ways of effecting a comparison. I’d make a distinction between the ‘big idea’, of comparison, and the choice of technique used to effect it. The analogy doesn’t work exactly, but think of the comparison as a structure that can be erected using different building materials; the choice of technique corresponds to choice of material. And depending on design and intended function, not all materials are suitable for all structures.
2.2.3 Pattern and SL distributions

This example may look similar to the previous one, but there are important differences, some of a statistically technical nature. In the last section the elements of which the pattern was composed were film ASLs, the interest being in comparison across national cinemas (genres or historical periods could equally well have been used for illustration). If the elements are SLs and the units of comparison are films identical methods of graphical display are available. SL distributions for two films, *Brief Encounter* (1945) and *Harvey* (1950), are illustrated using histograms in the left-hand panel of Figure 2.4.

Salt (1974, p.15) observed, of SL distributions, that ‘a considerable similarity of overall shape is apparent’. In the language I’m using, shape can be equated with pattern and what is being observed is that a lot of films have a qualitatively similar pattern. Many SL distributions have a single obvious peak (or mode)\(^2\) and are skewed with a long right tail, the bulk of the SLs being bunched up nearer the lower bound of zero. Salt went further.

Descriptive statistics, such as the ASL or median SL (MSL), can be used to differentiate between films with qualitatively similar SL distributions. The tack Salt took was to suggest that many distributions could be modelled (approximately) by the same kind of ‘mathematically’ defined curve, eventually settling on the lognormal distribution (Salt, 2006, p.391; and see the Appendix). The important idea is that the (ideal) shape can be reconstructed perfectly if you know two numbers (parameters), which define the form of the distribution. Several consequences follow.

1. Given a large body of films it can be asserted that those which have a lognormal SL distribution exhibit the same shape, or pattern. That is, they are single peaked, skew and have (usually) a longish right tail.

2. A much stronger statement about pattern, a quantitative one, that they have the same mathematical form is possible. Differences between distributions that can be so categorized can be summarized solely in terms of the two parameters that define them\(^3\).

\(^2\)Strictly speaking, ‘modal class’ if talking about grouped data, as in a histogram.

\(^3\)The dependence on two numbers is why discussions about whether the ASL or MSL is the better descriptor of ‘film style’ miss the point (see Baxter, 2012a). Call the parameters that define the exact shape of the distribution \(\mu\) and \(\sigma\) – they govern the position of the bulk of the data and its spread. The mathematics is in the Appendix, but...
3. Identifying a body of films that exhibit the same strong pattern establishes a norm against which films that deviate from the norm can be identified. It then becomes interesting to ask what form the deviation takes and why.

A technical issue raises its head at this point, which is the use of logarithmic transformation of the SLs. This is discussed in more detail in the Appendix; the practical value of such a transformation is that it can make it a lot easier to see what’s going on. It does this by eliminating what might be called visual biases that arise in comparing SL distributions, attributable to the uneven distribution of mass along the curve. Taking the logarithm of lognormal data produces new data that have a normal distribution, and it is much easier to visually assess normality, and departures from it, than lognormality from the untransformed data. This is best illustrated by returning to Figure 2.4.

*Brief Encounter* and *Harvey* were chosen for illustration because, according to some formal methods of making the judgment, they are among the most and least lognormal SL distributions you could wish to meet (Redfern, 2012a). Is this apparent from the histograms to the left of the figure? It depends on how experienced you are at looking at this kind of thing; close inspection reveals differences, but their importance is hard to judge, particularly given that appearances of histograms can be affected by the manner of their construction.

That there is a radical difference is immediately apparent if the data are transformed logarithmically, and displayed using KDEs as in the right of the figure. The log-transformation evens up the visual weight given to longer and shorter SLs. *Brief Encounter* looks comfortably normal, and hence lognormal on the original scale; *Harvey* doesn’t, the main departure from normality being two modes. The larger of these occurs at just over 3 on the log-scale which translates to about 25 seconds on the original scale. Reference back to the histogram reveals a bump in this region that the logged analysis suggests is genuine. The histogram for *Brief Encounter* is also uneven in this region but the logged analysis suggests it is not important variation. The statistical analysis here indicates a clear difference in patterns for the two films, that for *Brief Encounter* being representative of the norm. Interpretation of what these difference mean in ‘filmic’ terms, briefly, the median is \( \exp(\mu) \), where \( \exp() \) is the exponential function. It involves only one parameter. The ASL is \( \exp(\mu + \sigma^2/2) \), which combines both parameters but doesn’t allow them to be separately distinguished. Using the ASL and median jointly is preferable to using either separately if a numerical summary is required.
if anything, is a substantive matter, outwith the remit of statistics\(^4\).

### 2.2.4 Internal pattern - individual films

The examples so far focus on what I shall call ‘external’ pattern or structure. Statistics or graphs that characterize individual films can be studied for their own interest or, more interestingly, compared, with the aim of seeing if broader pattern, treating the individual patterns as elements to be ordered, can be discerned. This might be thought of as an attempt to discern ‘global’ patterning on the basis of external structure. Internal patterning or structure, the arrangement of shots within a film, is of equal interest, but quantitative approaches to this have perhaps received less systematic attention than external patterning. Chapters 7 and 8 explore possibilities and the present example only touches on some of these.

Data on SLs submitted to the Cinemetrics database come in two versions, ‘simple’ where essentially only the SLs are recorded, and ‘advanced’ where shot type is recorded as well according to whatever categories the analyst chooses. The Paramount version of von Sternberg’s *The Blue Angel* (1930), submitted by Charles O’Brien, is used for illustration in Figure 2.5.

Figure 2.5: Smoothed SL data for *The Blue Angel*, superimposed on backgrounds showing the actual SLs (left) and ‘wallpaper’ with color-coded shot type (right), blue being ‘singing’, green ‘dialog’; and salmon ‘action’. (see text for more detail).

The left-hand panel shows the SL data plotted against cut-point. As is typical the data are rather ‘noisy’ and a (loess) curve has been fitted through the data to smooth out the detail (Chapter 7). An over-smoothed curve has been used to highlight the general trend in SLs where, for example, the faster cutting over the first 40 minutes or so can be contrasted with the slower and more variable cutting rate over the following 30 minutes. The range of SLs dictates the scale of the plot which can sometimes make it difficult to see the finer detail of the trend.

In the right-hand plot the full scale is dispensed with; a less smoothed curve is used to show more of the detailed variation; and shots are coded by color according to the categorization used by O’Brien. Shots are colored at the cut-points, so white spaces correspond to the longer shots.\(^5\).

\(^4\)Lognormality is the dominant paradigm, but it has been challenged. See Chapter 12 for a detailed discussion of the issues involved.

\(^5\)Choice of background color is interesting. If you choose (in)appropriately and omit the smoothed curve then, depending on the film, op-art effects that sometimes resemble the paintings of Bridget Riley can be achieved.
The smoothing chosen shows the same distinction between the earlier and later parts of the film as the smooth to the left, while also indicating more of the nature of variation in the first half of the film. The degree of smoothing chosen depends on the purpose of illustration and need not be confined to a single choice. ‘Over-smoothed’ curves are useful for showing very general trends in the patterning of SLs; considerably less smoothing is needed if the aim is to illustrate variation between ‘scenes’ (Salt, 2010). Cutting et al. (2011) have used SL data to investigate whether internal patterning is generally consistent with a four-act structure, following ideas of Thompson (1999), but have conceded, following an intervention by Salt, that their initial positive confirmation of the hypothesis was methodologically flawed (Cutting et al. (2011)).

The example illustrated above, and many of those in the relevant chapters to follow, are concerned with the detection and illustration of internal structure for individual films. What Cutting et al. attempt is something altogether more ambitious, the detection of global patterning on the basis of internal structure. This is challenging and some aspects of the challenge are illustrated in the next example. (The explanation is also more complex that other examples, so some readers might prefer to skip it on a first reading.)

### 2.2.5 Internal pattern - an example of global analysis

The problem with global comparison based on internal SL patterns is that the latter are not easily reduced to a form readily admitting synthesis of some kind. It is possible to look at the patterns for a small body of films and say how and why they are different, but how to express this in quantitative form that allows further statistical analysis is not obvious. The process is complicated by the fact that internal patterns may be viewed at different scales.

Shot-lengths measured at a cut-point are an example of a time-series. Aspects of time-series can be quantified and compared across series. This is the kind of approach adopted in Cutting et al. (2010), one of whose analyses is described here by way of illustrating the complexity involved.

Salt (2010) expresses the view ‘that variation in cutting rate is ordinarily used as an expressive device of a conventional kind - more cuts for sections where there is more dramatic tension or action, and less for less of the same’, and that ‘in general there is a conventional idea about alternating scenes with different dramatic character in plays and films, so that things like cutting rate and closeness of shot which are used expressively should change on the average from scene to scene’. That is, if correct, short SLs will tend to occur in ‘clusters’ with other short SLs, and similarly for long SLs. There are several ways in which this might be investigated, one such being the partial autocorrelation function (PACF). The left-hand panel of Figure 2.6 emulates, with embellishment, an example given in Figure 1 of Cutting et al. (2010) for King Kong (2005).

The solid vertical lines indicate the size of the partial autocorrelation at lags of up to 20 – that is, they measure the strength of relationship between the SLs of shots, 1, 2, ..., 20 positions apart in the shot sequence. It is expected, if Salt’s view is correct, that the first few of these will be ‘significantly’ positive. The statistical significance is judged by the horizontal dashed line (at $2/\sqrt{n}$) – if this line cuts a vertical line the partial autocorrelation is significant at that lag. What Cutting et al. (2010) call the autoregressive index (AR index) is determined by the position of the first lag that fails to be significant. This is at lag 6 (just) for King Kong so its AR index is 5. The larger the index, as Cutting et al. interpret it, the more clustered a film is into packets of shots of similar length.

As described above the AR index can only take on the values 0, 1, 2, .... The cut-off point for determining significance depends on the sample size (via $2/\sqrt{n}$) and Cutting et al. express the view that ‘films with fewer shots are penalized; their bounds are higher, which tends to generate smaller AR indices’. They thus prefer to work with a modified autoregressive (MAR) index. This is obtained by fitting a smooth decay curve through the partial autocorrelations – the solid red curve in the figure – and defining the MAR index as the point of intersection with the solid horizontal line at 0.065, which is based on the mean number of shots in all films. For King Kong an MAR of 4.59 results.

---

I used the nls function in R with the same model described in Cutting et al. Their results are used in the right-hand panel of the figure.
Figure 2.6: To the left the partial autocorrelation function for the SLs of King Kong (2005), with a fitted curve. This is used to estimate a modified autoregressive index, plotted for 150 films in the right-hand plot. See the text for further details.

I have technical reservations about the procedures used that are discussed in detail in Baxter (2013c) and think the analytical lily is being rather over-gilded. The most important concern is that the MAR appears to be a surrogate measure for the (partial) autocorrelation at lag 1 and cannot be interpreted as an index of clustering in the sense understood in Cutting et al. (2010). This invalidates their interpretation of their results. For illustration, however, discussion of the right-hand panel of Figure 2.6 is based on the results in the supplement to Cutting et al. (2010) and on their terms.

The MARs are plotted against date in the right-hand panel of Figure 2.6; the dashed (blue) straight line is that fitted to the data in Cutting et al. (2010) (the upper-left panel in their Figure 2). From it they conclude that the results suggest ‘that Hollywood film has become increasingly clustered in packets of shots of similar length’ and that in this and other matters ‘film editors and directors have incrementally increased their control over the visual momentum of their narratives, making the relations among shot lengths more coherent over a 70-year span’. A more nuanced interpretation is possible.

The solid (red) curve (a loess smooth) does not presume linearity. There are subtle differences, disguised a little by the scale on which the curves are plotted. There is very little change over about 20 years from 1935 to about 1955. For the next 35 years or so there is a gradual increase, but also a point of inflection at about 1975 where the rate of increase seems to decrease, with a flattening out at about 1990. Either interpretation shows an increase over time, albeit a fairly modest one; that based on the loess smooth suggests that the increase mainly occurred over a period between about 1955 to 1990 and may have flattened out. You really need more, and more representative data7, and a longer time period, to be assertive about any of this; the conclusions are interesting, but the more ‘nuanced’ interpretation possibly has less of a ‘wow’ factor than that in the original publication.

There is an interesting interchange between Salt and Cutting about the paper under discussion here on the Cinemetrics website. Start with http://www.cinemetrics.lv/salt_on_cutting.php and follow the thread8. Cutting’s second contribution explains what the PACF is in more detail than

---

7 Of the 150 films used those from 1980 on were selected to be high-grossing ones, and earlier films selected from those with high ratings on the Internet Movie Database.

8 Since I first wrote this further debate about the validity of the analysis in Cutting et al. (2010), to August
attempted here\textsuperscript{9}, and both commentators discuss the relationship of AR and MAR indices to other forms of representing internal structure covered in Salt (2010). Salt usefully discusses how structure of the kind both authors deal with, and changes over time, can emerge from film-making practice at the level of individual films.

2.2.6 Shot-scale analysis

Table 2.1 is an example of shot-scale data, for a sample of Fritz Lang films, extracted from Barry Salt’s database on the Cinemetrics website. The analysis of such data was introduced in Salt (1974), and such analyses are a dominant feature of Chapters 12, 16, 19, 24 and 26 of Salt (2009). Shots are classified by scale as big close-up (BCU), close-up (CU), medium close-up (MC), medium shot (MS), medium long shot (MLS), long shot (LS) and very long shot (VLS). Illustrations of what is to be understood by these terms are provided in Salt (2009, p.156) and Salt (2006, p.253).

The row numbers in the tables are scaled to add to 100%; Salt scales numbers to add to 500.

<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>BCU</th>
<th>CU</th>
<th>MCU</th>
<th>MS</th>
<th>MLS</th>
<th>LS</th>
<th>VLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinnen, Die (1)</td>
<td>1919</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>20</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>Spinnen, Die (2)</td>
<td>1920</td>
<td>6</td>
<td>13</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>20</td>
<td>45</td>
</tr>
<tr>
<td>M&quot;ude Tod, Der</td>
<td>1921</td>
<td>4</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>22</td>
<td>43</td>
<td>4</td>
</tr>
<tr>
<td>Dr. Mabuse der Spieler (1)</td>
<td>1922</td>
<td>6</td>
<td>12</td>
<td>14</td>
<td>14</td>
<td>25</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>Dr. Mabuse der Spieler (2)</td>
<td>1922</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>26</td>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td>Kriemhild’s Rache</td>
<td>1924</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>9</td>
<td>17</td>
<td>60</td>
<td>4</td>
</tr>
<tr>
<td>Siegfried</td>
<td>1924</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>17</td>
<td>56</td>
<td>5</td>
</tr>
<tr>
<td>Metropolis</td>
<td>1926</td>
<td>1</td>
<td>10</td>
<td>12</td>
<td>15</td>
<td>14</td>
<td>39</td>
<td>9</td>
</tr>
<tr>
<td>Spione</td>
<td>1928</td>
<td>6</td>
<td>9</td>
<td>15</td>
<td>21</td>
<td>25</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>1931</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>20</td>
<td>24</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>Testament des Dr. Mabuse, Das</td>
<td>1933</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>16</td>
<td>28</td>
<td>33</td>
<td>4</td>
</tr>
<tr>
<td>Fury</td>
<td>1936</td>
<td>3</td>
<td>13</td>
<td>15</td>
<td>23</td>
<td>24</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>You Only Live Once</td>
<td>1937</td>
<td>9</td>
<td>16</td>
<td>14</td>
<td>22</td>
<td>24</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>You and Me</td>
<td>1938</td>
<td>4</td>
<td>15</td>
<td>22</td>
<td>23</td>
<td>19</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Hangmen Also Die</td>
<td>1943</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>19</td>
<td>26</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>Ministry of Fear, The</td>
<td>1945</td>
<td>3</td>
<td>1</td>
<td>14</td>
<td>22</td>
<td>22</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>Woman in the Window, The</td>
<td>1945</td>
<td>3</td>
<td>12</td>
<td>19</td>
<td>28</td>
<td>16</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Cloak and Dagger</td>
<td>1946</td>
<td>2</td>
<td>10</td>
<td>17</td>
<td>30</td>
<td>17</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Secret Beyond the Door, The</td>
<td>1948</td>
<td>3</td>
<td>5</td>
<td>17</td>
<td>26</td>
<td>24</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>Human Desire</td>
<td>1954</td>
<td>1</td>
<td>16</td>
<td>14</td>
<td>22</td>
<td>22</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>Beyond a Reasonable Doubt</td>
<td>1956</td>
<td>6</td>
<td>9</td>
<td>23</td>
<td>28</td>
<td>20</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>While the City Sleeps</td>
<td>1956</td>
<td>6</td>
<td>16</td>
<td>27</td>
<td>21</td>
<td>16</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Indische Grabmal, Das</td>
<td>1959</td>
<td>2</td>
<td>12</td>
<td>11</td>
<td>13</td>
<td>20</td>
<td>41</td>
<td>2</td>
</tr>
<tr>
<td>Tiger von Eschnapur, Der</td>
<td>1959</td>
<td>0</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>20</td>
<td>42</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2.1: Shot scale data (%) for films of Fritz Lang.

The data are represented as bar charts in Figure 2.7 and something like this is the standard way of presenting such data. Salt (2009) provides about 180 examples in the chapters cited above\textsuperscript{10}. Presented in this way, comparisons of whatever kind one wishes to effect – between specific films, directors, time periods etc. – requires visual comparison of the shapes of the bar charts, which is subjective. This can be (unconsciously) affected by the differential ‘weight’ given to absolute or relative differences in the heights of bars for different categories. Where a more ‘holistic’ appraisal of differences in the general shape of the charts is attempted this can be rendered cumbersome by what may be an unavoidable lack of concision of presentation; some of Salt’s

\textsuperscript{9}Notwithstanding what is in some of the cinemetric literature, ‘bar charts’, rather than ‘histogram’, is the appropriate form of diagram (and terminology) that should be used for the graphical representation of shot-scale data, and the bars should have gaps between them to make it clear that one is dealing with counts for categorized data, rather than grouped continuous data (e.g., SLs) for which histograms would be appropriate.

\textsuperscript{10}As it is relevant to other aspects of these notes, Cutting’s discussion of normality and lognormality (in his first contribution) is potentially misleading and needs mention. Data standardization, to zero mean and unit variance, and normalization (transformation to a normal distribution) are treated as the same thing. The confusion is common, but the distinction is an important one. Some of the examples Cutting uses to illustrate the prevalence of lognormality, based on histograms of log-transformed data, are also arguably misleading because the interval widths used for the histograms are too large to make useful judgments (Section 6.4.5).
Figure 2.7: Bar charts of shot-scale data for 24 films of Fritz Lang. Blue shading for early German silent films, green for early German sound, red for American and orange for late German.
illustrations are spread over five pages, and it can be difficult for a reader to follow an argument if reference to charts on several different pages is required.

None of this is to suggest that the conclusions derived from this kind of analysis are flawed – quite the opposite – but it can be time-consuming to arrive at them. What is now presented is an alternative approach to analysis, correspondence analysis, that operates in exactly the same spirit, but results in a much more concise representation of the data (one graph) and easier interpretation.

Figure 2.8: A correspondence analysis for the shot-scale data for films of Fritz Lang.

Details of how correspondence analysis works are left to Chapter 9. For current illustration it is enough to know that the rows of Table 2.1 can be reduced to points on a map. The distance between any two points on the map (i.e., their proximity) is an indicator of how similar their shot-scale profiles are. For example, in Figure 2.8 the two films furthest to the left in the upper-left quadrant, are very close, suggesting that they have very similar shot-scale profiles. The films in question are the *Die Niebelungen* films, *Siegfried* (1924) and *Kriemhild’s Rache* (1924). The reasons for their similarity, and comparative difference from other films, can be confirmed by inspection of the bar charts or, for preference, the numbers in the table. It can be seen that the two films have fewer close-ups (of any kind and as a proportion) and a rather higher proportion of long shots than other films.

In the figure, films are labeled by type (silent/sound) and origin (German/American) and the German silents and American Films have been ‘fenced in’ by joining up the dots that define the outer limits of their scatter\(^{11}\). The two sets of films are fairly clearly stylistically distinct – the only film corralled by both fences, and then only just, is *Hangmen Also Die* (1943). The cluster for the American films is a bit ‘tighter’ than that for the German silents.

The early German sound films, *M* (1931) and *Das Testament des Dr. Mabuse* (1933), sit more comfortably with the German silents than they do with the later American sound films. The late ‘Indian’ films, *Das Indische Grabmal* (1959) and *Der Tiger von Eschnapur* (1959), made after

\(^{11}\)Called a *convex hull* in respectable statistical parlance.
Lang returned to Germany in the late 1950s, are interesting in that they sit comfortably in the midst of the silent German films made 30-40 years earlier\textsuperscript{12}.

Having established that there is pattern in the data, a difference between the silent and American films in this case, it’s then of interest to ask what features of the data are responsible for it. The correspondence analysis provides information on this as well. Markers corresponding to the columns, in this case shot-scales, can be plotted on the same graph as the row markers (films). Think of them as landmarks, defining the nature of the terrain in which they sit. The films partake, as it were, of more of the terrain corresponding to the landmarks to which they are closest and less of that of distant landmarks\textsuperscript{13}. The closest marker to the \textit{Niebelungen} films, for example, is that for long shots, and the furthest markers are those associated with the various degrees of close-up. This confirms the observations about these films, three paragraphs ago, based on the table and bar charts.

As further illustration of interpretation, the six films closest to the BCU marker are among the eight with percentages of 6\% or more for BCUs in Table 2.1. These six are the two early German sound films, \textit{M} (1931), \textit{Das Testament Des Dr. Mabuse} (1933), three silents (from left to right) \textit{Die Spinnen} (1) (1919), Dr. \textit{Mabuse der Spieler} (1) (1922), \textit{Spione} (1928), and one American sound film \textit{You Only Live Once} (1937). The two American films with a relatively high proportion of BCUs that plot differently, late and the last Lang made before his return to Germany, are \textit{Beyond a Reasonable Doubt} (1956) and \textit{While the City Sleeps} (1956). They are the two films that plot furthest to the right in the top-right quadrant (on the same side of the graph as the BCU marker) and are distinguished from the other six films by having a noticeably larger proportion of MCUs, to which they plot closely and fewer MLSs and LSs.

Nothing new is claimed for any of the interpretation here; Salt (2009, pp 242-243) covers the main points, for example. This does provide confirmation (or reassurance) that the correspondence analysis is producing sensible results and I’d argue it’s a more effective way of seeing what the data have to tell you.

A second illustration of correspondence analysis is provided for 33 sound films of Alfred Hitchcock (14 British and 19 American) from 1929 to 1963, 13 early German and American sound films of Fritz Lang (1931-1956) and 18 films of Max Ophuls (1931-1955). The data are from Barry Salt’s database, and have been selected to make some additional interpretive points. labeled output from the correspondence analysis is shown in Figure 2.9.

In the previous illustration ‘fencing-in’ the Lang silents and American films was a fairly natural thing to do. The same might be done here, but is less helpful because the spread of some of the data, and consequent overlapping of the convex hulls, obscures the dominant patterns in the data. What has been done is to show boundaries that encompass the bulk of the Hitchcock American, Lang and Ophuls films, excluding 3, 2 and 3 films respectively, to emphasise that these bodies of films, in terms of their shot-scale distributions, are largely stylistically different. The British Hitchcock films have been left as individual points as they are quite scattered and overlap with each of the groups that have been isolated here.

The two unenclosed Lang films are the earliest and only German ones in the sample used, \textit{M} (1931) and \textit{Das Testament Des Dr. Mabuse}. This separation from Lang’s American films was

\textsuperscript{12}Barry Salt has previously commented (on a blog of Nick Redfern’s from June 2009) that ‘Lang reverted to his German scale of shot for \textit{Der Tiger von Eschnapur} at the end of his life’. Eisner (1976) notes that the films were based on a scenario by Lang and Thea von Harbou, also credited with the screen play for \textit{Das Indische Grabmal} (1921) when the idea was that Lang would direct it, before it ‘had been taken over by Joe May on the pretext that Lang was too young to direct this subject’. Eisner further notes, of the later films, that occasionally ‘the stylisation and deliberate abstraction recall the \textit{Niebelungen}’.

\textsuperscript{13}Anyone familiar with the statistical literature on CA will know that a lot of effort has gone into discussing appropriate ways of plotting column markers in relation to row markers. One style of plotting places the column markers on the periphery of the plot some way from the row markers. It is often emphasized that you \textit{cannot} interpret the difference between a row and column marker as a ‘distance’ in the way mathematicians understand the term. The interpretation offered here, therefore, is not a mathematically rigorous one. It is how interpretation of correspondence analysis plots is often done, however, and often works well, but its best to check conclusions make sense by reference back to the original data. For preference, partly to avoid this issue and partly to reduce over-crowding of plots, I’d usually plot row and column markers separately, but joint plotting is the default in much software.
also evident in Figure 2.8. Two of the three Ophuls films that sit closest to the pair, *Lachende Erben* (1932) and *Liebelei* (1932), are of comparable date. Their similarities can be confirmed by looking at the relevant bar charts in Figure 2.11, comparison across pages not being needed in this instance.

Three of the American Hitchcock’s sit outside the main body. Reading from most to least distant (bottom to top on the plot) the ‘outlying’ films are *Dial M for Murder* (1954), *The Trouble with Harry* (1954) and *To Catch a Thief* (1955). Why these films differ from the bulk highlighted in the plot can be investigated using the bar charts, but from the correspondence analysis itself it can be inferred that the first two films make rather more use of MLSs and rather less use of CUs and BCUs than films in the main body (see shortly). The correctness of this inference is readily enough confirmed.

The three Ophuls films not enclosed by the convex hull are among his four latest films in the sample. Two, the lowest of the Ophuls films on the plot, *La Ronde* (1950) and *Madame de . . .* (1953) are distinguished by having a somewhat higher ratio of MLSs to LSs and VLSs than the typical Ophuls film. At the other extreme, sitting outside the convex hull, and the furthest right film on the plot, *Le Plaisir* (1953) has much more emphasis on LSs and VLSs compared to MLSs (precisely twice as many, compared to 0.83 and 0.73 for the other outlying Ophuls).

Detailed discussion of the British Hitchcock films is deferred to Chapter 9. It suffices here to note that it is possible to see a stylistic development from early to late Hitchcock that is better appreciated if his films are examined in isolation from that of the other two directors. Salt (2009, p.243) comments on this in connection with a comparison with Lang, and their respective moves to Hollywood in the 1930s.

Detailed discussion of the interpretation of the shot-scale markers has been deferred until now. Correspondence analysis is widely used in quantitative archaeological applications (Baxter, 203, pp.136-146), often applied to tables of a form similar to Table 2.1 where rows correspond
Figure 2.10: Bar charts of shot-scale data for films of Hitchcock (dark red for British, red for American).
Figure 2.11: Bar charts of shot-scale data for films of Hitchcock (red), Lang (blue) and Ophuls (green).
to archaeological contexts, such as graves, and columns to artefact types (pottery, jewellery etc.) buried with the dead, table entries corresponding to counts of the artefact type, or simply 1 or 0 indicating presence or absence. Fashions change, so if the contexts range over any great time span early graves would not be expected to share any artefact types in common with later graves. Graves of a similar period are expected to share some artefact types in common.

One use of correspondence analysis is to re-order the rows and columns of the table to reflect both the temporal sequence of burials and the associated developments in fashion of the artefact types. If subjected to a correspondence analysis, and if successful, the resultant plot – for both row and column markers – typically shows a fairly clear ‘horseshoe’ pattern that, subject to checking against other forms of evidence, can be interpreted as a temporal gradient. This is called seriation in archaeology.

From what I’ve seen, cinemetric data doesn’t quite behave like this. You don’t get the tight clustering of row markers (films) about a horseshoe that is hoped for in archaeological seriation. The shot-scale markers in Figure 2.9 behave as you might hope, and can be interpreted as a stylistic gradient. This is made more evident if you join up the dots corresponding to the shot-scale markers in their natural sequence from BCU to VLS. The horseshoe is possibly not as close to ideal as one might wish, but is as satisfactory as often occurs in reality. The configuration in Figure 2.8, isn’t as satisfactory, which was why discussion was deferred at that point. It does, though, do a good job at separating scales in the MLS to VLS range from those closer up.

For those who prefer linearity (see the relevant footnote) it is possible to imagine straightening the horseshoe, the different ends carrying with them the points closest to them. Do this and you get a continuum with American Hitchcock dominating one end, Ophuls the other and Lang in the middle. British Hitchcock is spread around more, mainly in the middle and at the American Hitchcock end. Any evidence for a temporal gradient within the Hitchcock films, alluded to above, is not displayed with the labeling used here, and is better dealt with in a separate analysis.

---

14The horseshoe curve arises for mathematical reasons and is a non-linear gradient. This has troubled some, who would prefer their gradients linear (in vegetation science, for example). Techniques exist, such as detrended correspondence analysis, that attempt to ‘unbend’ the horseshoe, but this doesn’t usually trouble archaeologists, who are delighted to see the appearance of a horseshoe as vindication of their search for temporal structure.
Chapter 3

Getting R, getting started

3.1 Finding R

Either Google CRAN R (the Comprehensive R Archive Network) or use

http://cran.r-project.org/

which is where to begin. You are directed to pages that tell you how to install R on various platforms, and more information, if needed, is provided in the FAQs. R is updated frequently. Documentation on R is comprehensive and much of it is free. It is well worth looking at what is available in CRAN at an early stage.

3.2 Data entry

3.2.1 General

For other than very small data sets it is best to import data from an external source. This can be done in different ways. Although not the preferred method of the developers of R, many users may find it simplest, if starting from scratch, to create and import an Excel file. Use ?read.table in R to see documentation on what is available.

Create the data file; it is assumed below that headers naming the columns are given. Spaces in headers must be avoided (and if other illegal characters are used an error message in R will inform you). Next, highlight the data you want to import; copy it (to the clipboard) and go into R. You name the data file when reading it into R; for illustration the data for A Night at the Opera (1935), submitted by James Cutting, is used (see Section 3.2.2 for more detail), which will be named Night at the Opera in R. Type

Night at the Opera <- read.table(file = "clipboard", header = T)

and type Night at the Opera to see the result. Here <- is the assignment operator, and note that clipboard must be enclosed in double inverted commas or an error message results. If data are missing R requires that the offending cell be filled with NA. To use read.table a rectangular table of data is expected. Commands^ are preceded by the R prompt >.

It is best to keep headers informative but short (in writing-up an analysis or captioning a figure a key can always be provided). Headers beginning with a number are allowed, but column names in R will not be quite as you expect.

^You will notice, by the way, that R is directive- or command-driven rather than menu-driven. That is, you have to tell it what to do by writing things down. People (i.e. students I’ve taught over the years) can find this a confusing and challenging idea. Anyone who has ever word-processed a sentence, paying attention to spelling and syntax, is more than equipped to meet this challenge.
3.2.2 Using the Cinemetrics database

Some familiarity with the Cinemetrics database, http://www.cinemetrics.lv/database.php, is assumed here. Firstly, films recorded in the simple mode are used.

Within the database select A Night at the Opera; show the raw data, select it, and copy into an Excel file. The following protocol will be adopted here of editing the three column headers to read Id, SL, and Cut. This should leave you with an Excel file of three columns with these headers. Copy the file, go into R and read the data in as described above.

The shot length, SL, and cut point, cut, data are recorded in deci-seconds. For these notes it is convenient both to convert these to seconds and have a separate variable for shot lengths only. Either of the following commands will do this.

\[
\text{SL.Night.at.the.Opera <- Night.at.the.Opera$SL/10}
\]

\[
\text{SL.Night.at.the.Opera <- Night.at.the.Opera[,2]/10}
\]

The first of these picks out the variable of interest by name, the second by column number. The first version is generally neater. Conversion to seconds can be done before reading the data into R, in which case omit the division by 10.

Secondly, for films recorded in advanced mode, there is a fourth column corresponding to shot type that it is assumed will be named Type. For illustration Charles O’Brien’s submission of Lights of New York (1928) is used. The Type variable identifies whether the shot is ‘action’, ‘dialog’ or a title, ‘exp.tit’. If the information on the type is not needed just proceed as already described, naming the table of data LONY and creating a shot length variable SL.LONY, say. If the type variable is wanted create it using one of the following.

\[
\text{SL.LONY.Type <- LONY$Type}
\]
\[
\text{SL.LONY.Type <- LONY[,4]}
\]

How this might be used is taken a little further in Section 4.4

3.3 Packages

Packages are collections of functions that, together with arguments provided to them, control the analyses undertaken. Some packages are loaded automatically with R and the functions in them are immediately accessible. Others are bundled with R and must be loaded before use. Yet others need to be imported before they can be loaded.

For illustration the bundled MASS package, associated with the book Modern Applied Statistics with S: Fourth Edition (Venables and Ripley, 2002), is used. In the following code everything that precedes # should be typed; everything after is my comment. If writing your own code # can be used either to annotate it or comment out parts you don’t want to use for any particular analysis. Note the use of the library function for obtaining information about packages.

\[
\text{library(MASS)} # loads MASS}
\]
\[
\text{library(help = MASS) # lists available functions}
\]
\[
\text{?truehist # information on the function truehist in MASS}
\]

Importing packages is done in two stages. In R from the Packages menu select Set CRAN mirror to choose a site to import from then, from the same menu, select Install package(s) and then select the package you want to import. The package needs to be loaded before use with library(packagename) as above.

There is an overwhelming number of packages available. If you know what you want, all well and good. If not, Google is invaluable if the search terms are ‘R’ and the name of the technique of interest.
3.4 Reading

This is enough to get started, which is all that is being attempted here. Section 4.3 in the next chapter provides an introduction to writing functions in R, at a very basic level, which is useful to know about early. Subsequent chapters provide code to go with the analyses illustrated and introduce useful things like the labeling of graphs, in an incremental fashion.

There is, as noted, a vast amount of help around, much of it free. The CRAN site lists over a hundred books, in various languages and at all levels, many on specialized topics. I found the Venables and Ripley (2002) book and earlier editions invaluable when teaching myself about R and the related and earlier commercial package S-Plus. The title of the book reflects its earlier origins, but it works with R and I continue to find it invaluable. It is not an introductory text for non-statisticians and Dalgaard (2008) would be more suitable as an elementary-level introduction. A strength of R is its graphical capabilities, and Murrell (2011) collects together useful information on this, not always easily found elsewhere.
Chapter 4

Descriptive statistics

4.1 Introduction

This chapter provides a brief introduction to the calculation in R of the more commonly used descriptive statistics in cinematics, with a section (Section 4.5) that illustrates what you might do with some of these. That section could be regarded as a continuation of the examples provided in Section 2, but also introduces some of the practicalities of using R, as well as showing how R encourages the exploratory analysis of data. Definitions of some of the simpler statistics in wide use are defined and discussed in Section 4.6.

If interest centers solely on descriptive statistics commonly used in summarizing SLs, such as the average shot-length (ASL) or median shot-length (MSL), and the film is in the Cinematics database they can be found there. The following Section 4.2 shows how these are obtained in R. It’s easy enough to obtain the statistics for a large body of films in a single analysis, but it requires a different approach to the method of data entry described in the last chapter and assumed for the next few chapters on graphical analysis. Discussion of data entry for large data sets is provided in Chapter 10. What you might do with statistics so acquired is, as just noted, explored in Section 4.5.

Section 4.3 is useful. Functions are collections of commands that you create and give a name to that, among other things, allows you to execute a range of different analyses for a single film and/or repeat the same type of analysis for as many films as you wish. Once you’ve successfully created a function they save a lot of time if you are attempting more than infrequent ‘once off’ analyses.

The other thing to know about at an early stage is what I’ve called data manipulation in Section 4.4. Just because someone, possibly yourself, has gone to the trouble of collecting data on SLs for a film doesn’t mean you’re obliged to use all of it. Provided you know what the shot-type is, omission of those associated with the opening credits, or intertitles in silent films might be of interest, for example. The basics of this are discussed.

Among the simpler statistics the relative merits of the ASL and MSL have attracted debate in the literature. This is touched on in Section 4.5, which can also be regarded as an hors d’oeuvres for the graphically oriented chapters to follow. The debate is the basis for the case study in Chapter 11.

4.2 Basics

Assuming the data for A Night at the Opera has been imported into R and SLs, in seconds, saved in SL.Night_at_the_Opera it is convenient, to save typing, to define

\[ z \leftarrow \text{SL.Night_at_the_Opera} \]

then
mean(z) # ASL
median(z) # MSL
median(z)/mean(z) # MSL/ASL
sum(z)/60 # LEN (length in minutes)
length(z) # NoS (number of shots)
max(z) # MAX
min(z) # MIN
max(z) - min(z) # Range
ds(z) # StDev
sd(z)/mean(z) # CV (coefficient of variation)

gives the statistics reported with a Cinemetrics graph. Some of these are, perhaps, obvious, but they are defined and aspects of their use discussed in Section 4.6. The # character indicates a comment giving the names used in Cinemetrics.

Some of these statistics can be obtained with a single command. Thus, summary(z) returns

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.500 2.300 4.000 8.978 8.500 207.200
```

Here 1st Qu. and 3rd Qu. are the first and third quartiles, the difference between them being the interquartile range (IQR). They can be obtained using

```
quantile(z, .25) # Q1
quantile(z, .75) # Q3
IQR(z) # IQR
```

the names being those given to the quantities when they are discussed later1. The first quartile is the 25th percentile and the quantile function can be used to find percentiles other than the 25th and 75th illustrated above.

### 4.3 Functions

Googling the terms descriptive statistics and R brings up quite a few packages with functions that produce a range of statistics. None of these ever seem to do quite what you want, and it’s tedious to keep typing the commands of the previous section if doing a lot of analyses.

It’s easier to write a function to do exactly what you want. This sort of thing is often introduced at a later stage in introductory R notes, but worth engaging with early. Begin with something like

```
filmstats <- function(film) {}
filmstats <- edit(filmstats)
```

which brings up an edit screen where you can type in what you wish, thus ending with something like

```
function(film) {
z <- film
ASL <- mean(z)
MSL <- median(z)
ratio <- MSL/ASL
list(ASL = ASL, median = MSL, ratio = ratio) # List of statistics
}
```

---

1I’m resisting the idea of giving a precise definition of ‘quartile’. There are nine different algorithms available in R which can give slightly different results that don’t matter, given the size of data set usually available for analysis in cinemetrics.
listing just three statistics for brevity. The ASL and MSL are obtained using the functions \texttt{mean} and \texttt{median}; their ratio has to be calculated. Once done\footnote{On exiting from \textit{edit} mode and saving the work you will be notified of any errors. Sometimes these result in the work not being saved, which can be a pain. It is safest to copy work before you exit so it can be pasted back for correction if needed. Error messages are not always as helpful as you might wish.} \texttt{filmstats(SL.Night\_at\_the\_Opera)} will return

\begin{verbatim}
$ASL
[1] 8.978319
$median
[1] 4
$ratio
[1] 0.4455177
\end{verbatim}

As with much statistical software, numbers are reported with too many decimal places. This can be cured using \texttt{list(ASL = round(ASL,2), median = MSL, ratio = round(ratio,2))} in the function, rounding to two decimal places.

The obvious advantage of doing things in this way is that when invoking the function the data for any other film can replace \texttt{SL.Night\_at\_the\_Opera}. It's possible to write functions to do the computations for several films in one go; this is covered in Chapter 10.

### 4.4 Data manipulation

An initial inspection of SL data may suggest either that analysis is better conducted after some omissions (e.g., opening credits), or that investigation of the effects of omitting some shots (e.g., suspected outliers) is desirable. This is straightforward if you know what you want to omit. With no particular film in mind, and for example

\begin{verbatim}
mean(z[-1])  # Omit the first shot
mean(z[-c(1:4)])  # Omit the first 4 shots
mean(z[-347])  # Omit shot 347
mean(z[-c(5,108,347)])  # Omit the three shots listed
\end{verbatim}

The 69th SL for \textit{A Night at the Opera}, at 207.2, is somewhat longer than the next longest at 111.5. The command \texttt{filmstats(SL.Night\_at\_the\_Opera[-69])} returns an ASL of 8.64 compared to 8.98 using all the data.

The command \texttt{z <- sort(z, decreasing = T)} will reorder the SLs in \texttt{z} from largest to smallest. If the effect of possible outliers on, for example, ASL calculations is a concern this can be investigated systematically basing analyses on \texttt{z, z[-c(1)], z[-c(1:2)], z[-c(1:3)]} etc.

Suppose that a film has been recorded in advanced mode, as discussed in Section 3.2 for \textit{Lights of New York}, where the type of shot is recorded, one of the categories being titles, \texttt{exp.tit}. The SL and type variables were named \texttt{SL.LONY} and \texttt{SL.LONY.Type} in \texttt{R}. Suppose an analysis is to be undertaken omitting title shots. The necessary data is most simply created using

\begin{verbatim}
SL.LONY.NoTitle <- SL.LONY[SL.LONY.Type != "exp.tit"]
\end{verbatim}

The \texttt{!=} part of the command is to be read as 'not equal to' so what is being selected are shots that are \textit{not} categorized as titles, \texttt{exp.tit}, in the type variable\footnote{You need to be careful to both enclose the type category selected in double inverted commas and use square rather than round brackets. This can be tedious and it’s easy to slip-up, but you get used to it.}. If you wanted to select titles only use \texttt{==} rather than \texttt{!=}.

26
4.5 Illustrative graphical analyses

There is, as the *On Statistics discussion Median or Mean?* shows, some argument about the appropriate choice of statistics for summarizing SL distributions, most obviously whether the ASL or MSL is a better descriptor of ‘film style’. A similar, if less publicized, choice exists in terms of choosing a measure of dispersion, but usually the standard deviation goes with the ASL and the interquartile range (IQR) with the median. Redfern (2010a) has noted some alternatives to the IQR, all robust measures of dispersion considered below.

Rather than arguing from principle about what should be the preferred choice, a pragmatic approach is to see if the choice makes much difference when a large body of films are considered together. Those used here are 134 of the 150 described in Cutting et al. (2010). These are (mostly) Hollywood films from 1935-2005, selected as ‘high-profile’ in some way, at least in their day; 16 of the 150 are not used because they have recorded zero or negative SLs that are presumably recording errors (Redfern, 2012a) which preclude the analysis after log-transformation carried out elsewhere in these notes.

Four variables have been selected for illustration, the ASL, MSL, standard deviation (SD) and IQR. Figure 4.1 shows examples of what are called pairs plots or scatterplot matrices – just a collection of all the two-way plots it is possible to get for the variables. Each pair is plotted twice, with the axes reversed. Thus the first plot along at the top is of ASL against MSL; the first one down on the left is of MSL against ASL etc. It is assumed that four variables have been created for use in R named ASL, Median, SD and IQR. This is assumed here for ease of exposition.

The two plots shown are, enhancements apart, the same. The basic plot to the right is obtained from the following code where `cbind` combines the statistics into a table with four columns and `pairs` produces a basic scatterplot matrix based on the columns.

```r
stats1 <- cbind(ASL, Median, SD, IQR)
pairs(stats1)
```

Some broad features are immediately obvious. One is that there is a reasonably good linear relationship between most pairs of variables. The strongest, interestingly, is that between the...

---

4This is actually what I did here, operating on data for individual films already entered into R. If the data are created externally, in the form of a table put together and published by someone else, for example, and if they are in Excel, the table can be read in and individual variables extracted and named as described in Section 3.2.2.
ASL and IQR, the weakest that between the SD and IQR. One implication of this, with caveats to be entered, is that for any study that looks for patterns in a large body of data, it probably doesn’t matter much whether the ASL or MSL is used as a measure of location. Baxter (2012a) has suggested this and, equivalently, Salt (2011) has noted ‘a relatively fixed ratio between the ASL and Median’, with 82% of films in his sample of 1520 having an MSL/ASL ratio in the range 0.5-0.7, a ‘remarkable fact’ that ‘demands explanation’.

The first caveat is that for films with the larger ASLs (greater than about 12) and MSLs (greater than about 7-8) the linear relationships are not so strong. Salt has suggested in a number of places that general patterns can be expected to break down for films with large enough ASLs, 15 seconds sometimes being suggested as a rule of thumb. The second caveat is that there is a significant minority of films for which none of these statistics are particularly suitable summary measures (Baxter, 2012a), and these may or may not show up as departures from the general trends on the plots.

As well as noticing broad trends, and general departures from them, one can begin to pick out films that look statistically unusual for some reason. For example, three or four films stand out in the top-right corner of those plots involving the ASL and MSL. These are, chronologically, Detour (1945), Harvey (1950), The Seven Year Itch (1955) and Exodus (1960). Detailed examination of these, of the kind discussed in Section 6.4.4, suggests that Harvey and The Seven Year Itch are not suitably summarized by the ASL or the MSL, whereas the other two films are among the few that Redfern (2012a) (using the same sample as here) finds to have acceptably lognormal distributions5.

It’s noticeable that these films are relatively early in the 1935–2005 date range. No analysis takes place in a vacuum, and it’s well-known that ASLs (and MSLs) have declined steadily over the last 50 years and more. This is exploited in the left-hand panel of Figure 4.1 where films are coded into three date ranges, 1935–1955 (black open circles), 1960–1975 (red open triangles) and 1980–2005 (green crosses)6. The relationship between date and size of the ASL and median is immediately apparent, with the later films concentrated in the lower left for each plot and showing less obvious departure from a linear scatter.

The date codes were (1, 2, 3) and a variable DateGroup was created for this in R. The plot was produced using the scatterplotMatrix function from the car package, which needs to be installed (Section 3.3). Thus

```r
library(car)
scatterplotMatrix(~ ASL + Median + SD + IQR, smooth = F, by.groups = F,
                  groups = DateGroup, legend.plot = F)
```

does what’s required. This is quite a useful function, so worth discussing in a little detail. The variables you want plotting are listed here separately, preceded by ~. By default a straight line is fitted through the data, which can be suppressed but does no harm here and is left in. A smooth (loess) curve is also fitted through the data by default (Section 7.5), but this seems superfluous here, given the linearity, and has been suppressed by smooth = F. The groups = DateGroup argument tells R where to get the labelling information from; lines are fitted through the separate groups by default, which may be useful or may complicate the plot too much and has been suppressed here by by.groups = F. A legend is supplied by default, which can sometimes obscure things, so has been omitted here by legend.plot = F.

By default kernel density estimates (KDEs) (Section 5.2) are shown on the diagonal, for each variable separately, with a ‘rug’ showing individual data points. This is optional and can be replaced by other graphical displays, or omitted (use ?scatterplotMatrix to see the options). Here, they’re harmless and serve mainly to show the similarly skewed nature of the variables.

One of the great strengths of R is the ease with which rapid data exploration can be undertaken including experimentation with different date groups. What’s eventually selected for presentation,

---

5Redfern’s criteria for judging lognormality are stringent (Chapter 12); it means that any film accepted as lognormal by his methods is likely to be accepted as such by other approaches.

6Cutting et al. (2010) selected films at five year intervals.
if that is the end result, is usually arrived at iteratively. For many of the illustrations in these notes, tidying up figures for presentation took longer than the few minutes needed for the several analyses needed before a choice was made. That is, for exploratory analysis, R is very quick once the data are set up.

Figure 4.1 is perhaps on too small a scale to see fine detail but broad patterns are evident. On screen, and depending on its size, the complexity of any pattern and your eyesight, up to 8–10 variables can be examined usefully at one go, it always being possible to separately ‘home-in’ on any plots of particular interest. It’s also possible to concentrate on particular areas of interest by ‘magnifying’ the plot. Figure 4.2 illustrates, where only films with an ASL of less than 7.6 seconds are shown. This was designed to include all the films from 1980 on, with the exception of Coal Miner’s Daughter (1980) and Cast Away (2000) whose ASLs, 10.1 and 9.5, are somewhat bigger than the next largest of 7.5 seconds.

Figure 4.2: A ‘magnified’ version of the right-hand panel of Figure 4.1 for films with ASLs less than 7.6 seconds.

The patterns previously noted ‘hold-up’ pretty well – there’s always the danger that interesting detail can be obscured by inclusion of the more extreme values in a data set (now omitted) which can ‘squash-up’ the smaller values to the point that detail is lost. Not evident from the previous figure, and standing out a little, are four films with SDs greater than 10 that are larger than expected given other pattern, and in relation to the ASLs and MSLs. These are all from the 1960–1975 period and are Those Magnificent Men in Their Flying Machines (1965), Five Easy Pieces (1970), Jaws (1975) and The Rocky Horror Picture Show (1975). The first and last of these have one obvious and extreme outlier and Jaws has, perhaps, three less obviously extreme. Five

---

7One of the reasons for sometimes using logarithms, incidentally, which can remove this problem.
*Easy Pieces* has what I would characterize as an unusual right-tail, rather than a small number of obvious outliers.

The intriguing consistency of the relationship between the ASL and IQR remains, but is not pursued here. It is however, a useful peg on which to hang a discussion of measures of dispersion. Redfern (2010a) has discussed a number of robust alternatives to the SD, in addition to the IQR. One is the median absolute deviation (MAD) defined as

$$\text{MAD} = \text{med}_i|x_i - \text{med}_j x_j|$$

or, in words (a) calculate the median of all the SLs; (b) difference each SL from this median; (c) calculate the median of these differences.

Following Rousseeuw and Croux (1993), whose notation is used here, the other two are $S_n$ and $Q_n$. The first of these is defined as

$$S_n = k_s \text{med}_i(\text{med}_{j|x_i - x_j|})$$

where, in words, (a) for the first shot calculate the difference between its SL and that of each of the other shots; (b) find the median of these differences; (c) repeat this process for each shot in turn so you end up with $n$ medians; (d) find the median of these $n$ medians.

For large samples $Q_n$ is defined as the first quartile of

$$k_q(|x_i - x_j|; i < j)$$

Figure 4.3: A scatterplot matrix of four robust measures of dispersion descriptive of 134 Hollywood films (1935-2005).
which are just the absolute differences in SLs between all possible pairs of shots. In these definitions $k_s$ and $k_q$ are constants that can be ignored for present purposes.\footnote{The constants given in Rousseeuw and Croux (1993) are designed to achieve consistency, meaning that for a very large sample from a normal distribution you get pretty much the same values of $S_n$ and $Q_n$ as for the estimate of the SD. They don’t matter here because only their relative values are of interest.}

For the SLs of an individual film, $z$ say, the statistics are easily obtained as

- \( \text{mad}(z) \)
- \( \text{library(robustbase)} \)
- \( \text{Sn}(z) \)
- \( \text{Qn}(z) \)

the package, \texttt{robustbase}, having been installed to obtain the last two statistics. How these can be obtained for a body of films in a single analysis is discussed in Chapter 10.

The theory that establishes the properties of the statistics is quite complex, but the idea behind them is simple enough. The statistics differ in their ‘degree of robustness’, with the IQR at the bottom, MAD in the middle and the other two at the top. For practical purposes, and typical SL data, do these theoretical properties much matter? Figure 4.3 suggests not.

The results for MAD, $S_n$ and $Q_n$ are almost indistinguishable. There is some difference in results for the IQR, most obviously at the larger values of the statistics. This fairly closely mirrors what was the case for the ASL and MSL in Figure 4.1. The impression left by this last figure, and Figure 4.2 is that any discrepancy in conclusions that might be drawn, would arise in choosing between the SD and IQR (or other robust measures of dispersion).

The general impression, though it merits more systematic investigation, is that for later films from about the mid-1970s, and for comparing a large number of films, the choice of summary statistics to use may not matter much. The same patterns emerge. For earlier films there is sufficient evidence (Baxter, 2012a) that a significant minority exhibit characteristics that render them unsuitable for summary using just one or two statistics, so that generalizations about pattern based on just the ASL and SD, or MSL and IQR, should be approached with caution.

Individual films are the building blocks used when constructing pattern, and close inspection of them means that some may need to be rejected for the purposes of erecting structure. It is arguable, if not undeniable, that in looking at individual films, or comparing small numbers of films, a much richer analysis is possible using graphical rather than summary descriptive methods of analysis. Graphical methods are the subject of the next few chapters.

### 4.6 Definitions and comments on some simple statistics

Salt (2012a), commenting on his original use of the term ‘average shot length’ (ASL) in Salt (1974), said of his envisaged audience that they ‘might be just be able to cope with the notion of the arithmetic mean of a collection of numbers, as long as it was referred to as an average’ (my emphasis). The arithmetic mean is just the sum of a set of $n$ numbers divided by $n$.

Formally (and presumably the kind of thing Salt assumed his audience would not understand) let the $n$ shot-lengths (SLs) be written as \((x_1, x_2, \ldots, x_n)\) then the \textit{arithmetic mean} is defined as

$$
\mu_L = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum x_i/n
$$

where what is called the summation notation, \(\sum\), instructs you to add up the terms that follow.\footnote{It is convenient here to discuss an issue that has attracted some comment in the cinemetrics literature. I have used the notation $\mu_L$ to represent the mean. The subscript $L$ is being used to distinguish between it and the parameter $\mu$ used in defining the lognormal distribution (see the Appendix). The more important point to note is the use of Greek letters. Statisticians typically use such letters from the Greek alphabet for quantities that are descriptive of a \textit{population}. A \textit{population} consists of all the data there is to be observed. The mean is the mean and that’s it; you know it with absolute certainty – it is a \textit{parameter} that describes one aspect of the population. Commonly, only a \textit{sample} from a population is available. A mean can be calculated in the same way, but now it is a sample \textit{estimate} of an \textit{unknown} population parameter. Statisticians often use letters from the Latin alphabet to represent quantities that are descriptive of a \textit{sample}. The mean is still the mean and that’s it; you know it with relative certainty – it is a \textit{statistic} that describes one aspect of the \textit{sample}.}
The median is the central value in a set of ordered data; if \( n \) is an even number there are two central values, and the median is the arithmetic mean of these two values. The idea is that the median splits the data set into two sets of ‘small’ and ‘large’ values with equal numbers of observations in each. Mathematically the notation \( (x_1, x_2, \ldots, x_n) \) is sometimes used to denote the order so that, for example, \( x_1 \) is the smallest value, \( x_2 \) the second smallest, and so on.

The minimum and maximum are, fairly obviously, the smallest and largest SLs, or \( x_1 \) and \( x_n \) in the notation just introduced; the range is the difference between the maximum and minimum.

The arithmetic mean and median are measures of location. The idea, discussed in detail in Chapter 11, is that they are, in some sense, ‘typical’ of a set of data. The range is an example of a measure of dispersion or spread of a set of data.

Other measures of location and dispersion can be defined, an obvious example of the former being the mode, the most commonly occurring value in a set of data. This is a deceptively simple statistic that merits some discussion. If SLs were measured extremely accurately, to (unrealistically) ten-thousandths of a second, they would (usually) all differ and the mode is undefined. In the Cinemetrics database deci-seconds are used; with a fairly large number of shots some SLs will coincide, but there is no reason to expect that they will be ‘typical’, so the mode can be defined but may be of little use. If SLs are grouped into intervals of somewhere between, say, 0.5 to 2 seconds it makes useful sense to talk of a modal class, and such a class will often be evident in histograms of SLs based on these intervals. Salt (2013a) provides some examples that illustrate the effect of varying the interval width.

If one thinks in terms of histograms of SL data, possibly smoothed using something like a kernel density estimate (KDE) (Sections 5.2, 6.4.4), then the mode (or modal class) is where the peak occurs. If there are several clear peaks of differing heights the data are multi-modal and the terms ‘major’ and ‘minor’ modes can be used.

The acronym ASL (where A = ‘average’) is understandable, and established, but it is as well to be aware of precisely what ‘average’ means in this context (i.e. the arithmetic mean). There are at least two reasons for this. One is that, in common parlance, ‘average’ is used very loosely and often understood – if it is understood at all – to be the median or mode. For example, national income distributions are (like SL distributions) typically skewed. Official government statistics for the U.K. for 2011–12 show a mean per-weekly income of £528, a median of £427 and a modal region (it’s rather flat) centered on about £350.\(^\text{10}\) The report from which these figures are taken uses the term ‘average’ but explicitly equates it with the median.

Such scrupulous clarity is not always evident. To paraphrase Humpty Dumpty from Alice in Wonderland, politicians, journalists and others sometimes use the word ‘average’ to mean just to distinguish sample estimates from the population parameters; thus \( \bar{x} \) is in common use for the sample mean. Another convention is the use of ‘hat’ notation, for example \( \hat{x} \), where the ‘hat’ signifies you are dealing with a sample estimate.

Once you start dealing with sample estimates questions arise about how reliably they estimate population quantities. The development of statistical inference in the first half of the twentieth century – one of the great, possibly unsung, intellectual achievements of that period – enabled such questions to be posed and answered in a precise way, giving rise to the machinery of hypothesis/significance testing etc.

There is a view in the cinemetrics literature that SL data for a complete film constitutes a population and that methods of statistical inference are irrelevant. This has been most trenchantly expressed by Salt (2011) who robustly asserts that ‘in Cinemetrics we are dealing with all the shot lengths for the entire film, and this is the whole population, not a sample from it, so the question of sample reliability does not arise, and notions like confidence levels, rejection of the null hypothesis and robustness are both irrelevant and misleading’.

By contrast those who do use these supposedly ‘irrelevant’ methods are behaving as if the SL data are a sample from an infinite population. The use of this kind of as if thinking can be rationalized (using the idea of superpopulations for example) and is an established part of the statisticians repertoire. Not everyone likes it, and Salt is not alone in his thinking, but the argument is not as simple as it’s made out to be.

These differing standpoints go some way to explaining some of the more ‘confrontational’ arguments that have occurred in the cinemetrics literature, where the underlying reasons for disagreement are not always spelled out. I am making no attempt at adjudication since I think the issue is more complicated than is usually admitted. At a more mundane level it helps to be aware of notational niceties that occur in the statistical and cinemetric literature; usage in the latter is inconsistent and sometimes confusing – possibly because the idea of samples is not acknowledged and the Greek and Latin gets mixed up – but that’s another story.

\(^\text{10}\)https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206850/first_release_1112.pdf (Figure 2 – accessed 21/06/2013)
what they choose it to mean – depending, for example, on whether the wealth or poverty of the nation is what they wish to emphasize. The answer to Alice’s question ‘whether you can make words mean so many different things’ is, regrettably, ‘yes’. The academic literature dealing with SLs and their interpretation is not, alas, free of terminological confusion. Both Adams et al. (2002) and Kang (2003) express the view that the median is a better estimate of the average shot length in the presence of outliers, which ignores the way in which ‘average shot length’ has been defined elsewhere in the literature\(^\text{11}\).

The other reason for a clear awareness that the ASL is the arithmetic mean is that different kinds of mean, other than the arithmetic, can be defined, and one of them – the geometric mean – has had a role to play in a cinemetric debate that is the subject of Chapter 12. The geometric mean is defined as the \(n\)th root of the product of the observations or

\[
GM = \left( x_1 x_2 \ldots x_n \right)^{1/n} = \left( \prod x_i \right)^{1/n}
\]

where \(\prod\) is the product operator. Using rules that govern the manipulation of logarithms\(^\text{12}\) we have

\[
\log GM = \sum \log x_i/n
\]

or, in words, the logarithm of the geometric mean is the arithmetic mean of the logarithms.

The relevance of this emerges in analyses concerned with the distribution of SLs, where their lognormality is an issue (see Chapter 12 and the Appendix). One of two parameters, \(\mu\), that define the lognormal distribution can be estimated by log GM, and GM itself is one possible estimate of the median of a lognormal distribution. A demonstration that this particular estimate differs from other possible estimates is a strong plank in the argument that SLs distributions are not, as has been claimed, typically lognormal.

It is generally accepted as good practice that a measure of location should be accompanied by a measure of dispersion, and comparisons between measures of location can be meaningless without this. The mean\(^\text{13}\) is typically accompanied by the standard deviation which is the square-root of the variance or average squared deviation from the mean, defined mathematically as

\[
\sigma_L^2 = \frac{\sum (x_i - \bar{x})^2}{n}.
\]

The ASL and standard deviation are scale-dependent; that is, their values depend on the units of measurement (e.g., seconds or deci-seconds). The coefficient of variation (CV), the ratio of the standard deviation to mean or, in population terms,

\[
CV = \frac{\sigma_L}{\mu_L}
\]

removes the scale dependence. Salt (2011) discusses the interpretation of the CV, and the way it varies in sound films, at a little length.

The median is also the 50th percentile – it splits the ordered data into two groups with 50% on either side. The 25th percentile produces a 25:75% split of ordered data and is also called the first quartile, \(Q_1\); the 75th percentile, or third quartile, \(Q_3\), produces a 75:25% split. The difference between the quartiles, \(Q_3 - Q_1\), is the interquartile range (IQR), often used as a measure of dispersion to accompany the median\(^\text{15}\).

\(^\text{11}\)That is, it can be read as saying ‘the median is a better estimate of the (arithmetic) mean!’
\(^\text{12}\)Specifically, the log of a product is the sum of the logs of the individual terms and \(\log a^b = b \log a\).
\(^\text{13}\)From now on to be understood as the arithmetic mean or ASL unless otherwise qualified.
\(^\text{14}\)Readers familiar with statistics will recognize that this definition is appropriate for a population. See the earlier discussion of the arithmetic mean for more on this and notational use (i.e. \(\sigma_L\) is being used for the standard deviation, rather than \(\sigma\) which is the other parameter, with \(\mu\), that defines the lognormal distribution). If the data are treated as a sample from a large population then the divisor of \(n\) is replaced by \((n - 1)\) and the statistic denoted by \(s^2\) or \(\hat{\sigma}^2\) to emphasise that these are (unbiased) sample estimates of the unknown population variance. For descriptive purposes the distinction doesn’t usually matter for typical SL data, when \(n\) is large enough for differences to be unimportant.
\(^\text{15}\)Unless \(n\) is divisible by 4, how to assign numerical values to the quartiles is not obvious. It was noted in Section 4.2 that \(R\) has nine different algorithms for this purpose. If \(n\) is at all large, as with most SL data, the differences between them will be unimportant.
The median and IQR have been promoted in both the statistical and cinemetric literature as robust alternatives to the mean and standard deviation that are preferable in the presence of outliers in the data. The merits of this argument, in relation to cinemetric data, are the subject of Chapter 11. The rough idea – which involves some ‘heavy’ mathematics if treated rigorously and theoretically – is that you only need to change one observation in a data set by an arbitrarily large amount to change the mean (or standard deviation) by an arbitrarily large value. To change the median in a similar way you need to change half the observations. These are the extremes of what are called breakdown points – theoretically 0 for the mean and 0.5 for the median. A lot of ingenuity has been invested in devising robust estimators of measures of location and dispersion with high breakdown points, some of which have found their way into the cinemetric literature (Section 4.5).

Taken together the minimum, maximum, quartiles and median are sometimes called a five-number summary of a set of data, and can be obtained using the \texttt{fivenum(z)} function in R. They form the basis of the boxplot, a graphical method of displaying the data illustrated in Sections 5.3.1 and 6.2.

\footnote{You have to change $1/n$ observations to effect an arbitrarily large change in the mean, which is not 0, but as $n$ becomes arbitrarily large $1/n$ approaches 0.}
Chapter 5

Graphical analysis – basics

In this chapter the sequence of SLs in a film is ignored in the analysis. A variety of different kinds of plot have been used in the cinemetric literature, of which the histogram is the most common. Commands needed for the basic plots are illustrated first, followed by more detailed discussion. A Night at the Opera is used for illustration until further notice, copied to z as in the previous chapter.

5.1 Histograms

5.1.1 Basics – an example

Histograms are perhaps the most widely used method for graphical presentation in cinemetrics. There are issues with their construction and interpretation, an initial illustration of which is provided in Figure 5.1. The histograms were obtained from the following commands\(^1\).

\begin{verbatim}
hist(z)
hist(z, 50, xlab = "SL", main = " ")
hist(z[z < 120], 50, xlab = "SL (< 120)", main = " ")
hist(z[z < 37], 100, xlab = "SL (< 37)", main = " ")
\end{verbatim}

The top-left histogram uses the defaults in R. It is not that useful; the main problem is that, to accommodate the extreme value, rather large bin-widths of length 25 are imposed by the default.

The top-right histogram specifies the number of bins (or cells) to aim for, 50 in this case, and has the effect of reducing the bin-width to 5. The result is better but the scale is still affected by the extreme value\(^2\). The xlab and main arguments show how to control the labelling of the x-axis and title.

An obvious expedient to remove the effect of the outlying value is to omit it. In the bottom-left histogram this was done by selecting all SLs less than 120 using \(z[z < 120]\). A bin-width of 2 results.

Something similar is done in the final plot, where a subset of the histogram is magnified by selecting only SLs less than 37 and increasing the number of bins so that the bin-width is 0.5. This choice was made in order to try and emulate Figure 3 in DeLong et al. (2012)\(^3\).

The general appearance of the histogram for A Night at the Opera is characteristic of SL distributions. It is skewed with a single main peak and a long-tail. This kind of regularity,

\(^1\)The layout shown was obtained within R by sandwiching the code between `par(mfrow = c(2,2))` and `par(mfrow = c(1,1))`. The first command switches to a 2 × 2 plotting grid, and the second switches back to the defaults. It is usually most convenient to place the commands in a function (Section 4.3).

\(^2\)You don’t necessarily end up with exactly the number of bins specified, since R ‘works’ at providing ‘pretty’ (and sensible) boundaries to avoid the large number of decimal places a slavish adherence to a specified number of bins might impose.

\(^3\)The figure shown here doesn’t exactly reproduce theirs, the exact appearance depending on both interval boundaries and the plotting algorithm used.
noted by Salt (1974), led to the suggestion that many SL distributions might readily be modeled using a standard probability distribution. Later research led to the suggestion that the lognormal distribution was a fairly simple and appropriate model for, perhaps, 50% or more of films with average shot lengths (ASLs) of less than 15 seconds (Salt 2006, 2011).

This contention seems, variously, to have been both uncritically accepted, and strongly disputed and is discussed in detail in Chapter 12. Given the generally similar qualitative appearance of many SL distributions the temptation to try and describe them with a common underlying model is understandable. The main point is that if the suitability of the lognormal model is assessed on the basis of the appearance of a histogram, an element of subjective judgement is involved. This judgement is, in turn, influenced by the appearance of the histogram which is conditioned by the choices made in its construction. Initial ‘choices’ are usually made by the software. The user can usually modify them, with varying degrees of ease. It is actually much easier to do this in R than in some widely used spreadsheet packages. The next section explores this.

5.1.2 Technicalities

A histogram is constructed as follows.

1. The range of the data is subdivided into a set of contiguous intervals (‘cells’ or ‘bins’ is other terminology used). The default, assumed here, is that the intervals are of equal width. The interval boundaries are determined by the anchor point (the position of the bottom-left corner of the histogram) which may extend slightly beyond the minimum data value.

   The choice of interval width (bin-width) and anchor point determines both the number of intervals and their boundaries. Equivalently, specifying the number of bins and an anchor point will determine the width.\(^4\)

2. The number of observations in each bin is counted. If observations fall exactly on a boundary a consistent rule should be applied to determine which side of the boundary they go.

\(^4\)In R the desired number of bins will be modified to ensure a sensible choice of bin-width.
3. The count, or frequency, in each bin is represented in a chart by a bar whose area is proportional to the frequency. If equal bin-widths are used the height of the bar is also proportional to frequency and this is what most users are familiar with. Adjacent bars should touch; charts purporting to be ‘histograms’ with visible gaps between bars are simply wrong.

In R Sturge’s rule is the default for determining bin-widths in hist(z). It is usually inadequate for SL data, where shorter intervals than it provides are desirable. Other rules are available but it is simplest, and legitimate, to experiment with the number of bins. Commands such as hist(z, 50) do this. More fully, this can be written as hist(z, breaks = 50); the breaks = bit isn’t needed if only a number is specified in the second position, but can be used if exact control over the interval boundaries is required.

It can be convenient to represent the histogram using a probability (density) rather than frequency scale. Using hist(z, 50, freq = F), for example, does this (where F is shorthand for FALSE).

Try typing something like NatO.hist <- hist(z, 50); the requested histogram appears. Type NatO.hist and a lot of information you mostly don’t want appears. NatO.hist is an object that holds information about the histogram that can be extracted and manipulated if the urge to do so overtakes you.

Type names(NatO.hist) to see what’s available; you get

```
[1] "breaks"   "counts"   "intensities" "density"   "mids"
[6] "xname"    "equidist"
```

Typing NatO.hist$breaks brings up the break points (interval boundaries); NatO.hist$count the counts in the bins; NatO.hist$mids the interval mid-points. These can usually be ignored but are there if needed.

5.1.3 Example continued - log-transformation

One problem I have with histograms like those in Figure 5.1 is a difficulty in making assessments about how lognormal the data look (if this is an issue). I disagree, for example, with Salt’s (2006, 393) statement that for some of the examples he shows on p.392 the ‘observed results fit well with a Lognormal distribution’. Similarly, as is about to be illustrated, the assertion of DeLong et al. (2012) that A Night at the Opera has a lognormal distribution is not sustainable.

If the SL data are lognormal (approximately) then they should be approximately normal after a logarithmic transformation (see the Appendix for the maths and the next section for R code). I suspect that most people find it easier to make assessments about normality than lognormality. Without worrying too much about the mathematics, a normal distribution should be symmetrical about a single peak – this is not a sufficient requirement for the data to be normally distributed, but forms a starting point for seeing if it isn’t.

Figure 5.2 show histograms for logarithmically transformed SL data. The left-hand plot is the R default; the right-hand plot specified 30 as the number of bins. Part of the point here is to simply illustrate how the choice of numbers of bins (equivalently, bin-width) can affect the appearance of the histogram.

Neither plot looks convincingly normal, particularly that to the left. That is, the assertion that A Night at the Opera has a lognormal distribution looks clearly wrong. Kernel density estimates (KDEs) are superimposed on the histograms and are a more elegant way of showing this. KDEs, and the code required to obtain the figures, are discussed in Section 5.2 below.

---

5Either because the histogram is incorrectly drawn – this is what was the default in Excel for many years – or because you are dealing with an incorrectly named ‘bar chart’, terminology best reserved for representations of discrete data.
5.2 Kernel density estimates

A histogram is a particular form of density estimate. It has a number of limitations. One is that the appearance depends on the bin-width used and, the availability of ‘rules’ notwithstanding, the choice of what to present for publication purposes is ‘subjective’. This is unavoidable, and true for other forms of density estimate, though the suspicion exists that software defaults are sometimes used without too much thought being given to the choice.

More seriously, and something avoided by other forms of density estimate, is that the appearance can be affected by the choice of anchor point. Histograms are also unwieldy for comparing the SL distributions of more than a small number of films. Of alternatives to the histogram only kernel density estimates (KDEs) are considered in detail here.

Statistical texts and papers that deal with KDEs can look mathematically forbidding, but the basic idea is very simple. One way of representing a single SL is as a point on a line between the maximum and minimum SLs. The point is replaced by a symmetric bump, centered on the SL. The spread of values covered by the bump are associated with different heights. This is done for each SL in turn. The KDE is then simply obtained by adding up the heights of the different bumps at each point along the line.

For *A Night at the Opera* the default KDE obtained using `plot(density(z))` is shown in Figure 5.3. This can be broken up into two commands, `NatOdens <- density(z)` followed by `plot(NatOdens)`.

In the same way that the appearance of a histogram is controlled by the choice of bin-width, the appearance of a KDE is controlled by the spread of the bumps used in its definition. This is determined by the bandwidth (or window width). The form of the bump (or kernel) can be chosen in various ways, but commonly a normal (or Gaussian) distribution is used, in which case the bandwidth depends on the standard deviation of the distribution. Large bandwidths give over-smoothed KDEs; smaller bandwidths give under-smoothed KDEs. As with histograms a variety of rules exist for the automatic choice of bandwidth and, as with histograms, it is usually sensible to experiment with different choices.

Looking at the default is usually useful in that it provides a starting point for such experimentation. In fact, because of the impact of the extreme point and the skewness with a long tail, the KDE in Figure 5.3 has the same problems for interpretation as the histogram.

If attention is confined to the region with SLs less than 40 a default bandwidth of 0.98 is

---

6A technical issue is that because the bumps are spread around a point the KDE will have non-zero values lying outside the range of the observed SLs. There are ways of dealing with this, but it is simplest not to worry about it unless behaviour in the extreme tails is of especial interest.

7It is sometimes useful to do this, since `NatOdens` is an object containing information about the KDE than can be manipulated for other purposes. This is not pursued here.
obtained. Reducing this to 0.8 and tidying the graph up a bit for presentation using the following code results in Figure 5.4. The main feature is that the tail of the distribution is somewhat bumpier than would be expected for a lognormal distribution.

\[ \text{plot(density(z[z < 40], bw = .8), xlab = "SL (< 40), bandwidth = 0.8", main = ")} \]

This is a situation where looking at the SLs on a logarithmic scale is useful. Figure 5.5 illustrates the effect of different bandwidth choices; there is no need to select a subset of the data as done in previous figures. The code used was

\[ \text{plot(density(log(z)))} \]
\[ \text{plot(density(log(z), bw = .20), xlab = "log(SL), bandwidth = 0.20", main = "} \]
\[ \text{plot(density(log(z), bw = .12), xlab = "log(SL), bandwidth = 0.12", main = ")} \]
\[ \text{plot(density(log(z), bw = .35), xlab = "log(SL), bandwidth = 0.35", main = ")} \]

The upper-left plot is the default and used a bandwidth of 0.2447. With this as the starting point the upper-right figure reduces the bandwidth to 0.20. It doesn’t make too much difference here; some of the features of the default are emphasized slightly more. The under-smoothed lower-left plot, with a bandwidth or 0.12 , shows too much spurious detail; the over-smoothed lower-right, with a bandwidth of 0.35, removes some of the potentially informative detail in the upper plots.

The clear skewness of all these plots is sufficient evidence of non-normality to demonstrate that the assertion that *A Night at the Opera* has a lognormal SL distribution is wrong. The upper plots (and under-smoothed plots) add further credence to this assertion by highlighting regions of ‘lumpiness’ of a kind inconsistent with a normal distribution.

If, for whatever reason, your preference is for histograms, it is easy enough to superimpose a KDE on a histogram. This was done in Figure 5.2, where the default KDE was superimposed on the default histogram, and the KDE with a bandwidth of 0.20 superimposed on the second histogram, using

\[ \text{hist(log(z), freq = F, xlab = "SL", main = ")} \]
\[ \text{lines(density(log(z)), lwd = 2, col = "red")} \]
Figure 5.4: *The KDE for A Night at the Opera, using SLs less than 40 and subjectively chosen bandwidth of 0.8.*

Figure 5.5: *The default KDE for the logged SLs of A Night at the Opera in the upper-left, and KDEs for different choices of bandwidth.*
hist(log(z), breaks = 30, freq = F, xlab = "SL", main = "")
lines(density(log(z), bw = 0.2), lwd = 2, col = "red")

The arguments \texttt{lwd} and \texttt{col} control the line width and colour and are used here for presentational purposes. Though not used here \texttt{lty} can be used to control the line type; for example, \texttt{lty = 2} produces a dashed line.

5.3 Boxplots

5.3.1 Basics

The \textit{boxplot} (or box-and-whisker plot) is a useful descriptive display. That for \textit{A Night at the Opera}, shown in Figure 5.6, is obtained by \texttt{boxplot(z, horizontal = TRUE)} (the violin plot is discussed in the next section). The \texttt{horizontal = TRUE} argument produces the horizontal orientation of the plot preferred here; if a vertical orientation is desired it can be omitted, as this is the default.

The command \texttt{fivenum(z)} will produce the following ‘five-number summary’

\begin{verbatim}
[1] 0.5 2.3 4.0 8.5 207.2
\end{verbatim}

The upper and lower values are the minimum and maximum, which are the extremes of the boxplot. The second and fourth values are the lower and upper quartiles (\(Q_1\) and \(Q_3\)), and define the limits of the box shown. The width of the box is \(\text{IQR} = (Q_3 - Q_1)\), where \(\text{IQR}\) is the \textit{interquartile range}; the box thus contains the central 50% of the data. The middle value in the five-number summary, 4 in this instance, is the \textit{median}, and is highlighted by the line within the box.

The ‘whiskers’, dashed lines that extend from the boxes, go by default as far as \(1.5 \times \text{IQR}\) from the limits of the box, beyond which ‘unusual’ points are shown individually. For typical SL data nothing ‘unusual’ will be highlighted in the left tail; for skewed distributions rather a lot of ‘unusual’ values may typically be shown in the right-tail. Here the IQR is 6.2 (i.e. 8.5 - 2.3). Thus
the upper whisker is broken at 17.8 (i.e. $8.5 + 1.5 \times 6.2$). There are 65 SLs greater than this, but only that at the maximum, 207.2, really stands out.

The definition of ‘unusual’ is arbitrary, with 1.5 as the default. It can be controlled by the range argument, and range = 0 will extend the whiskers to the extremes. There has been some confusion in the cinemetrics literature about the identification of ‘unusual’ SLs, as defined above, with outliers. This, in turn, has fed into debates about the appropriate use of summary statistics for SL distributions discussed further in Section 5.3.3.

5.3.2 Interpretation

The boxplot is a nice presentational device, particularly for comparative purposes, under certain conditions. Different software packages may present boxplots in different ways; for the control that can be exercised in R see ?boxplot. Some software shows the mean as well as the median of the data; where ‘unusual’ values are indicated the default is often that used in R.

If a set of data is sampled from a symmetric unimodal (single-peaked) distribution (e.g., the normal) it should be approximately symmetrical, with the median near the middle of the box. Departures from symmetry in the form of skewness, can be manifested by the non-central placement of the median and whiskers of unequal length. For a reasonably symmetric underlying distribution the highlighting of ‘unusual’ values can draw attention to potential outliers; for very skewed distributions, typical of SLs, a lot of ‘unusual’ data may be highlighted that is, in fact, typical of what to expect from the underlying distribution.

![Boxplot and Violin Plot](image)

**Figure 5.7**: A histogram for the SLs of Harvey and a KDE, boxplot and violin plot for the logged SLs of the film.

The boxplot is not suitable for representing distributions with more than one mode, whether symmetric or not\(^8\). This is better illustrated using SLs after log-transformation, and this is done for Harvey (1950) in Figure 5.7.

I will leave it to readers to judge if they think the histogram for the untransformed data looks lognormal. Harvey has, in fact, one of the least lognormal distributions it is possible to

\(^8\)We will not worry here about the difference between ‘major’ and ‘minor’ modes.
find; looking at the KDE on a log-scale indicates one reason why – the central appearance is very
different from that which is to be expected from normality, which it should be if lognormality
obein.

It is good practice to look at several kinds of plot rather than going straight to a boxplot; if you
do the latter the near-perfect symmetry of the plot(bottom-left) may mislead you into thinking
that the distribution is symmetric (and possibly normal). The problem is that the boxplot is
constructed on the basis of the five-number summary only and takes no account of the density
of the distribution, and hence cannot represent modes. The violin plot attempts to improve on
this by showing the density as well as the boxplot. For the logged SLs of Harvey this produces a
graph with a ‘waist’ indicating the absence of a clear single mode. For a symmetrical unimodal
distribution it would be fattest in the middle, and tail off symmetrically on either side.

The R commands used for the figure were

```r
hist(SL.Harvey, 50, xlab = "SL", main = "Harvey")

logH <- log(SL.Harvey)
plot(density(logH, bw = .3), xlab = "log(SL), bandwidth = 0.3", main = "")

boxplot(logH, horizontal = TRUE, xlab = "log(SL)")

library(vioplot)
vioplot(logH, horizontal = TRUE)
```

### 5.3.3 Boxplots and outliers

There has been discussion in the cinemetrics literature about the appropriate way to summarize
SL distributions, in the context of stylistic analysis. The relative merits of the median shot length
(MSL) and average shot length (ASL) has attracted particular attention and is discussed further
in Chapter 11.

One argument against the use of the ASL, quite widely asserted, is that outliers are common
in SL data, and that the ASL is sensitive to this (i.e. it is not ‘robust’). Taken to extremes, this
has led to suggestions that the ASL should not be used as a measure of ‘film style’, despite its
widespread use for this purpose for over 30 years. The prevalence of outliers in SL data has been
exaggerated, as has their effect on ASL calculations. If true, it follows that the arguments against
the ASL, based on the supposed effect of outliers, have little weight for practical purposes\(^9\).

The opposite position is enunciated by Redfern (2012d). The use of boxplots to demonstrate
the prevalence of outliers is central to part of his argument, and one of the films investigated, The
Scarlet Empress (1934), is used as the peg on which to hang further discussion. The histogram
for the SLs is shown in Figure 5.8. The boxplot in the upper-right is the R default, and on the
basis of it Redfern claims that there are 39 outliers. This is a serious misinterpretation.

The lower-left plot is for the logged SLs. A KDE or histogram of the logged data is unimodal,
so use of the boxplot is justified. Discounting the two marginal outliers in the lower tail (induced
by the log-transformation) there is no evidence of any unusual data (outliers). This is exactly what
we would expect if the SLs followed a lognormal distribution, or at least a qualitatively similar
skewed distribution. That is, the 39 supposed outliers in the boxplot for the untransformed data
are no such thing; they are just what the boxplot is expected to show if the data are reasonably
skewed with the features to be expected of a skewed distribution.

Subsequent to the discussion Redfern\(^10\) has acknowledged that his original use and interpre-
tation of the boxplot, in terms of outlier detection, was inappropriate – if untransformed SLs are
used an allowance has to be made for their skewed distribution. An alternative version of the
boxplot, for The Scarlet Empress, gives rise to the final plot in Figure 5.8.\(^11\) There is one very

---

\(^9\) Other arguments can be adduced; the focus here is only on outliers and the use of boxplots for identifying them.
\(^11\) The blog uses Lights of New York for illustration; this does not affect the points made here.
marginal unusual value in the right tail, so the original misinterpretation in terms of 39 outliers is clear.

The original argument had the merit of making it clear how an outlier was defined; this is not usually the case in publications where the prevalence of outliers in SL data is asserted. There is the suspicion that many commentators are bad at judging what is and isn’t to be regarded as unusual in the tail of a lognormal distribution, and hence at judging what is an outlier\textsuperscript{12}. A technical discussion on the ideas that underpin the final plot in Figure 5.8 may be useful in understanding why. The plot is obtained using

\begin{verbatim}
library(robustbase)
adjbox(SE, horizontal = TRUE, xlab = "SL")
\end{verbatim}

remembering that the package \texttt{robustbase} has to be installed first (Section 3.3).

Hubert and Vandervieren (2008) (HV) describe the methodology (Googling brings up more than one earlier version of their paper). The limits beyond which unusual data are flagged in a boxplot are defined as

\[(Q_1 - k \times L), (Q_3 + k \times U)\]

where \(k\) is a constant and \(L\) and \(U\) are lower and upper values used in the definition. For both the default boxplot in R and that used in the \texttt{adjbox} function from the package \texttt{robustbase}, \(k = 1.5\). For the standard boxplot \(L = U = \text{IQR}\).\textsuperscript{13}

What the HV paper does is to define an interval that is asymmetric about the median, using different values for \(L\) and \(U\). These are defined after considerable experimentation, and are based

\textsuperscript{12}The lognormal assumption is convenient, but not essential here. Some authors who have asserted the prevalence of outliers do, however, also appear to accept the generality of lognormality (e.g., DeLong \textit{et al.} (2012); Cutting (http://www.cinematics.lv/cutting_on_salt.php)).

\textsuperscript{13}The choice of \(k\) is arbitrary. For data sampled from a normal distribution with sample size 600, 3-4 unusual values are expected to be flagged using the standard boxplot. For \textit{The Scarlet Empress} with \(n = 601\) SLs this is much what is observed on the log-scale. Note that this does \textit{not} imply that the SLs are lognormally distributed, though it is consistent with the possibility.
on a robust measure of skewness that is one of several possibilities that might have been used. More experimentation with SL data would be useful, and it would be interesting to compare the results with those obtained using boxplots with log-transformed SLs. It is obvious that this ought to change perceptions about the extent of problems with ‘outliers’. They do exist, the obvious one for *A Night at the Opera* being a case in point. Once identified in ‘sensible profusion’, or the lack thereof, their effect on ASL calculations – if this is of interest – can then be investigated, rather than it simply being asserted that they are problematic. As Salt (2012) suggests, where outliers do exist it may be of more interest to examine the role they play in the ‘style’ of a film than to focus on numerical measures of ‘style’ that ignore them.
Chapter 6

Comparative graphical analysis

6.1 KDEs and Histograms

The view adopted here is that if comparison between just a small number of films is required this is better pursued using graphical methods, with numerical descriptive statistics in a supplementary role. That is, given two films to compare and for example, it may not be that interesting to compare their ASLs or MSLs, even if accompanied by measures of scale and shape. It is likely that anything to be inferred from such comparisons will be evident from graphical analysis, and they may miss a lot if the SL distributions are at all ‘non-standard’.\(^1\)

The view is also taken that the comparison of SL distributions using separate histograms is unwieldy; they are not easily overlaid, for example. Comparison using KDEs is more effective. The skewed nature of SL distributions also makes comparison on the untransformed scale quite awkward, so many of the illustrations below will be based on log-transformed SLs. The code illustrated is easily adapted to untransformed SLs if this is what is really needed. For initial illustration a late Hitchcock silent, Easy Virtue (1928), and early ‘talkie’ The Skin Game (1931) are used. This is for comparison with Redfern (2009).

Figure 6.1 compares the histograms on the untransformed SL scale. The code used was

```r
par(mfrow = c(2,1))
hist(SL.Easy_Virtue, 50, xlim = c(0,175), xlab = "SL", main= "Easy Virtue")
hist(SL.The_Skin_Game, 200, xlim = c(0,175), xlab = "SL", main = "The Skin Game")
par(mfrow = c(1,1))
```

The maximum SL across the two films is 174.7 and xlim = c(0, 175) defines the range of the scale to go just beyond this for both histograms. The numbers of bins, 50 and 200 were selected, with a bit of experimentation, to ensure a reasonable appearance with equal bin-widths of 1 second. This is all in the interests of comparability. It is possible to program R to do this automatically.

The most obvious features are that the bulk of SLs are concentrated in much the same region for the two films, and The Skin Game has a much longer tail. A lot of finer detail is not that obvious, as later analyzes will show.

The same comparison is undertaken using KDEs in Figure 6.2. This is neater, and more readily interpretable since the two KDEs are easily overlaid on the same graph. The code used was

```r
plot(density(SL.The_Skin_Game), ylim = c(0, .14), lwd = 2, col = "blue",
xlab = "SL", main = "")
```

\(^1\)In dealing with a large body of films, summary statistics are useful for getting an overview of possible patterns in the data. The decline in ASLs since the 1950s is a commonly cited example of this sort of use. Where films have an approximate lognormal distribution it is possible to summarize them well using just two statistics, but the lognormality needs to be demonstrated first.
Figure 6.1: A comparison of SL histograms for Easy Virtue and The Skin Game.

```r
lines(density(SL.Easy_Virtue), lty = 2, lwd = 2, col = "red")
legend("topright", legend = c("The Skin Game", "Easy Virtue"), lty = c(1,2),
       lwd = c(2,2), col = c("blue", "red"), bty = "n")
```

The effect of the `lines` function is to overlay whatever is specified on the structure dictated by the `plot` command. Most of the arguments are for presentational purposes and can be omitted in exploratory analysis. The `lwd` argument, for example, controls the line-width, and is chiefly there to ensure better visibility when looking at the graph either on screen or the printed page.

The horizontal scale of the plot is dictated by the film named in the `plot` function; The Skin Game has the greater range of SLs and its placement here avoids, in this instance, having to use the `xlim` argument. The `ylim` argument is needed to control the vertical scale and accommodate the KDE for Easy Virtue; you only discover this after an initial analysis and need to experiment, or look at the two KDEs separately first, to get suitable limits. The `lty` and `col` arguments control the line type and color.

The `legend` function for – no surprise – adding legends is introduced here. Using `?legend` in R, or examples in texts such as Murrell (2011), are helpful guides to their construction.

As far as interpretation goes the greater density of Easy Virtue at the shorter SLs is more apparent than with the histogram. The rather different tail behavior for SLs greater than about 20 is also more apparent. This becomes even more obvious, along with other features, once KDEs on a log-scale are compared.

Figure 6.3 is similar to the previous figure except that a logarithmic scale is used. Commands identical to those for that figure were used, except that film names were replaced by their logarithms (e.g., `log(SL.Easy_Virtue)`) and the vertical scale was determined by `ylim = c(0, .55)`. Bandwidths of 0.17 and 0.30 were used for Easy Virtue and The Skin Game respectively.

Visually the plot for Easy Virtue doesn’t look too unlike a normal distribution. That for The Skin Game is clearly non-normal, and the rather lumpy nature of the tail should caution against any simple summary in terms of either the MSL or ASL. Apart from the different distributions the different tail behavior is evident with SLs for The Skin Game extending beyond those for Easy Virtue in either direction. This can be seen in other figures but is much more prominently...
Figure 6.2: *A comparison of KDEs for the SLs of Easy Virtue and The Skin Game.*

Figure 6.3: *A comparison of KDEs for the logged SLs of Easy Virtue and The Skin Game.*
displayed for the shorter SLs on the log-scale.

This analysis can be taken further if desired. The KDEs for the two films intersect to the left at about 0.7, corresponding to an SL of 2. For *The Skin Game* 24.6% of the SLs are less than this, compared to 5.5% for *Easy Virtue*. To the right the intersection is at about 2.77, corresponding to an SL of 16; 22.8% of SLs for *The Skin Game* exceed this, compared to 5.2% for *Easy Virtue*.\(^2\)

### 6.2 Boxplots and violin plots

For comparing SLs for two films, as in the previous section, graphical display may be all that is necessary. Graphs can be accompanied with numerical summaries if these are thought to be helpful, but this may be gilding the lily, and possibly misleading if SLs are at all non-lognormal. Redfern (2012d) has advocated the use of five-number summaries, which are enshrined in the boxplot (Section 5.3).

Comparative boxplots and violin plots, on a log-scale, are illustrated in Figure 6.4. The boxplots show, well-enough, the similar MSLs for the two films and the greater spread of *The Skin Game*. The rather lumpy nature of this film is not captured, so the plot is not as effective a method of comparison as the KDEs. To correct for this violin plots might be used. The default is not especially useful and in the figure the degree of smoothing, controlled by \( h = 0.2 \), was obtained after some experimentation. It shows the lumpier distribution of *The Skin Game* better, the appearance of the plot for *Easy Virtue* being little affected by the choice. It is a happy accident, and meaningless, that the plot for *The Skin Game* more resembles a violin.

![Boxplots and violin plots](image)

**Figure 6.4:** Comparative boxplots and violin plots, on a log scale, for *Easy Virtue* and *The Skin Game*.

The code used for the figures is

```r
boxplot(log(SL.Easy_Virtue), log(SL.The_Skin_Game), names = c("Easy Virtue", "The Skin Game"))
library(vioplot)
vioplot(log(SL.Easy_Virtue), log(SL.The_Skin_Game), h = 0.2, names = c("Easy Virtue", "The Skin Game"))
```

In addition to the use of the \( h \) argument, note the use of `names` to control labeling.

### 6.3 Cumulative frequency diagrams

Empirical cumulative frequency diagrams (CFDs) are another way of comparing distributions. They plot, in the examples used here, the proportion of SLs less than or equal to an SL against

\(^2\)In R, with a graph of the KDEs on the screen, vertical lines can be placed on the plot using, for example, `abline(v = 0.7)`. Play around with the value of \( v \) to locate positions of interest. Replace \( v \) with \( h \) if horizontal lines are of interest. Given a logged SL, 0.7 say, it can be converted to an SL using `exp(0.7)`.  

49
SL. If interest centers on comparing SLs for films in terms of the proportion of shots above/below some value they are a more natural tool to use than histograms or KDEs. I suspect non-statistical users find them more difficult to read. The right-hand plot in Figure 6.5 emulates that in Figure 4 of Redfern (2013a).

![Figure 6.5: A comparison of CFDs for the SLs of Easy Virtue and The Skin Game to the left and logged SLs to the right.](image)

Other than emphasizing what we already know, that The Skin Game has a longer tail than Easy Virtue, the plot for the untransformed SLs hides a lot of detail. A log-transformation is indicated once again, shown to the right of the figure. The median, which happens to be the same for both films is indicated by the dotted vertical line. The effect of the log-transformation is to ‘stretch out’ the lower tail while compressing the upper tail. As with the KDEs the difference in SLs at the lower values is much more obvious.

The plot for the logged data was obtained with the `ecdf` function using

```r
1SG <- log(SL.The_Skin_Game)
1EV <- log(SL.Easy_Virtue)

plot(ecdf(1SG), verticals = T, lwd = 2, col = "blue", pch = ",", xaxt = "n", xlab = "SL", ylab = "proportion", main = "")

lab <- c(1,5,10,20,50,100,200)
axis(1, at = log(lab), labels = lab)
lines(ecdf(1EV), verticals = T, lwd = 2, lty = 2, col = "red", pch = ",")
abline(v = median(lSG), lty = 3, lwd = 1.5)

legend("topleft", inset = c(0, .02), legend = c("The Skin Game", "Easy Virtue"), lty = c(1,2), lwd = c(2,2), col = c("blue", "red"), bty = "n")
```

where `xaxt = "n"` removes the default x-axis, and `lab` and `axis` replace it with an axis where the labeling is on the untransformed rather than SL scale.\(^3\)

\(^3\)This is just to show it can be done if you want to; it’s not needed if labeling with the logged values is acceptable. The plot is on the log-scale; `lab` selects those SLs you want to show as labels on the axis; in `axis` the 1 specifies the x-axis as the axis to draw; `at = log(lab)` generates the tick marks for the logged values at which the labels are to be placed; and `labels = lab` adds the labels specified in `lab`. This can be done in other ways, probably more neatly, but it worked here. In `legend`, placed by the first argument in the top-left corner, `inset = c(0, .02)` shifts it down slightly to avoid overlap with the guideline at the top; `?legend` provides the detail.

50
6.4 Comparison with reference distributions

6.4.1 Comparisons with the lognormal distribution

It is sometimes of interest to see how well an SL distribution conforms to a lognormal distribution and, perhaps more interestingly, to see what any departure from lognormality looks like. Lognormality does not imply that this is what the makers of a film consciously aim for. For a variety of reasons SLs are expected to vary (e.g., shorter ASLs for ‘action’ scenes, longer ASLs for ‘romantic’ interludes is mentioned from time to time). Patterns of SLs may vary more or less rhythmically within a film (Chapter 7) but this kind of variation is lost if one looks at the ‘external’ patterning of SLs, as we have been doing so far. It is worth quoting Salt (2011) at length, on this.

The Lognormal distribution results when the quantity under consideration, in our case shot length, is determined as a result of the probabilities associated with a large number of independent causative factors being multiplied together. . . . the amount of conscious decision-making in making films may be less than people think, and many film-makers may often be doing what they do unthinkingly, because that is what everybody does. In films what is presumably concerned in determining the length of a shot is the simultaneous interaction of such factors in the scene being filmed as how the actor moves in the shot with respect to the closeness of the camera, the length of the lines he speaks, and how the other actors react, and so on. The fact that different individuals are usually responsible for these various components of a film, from the scriptwriter to the director to the editor, assists the independence of these causes. However, once in a while a film-maker may decide to do something unorthodox on purpose (this is the aesthetic impulse, after all), and this will upset the regularity of features like the Lognormal distribution.

Elsewhere, in his contribution to the On Statistics pages on Cinemetrics, and with specific reference to films made in the early sound era, Salt cites technological factors that might be expected to militate against the appearance of lognormality. The extent to which films do have lognormal distributions has been disputed. The issue is addressed in Chapter 12 but is not that important for what is to follow in this section. The lognormal distribution is a useful approximate base, even if strictly wrong, for judging serious departures from regularity that invite explanation.

Lognormality implies normality of the log-transformed SLs. It is easier to visually assess normality than lognormality and the following exploration of the use of probability plots and KDEs is based on log-transformed data where the normal distribution can be used as a reference distribution. A brief excursion into the normal distribution is provided first (also, see the Appendix).

6.4.2 Aspects of the normal distribution

Call the SLs $X$ and denote the particular SLs for a film with $n$ shots as $(x_1, x_2, \ldots, x_n)$. Unless it is important to identify a specific shot it is simplest to ignore the subscript and refer to observed SLs as $x$.

In the same vein let $Y = \log X$ with $y = \log x$ the observed logged values. Interest for the purposes of this section lies in whether or not the $y$ look reasonably like a sample from a normal distribution. A normal distribution is completely defined by knowledge of its mean, $\mu$, and standard deviation (SD), $\sigma$. This is conventionally written as

$$Y \sim N(\mu, \sigma^2)$$

read as $Y$ is normally distributed with mean, $\mu$, and variance, $\sigma^2$, the variance being the square of the SD.

It is often convenient to standardize the data so it has a mean of zero and variance (and standard deviation) of 1. This is done by defining

$$z = \frac{y - \mu}{\sigma}.$$ 

If the data have a normal distribution then $z \sim N(0, 1)$ (the standard normal distribution) but this is not assumed here as the question of whether or not the transformed data are normal is that
In the context of SL distributions, it is possible for two films with lognormal distributions to have rather different ASLs and scales. Even though qualitatively similar they will look different on the untransformed scale; after log-transformation they will have the same normal ‘bell-shape’, but if placed on the same graph will have different locations and scales (spread). What the z-transform does is remove location and scale differences so that, if normal, the plots of the distribution will be identical. Any differences between films in terms of departures from lognormality may be more apparent if compared in this way and, in particular, they can be compared to the standard normal distribution to see where any differences from the normal lie.

In practice, it is necessary both to assess whether observed differences are of potential substantive significance, to see how these are manifest on the untransformed scale and, ultimately, to see if useful explanations can be offered for departures from lognormality (or, more generally, any commonality of distribution exhibited by SL distributions). These, sometimes complex, issues are largely avoided in the discussion to follow of how to effect comparisons.

### 6.4.3 Normal probability plots

Plots of the kind discussed so far are good at identifying clear departures from (log)normality in the middle of an SL distribution, or for identifying obvious outliers. More subtle features of tail behavior, particularly those associated with the smaller SLs, are harder to discern, and probability plots may be better for this purpose (though they can be difficult to interpret).

One way of thinking about this is that the ordered values of a sample are plotted against values they are expected to have if they follow a normal distribution. If they do, an approximate straight line should result. It’s simplest to think in terms of standardized data. The median of the observed data (the 50th percentile) should plot against the median, zero, of the standard normal; the lower quantile of the data, \( Q_1 \), or the 25th percentile, should plot against that for the standardized normal, -0.67, and so on. To illustrate, *A Night at the Opera* is used. The SLs are given by \( z \), \( \text{scale}(\log(z)) \) standardizes the logged data, and \( \text{qqnorm} (\text{scale}(\log(z))) \) produces Figure 6.6.

The plot is of the ordered data against the values they should have if they are exactly normal. The dashed line shows exactly where the points should lie in this case. It is obtained by using \( \text{abline}(a = 0, b = 1) \) which generates a line with zero intercept and unit slope. The solid line is obtained from \( \text{qqline}(\text{scale}(\log(z))) \) and serves a similar function except that it is designed to pass through the upper and lower quartiles. The way the \( \text{qqline} \) is defined avoids the effects of any outliers in the data. (The \( \text{ltty} \), \( \text{lwd} \) and \( \text{col} \) arguments that control the appearance of the lines have been omitted in the above commands.)

As applied here \( \text{abline} \) is only useful if the data are standardized. If not, the plot looks exactly the same but the scale on the vertical axis is different and this line is not appropriate. In this latter case if a line is fitted to the plot the intercept and slope can be used as estimates of the mean and SD.

The curved nature of the plot is a clear indicator of non-normality, related to the kind of skewness evident in Figures 5.1 and 5.5. The obvious outlier previously detected stands apart in the upper-right of the plot. Interpretation of these kinds of plot is not always straightforward; further examples are provided in the next sections.

### 6.4.4 Using KDEs for SL comparisons

To illustrate some of the issues concerning SL comparisons with KDEs, touched on in Section 6.4.1, *Exodus* (1960) and *Aristocats* (1970) will be used. They have been chosen because they have approximately lognormal distributions and markedly different ASLs (22.4 and 4.0) and MSLs

---

4Unfortunately the term *normalization* is sometimes used for the transformation; the term *standardization* is preferred here as it avoids the implication that the data are converted to normality. If the distribution is non-normal to begin with it will be non-normal after transformation to z.
Figure 6.6: A normal probability plot for standardized logged SL data for A Night at the Opera. See the text for an explanation of the reference lines.

(13.8 and 3.3), so that comparing their underlying distributions presents challenges not arising from different underlying distributional forms.

Figure 6.7: Different comparisons of SLs for Exodus and Aristocats using KDEs. See the text for discussion.

This is illustrated in the left-hand plot of Figure 6.7, using untransformed SL data. The differences in location and scale and shape make assessment of the underlying distributional similarity difficult. After log-transformation the middle plot removes this problem to some extent, but location and scale differences remain. By standardizing the log-transformed data, as in the final figure, their distributional similarity to each other and to the reference standardized normal distribution (hence implying lognormality) is apparent.

It should be emphasized that this way of comparing SL distributions focuses on only one aspect, the underlying probability distribution (if any). In other contexts location and scale differences, of the kind evident in the left-hand plot, may be of more direct interest.
Code to obtain the plots is of the basic form

```r
plot(density(f1))
lines(density(f2))
```

where \( f1 \) and \( f2 \) are the appropriate transformations for the films. That is, successively, \( f1 \) is \( \text{SL.Exodus}, \log(\text{SL.Exodus}), \text{scale}(\log(\text{SL.Exodus})) \), with \( \text{Aristocats} \) replacing \( \text{Exodus} \) for \( f2 \). Use the film with the largest SL for \( f1 \). Arguments governing appearance of the plots have been omitted; \( \text{ylim} \) may need experimenting with to fit everything on the plot; thus \( \text{ylim} = c(0, 0.7) \) and \( \text{ylim} = c(0, 0.4) \) were needed for the first two plots.

The normal curve is added using

```r
x <- seq(-3.5, 3.5, .01) # Plotting positions
y <- dnorm(x) # Normal densities
lines(x,y, lwd = 3) # Add to KDE plot
```

where 3.5 in the definition of \( x \) will normally be adequate, but can be increased if the range of the KDEs exceeds these limits.

### 6.4.5 Examples

**Walk the Line (2005)**

James Cutting\(^5\) in a debate with Barry Salt about various aspects of Cutting *et al.* (2010) presents a histogram very similar to that to the left of Figure 6.8, for *Walk the Line* (2005). It is used in support of the contention that many films, over a period of about 70 years and on a log-scale have distributions that ‘look normal’. Of the 134 films assessed for lognormality in Redfern (2012a), a subset of the 150 used in Cutting *et al.* (2010), *Walk the Line* is one of the least (log)normal according to several criteria used.

![Figure 6.8: A histogram, KDE and normal probability plot for the logged SLs of Walk the Line, also standardized for the last two plots.](image)

The KDE indicates why. There are several aspects of it that are clearly non-normal, most noticeably the left-tail behavior, but also a degree of skewness. These are emphasized even more in the probability plot where, apart from the curvature over much of the range, the aberrant left-tail behavior is clearly emphasized. The transformed data are clearly not normal and the film does not have a lognormal SL distribution.

The histogram is a blunt instrument for assessing normality at best. The problem is compounded here by the choice of large bin-widths, producing a rather uninformative histogram

---

\(^5\)http://www.cinemetrics.lv/cutting_on_salt.php
where the tail behavior is nearly invisible. The main point of the example, though, is to illustrate how sensitive the probability plot can be to distributional features of the tails, indicative of non-(log)normality.

Whether or not departures from lognormality are of any importance in whatever one considers to be the ‘stylistic’ aspects of a film is an entirely separate judgment. The probability plot shows that about 15 SLs (of 1310) are noticeably smaller than expected if lognormality applies. They are shots measured as having a duration of 1 second or less, down to 2 deci-seconds.

Comparing probability plots

Probability plots can be difficult to interpret, and direct comparison of plots for different films, on the same graph, may be useful in getting some feel for what patterns and differences tell you. Figure 6.9 attempts such a comparison for Walk the Line and A Night at the Opera, mainly to show how this might be done. It is necessary to use standardized logged SLs for this plot.

![Figure 6.9: To the left a KDE of standardized logged SL data for A Night at the Opera; to the right probability plots for this film and Walk the Line compared.](image)

The left-hand plot showing the KDE for standardized logged SL data for A Night at the Opera may be compared with the comparable plot for Walk the Line in Figure 6.8. The point to note for present purposes is that the left-tail of the KDE for the latter film trails beyond that for the standardized normal distribution; the opposite is the case for the former film. You need to look carefully to see this – it doesn’t ‘hit you in the face’. The difference is starkly evident in the tail behavior exhibited in the probability plots to the right. These were obtained from the following code. The scale function, without other arguments, generates data standardized to have zero mean and unit variance.

```r
slWTL <- scale(log(SL.Walk_the_Line))
slNATO <- scale(log(SL.Night_at_the_Opera))
x1 <- qnorm(slWTL)$x
y1 <- qnorm(slWTL)$y
x2 <- qnorm(slNATO)$x
y2 <- qnorm(slNATO)$y
plot(x1, y1, col = "blue", xlim = c(-3.5, 3.5), ylim = c(-4, 4), xlab = "theoretical", ylab = "observed", main = "")
par(new = T)
plot(x2, y2, pch = 16, col = "red", xlim = c(-3.5, 3.5), ylim = c(-4, 4), xlab = "", ylab = "", main = "")
abline(0,1, lwd = 2)
```
The objects \( x_1 \) to \( y_2 \) hold the coordinates for plotting the two probability plots, executed using the \texttt{plot} commands; \texttt{par(new = T)} overlays the plots, with the \texttt{xlim} and \texttt{ylim} arguments set to ensure the plot limits are the same. The \texttt{pch} argument determines the plotting symbol, the default (1) being an open circle. Googling ‘R pch’ is enough to see what the choice is, and the symbols are listed in books such as Venables and Ripley (2002) and Murrell (2011).

**Comparing more than two KDEs**

One obvious advantage of KDEs over histograms is the ease with which they can be overlaid, for comparative purposes. How many plots can be usefully overlaid depends a bit on what they look like and what you might want to say about them. The next example started out in a ‘for fun’ spirit to see how far things might be pushed but turns out to embody a possibly useful idea.

![Figure 6.10: A comparison KDEs for the logged SLs of three named films against a background of eight others with approximately lognormal distributions.](image)

Figure 6.10: A comparison KDEs for the logged SLs of three named films against a background of eight others with approximately lognormal distributions.

In Redfern’s (2012a) study of the lognormality of SL distributions only 8/134 films satisfied the quite stringent statistical hypothesis tests (of normality of log-transformed data) he applied. In chronological order these were *Captain Blood* (1935), *Fantasia* (1940), *The Great Dictator* (1940), *Brief Encounter* (1945), *Night of the Hunter* 1955, *Exodus* (1960), *Aristocats* (1970) and *Barry Lyndon* (1975). They might be taken as a standard against which the lognormality of other films can be compared. Their KDEs are plotted as ‘background’ in Fig 6.10, as dotted lines and without labeling to play down the visual impact. The use of hypothesis tests is discussed more fully in Chapter 12; here more informal graphical methods of assessing (log)normality are used.

Fig 6.10 includes KDEs for the logged SLs of *Harvey* (1950) and *Pursuit to Algiers* (1945), about which I think there can be no argument about their non-normality. *Madagascar* (2005), with 1,131 shots, looks reasonably normal to my eye, more so, in fact, than some films in the ‘reference’
collection. Close inspection and further analysis shows that the clear rejection of normality in Redfern (2012a) is attributable to two outliers in the left tail produced by the log-transformation, and omitting these results in a judgment of normality more in accordance with the visual impact.\footnote{One of the tests used in Redfern (2012a) is the Shapiro-Francia test of normality, and the decision rule used is a 5\% significance level. The unmodified data for Madagascar is, on the log-scale, rejected as normal at the 1\% level also. On omitting the two outliers, associated with SLs of 1 deci-second, normality is acceptable at the 3\% level.}
Chapter 7

Time-series analysis of SLs

7.1 Introduction

So far the concern has been with what might be called the ‘external’ or ‘global’ analysis of SL data. That is, statistics or graphs that characterize the SL distribution as a whole. This chapter and the next look at methods relating to the ‘internal’ structure of SL data; that is, the distributional patterns within a film, taking into account the order of shots. This can be characterized by either the sequence order, a number between 1 and n, or the time at which a transition between shots, a ‘cut’, is recorded as occurring.

This is the province of time-series analysis. A large body of statistical theory has developed around this area, some of it complex, and some of which has been used in the cinemetric and related literatures. In this chapter methods of graphical analysis, mostly simple, are reviewed. Interest in this is evidenced in the Cinemetrics software, in the form of fitting polynomial regression models (called trendlines) to SL data, or moving averages.

Moving averages are a form of smoothing method, and unless otherwise stated the methods illustrated in the chapter are of this type. Smoothing methods have underlying features in common. A region (neighbourhood or window) of SLs is associated with each shot; the SLs in the region are used to calculate a ‘representative’ value for that region; the points so defined are ‘joined-up’ to produce a ‘curve’ smoother than the plot of the original data.

The idea of moving averages, pursued in later sections, can be illustrated using an example for *Ride Lonesome* (1959) from Salt (2010). Here an outline of the way these specific moving averages (MAs) are constructed is given; generalization is fairly obvious, but dealt with later. For the upper plot:

1. A window of 21 shots is selected; call this MA(21).
2. Centered MAs are calculated, beginning with the 11th shot, that are the averages of the 21 SLs centered on the shot.
3. The centered MAs are plotted against the sequence numbers of the central shots.

*Ride Lonesome*, with $n = 520$ shots, and an MA(21) ‘smooth’, gives plotting positions for the 11th to 510th shot (i.e. you ‘lose’ 10 shots either side of the central value and cannot, for example, calculate the MA for the 6th shot). Plotting positions might be obtained for shots outside this range in various ways, but this is not attempted here. Note that Salt uses deci-seconds for SLs (seconds are mostly used elsewhere in these notes); that the plot could use cut-point rather than sequence number; and that the background plot of SLs against sequence is ‘truncated’.

Further smoothing is achieved in the second plot by applying a second moving average to the first set of moving averages. The result of an MA(11) smooth is further smoothed in an MA(21) analysis. Note that if corrective action is not taken this involves more loss of ‘information’ at either end of the sequence. Salt (2010) contrasts these kind of plots with polynomial smoothing.
The latter is an option in the Cinemetrics software and is discussed first before moving on to moving averages and other methods of smoothing.

In most of remaining sections an attempt has been made to cover those methods I’ve seen in the literature, without attempting to be selective, and without too much critical comment. The reader may think that a lot of the plots, for Top Hat used for many of the illustrations, look very similar, and they would be correct. The concluding Section 7.8 provides a more comparative and critical assessment of the different methods.

### 7.2 Polynomial models (trendlines)

In the present context a polynomial model is just a curve fitted through a plot of SLs against either their sequence number in a film or cut-point. They can be used either to get some idea of the general trend in SLs in the course of a film (hence ‘trendline’) or, more ambitiously, to recover the detailed temporal structure within a film, making it more visible by ‘smoothing out’ some of the variation inherent in SLs. Salt (2010) for example, looks at the possibility of recapturing structure at the level of scenes. Other smoothing methods, discussed later, are possibly better at this, but it is convenient to begin with polynomial models to introduce ideas used in the sequel.

For notational convenience equate the term $y$ with the SLs. Mathematically, a polynomial is defined as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \ldots$$

where $x$ can be the cut-point associated with (the end of) a shot or its position in the sequence of shots in the film. The basic idea is to see how SLs vary over the course of a film.

Where you truncate the series depends on what you want to do with it. The power of $x$ at which you choose to truncate it is the order of the polynomial. Thus

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$
is of order 2 and fits a \textit{quadratic} polynomial to the data. If the term $\beta_3 x^3$ is added a cubic polynomial is obtained. Special cases are the models $y = \beta_0 + \beta_1 x$ and $y = \beta_0$. The first of these fits a \textit{linear} model (straight line) through the data, which shows if the ASL tends to increase or decrease between the beginning and end of a film. The second simply fits the ASL.

Figure 7.2: Fifth and first order polynomial fits to the SL data for Intolerance plotted against cut-point. See the text for more detail.

All this is best illustrated by an example, which is based on Tsivian’s (2005) study of \textit{Intolerance} (1916) (see, also, http://www.cinemetrics.lv/tsivian.php). The left-hand plot in Figure 7.2 emulates the first figure in the web version, plotting the curves against cut-point. The solid line is a fifth-order polynomial fit, the dashed line is the fit from a linear model. The vertical scale is constructed to match that in the article and is the reverse of that used in other sections of this chapter.

For a detailed discussion see Tsivian’s article. Briefly the linear fit shows that SLs at the end of the film tend to be shorter than those at the start. The fifth-order polynomial produces a more nuanced picture. The ‘bumps’ are associated with periods of faster cutting or ‘action’, relative to their surroundings, with a bit of a rest in the middle. The tempo increases fairly steadily from just before the middle of the film with what Tsivian calls a ‘peaceful apotheosis’ associated with the slowing down at the end.

The right-hand plot is as the left, with a bit of embellishment, which includes the ASL for the film as a whole, shown as a horizontal dotted line. The coloring is an attempt to indicate where the four interwoven stories in \textit{Intolerance} occur. The dominance, in terms of time occupied, of the Babylonian and Modern stories is evident, as is faster cutting between stories towards the end.

Tsivian also looks at SL dynamics within stories. Figure 7.3 emulates this, where the fits and plots are based on shot sequence number within the story. Other than the Judean story the tempo is seen to increase during the film, with the pattern for the Babylonian and Modern stories very

\footnote{Some comment on scaling may be merited here. The visual impression of changes in the SL dynamics over the film may be exaggerated or under-emphasized if the range of the vertical scale is too small or too large. There is white space at the bottom of the left-hand graph which helps ‘play down’ the magnitude of SL variation. This has been removed in the second plot. Where data are naturally bounded below by zero, including this as one end of the scale is sometimes recommended in texts that deal with this sort of thing. It’s not always sensible if the ‘business’ part of a graph is at a great distance from the origin, but is reasonable enough here. Truncating at 4, to eliminate white space in the upper part of the left plot, would exaggerate the nature of the variation in SLs. The scale in Figure 7.3 is truncated, but this is less important as the aim is comparison.}
similar. The rather different pattern for the Judean story is evident; possible reasons for this are discussed by Tsivian.

Figure 7.3: Linear and polynomial fits for the SLs of the different stories within Intolerance. For the Babylonian and Modern stories 7th order polynomials are fitted, 5th order for the French story and 4th for the Judean.

Code for the left-hand plot in Figure 7.2 is given below, where SLs for the film, in seconds, are stored in SL

```r
minutes <- cumsum(SL)/60 # cut-point in minutes
fit1 <- lm(SL ~ poly(minutes, 1))
fit5 <- lm(SL ~ poly(minutes, 5))
Ylim <- c(14,0)
pred <- 1:max(minutes)
plot(pred, predict(fit5, data.frame(minutes = pred)), "type = "l", ylim = Ylim, xlab = "minutes", ylab = "SL", main = "'Intolerance"")
lines(pred, predict(fit1, data.frame(minutes = pred)), type = "l")
```

The arguments 1 and 5 in the `poly` function specify the order of the polynomials selected for plotting. Defining `data <- 1:n` and replacing all occurrences of `minutes` with `data` will result in an analysis based on sequence number rather than cut-points. The plots in Figure 7.3 were obtained in this way, using data for the individual stories. The value of `Ylim` is passed to the `ylim` argument in `plot`, defining the scale of the vertical axis and, in this instance, reversing the ‘natural’ ordering of plotting positions. Code for the right-hand plot of Figure 7.2 is discussed in the next chapter (Section 8.1).

The `cumsum` function generates the cumulative sums of the SLs (i.e. the cut-points); `lm` (for linear model), fits the polynomial model specified in its arguments.
7.3 Simple moving averages

Figure 7.4 shows moving averages (MAs) for *Ride Lonesome* obtained in R. The lower MA is on the same scale as the SL data and shows the pattern in the data well enough, as can be seen from the accompanying plot of SLs against sequence number. The upper MA is the same as the lower, magnified by an arbitrary factor (of 5) to show the detail better. This is just an idea that may help if the need is felt for a direct comparison between the MA and the raw data when the scale is such that the MA on the proper scale is too ‘squashed up’.

![Ride Lonesome – moving average plots](image)

Figure 7.4: Moving averages for *Ride Lonesome*. The upper (blue) MA is simply a magnified version of the lower (red) MA, to show more detail, and the scale should not be referred to the vertical axis.

Moving averages in R are available in several packages and the `rollmean` function from the `zoo` package was used here. It’s simplest to set this up as a function (Section 4.3) as follows.

```r
MAplot <- function(film, MA = 21) {
  library(zoo)
  SL <- film
  SLcut <- cumsum(SL)
  n <- length(SL)  # Number of shots
  minutes <- SLcut/60
  ppMA <- rollmean(SL, MA)
  ppMin <- minutes[((MA + 1)/2):((n - ((MA -1)/2))]
  ppSeq <- c(1:n)[((MA + 1)/2):((n - ((MA -1)/2))]

  plot(1:n, SL, type = "l", xlab = "sequence number", ylab = "SL")

  lines(ppSeq, ppMA, lwd = 3, col = "red")
  lines(ppSeq, 5*ppMA, lwd = 3, col = "blue")
}
```

62
Once the function is written `MAplot(SL.RideLonesome)` produces the figure. The obvious advantage of setting it up like this is that SL data for any other film can be substituted for *Ride Lonesome*. The default window for the plot has \( MA = 21 \) shots. This can be changed if you want a larger or smaller window\(^3\). Salt (2010) suggests that this default possibly ‘works best’. Having \( MA \) as an odd number ensures that the centre of the window corresponds to an observed shot. If the ‘magnified’ moving average is surplus to requirements omit the final line, most simply done by commenting it out by placing \( \# \) at the start of the line.

The code `SLcut <- cumsum(SL)` generates the cut-points in seconds; \( \text{minutes} \) is the cut-point in minutes; \( \text{ppma} \) provides the MAs; \( \text{ppSeq} \) the plotting positions if sequence order is used; and \( \text{ppMin} \) the plotting position, in minutes, if cut-point is used. The objects \( \text{SLcut, minutes and ppMin} \) have not been used for the present figure, but are needed if \( \text{SLcut replaces 1:n in the plot command, and ppMin replaces ppSeq in the lines commands to get plots against cut-point.} \)

### 7.4 Smoothed moving averages

Although often giving a usable picture, MAs can be ‘rougher’ than one would like, with distracting detail. An obvious idea, once you have it, is to apply further smoothing to the MA. Salt (2010) suggested using a second MA smooth. Baxter (2012b) suggested using *loess smoothing*. The approaches are compared here. Loess smoothing is discussed further in Section 7.5.

![Ride Lonesome – smoothed MA plots](image)

**Figure 7.5:** *Double MA (blue) and loess smooths (black dashed line) for Ride Lonesome, with the original MA(21) smooth (red) also shown.*

Apart from being smoother, by design, the original and doubly-smoothed MAs in Figure 7.5 are almost indistinguishable (particularly if you are not looking at the figure in color). Following Salt (2010) the doubly-smoothed version was obtained by what may be called an \( \text{MA(11) + MA(21)} \) smooth. There are \( n = 520 \) shots in *Ride Lonesome*. For the original \( \text{MA(21)} \) smooth 10

---

\(^3\)That is, if you want an MA of 15 the call to the function is `MAplot(SL.RideLonesome, MA = 15)`. 

63
shots are ‘lost’ at either end, so the plot sequence is from the 11th to 510th shot. If preceded by an MA(11) smooth an additional 5 shots are ‘lost’ at each end so the sequence is from 16 to 505. Ignoring labels and style arguments, the single- and double-smooth shown can be obtained by

```r
library(zoo)
Seq1 <- 11:510
Seq2 <- 16:505
MA <- rollmean(SL.Ride_Lonesome, 21)
MA1 <- rollmean(SL.Ride_Lonesome, 11)
MA2 <- rollmean(MA1, 21)
plot(Seq1, MA, type = "l")
lines(Seq2, MA2, type = "l")
```

This could be written in a more general form as a function for repeated analyses. Fitting loess curves is discussed in detail in the next section. That shown was chosen to mimic the other two curves as much as possible. Deferring discussion of details, it was added using

```r
lines(loess(11:510, rollmean(SL, 21), span = 1/9, family = "g", degree = 2),
      lwd = 3, lty = 2, col = "black")
```

### 7.5 Loess smoothing

#### 7.5.1 Examples

Given that MAs are mathematically simple, easy to understand and seem to satisfactorily do the job of smoothing the data in a comprehensible way, it’s legitimate to ask why loess smoothing needs to be considered. The mathematics is fairly horrible\(^4\). Conceptually the ideas involved aren’t too demanding and, in R at least, it is arguably computationally simpler to get a loess curve than an MA.

There is greater flexibility in the degree and type of smoothing, allowing some choice in highlighting features of the data one might wish to emphasize. Smoother curves that those obtained using MAs can be obtained which, aesthetic appeal apart, can also assist comparison between films. Additionally, a single loess fit can be used to smooth the data, rather than applying it to an MA or using a doubly-smoothed MA without (given the way the algorithm works) ‘losing’ plotting positions.

The ideas involved may best be illustrated by example, before looking at the detail. *Top Hat* (1935) will be used. In a sense this is quite an ‘easy’ film to study, having a reasonably clear structure that makes it useful for methodological comparison, without having to worry if one method is more ‘right’ than another. The film has been studied using other approaches in Redfern (http://nickredfern.wordpress.com/category/top-hat/) that are discussed in Section 7.6.

Figure 7.6 shows the plot of SL against cut-point for *Top Hat*. An MA(21) smooth and two loess curves are superimposed. Details of the latter are discussed shortly.

For the moment note that, after smoothing and whatever method is used, the SLs bounce along very nicely at intervals (very roughly) of 5-10 minutes. One of the loess smooths produces a result barely distinguishable from that for the MA; as noted above it looks nicer, doesn’t lose information at the ends, and is as easy to produce as an MA. The choice available in loess smoothing is illustrated by the contrast between the two shown. They bounce in the same places but on two occasions, a bit after 20 minutes and in the mid-60s, one bounces somewhat higher than the other.\(^5\) It is giving more weight to the proximity of some reasonably long SLs in the intervals involved.

---

\(^4\)See Cleveland (1979); statistical texts centered on the use of R that deal with the method tend to spare readers the gory detail (Venables and Ripley, 2002; Faraway, 2006).

\(^5\)These are the scenes with the musical numbers ‘Isn’t This A Lovely Day (To Be Caught In The Rain)’ and ‘Cheek to Cheek’.
Figure 7.6: Plots of SL against cut-point for Top Hat with a moving average and two loess smooths superimposed.

Figure 7.7: Different loess smooths for the SLs of Top Hat. See the text for details.
7.5.2 Some technicalities

These two loess smooths and two others can be obtained, and shown on the same plot, Figure 7.7, without the distraction of the background SL plot and MA, using the following code (omitting labelling information).

```r
lfitg2 <- loess(SL ~ minutes, span = 1/9, family = "g", degree = 2)$fitted
lfitg1 <- loess(SL ~ minutes, span = 1/9, family = "g", degree = 1)$fitted
lfits2 <- loess(SL ~ minutes, span = 1/9, family = "s", degree = 2)$fitted
lfits1 <- loess(SL ~ minutes, span = 1/9, family = "s", degree = 1)$fitted
plot(minutes, lfitg2, type = "l", lty = 2, lwd = 3, col = "red")
lines(minutes, lfitg1, type = "l", lty = 1, lwd = 3, col = "green4")
lines(minutes, lfits2, type = "l", lty = 2, lwd = 3, col = "purple")
lines(minutes, lfits1, type = "l", lty = 1, lwd = 3, col = "blue")
```

The degree of smoothing is controlled by the size of the neighbourhood surrounding an SL. This, in turn, is controlled by the `span` argument. The default is 3/4, which is too large for present purposes, and over-smoothes. A smaller value of 1/9 has been selected here and, as MA(21) seems to do for moving averages, often works well as a starting point.

The representative point is calculated by predicting it at the SL defining the neighbourhood, from a polynomial regression model fitted to the data in the neighbourhood (polynomial fitting is covered in Section 7.2). Most simply a linear model (straight line) is fitted to the data, specified by `degree = 1` in the above code. The default is to fit a quadratic regression. This is `degree = 2` in the code but it need not be specified explicitly.

The other choice that affects the result is the choice of fitting procedure, controlled by `family`. The default, `family = "g"`, assumes normal (Gaussian) errors – essentially this is the standard method for least squares fitting found in introductory textbooks\(^6\). The choice `family = "s"` applies an iterative estimation procedure that downweights outliers – in this context ‘unusually’ long SLs relative to other values in a neighbourhood. That is, it is intended to be robust.

The degree of smoothing, via `span`, can be experimented with, as with the choice of window size in MAs. The sensible approach to the other choices is to experiment with them rather than making a priori judgments about what is best. Top Hat has fairly obvious structure, and Figure 7.7 suggests that the ‘robust’ fits, the purple and blue lines, are not that great at picking this up. That for the linear fitting, the blue line, is borderline useless. The problem would seem to be that the effect of the longer SLs, arguably responsible for the ‘character’ of some scenes, is being downweighted too much (and to the point of non-existence in some cases).

The procedures with `family = "g"` selected , where no attempt is made to downweight longer SLs, produce much more satisfactory results here. That with a quadratic fit emphasizes two important scenes somewhat more than the linear fit, as it accommodates itself more to the longer SLs. Both fits show similar general patterns in the data\(^7\).

The above is not to be read as a general recommendation of any sort. If one is interested in identifying broad trends in the SL dynamics, as in the polynomial smoothing of Section 7.2, a reasonably large value for `span` would be appropriate. A case can be made for reducing the effects of extremes by using `family = "s"` and `degree = 1`. As ever, trying out more than one choice and trying to explain any radical differences, if they emerge, is the sensible course.

A final thought, prompted by the above discussion, is what might happen if smoothing was applied to the logged SLs. It might be expected with the untransformed SLs that peaks would be emphasized at the expense of troughs, and that log-transformation would endow them with more equal status\(^8\). Figure 7.8 investigates this.

---

\(^6\)I’m lying here, to keep ideas simple. What is actually applied is a weighted least squares procedure, summarized in Cleveland et al. (1993, p.314). I’m taking the attitude that the user need not worry about this.

\(^7\)In Baxter (2012b) where I discussed loess smoothing it was illustrated using the `scatter.smooth` directive. This is a quick way of adding a loess smooth to a plot, with similar arguments. The defaults differ, however, with `span = 2/3, family = "s"` and `degree = 1`. The last two are the ‘opposite’ to what you get with `loess`, so explicit declaration of what you intend is the safest approach. I now prefer to use `loess` directly, as illustrated in the section, as it allows much more control over style (line type, line width, and color) than `scatter.smooth`.

\(^8\)One of the effects of log-transformation is to downweight the ‘importance’ of larger values, so in this sense it has some of the flavor of ‘robust’ methodologies.
Figure 7.8: Different loess smooths for the logged SLs of Top Hat. See the text for details.

With minor variation in the troughs the patterns identified using the degree = 2 option are now very close. Similarly, with some variation for two of the peaks, the degree = 1 smooths look similar. The implication is that choosing between the "s" and "g" options for the family argument is less critical; probably to be expected because of the log-transformation.

Qualitatively – that is, in terms of the ‘up-and-down’ nature of the plots, ignoring the magnitude of ‘up-ness’ and ‘down-ness’ – the different smooths give a similar picture, those for the degree = 2 option being a bit sharper. As expected, compared to Figure 7.7, troughs are emphasized more, chiefly because the impact of the peaks is diminished and equalized. Inevitably, whether these representations or those on the untransformed scale are preferred depends on how comfortable one is with using logarithms; and whether there is a desire to retain the information in the untransformed SLs, rather than ‘rebalancing’ the emphases on longer and shorter SLs.

7.6 Robust methods

7.6.1 Introduction

Polynomial (trendline) fits to SL data are an example of parametric regression models meaning, in this instance, that they have a predetermined mathematical form. Because of variation in the data this form has to be estimated, commonly by fitting a curve than minimizes the sum of squared distances between it and the observed SLs. This is the method of least squares. The exact form of the fitted curve determined in this way depends on all the SLs in a film, and can be sensitive to isolated very large SLs. That is, some perceive it as an estimate that is not robust. Robust methods exist that downweight the importance given to more extreme SLs.

The loess smoother belongs to a class of methods that are sometimes called non-parametric regression. No predetermined form is specified for the curve; the data ‘decides’ what form the curve will take, except that you have to tell it how. The representative points determined for
the neighbourhoods are affected only by the SLs in the neighbourhood. A parametric (linear or quadratic) regression fit is used within neighbourhoods, and may be estimated in a robust or non-robust manner, depending on whether the family = "s" or family = "g" argument is used.

Two points can be made here. One is that the distinction between what is parametric and non-parametric isn’t always as clear as one would like. Loess smoothing is characterized as non-parametric, but the fitting method used depends on parametric ideas.

The second point is that a distinction needs to be made between non-parametric and robust methods. Non-parametric methods can sometimes be viewed as robust alternatives to parametric methods, but there are methods for which parametric analogues don’t exist. The loess smoother is an example of a non-parametric method that is not necessarily robust.

Another point to emphasize is that the idea that robust methods are somehow ‘better’ than non-robust methods (some of the cinematic literature can be read in this light) ought to be rejected. They are simply different ways of looking at data. What is appropriate depends on what you are attempting to do, and it is often fruitful to examine data using both kinds of approach.

7.6.2 Moving medians

One plank in the argument for preferring the MSL to the ASL as a measure of ‘film style’ is that it is robust, unaffected by outlying large SLs. Regardless of the merits or otherwise of this argument it suggests, in the context of graphical analysis of internal SL structure, the possibility of using moving medians (MM) as a robust alternative to moving averages. To illustrate, MM(21) and MA(21) smooths for Top Hat are contrasted in Figure 7.9 (the latter, as noted from Figure 7.6, is similar to the loess (1/9, g, 1) smooth, also shown).

![Figure 7.9: Moving average, moving median and loess smooths for Top Hat.](image)

The loess and MA(21) smooths identify the same peaks and troughs, the former giving a little more emphasis to the two musical numbers that some would regard as among the highlights of the film. The MM(21) smooth almost, but not quite, identifies similar peaks and troughs, but with
different emphases that are nothing like as strong as those for the other two smooths. The failure, if that is what it is, of the moving median to highlight two of the major features suggested by the other smooths is because they include some of the longest shots in the film, which the ‘robust’ median effectively discounts.

The plot may be constructed in the same way as others previously discussed. The only new feature is the use of the \texttt{rollmedian} function from the \texttt{zoo} package to get the moving median. It requires that the window size used be an odd number.

### 7.6.3 Moving average of ranked SLs

Non-parametric methodology is often associated, at least at the level of introductory statistics texts, with the analysis of ranked data. Since some kinds of data exist only in ranked form this does not necessarily make such analyses ‘robust’, but they can be regarded as such when applied to measured data, like SLs, that are converted to ranks. This is because differences of magnitude, conveyed by the SLs, are eliminated by their conversion to ranks, so reducing any effect on the analysis that extreme values might have.

To illustrate, there are $n = 595$ shots in \textit{A Night at the Opera}, the four largest of which are 101.7, 106.4, 111.5, 207.2, with differences between them of 4.7, 5.1 and 95.7, the last of these obviously being very much the largest. If ranked by size they are converted to 592, 593, 594, 595 with the same difference, 1, between them. This means that any analysis based on the ranks is not affected by the fact that the actual SL for the largest value is very different from everything else, so the analysis can be considered ‘robust’. The mean and median of ranked data are the same, so issues of choosing between them go away.

It is in this spirit that Redfern (http://nickredfern.wordpress.com/2011/06/16/time-series-analysis-of-top-hat-1935/), with some statistical embellishment, presents \textit{time-series Analysis of Top Hat (1935)}. The outcome of the analysis, from that paper, is shown in Figure 7.10. It can be compared with the MA analyses in Figures 7.6 or 7.9.

![Figure 7.10: A time-series analysis of ranked SL data for Top Hat from Redfern. See the text for details. (Source: http://nickredfern.wordpress.com/2011/06/16/time-series-analysis-of-top-hat-1935/)](image)

Ignoring the statistical flourishes, but only for a moment, the only difference from previous comparable analyses is that ranked data are used, and for the R code all that is needed is to replace \texttt{SL} with \texttt{rank(SL)} in the relevant places. That is, it is very easy to get the plot (or as easy as any of the other plots). Before looking at interpretation, technical details explaining what the reference lines in the figure do are provided. I’m going to suggest, in Section 7.8.3, that if interest lies in an ‘holistic appreciation’ of pattern, rather than identifying ‘significant’ peaks and troughs, the statistical flourishes aren’t needed. Anyone uncomfortable with the maths who wants to be persuaded can skip the next section, but it does give the detail needed to emulate Figure 7.10.
Technical details - the Mann-Whitney statistic

The basic idea behind the reference lines in Figure 7.10 is to provide ‘objective’ guidelines to identify ‘significant’ troughs and peaks in the plot. An attempt will be made to describe what is going on in words. The source of the ideas is Mauget (2003, 2011) and the mathematics is also given in Redfern (2011, 2012b) as well as the paper from which Figure 7.10 comes.

For the purposes of this section let \( n \) be the number of shots, \( n_1 \) the size of the window, and \( n_2 = (n - n_1) \). If \( S \) is the sum of the SLs in a window then \( MA = S/n_1 \) is the moving average. Previous figures have plotted this against either the cut-point or shot sequence number as the window is moved along. The latter is assumed in this section.

Let \( r \) be the ranked value of an SL (1 for the smallest, \( n \) for the largest, and tied ranks being averaged), \( R \) their sum, and \( MAR = R/n_1 \) the moving average of the ranks. Figure 7.10 is essentially just a plot of \( MAR \) against shot sequence number. The ‘complications’ are driven by a wish to identify the ‘important’ peaks and troughs.

This is done by converting the statistics \( R \) into \( z \), a standardized normal statistic, plotting this and showing guidelines at -2 and +2 (actually ±1.96). The idea, very, very roughly is that ‘real’ peaks and troughs will tend to stray beyond these limits, but unimportant bumps and depressions trying to masquerade as something a bit more imposing in the ‘physical landscape’ have only about a 5% chance of doing so. That is, you will be fooled from time to time, but not that often.

Before outlining some problems with the idea the mathematics will be disposed of (this is the bit to ignore if you wish). The Mann-Whitney \( U \) statistic is defined as

\[
U = R - \frac{n_1}{2} (n_1 + 1)
\]

and the normal standardized value as

\[
z = \frac{U - \mu}{\sigma} = \frac{U - \frac{n_1n_2}{2}}{\sqrt{n_1n_2(n_1 + n_2 + 1)/12}}
\]

which is an asymptotic approximation\(^9\). Mathematically, \( (U - \mu) \) is defined such that extreme negative and positive values occur if the window contains the \( n_1 \) smallest or largest SLs in a film. Windows with less exceptional collections of SLs, closer to and either side of the median will produce values of \( (U - \mu) \) ‘close’ to zero.

The meaning of ‘close’ now needs to be defined ‘objectively’. In defining \( z \) the division by \( \sigma \) converts \( R \), via \( U \), to a scale where this can be done and ‘close’ is defined to mean anything in the range (-1.96, +1.96). This is objective but also arbitrary since it is tied to the ‘5% chance’ mentioned above\(^10\). You could equally validly change the rules and use 1% or 10% and come up with a different assessment of what constitutes a significant peak or trough. This is illustrated in Figure 7.11, based on Figure 7.10, using a window of size 21 and with additional 1% and 10% guidelines\(^11\).

In the figure the solid horizontal red lines are those used by Redfern (placed at ±1.96 to isolate ‘significant’ peaks and troughs using the 5% rule described earlier). The dashed blue and dotted green lines use 1% and 10% rules. Using the former means that you need more convincing that a feature is ‘significant’; using the latter means you’re more relaxed about what looks important. The obvious thing, using Figure 7.11 for illustration, is that the judgment depends on the

\(^9\)Meaning it’s valid if \( n_1 \) is large enough; defined as more than 20 in some texts, but you can get away with a bit less.

\(^10\)Introductory statistics texts and expository papers are replete with stern injunctions about this sort of thing, then proceed to use the 5% convention anyway. Use of the convention, without comment, is common in the non-statistical scientific literature.

\(^11\)There are some differences between Figures 7.10 and 7.11, most obviously between about the 400th and 450th shot region. An MA(21) smooth as opposed to Redfern’s MA(20) is used, but more importantly there are differences in the data. That used in these notes, and here for consistency, is James Cutting’s submission to the Cinemetrics database, whereas Redfern uses his own analysis, providing a link to it in his article. The interpretive issues raised here apply to both analyses.
Figure 7.11: A time-series analysis of ranked SL data for Top Hat in imitation of Figure 7.10, with differences explained in the text.

rule you use. With a 1% rule you end up with five ‘significant’ peaks (two very marginal), and eight troughs (two marginal, but another not quite making it). With a 10% rule you’d be happy with eight peaks (two a bit marginal), and nine troughs.

Code for constructing Figure 7.11 is straightforward, if a little tedious to set up. It is given for a window size of 21, with labelling arguments in the plot command omitted. The rank and sqrt functions are used.

\begin{verbatim}
n1 <- 21
n2 <- n - n1
SLrank <- rank(SL)
MW <- n1 * rollmean(SLrank, n1) - (n1 * (n1 + 1)/2)
mu <- n1 * n2/2
sigma <- sqrt(n1 * n2 * (n1 + n2 + 1)/12)
zMW <- (MW - mu)/sigma
ppSeq <- ((n1 + 1)/2):(n - ((n1 -1)/2))
plot(ppSeq, zMW, type = "l", lwd = 1.5)
abline(h = 0)
abline(h = 1.96, col = "red", lty = 1, lwd = 1.5) # 5% line
abline(h = 2.58, col = "blue", lty = 2, lwd = 1.5) # 1% line
abline(h = -2.58, col = "blue", lty = 2, lwd = 1.5)
abline(h = -2.58, col = "green4", lty = 3, lwd = 2) # 10% line
abline(h = -1.64, col = "green4", lty = 3, lwd = 2)
\end{verbatim}

7.6.4 Order structure matrices

Another graphic explored in Redfern\textsuperscript{12} (see also, Redfern, 2013a), based on ideas in Bandt (2005), is that of the order structure matrix. Figure 7.12 shows a visual representation of such a matrix for Top Hat. To quote Redfern, the dark patches in Figure 7.12 correspond to the peaks in Figure 7.11 and exhibit clustering of longer shots in the film, while the light patches correspond to the troughs.

\textsuperscript{12}http://nickredfern.wordpress.com/2011/06/16/time-series-analysis-of-top-hat-1935/
of Figure 7.11 and show where the clusters of shorter shots are to be found. ‘Although this plot looks complicated, once you get used to the method and are familiar with the events of the film you can simply read the changes in cutting style from left to right’. The emphasis is mine; it suggests other styles of presentation explored in the next chapter. Order structure matrices are exploited more extensively, emulating closely the style of presentation in Bandt’s (2005) Figures 1-4, in Redfern’s (2012c) exploration of the structure of ‘slasher’ films.

The basis of the construction is as follows. Compare the first SL to every SL in sequence (including itself) and code it 1 if it is less than the SL it’s compared to and 0 otherwise. This generates a row of \( n \) numbers that are either 1 or 0. Do this for each shot in turn and stack the rows up to get an \( n \times n \) table of 1s and 0s. Figure 7.12 is just a picture of this table, colored with black for a 1 and white for a 0. It can be thought of as a screen with \( n \times n \) pixels colored black or white.


The way this has just been described means that the longest shot corresponds to a column of 1s in the table, coded black, so its represented as a black ‘line’ on the screen. The longer shots will be mostly black lines broken up by bits of white. There is an obvious gradation, with the shorter shots tending towards white. If, for example, there is a cluster of longer shots, this will appear as a dark vertical band on the plot.

This is a robust method in a similar way to the MA analysis of ranked data in Figures 7.10 and 7.11. In the MA analyses the information provided by the absolute values of the SLs is removed in two ways, by plotting against the shot sequence number, and by using only relative magnitudes of length as reflected in the rank. In a sense the order structure matrix analysis goes even further, discarding information about differences in rank order, using only information about whether one shot is shorter or longer than another.\(^\text{13}\).

\(^\text{13}\)It’s tempting to suggest that the time and effort devoted to collecting SL data has been matched by that used
Code to produce the plot follows, where n is the number of shots and labelling information, which can be included in the `image` function, is omitted (SL holds the SL data).

```r
mat <- matrix(0, n, n)
for(i in 1:n) {
  for(j in 1:n) {
    mat[i,j] <- ifelse(SL[i] >= SL[j], 1, 0)
  }
}
image(1:n, 1:n, mat, col = c("white", "black"))
```

The table of 1s and 0s is contained in `mat`, Figure 7.12 being produced by `image`. Users may choose colors via the `col` argument; to see what the choice is type `colors()` or `colours()` when in R.

### 7.7 Kernel density estimation

Kernel density estimates (KDEs) were discussed in terms of the analysis of SL distributions, ignoring their time-series structure, in Sections 5.2 and 6.1. They can be used to investigate internal structure by applying the methodology to cut-points rather than SLs. Using *Top Hat* once again, an analysis from Redfern is reproduced in Figure 7.13.

![A KDE of the time-series structure of Top Hat, from Redfern.](http://nickredfern.wordpress.com/2012/03/15/using-kernel-densities-to-analyse-film-style/)

A little explanation is needed here. The scale on the horizontal axis is of cut-point divided by total running time, used in order to effect comparisons with other films, and not essential if only a single film is being examined. The ‘rug’ at the bottom of the plot shows the individual cut-points. It can be seen from this the the troughs corresponding to lower densities occur in those regions where the distance between cut-points is longer (i.e. to where SLs are longer). This is the opposite to previous comparable plots where regions with the longer SLs show as peaks.

With minor and unimportant presentational differences the plot can be obtained from

```r
SL <- SL.Top_Hat
SLcut <- cumsum(SL)
par(mfrow = c(2,1))
D <- density(SLcut/max(SLcut), bw = 0.025, from = 0, to = 1)
plot(D, xlab = "cut-point (scaled)", ylab = "density", type = "l", main = "")
rug(SLcut/max(SLcut))
```

to devise methods of statistical analysis that ignore much of the information so collected. Some commentators (e.g., Salt, 2012) have expressed the view that the unmolested SLs carry information that might be of interest. The issue – a real one in the context of graphical presentation – is discussed in the summary to the chapter.
where the `rug` function adds the data points at the bottom of the plot.

To ease comparison with smooths produced by other methods Figure 7.14 is shown. The smooths are based on a fit against cut-point (i.e. without scaling). A ‘mirror image’ of the KDE is shown, so that the peaks are associated with the longer shots in both plots. The loess (1/9, g, 1) smooth is that shown earlier to produce very similar results to an MA(21) smooth. The bandwidth for the KDE was chosen to get a good match and, end effects apart, there are no differences between the two plots of substantive import.

![Top Hat – KDE for cut-point (mirror image)](image1)

![Top Hat – loess (1/9, g, 1) smooth](image2)

Figure 7.14: A KDE of the time-series structure of Top Hat, using cut-point and showing the ‘mirror image’ contrasted with a loess smooth.

### 7.8 Summary

#### 7.8.1 General similarities

This section should be preceded by the caveat that, with ‘strong’ structure in a data set it might be expected that, conditionally on the same data transformation being used, any sensible method would be expected to produce similar results. Top Hat was chosen because it arguably has strong structure that should be relatively easy to unveil. Any method failing to do this could immediately be regarded as suspect. The tentative conclusions now offered need to be confirmed with a range of films having less obvious patterns in their internal SL distributions.

That MAs and smoothed MAs produce similar results for untransformed SLs, the latter having more aesthetic appeal, is to be expected. The results of either were essentially reproduced using loess smooths that did not explicitly downweight extreme SLs (Figures 7.5 and 7.6). The ‘robust’ versions of the loess smooth were less successful at identifying the ‘obvious’ structure – more careful discussion of this follows shortly. The loess and KDE smooths in Figure 7.14 were very similar.
Leaving aside the order matrix structure analysis and polynomial models for separate consideration, the same smoothing techniques were applied to obtain the other plots presented. The MM analysis just replaces the mean of an MA analysis with the median. I’d not seen this used before; used it in a spirit of exploration and adventure; and wasn’t that impressed. It just about captures the same pattern as other plots, but not very convincingly. This and the robust loess smooth with linear localised fitting were the least successful analyses, and I think for the same reason. This is that not only do they, intentionally, downweight the influence of the longest SLs, but do so to the extent of largely ignoring them. From a purely statistical point of view this seems wasteful of potentially useful information. In context, one assumes that the greater length of these SLs is intentional; using methods of analysis that not only downplays but possibly ignores them could be considered perverse.

The use of ranked data can be viewed as a robust methodology that downweights the importance of extreme SLs without discounting them. As for log-transformation, it is natural (for a statistician) to think in terms of this when data are highly skewed and differ by orders of magnitude. That there are differences in magnitude between SLs is recognised; these are downweighted, but differently from, and less so, than with the use of ranked data.

The use of untransformed, logged or ranked SL data, whether using MAs or some of the loess smooths, produces qualitatively similar results. That is, they all bounce along at reasonably ‘regular’ intervals, with about 10 fairly clear peaks and corresponding troughs in most cases (you can quibble about some very minor hillocks). If qualitative similarity is what is mainly of concern – quite legitimate if the aim is comparison with other films to see if the ‘rhythms’ are the same – there is no need to choose. The heights of the bounces are different, however, for different methods.

7.8.2 Specific differences between methods

It is convenient, without necessarily privileging it, to take the loess (1/9, g, 2) smooth as a base for comparison. The difference between peaks is more prominent than for other analyses. The five main peaks in order of occurrence are associated with musical/dance numbers, No Strings (NS), Isn’t This A Lovely Day (ITLD), Top Hat, White Tie and Tails (THWTT), Cheek to Cheek (CC) and The Piccolino (P). Their relative importance (height) is ITLD (very obviously), CC (also obviously), P, THWTT and NS.

The loess (1/9, g, 1) smooth reproduces this pattern, except that the prominence of ITLD is less evident and similar to CC. The degree = 1 smooths, particularly that with family = “s”, are much less satisfactory at discriminating between peaks (Figure 7.7). As noted in connection with Figure 7.14 a KDE analysis with the same pattern as the loess (1/9, g, 1) smooth is readily obtained, both being similar to the MA(21) smooth.

For the analysis of log-transformed data the loess (1/9, g, 2) and loess (1/9, g, 1) smooths are very similar and highlight the same peaks as the smooths for untransformed data. The chief difference is that in terms of relative importance there is little to choose between the three central peaks for THWTT, CC and P. The other loess smooths are better than for untransformed SLs, the chief difference, and particularly for the loess (1/9, s, 1), being that the prominence of ITLD is considerably diminished.

\[\text{14} \text{Since writing this Redfern (2013a) has also illustrated the use of the MM and takes a somewhat more positive view.}\]

\[\text{15} \text{Note that the nature of ‘robustness’ here differs from that employed in using the MSL rather than ASL to summarize an SL distribution, where the former is designed to completely ignore extreme (and not so extreme) values.}\]

\[\text{16} \text{Meaning, roughly that some SLs may be 10, 100 or 1000 times greater than others – 1 deci-second, and 1, 10 and 100 seconds, for example. On a log-scale these numbers have an equal distance between them.}\]

\[\text{17} \text{With untransformed data it is visually more ‘natural’ to focus on peaks as they stand out more. Troughs, by definition (in most of the methods used) are associated with short SLs that will be ‘squashed’ towards the bottom of a graph. That is, their depths are not free to range in the same way that the unconstrained heights associated with longer SLs are. This visual bias is not necessarily a ‘good thing’ since in films where ‘dramatic action’ is associated with clusters of short SLs their importance may be under-emphasized. Rank and log-transformations are ways of avoiding this.}\]

\[\text{18} \text{These are peaks A, C, D, E, G in Redfern’s analysis of ranks (Figure 7.10) whose discussion informs that here.}\]
For the analysis of ranks more peaks are given equal prominence (Figures 7.10 and 7.11, bearing in mind there are slight differences in the data used), with different rank ordering. The peak for THWTT emerges as most prominent, while that for the previously dominant ITLD does not stand out in any way, and other peaks, not previously noted and not all associated with musical/dance numbers attain some sort of parity.

That these differences emerge is a consequence of the transformation used and/or the treatment of outliers. The two can be related, but the difference should be borne in mind. The original loess smooths with family = "s" don't involve a transformation but do involve downweighting of outliers. Ranking is a transformation that removes the idea of an outlier from consideration. It can be thought of as robust, but does not need a justification on these grounds to be used. Log-transformation is a nice way of producing data with the same orders of magnitude, and this is often analytically useful. It preserves the rank order of the data, while retaining but downweighting information on absolute SL differences, often getting rid of what some perceive as ‘problems’ with outliers in the process.

It is pointless to claim that any of these methods is ‘best’. Where they do not show the same thing it is because the data are being treated differently. It is not ‘given’ that ‘outliers’ should be regarded as a problem, and may even be the case that you wish to emphasize them, which the loess (1/9, g, 2) seems to do quite nicely. Why ITLD stands out in some analyses is obvious from Figure 7.6: it is associated with 2/4 and 3/8 of the longest SLs. Remove these or seriously downweight them and the prominence of ITLD can be markedly diminished (as, arguably, is the way in which the film is being represented). That THWTT stands out but not exceptionally so in some analyses is because it has a reasonable number of reasonably long SLs but nothing extreme. The sequence in which the number occurs extends over more shots than other highlights, so it stands out more once the effect of longer SLs is diminished. The rank and log-transforms have potential appeal in representing films where the focus is directed to shorter SLs.

In short there is a choice, and nothing militates against trying more than one. A sensible strategy, if these sort of analyses are of interest, would be to look at a loess (1/9, g, 2) smooth, before and after log-transformation and a rank transformation. For publication these don’t always need to be used, but take an interest if there are marked differences. If writing for an audience that you think can cope with a moving average but not much else, see if it can be reproduced with a loess (1/9, g, 1) smooth or something similar (vary the span) and publish that, just calling it a smoothed moving average, omitting detail. It looks nicer.

### 7.8.3 Additional methodological comments

**Polynomial smoothing**

Relatively little space has been devoted to polynomial (trendline) smoothing (Section 7.2). There is nothing wrong with it, but for complex patterns choosing an appropriate order can take time and it may need to be quite high (Salt, 2010). Venables and Ripley (2002, p.228) note the use of polynomial smoothing methods, but also that there are ‘almost always better alternatives’, loess smoothing among them.

It is, though, worth making a general point here, which is that whatever method is used, it can be worthwhile examining smoothing at different scales. For example, in the analyses of *Intolerance* in (Section 7.2) first order linear smooths capture the general change in SLs over the course of the film, and fairly low order polynomials reveal the similarities and differences between the different stories without the confusion of too much detail.

That analyses at this kind of scale are of interest is evidenced in several places in the cinematics literature. Salt (2010), for example, begins by stating that his ‘belief had been that the cutting rate nearly always speeded up over the whole length of most ordinary commercial American features’ and that examples from a sample of 20 American feature films from 1959 contradicted this. In particular, the first degree trendline for his sample showed that nearly half of them had slower cutting in the second half of the shots compared to that in the first half of the shots. O’Brien (2005, p.91), in his study of French films during the period of conversion to sound,
found a ‘surprising’ difference for some films between ASLs for the first 30 minutes and ASLs for the rest of the film, that was in some cases ‘extreme’, with the direction of change ‘unpredictable’. Smoothing methods are ideal for examining changes at this sort of scale, and are less restrictive than the use of trendlines.

Enhancing MA plots based on ranked data

In Section 7.6.3 the application of moving averages to ranked data was illustrated. Such plots, as illustrated in Figure 7.10, can be enhanced by adding reference lines, based on the Mann-Whitney statistic, to identify ‘significant peaks and troughs. It is arguable, and I’d so argue, that this approach to interpretation of individual films may not be that helpful. If taking an ‘holistic’ view, ignoring statistical considerations and looking at the pattern as a whole, most reasonable observers would probably concede that there are about 10 fairly clear peaks and troughs, and it is the overall pattern that is of interest. The valleys are defined by the hills that surround them, and vice versa, and their depths and heights are of secondary interest. The picture may change with different window sizes, but that is a different issue.

The term ‘individual’ is emphasized above, since ‘cut-off’ rules of the kind shown may be useful for comparative purposes. They may be viewed as indentifying clusters of long and of short shots which can be plotted on suitably structured graphs to compare patterns between films. Redfern (2011, 2012b) illustrates this idea with application to BBC and ITV news bulletins.

Order structure matrices

That order structure matrices achieve anything not possible, and more comprehensibly, with other methods of display is something I remain to be convinced of. As quoted in Section 7.6.4 Redfern admits that the plot looks ‘complicated’ but asserts that it is ‘simply read’ given familiarity with the events of the film. Given such familiarity I suspect it would usually be possible to construct a moving average or loess smooth reflecting the pattern of events that most observers would find easier to read.

A disinterested observer has, in my hearing, referred to plots such as Figure 7.12 looking ‘like tartan’. This is fine if you like tartan, but the plot looks unnecessarily ‘busy’ to me, and is distracting. The bottom-to-top as opposed to left-to-right view of the plot is the same with a color reversal, so adds nothing to it. The remark that you ‘read the changes in cutting style from left to right’ would seem to acknowledge that the bottom-to-top view adds nothing.

The essential part of the plot, therefore, consists of vertical lines colored with varying degrees of black and white that reflect the rank of the SL (a sort of ‘pseudo-gray’ scale). This kind of thing can be achieved more directly, using other colors if needed, and other forms of coding, including the SLs themselves or categorised sets of SLs. Some of these alternatives are explored in the next chapter.

7.8.4 Discussion and conclusion

There isn’t really a definitive conclusion to be drawn. If you want to look at the internal structure of a film’s SLs several possibilities have been presented and others could be devised. A lot of them end up telling you the same thing; where they don’t it’s for a reason that may well be worth knowing about.

What one tries to do with the kinds of plot discussed is entirely up to the analyst. Tsivian (2005) explores the complexity of the structure of different stories within a film using polynomial models; Salt (2010) looks at the possibility of recapturing structure at the level of scenes using smoothed moving averages; Redfern (2012c) compares the editing structure of slasher films using order structure matrices. Another idea that has been explored, in essence, is to use plots of smoothed SLs to identify significant ‘change-points’ or ‘break-points’ in a film having a useful cinematic interpretation. Salt’s (2010) identification of structure at the level of ‘scenes’ can be
viewed in this light, though I suspect knowledge of the scenes is needed to select a smooth that identifies them.

More ambitiously, ‘global’ patterns across a body of films, based on statistical characterizations of their internal structure, have been investigated by James Cutting and his colleagues in a series of papers (e.g., Cutting et al., 2010; DeLong et al., 2012; and see Section 2.2.5). One of their investigations was of an ‘hypothesis’ of Thompson (1999) that many films have a four-act structure with acts of equal length. Their initial investigation (Cutting et al., 2011) can be thought of as a kind of ‘averaging’ of SL patterns after a quite complex ‘standardization’ to account for the fact that films have different numbers of SLs and markedly different ASLs. The original analysis found in favour of the four-act hypothesis but, after Salt pointed out that this was probably an artefact of the statistical manipulations involved, re-analysis led to a retraction of the original claim (Cutting et al., 2012). To be precise about what this means, the four-act ‘hypothesis’ is not ‘disproved’ as a consequence, it merely means that a demonstration of its validity based solely on statistical analysis of what Cutting et al. (2011) call the ‘physical properties’ of the films – SLs and transitions – has not been achieved.

Elsewhere, and in a similar kind of spirit, Adams et al. (2005), use rather complicated Bayesian statistical methods to investigate three-act structures in films. Their methodology is based on what they call a ‘tempo function’ based on aural as well as visual characterizations of shots, but could be applied to SL data only. I cannot claim to have examined their method in forensic detail. As far as I can tell they work with films where a three-act structure is assumed, and the focus is on identifying where the act boundaries lie. Prior assumptions – it is part of the Bayesian approach – are made about the region in which boundaries are likely to lie, and the methodology amounts to looking for evidence of what might be called change-points, equated with ‘dramatic climaxes’, in the smoothed tempo function in the regions where boundaries are assumed.

The main flaw in the original Cutting et al. (2011) paper seems to be that, unintentionally, the statistical manipulations employed impose upon the modified data precisely the ‘act-structure’ that the methodology is supposed, independently, to investigate without presumption. In Adams et al. (2005) a three-act structure is assumed at the outset, with the methodology tested on films where there are strong non-statistical reasons for believing they have such a structure. At best, the methodology identifies act boundaries with reasonable precision. At worst such identification is imprecise, which might indicate either that the presumed act-structure does not exist, or that the assumptions about what physical characteristics of a film define act-structure are wrong. The intent of the methodology, though, is neither to determine whether act-structure, revealed by physical characteristics of the film, exists at all, nor, if it does, how many acts are involved.

The papers just discussed are ambitious, and point in directions where the study of pattern in film – what I’ve termed global analysis based on internal structure – might go. None of the methodologies discussed are straightforward enough to invite enthusiastic emulation, even if their limitations can be dealt with. The merits of such exploration are exemplified in Cutting et al. (2010) who attempt to link the patterns they discern in their analyses to theories of human cognition. It is not necessary to agree with their claims to recognize the interest in pushing research in this kind of direction. Patterns often exist for a reason.

---

19I’ve refrained from comment on their headline grabbing analysis of $1/f$ spectral patterning – it made New Scientist – because I have yet to reproduce their analyses. The claim (Cutting et al., 2010) is that their study has ‘demonstrated that the shot structure in film has been evolving toward $1/f$ spectra’ with films ‘after 1980 generally approaching a $1/f$ profile’, matching patterns found in ‘the mind’, concluding that ‘we view $1/f$ film form as an emergent, self-organizing structure’. This conclusion is supported by a quadratic polynomial fit to results from their $1/f$ analysis in their Figure 2. If you fit a loess smooth rather than quadratic model (e.g., similar to our Figure 2.6) a different picture emerges that suggests ‘evolution’ between about 1970 and 1990, but thereafter is pretty flat. That is, ‘evolution’ is not occurring in the manner suggested, so the claims might usefully be re-evaluated.
Chapter 8

Enhanced plots

8.1 *Lights of New York* (1928) explored

In connection with the second plot for *Intolerance*, in Figure 7.2 of Section 7.2, discussion of its construction was deferred to here. Rather than plunging into the code, whose one page may look forbidding, the ideas involved will be introduced incrementally via a series of analyses for *Lights of New York*. This film is chosen simply because it was one of the first films I looked at when thinking about how shot-type information might be included in analyses of SLs (Baxter, 2012a).

It should be noted that if analyzing data using the *Cinemetrics* software with data submitted in advanced mode it is possible to both color-code the shots by type and overlay trendlines or moving averages where the degree of smoothing can be controlled. This is very much in the spirit of what is discussed in this chapter, though other methods of display, not readily available in *Cinemetrics*, are explored.

8.1.1 Color-coding shots

![Figure 8.1: Loess smooths for Lights of New York overlaid on color-coded SL data to the left (red = ‘action’, blue = ‘dialog’, green = ‘title’), and compared, without enhancement, to the left.](image)

Figure 8.1: *Loess smooths for Lights of New York overlaid on color-coded SL data to the left (red = ‘action’, blue = ‘dialog’, green = ‘title’), and compared, without enhancement, to the left.*
To the left of Figure 8.1 a plot of SLs against cut-point is shown; to the right two loess smooths are compared. I quite like the smoother (red, dashed) curve in the right panel which shows a nice reasonably rhythmic alternation of peaks and troughs over the first 35 minutes or so.

These plots are similar to ones already seen, except that shots are color-coded according to whether they are action (red), dialog (blue) or title (green) shots. The submission of the film to the Cinemetrics database, by Charles O’Brien, is in advanced mode with the ‘Type’ variable categorizing shots as ‘action’, ‘dialog’ and ‘exp.tit’. The data were imported to Excel, where column headers were edited as described in Section 3.2.2, then copied and imported into R with the name Lights_of_New_York (Section 3.2.1).

The R code for this and subsequent analyses can be built in a modular fashion. Initially the structures needed for plotting can be set up without actually doing any plotting. This is best done by creating a function that can be modified and added to as the need arises (Section 4.3). So begin with

```r
LONY.plot <- function() {
z <- Lights_of_New_York
SL <- z$SL/10 # SL in seconds
type <- z$Type # Shot type
minutes <- cumsum(SL)/60 # Cut-point in minutes
n <- length(SL) # Number of shots
order = c(1:n) # Shot sequence
SLrank <- rank(SL) # SL rank
#
# Extract and name SLs for shots of different types
# SLA <- SL[type == "action"]
SLD <- SL[type == "dialog"]
SLT <- SL[type == "exp.tit"]
#
# Do the same for cut-points
#
minA <- minutes[type == "action"]
minD <- minutes[type == "dialog"]
minT <- minutes[type == "exp.tit"]
#
# Loess smooths used for plots
#
LONYfit1 <- loess(SL ~ minutes, span = 1/9, family = "g", degree = 2)$fitted
LONYfit2 <- loess(SL ~ minutes, span = 1/6, family = "g", degree = 2)$fitted
}
```

This is intended to be reasonably self-explanatory. It won’t at this stage produce anything visible, unless there are obvious errors, which may show if you run it using LONY.plot(). The next bit of code, if added at the end of the function (before the final `}`), should produce the left-hand panel of Figure 8.1.

```r
plot(minutes, LONYfit1, type = "n", lwd = 2, col = "black", xlab = "minutes", ylab = "SL", main = "Lights of New York", ylim = c(0, max(SL)))
for(i in 1:n) lines(c(minA[i],minA[i]), c(0, SLA[i]), col = "red")
for(i in 1:n) lines(c(minD[i],minD[i]), c(0, SLD[i]), col = "blue")
for(i in 1:n) lines(c(minT[i],minT[i]), c(0, SLT[i]), col = "green")
lines(minutes, LONYfit1, type = "l", lwd = 3, col = "black")
legend("topright", "Loess(1/9, g, 2)", lty = 1, lwd = 3, col = "black", bty = "n")
```

In the `plot` function the `type = "n"` argument produces an empty plot, with labeling, that is a framework for the additions to follow which are, in order, vertical lines of different colors corresponding to the shot-types plotted at cut-points, and a loess smooth. The `lty` and `lwd` arguments control line type and width, and `col` selects the line color. An initial blank plot is needed since any attempt to plot at this stage will be over-plotted by what follows. The `ylim` argument is needed to ensure proper limits for plotting the SLs.
The bit where care is needed is in the `for` commands, where what is happening is that the beginning and end points of each line corresponding to an SL, and the appropriate color-coding, is being specified. Once added, get the plot using `LONY.plot()`.

The right-hand panel of the figure is obtained by adding the following to the function.

```r
plot(minutes, LONYfit1, type = "l", lwd = 3, col = "black", xlab = "minutes", ylab = "SL", main = "Lights of New York", ylim = c(0, max(LONYfit1)))
lines(minutes, LONYfit2, type = "l", lty = 2, lwd = 4, col = "red")
legend("topleft", c("Loess(1/9, g, 2)"), lty = c(1, 2), lwd = c(3, 4), col = c("black", "red"), bty = "n")
```

This kind of thing has already been illustrated in Section 7.5. As described so far, running `LONY.plot()` will only show the plot produced by the final block of code added. To show all plots insert `win.graph()` at the start of each block. Typing `graphics.off()` when the R prompt `>` is showing will clear the screen of graphical clutter; it can usefully be inserted at the start of the function, if it is to be run repeatedly.

### 8.1.2 Adding ‘wallpaper’

Once the idea of adding color-coding, and the means of doing it, is grasped, experimentation with a variety of different effects is possible. Adding ‘wallpaper’, as in Figure 8.2 is one possibility.

![Figure 8.2: Loess (1/9, g, 2) smooths for Lights of New York with color-coded ‘wallpaper’ (red = ‘action’, blue = ‘dialog’, green = ‘titles’). To the left untransformed SL data is used; to the right action shots have been coded 0 and other shots 1.](image)

The code

```r
plot(m, LONYfit1, type = "n", xlab = "minutes", ylab = "SL", main = "Lights of New York", ylim = c(0, max(LONYfit1)))
for(i in 1:length(minA)) abline(v = minA[i], lwd = 1.5, lty = 1, col = "red")
for(i in 1:length(minD)) abline(v = minD[i], lwd = 1.5, lty = 1, col = "skyblue")
for(i in 1:length(minT)) abline(v = minT[i], lwd = 1.5, lty = 1, col = "green")
lines(m, LONYfit1, lwd = 4)
legend("topleft", "Loess(1/9, g, 2)", lty = 1, lwd = 4, col = "black", bty = "c", bg = "white")
```
produces the left-hand panel of the figure. This is actually a bit simpler than adding the color-coding in Figure 8.1, since the vertical lines associated with the SLs don’t need to be bounded in any way.

The right-hand plot introduces what I suggested in Baxter (2012b) might be a novel idea; it has not, in any case, previously been discussed in these notes. If the type categories admit dichotomization in any way then the dichotomy can be represented by a 0-1 coding that can then be treated like any other numerical variable. In the code that follows SLtype codes action shots as 0, and dialog and titles as 1. There are relatively few titles so this will not affect the analysis much, even if one has doubts about amalgamating titles with dialog. The loess fit LONYfit3 is just the smoothing algorithm applied to these data with span = 1/6 chosen after a little experimentation. Other than that, what’s done is identical to the previous block of code and the wallpaper is the same.

```r
SLtype <- ifelse(type == "action", 0, 1)
LONYfit3 <- loess(SLtype ~ minutes, span = 1/6, family = "g", degree = 2)$fitted
plot(minutes, LONYfit3, type = "n", xlab = "minutes", ylab = "Coded SL",
main = "Lights of New York - (0,1) coding", ylim = c(0,1))
for(i in 1:length(minA)) abline(v = minA[i], lwd = 1.5, lty = 1, col = "red")
for(i in 1:length(minD)) abline(v = minD[i], lwd = 1.5, lty = 1, col = "skyblue")
for(i in 1:length(minT)) abline(v = minT[i], lwd = 1.5, lty = 1, col = "green")
lines(minutes, LONYfit3, lwd = 4)
legend("topleft", "Loess(1/6, g, 2)", lty = 1, lwd = 4, col = "black", bty = "o", bg = "white")
```

Some care is needed in interpreting these plots. The wallpaper codes for shot-type; action/dialog, if the relatively few titles are ignored. The loess curve in the left-hand panel is based on SLs so there is not a necessary relationship between any pattern it shows and that of the background, though one is expected if SLs and shot-type are associated. For Lights of New York (and other films) there is an expectation that ‘action’ shots will tend to be shorter than ‘dialog’ shots, confirmed by the left-hand panel of Figure 8.1, and this is reflected in a reasonable match between the patterns exhibited in the two displays embodied in the left-hand panel.

I’m not sure if mountain goats really do jump from peak-to-peak, but you can imagine one doing so in the right-hand panel, from the left, moving slowly but steadily down to lower altitudes. Here, because of the coding, the loess smooth directly models the variation between action and dialog. The peaks correspond, as it were, to bursts of inactivity, and things get more exciting as the goat progresses down the range. This is to be hoped for, since the two representations of pattern are designed to show the same thing and should be concordant. This they reasonably well are. The level of smoothing has been chosen, for the convenience of the goat, to be not too rocky – goats are only interested in the heights.

### 8.1.3 Order structure matrices revisited

Order structure matrices were discussed in Section 7.6.4. In Section 7.8.3 the opinion was expressed that these were over-complicated and might be improved upon. The gist of the argument was that the plots had redundant aspects that generated an appearance, unpleasing to some eyes, that might be modified to more effectively display what was intended of them. As a reminder, an order structure matrix plot for Lights of New York is shown to the left of Figure 8.3.

As expounded by Redfern (Redfern, 2012c; and elsewhere) changes in cutting style are read from left to right. This means that the horizontal axis, which repeats the same information with a color reversal, is redundant and (in my view) complicates the picture. As the vertical lines that are being ‘read’ can be viewed as on a ‘pseudo-gray’ scale, the use of a proper gray-scale, ‘discarding’ the horizontal axis, seems simpler. This is illustrated in the right-hand panel of Figure 8.3.

The plot tells exactly the same story as the order structure matrix plot but, to my eye, is easier to read and nicer to look at. It is easier to explain (shots are color-coded according to the

---

Figure 8.3: An order structure matrix plot Lights of New York to the left, and an alternative representation to the right with a superimposed loess (1/9, g, 2) smooth based on ranked data.

rank of their SLs) and it is easier (I think) to overlay plots, such as the loess smooth shown, that can aid interpretation\(^2\).

Code for getting the order structure matrix plot was given in Section 7.6.4, but is repeated below for the sake of completeness.

```r
mat <- matrix(0, n, n)
for(i in 1:n) {
  for(j in 1:n) {
    mat[i,j] <- ifelse(SL[i] >= SL[j], 1, 0)
  }
}
image(1:n, 1:n, mat, col = c("white", "black"), xlab = "shot number", ylab = "shot number")
```

The code for the right-hand panel of Figure 8.3 is as follows.

```r
LONYfitrank <- loess(SLrank ~ order, span = 1/9, family = Family, degree = Degree)$fitted
test.color <- gray((n - order)/n)
plot(order, LONYfitrank, type = "n", xlab = "shot number", ylab = "SL rank",
    main = "Lights of New York", ylim = c(100,260))
for(i in 1:n) abline(v = i, col = test.color[SLrank[i]], lwd = 1)
lines(order, LONYfitrank, type = "l", lwd = 5, col = "red")
```

Here, test.color sets up the gray-scale color palette to be used. It is defined in the way it is to mimic the same coloring as in the order structure matrix plot. The ‘reverse’ of the black/white coding used is obtained with the simpler test.color <- gray((order)/n). The scale has to have limits of 0 and 1. Other palettes are available and discussed a little further in Section 8.3. Some of these are a little simpler to define but, in some cases, less useful.

Where loess smooths have been shown in the above figures the general message is the same, but the precise picture depends, as discussed in the previous chapter, on both the degree of

\(^2\)To be clear about this, I have no problem with the idea behind using an order structure matrix plot, but they are cumbersome and the idea can be more easily implemented as shown. Adding loess smooths to the modified plot is a useful aid to selecting the degree of smoothing that captures the main patterns of variation in the data.
smoothing and methods of data transformation adopted. Illustrated in the figures are smooths of the untransformed SLs, ranked SLs and 0-1 coded SLs. The last of these attempts something different from the other two, but should produce results with a similar interpretation if SLs are strongly related to shot-type. It is, I think, possibly a matter of taste and choice of emphasis as to which kind of plot is to be preferred when confining attention to the analysis of a single film.

8.2 **Intolerance revisited**

It would be remiss not to include the code for the right-hand panel of Figure 7.2 promised in Section 7.2. Here it is.

```r
Intolerance.plot <- function() {
  #
  z <- Intolerance
  SL <- z$SL/10
  minutes <- cumsum(SL)/60  # cut-point in minutes
  ASL <- mean(SL)
  Type <- z$Type
  n <- length(SL)
  #
  # extract SLs and cut-points for individual stories
  #
  B <- SL[Type == "Babylon"]
  F <- SL[Type == "French"]
  J <- SL[Type == "Judean"]
  M <- SL[Type == "Modern"]
  Babylon <- minutes[Type == "Babylon"]
  French <- minutes[Type == "French"]
  Judean <- minutes[Type == "Judean"]
  Modern <- minutes[Type == "Modern"]
  #
  fit1 <- lm(SL ~ poly(minutes, 1))  # straight-line fit
  fit5 <- lm(SL ~ poly(minutes, 5))  # 5th order polynomial fit
  Ylim <- c(9,0)  # Limits for y-axis, reversed as in Tsivian (2005)
  pred <- 1:max(minutes)  # plotting positions (x-axis) for fits
  #
  plot(pred, predict(fit5, data.frame(minutes = pred)), type = "n", ylim = Ylim,
       xlab = "minutes", ylab = "SL", main = "Intolerance")
  for(i in 1:length(Babylon)) abline(v = Babylon[i], lwd = 1.5, lty = 1, col = "green")
  for(i in 1:length(French)) abline(v = French[i], lwd = 1.5, lty = 1, col = "skyblue")
  for(i in 1:length(Judean)) abline(v = Judean[i], lwd = 1.5, lty = 1, col = "red")
  for(i in 1:length(Modern)) abline(v = Modern[i], lwd = 1.5, lty = 1, col = "orange")
  lines(pred, predict(fit5, data.frame(minutes = pred)), lwd = 3)
  lines(pred, predict(fit1, data.frame(minutes = pred)), lty = 2, lwd = 3)
  abline(h = ASL, lty = 3, lwd = 4)  # adds line at SL
  #
  legend("topleft", legend = c("Babylonian", "French", "Judean", "Modern"), lty = c(1,1,1,1),
          lwd = c(2,2,2,2), col = c("green", "skyblue", "red", "orange"), title = "Story", bg = "white")
  legend("topright", legend = c("ASL", "Linear trend", "5th order polynomial"), lty = c(3,2,1),
          lwd = c(4,3,3), bg = "white")
}
```

It is built up in the same way as the code for *Lights of New York*. There are seven categories in the ‘Type’ variable, the majority of which identify which of the four stories that comprise *Intolerance* a shot is located in. Of the 1989 shots 56 are in three different ‘other’ categories, and were not used explicitly in constructing the figure. It is assumed that columns have headers labelled as discussed in Section 3.2.2 and that the file is imported into R with the name *Intolerance*. 84
8.3 Jeu d’esprit

In constructing plots like those in Figure 8.2 it is possible to be distracted by the fact that, stripped of labeling and the loess smooths, the picture looks something like a production of ‘modern’ art. The temptation to seek out matching Bridget Riley’s (it can be done reasonably well) or design your own wallpaper or Bridget Riley needs to be resisted. The purpose of constructing figures like those illustrated is, after all, to convey a message, and I’m not sure how much modern art aspires to this.

Similar temptations, and pitfalls, exist in constructing graphics like that in the right-hand panel of Figure 8.3. The free choice of color available in Figure 7.2 and 8.2 (where a discrete and unordered set of categories is represented) is constrained by the need to have a gradation of colors that reflects the coding (if SLs are ranked) from shortest to longest shot. A gray-scale seems sensible, but you get bored.

What follows are disjointed ideas that occurred at times when my concentration on the business to hand wavered a little. R code for everything that follows is discussed in Section 8.3.3.

8.3.1 Halloween (1978) – matching color to content

In looking at Redfern’s (2012c) analyses of slasher films, using order structure matrices, and seeing how the alternative representation explored in Figure 8.3 worked, I wondered (idly) whether a color scheme more suited to the films’ content might be devised. It turns out it can, and with surprising ease, as the left panel of Figure 8.4, for Halloween (1978), shows. Compared to the earlier analysis of Lights of New York the only change is to use the `heat.colors` rather than `gray` palette, which produces, without any tweaking, the satisfactory splatter of bright red towards the end.

Figure 8.4: Heat color representations of SL data for Halloween. The same number of shots, with the same coloring are used, but plotted against shot number to the left and cut-point to the right. The scales on the upper axes are minutes and shot number respectively. Loess (1/9, g, 2) smooths of the ranked SLs are shown.

A loess (1/9, g, 2) smooth to the ranked SLs is shown. This can obviously be modified if wished, but corresponds well enough with the wallpaper. The y-axis has been reversed so that peaks correspond to concentrations of shots with short SLs.

In Redfern (2010c) the visual representation is ‘validated’ by a reading of the film based on a close knowledge of it, overlaid with an interpretation of the way different passages in the film

---

3See Redfern’s research blog, http://nickredfern.wordpress.com/tag/horror-films/, for links to the data.
function. I suspect most interpretations of these and other forms of graphical representation are undertaken in a similar way; more interestingly, for present purposes, are attempts to abstract from individual patterning and recognize common structure across a range of films – of similar genre, or whatever comparison seems interesting. Redfern, for example, concludes on the basis of analysis of four slasher films that ‘the genre is characterized by a common editing pattern to such an extent that the films that followed Halloween may be described as stylistically formulaic’.

My usual approach when meeting a new methodology, or one new to me, is to prod it a little to see how it reacts, and how claims for it stand up. In trying to compare Halloween with other films some problems, or at least variants of the way it might be displayed, became apparent.

One issue is that the visual impact of a graphic can depend on the number of shots a film has. Films with a relatively small number of shots can look ‘washed-out’ by comparison with films with a significantly larger number when viewed on a large screen. Ways of getting round this are discussed shortly.

A second issue is revealed by looking at the scale at the top of the graph; this is in minutes and not linearly related to the lower scale of shot order. Thus, the dominant impression, the splodge of red at the right, is produced by over a quarter of the shots in the film, but only accounts for about a sixth of the running time. There is one shot in Halloween of over four minutes, occurring early, and five others of over one minute between shots 155 and 342. It is possible to pick out disjoint sets of 20 shots or more in the final sixth of the film (in terms of time) that occupy less than a minute. In the visual representation of Halloween in Figure 8.4 these blocks of shots receive over 20 times the weight of individual long shots that occupy a similar or longer period of time.

That is to say, the left-hand plot in the figure, and the related order structure matrix plot, privileges short shots. This is doubtless the intent; the fact that there is a concentrated flurry of short shots towards the end is emphasized.

The order structure matrix approach is promoted as a robust method; it achieves robustness, in the context of the analysis of SLs, by ignoring all the information contained in the measurements except that one shot is longer or shorter than another and occurs earlier or later. In graphical terms, and in effect, this ‘robustness’ manifests itself in two ways. The first is by using graded color-coding of shots that reflects their rank ordering only; the second is by plotting the lines that represent a shot against their equi-spaced order in the shot sequence, ignoring their length and hence the time of cut-points between the shots.

With the simpler form of plotting advocated here the constraint of plotting against shot order is removed. It is straightforward to retain the robustness inherent in the color-coding, while reintroducing other SL information by plotting against cut-points. The right panel of Figure 8.4 is the result. If the left-hand panel privileges the representation of the shorter shots, associated with what Redfern terms ‘the frenzied violence of body horror’, then the right-hand panel privileges longer shots ‘creating a pervading sense of foreboding’.

I will leave it to the reader to decide if these ways of displaying the data have any merit and, if they think they do, which of the two ways of displaying the data is to be preferred, if any and if a choice has to be made. What I’m doing here is evading any discussion of any ‘meaning’ the graphs might have beyond the statistical display of pattern in the data in two different ways. This is because I suspect it might be a difficult question I can’t answer. The next section concentrates on an interesting technical problem.

8.3.2 Controlling for shot numbers - an idea

This, touched on briefly above, is how you might go about comparing patterns in films with different numbers of shots, beyond the qualitative comparison of graphs of the kind illustrated in

---

4My original, unworthy, thought was to see if I could find films like Bambi, or others with seemingly innocuous titles, that, judged ‘cold’, had editing structures indistinguishable from that of Halloween. The idea was abandoned, temporarily anyway, at an early stage.

5This is obvious with screen output from R if the screen is large enough. The effect is hidden a bit on reducing graphs to the size needed for a normal-sized page - because of the compression involved - so is not attempted here.

6In the context of other genres these might be what I think I’ve seen referred to as contemplative passages, romantic interludes or, as the attention-deficit might have it, the boring bits.
this and previous chapters. The question is of more than passing interest. Arguably this difficulty
was one of the rocks on which the attempt by Cutting et al. (2011, 2012) to demonstrate the
generality of four-act structure in films foundered. One of their basic ideas, converting actual SLs
into an equal number of what they call ‘adjusted’ SLs for each film, is sound. The manner in
which this was achieved, however, introduced artifacts into the adjusted data that meant that the
subsequent detection of four-act structure was a self-fulfilling prophecy.

What follows presents an idea of how ‘adjusted’ SLs might be calculated, to then be put to
whatever purpose one devises them for. The idea has a simple core, which is to fit a KDE to the
data and then design the wallpaper using the KDE (which can be thought of as estimating SLs)
rather than the SLs. It is the detail concealed by this idea that is important, so it will be taken
in stages.

![Figure 8.5: Heat color representations of SL data for Halloween. The KDEs are the same, with
different scaling, in both plots. Wallpaper to the left is based on the SL ranks; wallpaper to the
right is based on the KDE – see the text for details of the latter.](image)

The left-hand panel of Figure 8.5 shows the default KDE for the cut-points of Halloween
overlaying wallpaper defined by the ranks of the SLs (see Section 7.7 for a discussion of this usage
of KDEs). There are several problems with the plot. The KDE is over-smoothing and, as is its
wont, strays well into areas beyond the limits of the film (negative time, and after you’ve stooped
watching). This also has the effect of compressing the wallpaper a bit, since this is confined to
the real duration of the film.

Leaving aside, momentarily, the over-smoothing issue, it is possible to do something about
other aspects of its appearance. Left to its own devices the `density` command in R computes
density estimates at 512 equally spaced points with limits defined by an internal algorithm. For
the purposes of plotting, both the limits and the number of plotting points can be controlled, so
in the right panel of Figure 8.5 the KDE is ‘forced’ to lie between the beginning and end of the
film and 1024 plotting positions are used\(^7\).

The KDE is just a magnified portion of that appearing to the left; the important difference is
that the wallpaper is designed using the heights of the KDE at each of the 1024 plotting positions.
Since the KDE smooths the data the wallpaper pattern, in contrast to that to the left where
ranked SL coloring is used, is also smoothed and more ‘solid’ (a function, also, of the greater
number of plotting positions). The index on the x-axis is the number of the plotting position; it

\(^7\)For algorithmic reasons the number of points needs to be a power of 2, so 512 = 2\(^9\); the choice of 1024 = 2\(^{10}\)
was a bit arbitrary, I don’t think it makes much difference.
could be converted to time but is left as it is. For later reference it means that the 512th plotting position corresponds to the middle of the film; the 256th position to the end of the first quarter, and so on.

All this, coupled with the over-smoothing, simplifies the picture of the film, too much so. Interpretation is sensible enough; a slow beginning; a touch of ‘frenzied violence’ to let you know you’re in a slasher movie; a long lull in which to anticipate what you’ve come to see; which is the prolonged outburst of frenzied violence at the end. This, of course, misses out on the subtleties of construction, where the calm is punctuated by bits of action to keep you on your toes. The obvious thing to do is reduce the smoothing by reducing the bandwidth from its default of 6.8. This is done in the left-hand panel of Figure 8.6.

The selected bandwidth is 1 minute. You can get close to reproducing something like wallpaper based on the actual SLs by reducing this to 1 second (bearing in mind an estimate with, in this case, rather more points than SLs is being used) but this loses the advantages of simplification. The picture is quite pleasing; before the leap into red violence at the end, and as these things go, there is a reasonably regular bounce both in terms of spacing and the heights/depths achieved (‘as these things go’ should be emphasized).

Rather than filling in Figure 8.6 with yet another graph for Halloween with a different bandwidth, the right-hand panel of Figure 8.6 is for Lights of New York with the same bandwidth of 1 minute. It was, in fact, initial problems encountered in effecting a comparison between the two films that motivated some of the above development. Lights of New York was made 50 years before the official advent of the slasher genre, so presumably can’t be admitted to the slasher canon. Readers can supply their own commentary on what aspects of the graphics alert you to this.

It is worth standing back a little to reflect on what has been done above. The story began with the stringently robust order structure matrix representation, replaced by an analogous but simpler plot of color-coded ranked SLs plotted against shot order. This led on to the idea of plotting against cut-point, thereby introducing the lengths of shots back into the fold, and thence to the earlier idea of using KDEs to smooth the data. Loess smooths could be used, but KDEs

---

8In conducting experiments on this what was at the forefront of my mind was what might be a suitable choice of bandwidth that could be used for comparing a body of films, so you need to lose detail if you want to compare broader underlying structure while maintaining a similar level of simplification. Or at least that’s the current idea.
have the advantage (in the R implementation) of allowing greater control over the number and location of plotting positions. Finally, the idea of color-coding by SL ranks was abandoned, to be replaced by color-coding of the estimated values of the KDEs at the plotting positions.

Two things have happened here. One is that the original idea of a robust representation of the data has been almost abandoned. The one vestige that remains is that KDEs might be regarded as robust in that they smooth out the effects of the more extreme SLs. That they are, in effect, estimates of SLs that downplay extremes incidentally avoids representational issues that would arise if actual SLs were color-coded. The second thing that has happened is that the end result is two equivalent representations of the data, the KDE of the cut-points, and the KDE at each plotting point coded by color. Either or both could be used, depending on the taste of the analyst and what they think best conveys the message they have to impart. These or any other graphical simplifications can be checked against the raw data (e.g., the left-hand panel of Figure 8.1).

It is then also legitimate to ask why bother, if one goes down the above route and opts for the KDE, with all the manipulation needed to get the color. An answer to this is attempted in Section 8.3.4.

8.3.3 Code

In the code to follow labeling and style arguments, including the code needed for the upper axes in Figure 8.4, have been omitted.

The left-hand plot of Figure 8.4 is obtained from

```r
SL <- film1
minutes <- cumsum(SL)/60
SLrank <- rank(SL)
n <- length(SL)
order <- 1:n
fitrank <- loess(SLrank ~ order, span = 1/9)$fitted
ylim = c(max(fitrank), min(fitrank))
test.color <- heat.colors(n)
plot(order, fitrank, type = "n", ylim = ylim)
for(i in 1:n) abline(v = i, col = test.color[SLrank[i]])
lines(order, fitrank)
```

The `test.color <- heat.colors(n)` command selects the heat color palette. I used this because it was convenient for what I wanted to do. A little bit of Googling will bring up a lot of possibilities I haven’t explored. For the right-hand plot, replace `order` with `minutes` in the `fitrank` definition and `plot` command; replace `v = i` in `abline` with `v = minutes[i].`

For Figure 8.5 the left-hand plot is produced by

```r
D <- density(minutes)
plot(D, type = "n")
for(i in 1:n) abline(v = minutes[i], col = test.colorx[SLrank[i]])
lines(D)
```

The right-hand plot needs a little more effort.

```r
D <- density(minutes, from = 0, to = max(SLcut), bw = "nrd0", n = 1024)
x <- D$x
y <- D$y
nx <- length(x)
orderx <- 1:nx
test.colorx <- heat.colors(nx)
SLrank <- rank(-y)
plot(orderx, y, type = "n")
for(i in 1:nx) abline(v = i, col = test.colorx[SLrank[i]])
lines(orderx, y)
```

9See `?heat.colors` for this and other options
The density estimate, \( D \), specifies the bandwidth, via \( bw \) where "nrd0" is the default, and the number of plotting points via \( n = 1024 \) (where 512 would be the default). The variables defined as \( x \) and \( y \) extract the plotting positions and the heights. The next four commands re-define variables needed for the plots, \( SLrank \) being defined as it is to ensure reds at the short end of the SL scale. For analyses with any other bandwidth simply substitute for "nrd0", so \( bw = 1 \) for the right-hand plot of Figure 8.6.

### 8.3.4 Coda

Before discussing how some of the development in Section 8.3.2 might be useful, some further aspects of the kind of plots introduced there are considered. Figure 8.7 compares plots of different kinds for *The House on Sorority Row* (1983) and *Pinocchio* (1940).

![Color-coded plots based on the KDEs for *The House on Sorority Row* and *Pinocchio* (1940), based on a bandwidth of 1, and gray-scale ranked data with superimposed loess (1/9, g, 2) smooths.](image)

Figure 8.7: *Color-coded plots based on the KDEs for The House on Sorority Row and Pinocchio (1940), based on a bandwidth of 1, and gray-scale ranked data with superimposed loess (1/9, g, 2) smooths.*

*The House on Sorority Row* is an accredited slasher film; *Pinocchio* (1940) is not usually admitted to the canon. Google brings up an interesting summary of a topic noting ‘Disney’s 1940 *Pinocchio*, and ... *House on Sorority Row*, the similarly themed and plotted 1983 flick’ but, alas, the ellipsis conceals the fact that mention of the two films is separated by an intervening distance in the full text. The films are rarely mentioned in the same breath.
Looking at the upper plots in Figure 8.7, and ignoring the KDEs for a moment, that the two film have distinctly different structures, one characteristic of a slasher film, is not something that screams out at me. If anything, *Pinocchio* looks bloodier in its later stages, but as this perception is affected by the choice of color, this observation can be dismissed. The more ‘rhythmic’ build-up in the *The House on Sorority Row* to the ending, between (about) plotting positions 400 and 900 may indicate an intentionally structured difference, and something of the sort was also evident for *Halloween*. There are, however and emotive coloring apart, other difficulties with this kind of plot. One is that the smoothing based on the KDE may either induce or conceal perceptions of structure in the data; another is that the nature of the gradation in color can affect perceptions of emphasis given to different portions of the graph.

The gray-scale coding of ranked SL data in the lower plots is also not without problems. Unaided by the plotted smooth, a loess fit in this case, the ‘rhythmic’ temporal pattern in *The House on Sorority Row* just alluded to is no longer so evident.

![Pinocchio graph](image)

![The House on Sorority Row graph](image)

**Figure 8.8:** Gray-scale plots for the ranked data of *Pinocchio* and *The House on Sorority Row*, with a different aspect ratio from that in Figure 8.7.

The generally ‘darker’ nature of the second half of the plot for *The House on Sorority Row*, with the, perhaps, more prominent darker banding is at least partly a perceptual artifact of the aspect ratio chosen for the figure. Stretch out the plots, as should maybe have been done in the first place, now remedied in Figure 8.8, and I, at least, find it hard to see major differences in the
structures.

This is by way of concluding that, detached from other methods of representation, graphics based on graded color-coding of shots have their limitations for the purposes of comparison. This applies to order structure matrix plots, color-coded ranked data (essentially the same thing) or color-coded KDEs. ‘Time’ is represented by its rank in the first kind of plot; by cut-points in KDE plots; and either can be used for the second kind of plot. SLs are coded by their rank in the first two kind of plots, and by estimates derived from KDEs in the final kind. These last might be thought of as ‘adjusted’ SLs, though defined differently from the procedure used in Cutting et al. (2011).

I find smoothed estimates, which can be selected with the aid of color-coded plots, more satisfactory for visual comparison. Looking at those in Figure 8.7 suggests that for any particular film much the same pattern is suggested.

The reason for embarking on the development leading to the idea of color-coded KDEs was, though, motivated more by an interest in numerical than graphical comparison. The patterns exhibited by the overlaid smooths in the upper and lower plots of Figure 8.7 are similar enough to suggest that if this mode of summary is preferred the choice of analysis using either cut-points or ranks, and hence KDEs or loess smooths, is not critical (see, also, Section 7.7). KDEs of cut-points are used in what follows.

![Image of two KDE plots](image-url)

Figure 8.9: KDEs for the cut-points of several films, standardised to eliminate differences in length. Bandwidths of 2 and 4 are used for the left and right plots; vertical lines divide the films into (temporal) quarters.

The KDEs of *The House on Sorority Row* and *Pinocchio* are compared in Figure 8.9 using bandwidths of 2 and 4. *Top Hat* (1935) and three other slasher films, *Halloween* (1978), *Friday the 13th* and *The Slumber Party Massacre* (1982), are thrown in for good measure\(^\text{10}\).

The method of construction, based on 1024 plotting positions running from the beginning to end of the film, ensures that a direct visual comparison unaffected by the differing lengths of the films is legitimate, but is unavoidably affected by the level of smoothing chosen. To my eye *Top Hat* looks to have distinctive qualities (more ‘rhythmic’ and ‘peakier’), while three of the four slasher films, *The House on Sorority Row* being the exception, are distinguished by the more intense ‘action’ – equating this with shorter shots – in the final quarter or so.

\(^{10}\)The slasher films are those analyzed in Redfern (2012c) and the data were obtained via his research blog. A common editing pattern is claimed for all four films though some of the singularities of *The House on Sorority Row*, whose KDE looks different from the other three, are noted.
The quarters were inserted with the ideas of Cutting et al. (2011) about ‘act structure’ in films in mind. The method of deriving ‘adjusted’ SLs is possibly simpler than that of Cutting et al. without the flaws subsequently admitted in their adjustment procedure (Cutting et al., 2012), but depends on the level of smoothing, and there is no obvious ‘right’ way of doing this. With the eye of faith, and ignoring Top Hat, one might discern in the smoother plot of Figure 8.9, a pattern where over the first quarter or so ‘action’ rises to a peak, eases off and flattens out for a bit, then begins to rise again about two-thirds to three-quarters of the way through, before declining at the end. The less smooth plot shows considerable variation within these broad divisions.

The main point to make is that the KDE for a film is a function, quantified by measurements at (in this instance) 1024 equi-spaced points. There is a body of statistical methodology, functional data analysis (http://www.psych.mcgill.ca/misc/fda/ - accessed 01/11/2012), directed at the analysis of such data that might be useful if more ‘targeted hypotheses’ than anything attempted here can be formulated, that would benefit from a quantitative comparison that goes beyond the graphical.

When I originally wrote the above I said that I could see considerable problems in implementing this kind of idea. One thought was that given two functions quantified as described, it is possible to measure the ‘distance’ between them and – in principle – given data for a body of films to identify clusters of films with similar profiles. Since writing this attempts have been made in this direction in Baxter (2013c,d) and Redfern (2013b) and a summary of these with illustrative applications is provided in Chapters 13 and 14, absent from the earlier version of these notes.
Chapter 9

Shot-scale data analysis

9.1 Correspondence analysis – initial example

In Section 2.2.6 the use of correspondence analysis for investigating patterns in shot-scale data was introduced (see, also, Baxter, 2012c). A more detailed exposition of the methodology, and its implementation in R is provided here, using 42 films of Alfred Hitchcock in Table 9.1 as a running example.

<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>BCU</th>
<th>CU</th>
<th>MCU</th>
<th>MS</th>
<th>MLS</th>
<th>LS</th>
<th>VLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleasure Garden, The</td>
<td>1925</td>
<td>13</td>
<td>47</td>
<td>96</td>
<td>98</td>
<td>104</td>
<td>153</td>
<td>18</td>
</tr>
<tr>
<td>Downhill</td>
<td>1927</td>
<td>36</td>
<td>25</td>
<td>89</td>
<td>93</td>
<td>98</td>
<td>154</td>
<td>5</td>
</tr>
<tr>
<td>Lodger, The</td>
<td>1927</td>
<td>29</td>
<td>94</td>
<td>74</td>
<td>78</td>
<td>106</td>
<td>119</td>
<td>0</td>
</tr>
<tr>
<td>Champagne</td>
<td>1928</td>
<td>37</td>
<td>65</td>
<td>78</td>
<td>90</td>
<td>116</td>
<td>108</td>
<td>5</td>
</tr>
<tr>
<td>Easy Virtue</td>
<td>1928</td>
<td>25</td>
<td>61</td>
<td>96</td>
<td>65</td>
<td>97</td>
<td>140</td>
<td>14</td>
</tr>
<tr>
<td>Farmer’s Wife, The</td>
<td>1928</td>
<td>10</td>
<td>42</td>
<td>97</td>
<td>86</td>
<td>115</td>
<td>145</td>
<td>6</td>
</tr>
<tr>
<td>Manxman, The</td>
<td>1928</td>
<td>21</td>
<td>79</td>
<td>66</td>
<td>77</td>
<td>117</td>
<td>123</td>
<td>16</td>
</tr>
<tr>
<td>Ring, The</td>
<td>1928</td>
<td>41</td>
<td>65</td>
<td>90</td>
<td>89</td>
<td>98</td>
<td>112</td>
<td>6</td>
</tr>
<tr>
<td>Blackmail (Silent version)</td>
<td>1929</td>
<td>39</td>
<td>49</td>
<td>67</td>
<td>85</td>
<td>137</td>
<td>113</td>
<td>10</td>
</tr>
<tr>
<td>Blackmail (Sound version)</td>
<td>1929</td>
<td>42</td>
<td>41</td>
<td>87</td>
<td>74</td>
<td>112</td>
<td>128</td>
<td>16</td>
</tr>
<tr>
<td>Juno and the Paycock</td>
<td>1930</td>
<td>9</td>
<td>30</td>
<td>53</td>
<td>84</td>
<td>164</td>
<td>159</td>
<td>0</td>
</tr>
<tr>
<td>Murder</td>
<td>1931</td>
<td>33</td>
<td>69</td>
<td>89</td>
<td>74</td>
<td>92</td>
<td>138</td>
<td>5</td>
</tr>
<tr>
<td>Skin Game, The</td>
<td>1931</td>
<td>28</td>
<td>78</td>
<td>50</td>
<td>60</td>
<td>136</td>
<td>125</td>
<td>22</td>
</tr>
<tr>
<td>Number Seventeen</td>
<td>1932</td>
<td>42</td>
<td>58</td>
<td>51</td>
<td>90</td>
<td>103</td>
<td>126</td>
<td>28</td>
</tr>
<tr>
<td>Rich and Strange</td>
<td>1932</td>
<td>24</td>
<td>75</td>
<td>75</td>
<td>75</td>
<td>88</td>
<td>131</td>
<td>33</td>
</tr>
<tr>
<td>Waltzes from Vienna</td>
<td>1934</td>
<td>19</td>
<td>49</td>
<td>42</td>
<td>79</td>
<td>126</td>
<td>186</td>
<td>0</td>
</tr>
<tr>
<td>Thirty-Nine Steps, The</td>
<td>1935</td>
<td>27</td>
<td>72</td>
<td>80</td>
<td>87</td>
<td>75</td>
<td>132</td>
<td>26</td>
</tr>
<tr>
<td>Secret Agent, The</td>
<td>1936</td>
<td>43</td>
<td>94</td>
<td>105</td>
<td>93</td>
<td>79</td>
<td>73</td>
<td>14</td>
</tr>
<tr>
<td>Sabotage</td>
<td>1937</td>
<td>63</td>
<td>87</td>
<td>104</td>
<td>96</td>
<td>64</td>
<td>79</td>
<td>8</td>
</tr>
<tr>
<td>Young and Innocent</td>
<td>1938</td>
<td>27</td>
<td>96</td>
<td>124</td>
<td>75</td>
<td>67</td>
<td>95</td>
<td>16</td>
</tr>
<tr>
<td>Jamaica Inn</td>
<td>1939</td>
<td>15</td>
<td>85</td>
<td>71</td>
<td>92</td>
<td>88</td>
<td>135</td>
<td>18</td>
</tr>
<tr>
<td>Lady Vanishes, The</td>
<td>1939</td>
<td>24</td>
<td>72</td>
<td>117</td>
<td>114</td>
<td>103</td>
<td>59</td>
<td>11</td>
</tr>
<tr>
<td>Foreign Correspondent</td>
<td>1940</td>
<td>22</td>
<td>126</td>
<td>103</td>
<td>79</td>
<td>84</td>
<td>77</td>
<td>9</td>
</tr>
<tr>
<td>Rebecca</td>
<td>1940</td>
<td>18</td>
<td>100</td>
<td>111</td>
<td>96</td>
<td>83</td>
<td>78</td>
<td>12</td>
</tr>
<tr>
<td>Saboteur</td>
<td>1942</td>
<td>29</td>
<td>90</td>
<td>74</td>
<td>74</td>
<td>90</td>
<td>114</td>
<td>28</td>
</tr>
<tr>
<td>Suspicion</td>
<td>1942</td>
<td>38</td>
<td>107</td>
<td>130</td>
<td>89</td>
<td>58</td>
<td>62</td>
<td>16</td>
</tr>
<tr>
<td>Shadow of a Doubt</td>
<td>1943</td>
<td>24</td>
<td>101</td>
<td>107</td>
<td>83</td>
<td>100</td>
<td>78</td>
<td>8</td>
</tr>
<tr>
<td>Lifeboat</td>
<td>1944</td>
<td>21</td>
<td>122</td>
<td>114</td>
<td>106</td>
<td>80</td>
<td>38</td>
<td>19</td>
</tr>
<tr>
<td>Notorious</td>
<td>1947</td>
<td>72</td>
<td>119</td>
<td>88</td>
<td>74</td>
<td>74</td>
<td>66</td>
<td>8</td>
</tr>
<tr>
<td>Paradine Case, The</td>
<td>1947</td>
<td>23</td>
<td>140</td>
<td>89</td>
<td>116</td>
<td>57</td>
<td>66</td>
<td>7</td>
</tr>
<tr>
<td>Strangers on a Train</td>
<td>1951</td>
<td>44</td>
<td>116</td>
<td>114</td>
<td>59</td>
<td>65</td>
<td>87</td>
<td>16</td>
</tr>
<tr>
<td>I Confess</td>
<td>1952</td>
<td>32</td>
<td>108</td>
<td>82</td>
<td>81</td>
<td>77</td>
<td>101</td>
<td>20</td>
</tr>
<tr>
<td>Dial M for Murder</td>
<td>1954</td>
<td>15</td>
<td>69</td>
<td>98</td>
<td>141</td>
<td>141</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>Rear Window</td>
<td>1954</td>
<td>17</td>
<td>75</td>
<td>113</td>
<td>74</td>
<td>72</td>
<td>121</td>
<td>29</td>
</tr>
<tr>
<td>Man Who Knew Too Much, The</td>
<td>1955</td>
<td>15</td>
<td>59</td>
<td>115</td>
<td>96</td>
<td>81</td>
<td>111</td>
<td>23</td>
</tr>
<tr>
<td>To Catch a Thief</td>
<td>1955</td>
<td>8</td>
<td>64</td>
<td>72</td>
<td>104</td>
<td>102</td>
<td>113</td>
<td>37</td>
</tr>
<tr>
<td>Trouble With Harry, The</td>
<td>1956</td>
<td>4</td>
<td>20</td>
<td>113</td>
<td>145</td>
<td>121</td>
<td>73</td>
<td>23</td>
</tr>
<tr>
<td>Wrong Man, The</td>
<td>1957</td>
<td>31</td>
<td>127</td>
<td>150</td>
<td>75</td>
<td>63</td>
<td>51</td>
<td>4</td>
</tr>
<tr>
<td>Vertigo</td>
<td>1958</td>
<td>15</td>
<td>113</td>
<td>104</td>
<td>84</td>
<td>56</td>
<td>92</td>
<td>35</td>
</tr>
<tr>
<td>North by North-West</td>
<td>1959</td>
<td>12</td>
<td>87</td>
<td>96</td>
<td>96</td>
<td>77</td>
<td>99</td>
<td>33</td>
</tr>
<tr>
<td>Birds, The</td>
<td>1963</td>
<td>59</td>
<td>127</td>
<td>111</td>
<td>69</td>
<td>57</td>
<td>58</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 9.1: Shot scale data (%) for films of Alfred Hitchcock.
The data are as given in Barry Salt’s database on the Cinemetrics site (i.e. scaled to sum to 500 rather than percentages as in Section 2.2.6). The first nine films, to the silent version of Blackmail, are British silent films; the 14 from the sound version of Blackmail through to The Lady Vanishes are British sound films; the remaining 19 are American. The conventional bar chart representation of the data for the silent films is shown in Figure 9.1; bar charts for the other films are in Section 2.2.6.

Before looking at the idea behind correspondence analysis (CA) it is convenient to have an example to hand. A basic CA is shown in Figure 9.2, and was obtained using the ca function in the ca package, which needs to be imported then loaded using library(ca). The data are read into R, named Hitchcock say, with shot-scales extracted as shot <- Hitchcock[, 3:9]. All that is then needed to get the figure is plot(ca(shot)).

The blue dots in the plot correspond to films and are labeled, by default, according to their order in the table. Films close to each other on the plot have similar shot-scale profiles; films at some distance apart have comparatively different profiles. The silent films are labeled 1-9, for example, and plot fairly close to each other so, as a broad statement – to which there are specific exceptions – it can be inferred that they tend to be more similar to each other than they are to most later films. Their similarity to each other is also evident in Figure 9.1.

The red triangles correspond to the different shot-scales and are labeled accordingly. Their interpretation was discussed at a little length in connection with Figure 2.9 and the salient points will be reiterated. For the moment it is sufficient to note that their positions allow useful inferences to be drawn about why points corresponding to films on the plot are similar or distant. For

---

1The scaling chosen has no effect on the analysis.
example, it can be inferred that films to the left, plotting in the same general region as LSs and MLSs, will feature comparatively more of these kind of shots than films to the right; conversely, those to the right are more likely to favor the various degrees of close-up.

Figure 9.2: *A basic correspondence analysis of the shot-scale data from Table 9.1.*

There are some caveats about interpretation, discussed shortly, that need to be understood but, fundamentally, this is how a large number of applications of CA are presented and interpreted. Here, once the machinery has been set up, the analysis is executed by – in this example – typing one ‘word’ of 14 characters. For exploratory purposes the default output obtained using the ca function is often adequate\(^2\). All CA is doing, as applied here, is reducing a table of data to a picture; given its utility and ease of application the reasons for its popularity in a wide range of disciplines is obvious.

For publication purposes, or communicating results with others, the default CA plots produced by most software can usually be improved. Figure 9.2 is too ‘busy’ for my taste; duplicating points for films and labels doesn’t seem necessary; information on whether the film is silent/sound, British/American needs to be recovered by referring the numbers on the plot back to the table, and so on.

Figure 9.3 is an alternative version of the plot. Colored symbols replace the solid circles and labels; it is immediately apparent that the silent films form a comparatively tight cluster compared to the British and American sound films. The latter are rather spread out but, with the exception of four British sound films, do not overlap with other films. It is easily enough established that the overlapping British sound films are four of the five latest.

\(^2\)Other packages that include functions for correspondence analysis include MASS, vegan, ade4, anacor, FactoMineR etc. They vary a little in ease of application and the way CA maps are presented, but the basics are similar.
Joining up the markers for shot-scales emphasizes that there is a reasonably good ‘stylistic gradient’ of the kind discussed in connection with Figure 9.1, the position of the VLS marker being anomalous. The left-hand side of the plot is associated with the longer shots, and the right-hand side with the closer shots and, in broad terms, tends to differentiate earlier from later films.

Figure 9.3: An enhanced correspondence analysis of the shot-scale data from Table 9.1.

9.2 What correspondence analysis does

While CA has earlier roots it is the (sometimes formidable) mathematical development of French mathematicians and statisticians, notably Jean-Paul Benzécri, that is usually cited as the important ‘popularizing’ development. Greenacre (1984), a student of Benzécri, published the first major English language text, also sufficiently mathematically demanding to prompt publication of Greenacre (2007), the second edition of a text first published in 1993, much cited and meriting the adjective popularizing without the use of inverted commas.

Since then CA has come to be widely used as a data-analytic tool across a range of disciplines in the sciences, social sciences and humanities. This has generated a large number of expository articles, often similar in kind but aimed at different audiences. My own contribution to this genre, Baxter and Cool (2010), was aimed at archaeologists; its relevance as far as this text is concerned is that there was a specific focus on R. The intended readers were assumed to have data they wanted to analyse, and knowledge that CA might be a useful tool, but no prior experience of R or carrying out their own CA\(^3\).

\(^3\)The paper can be found on the web, in more than one place, by Googling sensible terms.
The appeal of CA is down to both its utility and conceptual simplicity. As usually applied, ignoring the mathematics and ‘deeper’ theoretical issues, CA takes a (possibly) large table of non-negative data and reduces it to a picture with rows of the table represented by one set of points, and columns by another. Focusing on rows (films) to be specific, rows with similar profiles should be close to each other on the plot; rows with dis-similar profiles should be distant. (The row profile is given by the numbers in a row scaled to have a common value; the shot scale data collected by Salt are already in this form as they add to $500^4$.)

The two previous figures are in two dimensions and can be thought of as maps of the data. The dimensionality of the data set is defined by the number of columns, seven in the example. The distance between any two rows, for example, can be defined precisely and mathematically but if there are more that three columns, can’t be represented exactly using conventional plotting methods. The real data can be thought of as a ‘cloud’ of points, ‘living’ in seven dimensions that you can’t visualise; to get a look at it the data needs to be ‘squashed’ (mathematically, of course) onto a two-dimensional plane that preserves as well as possible distances between points, and hence any patterns that characterize relationships between them.

This means that the end result is an approximation to the reality. The quality of this approximation can be judged in various ways, which will not be dealt with here. The other technical point to note, but not worry about too much, is that mathematicians can define distance in different ways, and what is being approximated is chi-squared distance. It differs from what you are used to in daily life when you talk about physical ‘distance’, but not enough to worry about for the purpose of the present discussion.

Those who, quite rightly, are uneasy about accepting these blasé assurances and consult other literature need to be warned about some of what is around, particularly the more evangelical kind of exposition that can oversell CA. It is often emphasized that CA is appropriate for analyzing tables of counted data, which is true but ignores the fact that CA can be applied to any table of non-negative numbers (with care) – that is, it actually has a wider range of application than some advocates imply. This does no harm; more misleading is the not uncommon suggestion that one of the defining features of CA is its ability to jointly represent row and column points on the same plot, each informing the others interpretation. Two comments can be made here; one is that this feature is shared by other statistical methods (Greenacre, 2010); the other is that there is absolutely no reason why just the row or column plot can’t be reported separately. CA is a useful, descriptive, data-analytic tool to be applied as the user sees fit.

There are some other issues to be dealt with. These will arise naturally in what follows, where the plan is to show how Figure 9.3 can be constructed incrementally, by building separate plots for the rows and columns before overlaying them.

9.3 Producing enhanced CA plots

Remember that \texttt{plot(ca(shot))} was all that was needed to get Figure 9.2, the table of shot-scale data being held in \texttt{shot}. Proceed slightly differently as follows, where the first command carries out the CA without plotting and remaining commands extract the coordinates needed for plotting (note the use of the semi-colon to put several commands on the same line).

\begin{verbatim}
CAH <- ca(shot)
x <- CAH$rowcoord[,1:2]
y <- CAH$colcoord[,1:2]
x1 <- x[,1]; x2 <- x[,2]; y1 <- y[,1]; y2 <- y[,2]
\end{verbatim}

At this point \texttt{plot(x1, x2)} would give a reasonably uninformative plot of the rows (films). Adding symbols can be done in various ways, one of which is

\footnote{It is not necessary that data be collected in this format; an early mathematical step in CA, invisible to the user, is to effect this scaling.}

\footnote{Some authors are quite stern about this, and quantities called inertias can be used both to investigate how good the overall approximation is, and how well individual rows or columns are represented. The literature cited provides an entry point.}
Symbol <- c(rep(17,9), rep(16, 14), rep(15, 19))
color <- c(rep("green4",9), rep("red", 14), rep("blue", 19))

library(MASS) # Loads package MASS needed for eqscplot
eqscplot(x1, x2, pch = Symbol, col = color)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)

The \texttt{rep(a, b)} function produces \texttt{b} copies of \texttt{a}. The \texttt{c(x, y, z)} structure strings together the contents of \texttt{x}, \texttt{y} and \texttt{z}; 17, 16 and 15 are the plotting symbols for solid triangles, circles and squares. The \texttt{eqscplot()} requires \texttt{MASS} and ensures equal scaling of the axes, important for the distance interpretation of the map. The arguments \texttt{pch} and \texttt{col} specify plotting symbols and colors; axis labeling arguments have been omitted. The \texttt{abline} commands add dashed vertical and horizontal lines at 0, to provide a reference grid. The outcome is the left-hand plot in Figure 9.4.

For the column plot to the right, assuming \texttt{MASS} is loaded the following code was used, where labeling arguments are omitted again. Legends can be added to the plots, as desired.

Names <- names(shot)
eqscplot(y1,y2, type = "n")
text(y1, y2, Names)
lines(y1,y2, lwd = 2)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)

This is a perfectly good joint representation of the data, and you can readily interpret it in the same way as Figure 9.3 (which is where the grid lines are useful). Some authors (myself included) often prefer to present results in this way since it avoids some of the clutter that can arise with large tables where the overlaid plots can become difficult to read. Nevertheless code for overlaying the plots will be illustrated. This is partly to introduce features of \texttt{R} which may be of interest, but also because it raises an important interpretational issue.

In Section 2.2.6 an analogy was suggested whereby the film markers could be thought of as settlements in a landscape whose terrain was defined by the column markers. This is achieved by overlaying the two maps of Fig 9.4. The problem is that it is not obvious how to do this since, before overlaying the maps, it is legitimate to stretch one map relative to another. Suppose, for example, that the row map is fixed. Imagine that the column map is on a rubber sheet and pin it to the row map at the origin (the (0,0) point) with the grid (the dashed lines) aligned. Now grab

Figure 9.4: Separate row and column plots for a CA of Table 9.1.
the rubber sheet at each corner, pull out, and stretch the sheet uniformly, so that eventually the joint map, defined by the stretched sheet, will have column markers on its periphery, with row markers squeezed up in the middle. The role of the maps can be reversed, so you can get a plot with column markers so squeezed up.

There are other ways of stretching and how it’s done can be expressed quite precisely in mathematical terms. Greenacre’s texts may be consulted for technical details; the ‘correct’ way of stretching (in the view of some) leads to precisely the squeezing effect just described, and because of it the resultant map can be difficult to decipher. Accordingly, many practitioners opt for a kind of compromise, which is the symmetric map of the sort shown in Figures 9.2 and 9.3. Roughly, the overlaying is done to give equal emphasis to both row and column markers, with a more interpretable map resulting.

The better implementations of CA allow a choice of mapping. The symmetric map (which can also be defined in a precise mathematical way) is the default in the ca function, but this is not always the case for other implementations. Incidentally, plotting row and column maps separately avoids too much agonizing about the issue.

There is an important consequence of all this, which is that you cannot interpret the positioning of a row marker relative to a column marker in terms of ‘distance’ in the way that the positioning of two row markers can be interpreted as distance. For example, in Figure 9.3 the green triangle that sits almost on the LS marker is for *Downhill*. The film is ‘near’ or ‘close’ to the marker on this particular joint map, but if the column map was stretched this would no longer apply. With a symmetric map interpretation can often be carried out by thinking in terms of relatively ‘nearer’ or ‘further’, but do the thinking in inverted commas.

The plotting coordinates saved when using the ca function do not need modifying to get a symmetric map when the separate plots are overlaid. In the code to follow z combines plotting coordinates for both rows and columns, so that everything is included in the plot. The type = "n" argument produces a blank plot to which points and lines will be added. The numerical values on the axes, as in Figure 9.2, don’t add much to interpretation (and vary according to how scaling is effected in different packages and functions) so are eliminated by the axes = F argument. The subsequent axis commands add labels for the grid produced by the abline commands.

```r
z <- rbind(x,y)
library(MASS)
eqscplot(z[,1], z[,2], type = "n", xlab = "", ylab = "", axes = F)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)
points(x1, x2, pch = Symbol, col = color, cex = 1.2)
text(y1, y2, Names, cex = 1.2)
lines(y1, y2, lwd = 2)
axis(1, at = 0, labels = 0)
axis(2, at = 0, labels = 0)
legend("topleft", legend = c("British silent", "British sound", "American"), pch = c(17,16,15),
col = c("green4", "red", "blue"), bg = "n", cex = 1)
```

The points function adds the symbols and colors for the row markers at the coordinates defined by x1 and x2. Note that these points are added outside the original plot command; for the individual row plot they were specified within the plot command. The cex argument is the character expansion, with a default of 1 and used to control the size of plotting symbols. The text and lines functions add the column marker information, the lines being an optional add-on, and Names in the text command is the previously defined text Names <- names(shot) plotted at the coordinates y1 and y2. Figure 9.3 is the end result.

Time has been taken over this to illustrate the kind of control that can be exercised if you want to go beyond the default plot – and this is not compulsory. Identifying individual films is straightforward, either by labeling the row plot with numbers rather than symbols, or even adding film titles, though this produces an unreadable plot. To illustrate, with embellishments, Figure 9.5 shows a plot for rows only, labeled by date (year since 1900).

It is clear that the earliest of the British sound films have a shot-scale structure close to those of the silents, whereas some of the later ones have a structure more akin to the American ones. This is emphasized by constructing enclosures that separate out ‘early’ films, defined here as films
up to and including 1932, and ‘late’ American ones from 1940 on. Four of the five latest British
sound films from 1936-39 sit comfortably within the body of American films; Jamaica Inn (1939)
is the exception, which lies between the two groups highlighted. The two 1935 films are also
intermediate between the highlighted early and late groups – it should be emphasized that the
positioning of films is only influenced by date to the extent that date and shot-scale structure are
associated.

Figure 9.5: The row plot from a CA of Table 9.1, labeled by date; green for silent, red for British
sound and blue for American films. See the text for details of how the convex hulls are constructed.

The 1934 film Waltzes from Vienna is an obvious outlier in the context of patterns displayed
elsewhere in the plot. It jumps away from the main body of early films but in an opposite direction
to those coming not long after (in the sample used here). The other British sound film that behaves
in the same way is Juno and the Paycock (1930) , classified here as ‘early’ but an outlier within
that group (it is the film furthest to the left in the relevant plots, with Waltzes from Vienna a
bit above it). Both films are distinguished from all others by a relative paucity of close-ups, and
Waltzes from Vienna is additionally distinguished by an excess of LSs.

The plot with dates color coded by ‘provenance’ is very simply obtained.

date <- Hitchcock[,2] - 1900
eqscplot(x1, x2, type= "n", xlab = "axis 1", ylab = "axis 2")
text(x1, x2, date, col = color)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)

If Hitchcock, in his conversations with Truffaut (1985, pp. 85-87), is to be believed Waltzes from Vienna is
not, for other reasons, typical. He variously describes the film, made when his career was at a low ebb, as ‘made
cheaply’, bearing no relation to his ‘usual work’ and ‘very bad’. Later on (page 314) he does not demur when
Truffaut suggests the film was ‘not in your genre, you dismissed [it] as an out-and-out waste of time’. Truffaut fails
to ask why this affected the shot-scale structure. Hitchcock wasn’t too happy about Juno and the Paycock either
saying (page 69) that he ‘didn’t feel like making the picture’ because of the difficulty of narrating ‘in cinematic
form’ what he describes as an ‘excellent play’. He ends by saying that when the film got good notices he was
‘ashamed because it had nothing to do with cinema’.
Next, define two bodies of film silent, for the first 15 films in Table 9.1 that includes all the silents and British sound films to Rich and Strange in 1932; and american for all the American films from 1940 on. The eight British sound films from the 1934-39 period are accorded no special treatment.

\[
\text{silent} \leftarrow x[1:15,1:2]
\]
\[
\text{hpts} \leftarrow \text{chull(silent)}
\]
\[
\text{hpts} \leftarrow \text{c(hpts, hpts[1])}
\]
\[
\text{lines(silent[hpts, 1], col = "green4", lwd = 2)}
\]
\[
\text{american} \leftarrow x[24:42,1:2]
\]
\[
\text{hpts} \leftarrow \text{chull(american)}
\]
\[
\text{hpts} \leftarrow \text{c(hpts, hpts[1])}
\]
\[
\text{lines(american[hpts, 1], col = "blue", lwd = 2)}
\]

Added to the preceding code the bodies of early and late films defined above are enclosed, to produce Figure 9.5.

### 9.4 CA and supplementary points

If Figure 9.3 is compared with Figure 2.9 it will be seen that the configuration of those Hitchcock films common to both analyses is not identical. This is a consequence of the fact that the map produced is determined by all the data entered into an analysis. The main potential drawback of this, though it need not be serious, is that if there is an interest in comparing subsets of films to see if they are distinct, including them all in the CA analysis as has been done so far may blur distinctions.

There is a way round this. Suppose that there are two bodies of films one wishes to effect a comparison between. Construct the CA map using one body of data, then add points to the map, corresponding to films in the second body, as supplementary points. The idea is that they are brought along to a pre-existing landscape and then fitted into the part of the terrain that their shot-structure best suits them to. They cannot influence the terrain; they have to live with what’s there.

For illustration 18 films of Max Ophuls will be used (Table 9.2). These will be fitted onto the map defined by the Hitchcock films that have been the subject of this chapter so far.

<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>BCU</th>
<th>CU</th>
<th>MCU</th>
<th>MS</th>
<th>MLS</th>
<th>LS</th>
<th>VLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verliebte Firma, Die</td>
<td>1931</td>
<td>8</td>
<td>29</td>
<td>51</td>
<td>132</td>
<td>88</td>
<td>148</td>
<td>46</td>
</tr>
<tr>
<td>Lachende Erben</td>
<td>1932</td>
<td>8</td>
<td>14</td>
<td>45</td>
<td>114</td>
<td>145</td>
<td>125</td>
<td>49</td>
</tr>
<tr>
<td>Liebelei</td>
<td>1932</td>
<td>10</td>
<td>18</td>
<td>53</td>
<td>112</td>
<td>136</td>
<td>126</td>
<td>42</td>
</tr>
<tr>
<td>Verkaufte Braut, Die</td>
<td>1932</td>
<td>6</td>
<td>21</td>
<td>61</td>
<td>103</td>
<td>105</td>
<td>113</td>
<td>53</td>
</tr>
<tr>
<td>Signora di Tutti, La</td>
<td>1934</td>
<td>65</td>
<td>40</td>
<td>81</td>
<td>110</td>
<td>84</td>
<td>80</td>
<td>40</td>
</tr>
<tr>
<td>Komedie om Geld</td>
<td>1935</td>
<td>25</td>
<td>36</td>
<td>69</td>
<td>99</td>
<td>81</td>
<td>127</td>
<td>61</td>
</tr>
<tr>
<td>Tendre Ennemie, La</td>
<td>1936</td>
<td>12</td>
<td>46</td>
<td>58</td>
<td>119</td>
<td>90</td>
<td>108</td>
<td>66</td>
</tr>
<tr>
<td>Yoshiwara</td>
<td>1937</td>
<td>28</td>
<td>34</td>
<td>80</td>
<td>87</td>
<td>80</td>
<td>133</td>
<td>56</td>
</tr>
<tr>
<td>Werther</td>
<td>1938</td>
<td>10</td>
<td>34</td>
<td>43</td>
<td>72</td>
<td>91</td>
<td>180</td>
<td>71</td>
</tr>
<tr>
<td>Sans Lendemain</td>
<td>1939</td>
<td>20</td>
<td>69</td>
<td>86</td>
<td>124</td>
<td>90</td>
<td>85</td>
<td>26</td>
</tr>
<tr>
<td>De Mayerling à Sarajevo</td>
<td>1940</td>
<td>29</td>
<td>39</td>
<td>74</td>
<td>134</td>
<td>65</td>
<td>111</td>
<td>48</td>
</tr>
<tr>
<td>Exile, The</td>
<td>1948</td>
<td>7</td>
<td>15</td>
<td>49</td>
<td>109</td>
<td>140</td>
<td>143</td>
<td>37</td>
</tr>
<tr>
<td>Letter from an Unknown Woman</td>
<td>1948</td>
<td>16</td>
<td>50</td>
<td>71</td>
<td>116</td>
<td>116</td>
<td>108</td>
<td>24</td>
</tr>
<tr>
<td>Caught</td>
<td>1949</td>
<td>38</td>
<td>18</td>
<td>71</td>
<td>156</td>
<td>88</td>
<td>111</td>
<td>22</td>
</tr>
<tr>
<td>Reckless Moment, The</td>
<td>1949</td>
<td>22</td>
<td>37</td>
<td>76</td>
<td>118</td>
<td>86</td>
<td>113</td>
<td>49</td>
</tr>
<tr>
<td>Rondo, La</td>
<td>1950</td>
<td>5</td>
<td>66</td>
<td>41</td>
<td>163</td>
<td>127</td>
<td>71</td>
<td>25</td>
</tr>
<tr>
<td>Plaisir, Le</td>
<td>1952</td>
<td>0</td>
<td>7</td>
<td>46</td>
<td>87</td>
<td>120</td>
<td>153</td>
<td>87</td>
</tr>
<tr>
<td>Madame de...</td>
<td>1953</td>
<td>4</td>
<td>22</td>
<td>71</td>
<td>120</td>
<td>161</td>
<td>88</td>
<td>29</td>
</tr>
<tr>
<td>Lola Montès</td>
<td>1955</td>
<td>1</td>
<td>20</td>
<td>55</td>
<td>114</td>
<td>108</td>
<td>135</td>
<td>61</td>
</tr>
</tbody>
</table>

Table 9.2: Shot scale data (%) for films of Max Ophuls.

If Figure 2.9 is examined it will be seen that with two exceptions there is little overlap between the Ophuls films and American Hitchcock, but rather more with the British Hitchcock sound films.

---

7 The coding is not especially transparent. The ‘enclosures’ are what are called convex hulls and the coding is based on that given as an example in the help facilities `?chull`. 

102
Plotting the Ophuls films as supplementary points on a map determined by all the Hitchcock films produces Figure 9.6.

There is still overlap but the distinction between the works of the two directors is sharper than was previously the case. This could be described in different ways, but it would be possible, for example, to construct a convex hull for 16/18 of the Ophuls films that enclosed only one Hitchcock in addition. The cluster of eight Ophuls films (one slightly astray) in the middle of the lower-left quadrant includes the four earliest Ophuls films in Table 9.2, German from the years 1931-32 and clearly distinct from the Hitchcock films of that date. Also in the group are the three last, French, films in the table dating from 20 years or more later. The films are distinguished by the relatively low number of close-ups and large numbers of MLSs, characteristics shared by *The Exile*, the only other film in this cluster.

Figure 9.6: The row plot from a CA of Hitchcock films from Table 9.1 (open circles) with supplementary points (closed circles) for the films of Ophuls from Table 9.2.

The bones of the code needed are given below. It is assumed that a file *Ophuls*, similar to that of *Hitchcock* has been created; details of convex hull and legend construction are omitted.

```r
sup <- Ophuls[, 3:9]
shotsup <- rbind(shot, sup)
CAso <- ca(shotsup, subsetrow = 1:42, suprow = 43:60)
x <- CAso$rowcoord[,1:2]
y <- CAso$colcoord[,1:2]
x1 <- x[,1]; x2 <- x[,2]; y1 <- y[,1]; y2 <- y[,2]
Symbol1 <- c(rep(1,42), rep(16,18))
color1 <- c(rep("black",42), rep("darkorange",18))
eqscplot(x1, x2, pch = Symbol1, col = color1, xlab = "axis 1", ylab = "axis 2", cex= 1.2)
abline(v = 0, lty = 2); abline(h = 0, lty = 2)
```

---

8In his chapter on the stylistic analysis of the films of Ophuls Salt (2009, p.394) comments, in his discussion of *La Ronde*, on the similarity of shot-scale distribution of the late French films to the earlier sound films. Salt's analysis goes beyond what is attempted here, where the body of Ophuls films is being used simply to illustrate the idea of supplementary points.
9.5 Other plotting methods

9.5.1 Analysis of ranks

The utility of CA for looking at shot-scale data is unquestionable. It does what presentation in the form of bar charts does, does it more concisely, and – I would argue – more informatively. As with all techniques of this kind it is always possible to go back to the raw data, either in tabular form or as a bar chart, to check any inferences you may wish to draw. The CA directs attention to where it might be useful to look.

Redfern (2010b, 2010c), and in other posts on his research blog, has developed a method based on ranked shot-scale data. The first paper uses the same body of Hitchcock films used here. The focus is more on examining ‘hypotheses’ about differences (or their lack) between the periods of silent, British sound and American films, rather than descriptive graphical analysis. That is, whereas knowledge of period informs the interpretation of CA plots but not their construction, the same knowledge is integral to the construction of Redfern’s method.

The method is best explained by describing the construction. For each category:

1. Re-arrange the numbers in each row from largest to smallest.
2. Obtain the mean of each column in the new table so obtained.
3. Plot the means against the numbers 1 to 7 (the ranks of the data).
4. Fit a (regression) straight line through the data.

Once this is done compare the fitted lines. The row numbers need to be the same for this to work and, rounding error apart, they should be. These are converted to mean relative frequencies (MRFs) by dividing row numbers by row totals. Do all this for the data of Table 9.1 and Figure 9.7 can be constructed.

![Figure 9.7: Mean relative frequency plotted against shot scale ranks for three categories of Hitchcock films. See the text for an explanation.](image)

There is little difference in the fitted lines. Redfern concludes that ‘no one shot scale dominates Hitchcock’s films . . .’ and that ‘it is clear that Hitchcock’s style did not change with either the
introduction of sound technology or the move from Britain to America’. You need to read the papers carefully to work out what this means, since ‘shot scale’ seems to be used in two different ways. For example, in the sentence following the above conclusion it is stated that ‘if we turn our attention to the shot scales themselves we can see that there is a clear change in Hitchcock’s style’. On first reading this I thought it was a direct contradiction of the preceding sentence.

The second usage is straightforward. It is just saying something along the lines that the extent to which Hitchcock used some shot-scale types varied in his different periods. The first usage is based on the ranking of shot-scale types, and refers to the ranks, the data becoming divorced from types in the process. Thus, and hypothetically, any column of the rearranged table of data could consist of numbers for a single shot type, or of six examples for each of the seven shot types – the method doesn’t distinguish between these possibilities.

In plotting the mean of each column of the data after it has been rearranged by row rank order, ‘shot-scale’ simply refers to the rank order. I think the aim in comparison is to see if different bodies of data exhibit linear patterns in the plot of means of the ranked data and, if so, whether the slopes and intercepts of the fitted lines differ much. Departures from linearity are of particular interest; for example, and as I’ve understood it, if the mean for the data ranked 1 is noticeably greater than that predicted from a linear model based on all the means, that shot-scale is said to ‘dominate’. Hypothetically, had this occurred for one of Hitchcock’s periods, it would be saying that within that period individual films tended to have a higher proportion of one type that the linear model based on that film predicts. No information about the distribution of shot-scale types that dominate the films is used.

The procedure is – in my opinion – rather arcane and the interpretation is not easy to understand. The detachment of the analysis from shot-scale types is a (severe) limitation. As much is acknowledged by Redfern, both explicitly (without the use of the bracketed adjective) and implicitly in the use of analyses that do recognise type differences and produce more readily understood conclusions. It is questionable whether the analysis of the ranked data adds much to what can be done by simpler means.

The analyses that allow for shot-scale type differences are of two kinds. One is a statistical hypothesis testing approach where, for each shot-scale type in turn, the hypothesis tested is that there is no difference between periods in the typical proportion of each type. The broad conclusion is that the American films differ from the British films, with a move towards ‘tighter framing’, but that the British silent and sound films do not differ. Ultimately shot-scale types are examined at the level of individual films, leading to the suggestion that the later British sound films were evolving towards the American style (see, also, Salt, 2009, p.243).

There are limitations to the hypothesis testing approach since it ignores the possibility of temporal evolution of style within a period. That is, there may be a noticeable increase or decrease in the use of a particular shot-scale type within a period that does not sufficiently affect the ‘typical’ value (be it mean or median) for it to be statistically significantly different from the typical value in another period with a different pattern. An alternative approach is sketched in the next section.

The code to obtain a plot of the kind illustrated for a single body of data follows (style and labeling arguments omitted). Call the body of interest body and assume, as in Salt’s database, rows have been scaled to sum to 500.

```r
MRF <- apply(apply(body/500, 1, sort), 1, mean)
MRF <- sort(MRF, decreasing = T)
Rank <- 1:7
plot(Rank, MRF, ylim = c(0,.3))
abline(lm(MRF ~ Rank))
```

The first line produces the means that need to be plotted, but in the reverse rank order needed, which the second line corrects.
9.5.2 Analysis of individual shot-scale types

The idea is both simple and obvious. For each shot-scale type plot its occurrence (percentage in the plots to follow) against date. This is done in Figure 9.8 with the added embellishments of loess (3/4, s, 2) smooths complete with two standard error limits, and vertical lines at 1929.5 and 1939.5. The loess fit is designed to be a smooth one, ignoring very local variation, and the robust version that downweights extremes is also used. The vertical lines separate out the three periods except that the sound version of *Blackmail* is a 1929 film. The hypothesis testing approach effectively assumes ‘flat’ distributions within each period, apart from random variation, and tests the hypothesis that ‘flat’ lines fitted within each period do not differ significantly in their ‘level’.

![Figure 9.8: Plots of shot-scale occurrence (%) against date for individual films and each shot-scale type, with added loess (3/4, s, 2) smooths.](image)

The usual caveats concerning sample sizes, dependence of the loess curve on the degree of smoothing, and so on, can be entered. There is little doubt that Hitchcock’s style, as reflected in the shot scale distributions, does change over the time period involved, so interpretation can concentrate more on an ‘holistic’ examination of ‘how’.

Given the shortness of the period involved, not too much can be said about developments within the silent period. There is the clear suggestion of evolution within the British sound period, with the use of CUs and MCUs increasing and that of MLSs and LSs decreasing. The change for MCUs possibly starts a little later than for the others. For the American period MLSs continue to decline while the use of VLSs shows an increase, though this is one of the least commonly used types. Otherwise, for types where some evolution in use in the British sound period was suggested there tends to be a levelling out.

It needs to be re-emphasized that this kind of interpretation does not really pay attention to the fact that there is a lot of variation in the data, evident from the CA graphs, particularly as the robust loess smooth downplays this. Nevertheless, the temporal variation exhibited in the British sound period suggests that conclusions derived from hypothesis testing, that there are no
significant differences between the use of shot-scales in the two British periods, could be viewed as un-nuanced. The American period is hardly static – it shows both some change in the extent to which some shot-scale types are used and is highly variable – but in very broad terms it can be argued that the period was characterized by a fairly stable pattern of usage.

In the following code y is the shot-scale data for a single type and x is the date. This would produce a single one of the plots shown in Figure 9.8. I’ve left in the code that produces the shading between the 2-standard error limits, but not explained it (the shading is not really necessary, but it looks nice).

```r
y <- y/5
fit <- loess(y ~ x, family = "s")
pred <- predict(fit, se = T)

# plot 2-standard error limits
plot(x, y, xlab = "date", type = "n", ylab = "percentage of total")
lines(x, pred$fit - 2*pred$se.fit, lty = 2, col = "blue", lwd=2)
lines(x, pred$fit + 2*pred$se.fit, lty = 2, col = "blue", lwd=2)

# create polygon bounds and add to plot
y.polygon <- c((pred$fit + 2*pred$se.fit), (pred$fit - 2*pred$se.fit)[42:1])
x.polygon <- c(x, sort(x, decreasing = T))
polygon(x.polygon, y.polygon, col="lightblue", border = NA)

# add everything else in
lines(x, pred$fit, col = "darkred", lwd = 3)
points(x, y)
abline(v = 1929.5, col = "red")
abline(v = 1939.5, col = "red")
```

In previous uses of `loess` a different way of getting the fitted line was used. Here it’s done using the `predict` function which allows standard errors to be saved. These are used to plot 2-standard error limits with the first two `lines` commands. The blank plot has these limits and the fill between them added before the fitted line and points, to avoid over-plotting by the fill.
Chapter 10

Large data sets

10.1 Introduction

The main aim of this chapter is to discuss how ‘large’ data sets can be imported into R, allowing similar analyses to be done for a set of films at one go. ‘Large’ is defined to mean a sufficient number of films where the analyst regards it as tedious to enter data and do the analysis one film at a time.

The focus is on SL data, and for illustration 134 of the 150 films for which data were collected by Cutting et al. (2010) will be used. This data set has been discussed previously; data for 16 films aren’t used because the versions submitted to the Cinemetrics database contain zeroes and/or negative values – the latter clearly errors – that precludes the use of logarithms to be explored. These data were used, in this form, by Redfern (2012a) to explore the issue of lognormality of SL distributions, and are used to look at aspects of this question as a case study in Chapter 12. Sections 10.2 and 10.3 discuss one approach to getting larger sets of SL data into R with illustrations of the sort of things that can then be done with it.

10.2 Data entry and basic analysis

10.2.1 Data entry

What follows describes one way in which a data set for entry into R might be created from scratch. With previously created data not readily converted to the format used here it’s worth looking at the R Data Import/Export manual (http://cran.r-project.org/doc/manuals/R-data.pdf).

Previously, in Section 3.2.2, all the data for a single film was copied from the Cinemetrics database into Excel; column headers edited; then imported into R using the read.table function. The column corresponding to SL data was then extracted within R. For present purposes the data for a film was copied to Excel and columns other than that for SL, deleted. This was repeated for other films, copied to the same Excel file, so that the end product for the data used here was an Excel file with 134 columns of unequal length.

Columns can be named as you go along, most obviously with the name of the film. Something like Brief Encounter won’t work, because R doesn’t like the space; both Brief Encounter and "Brief Encounter" will, though in the last case R will change it to Brief.Encounter which could be entered directly.

On import into R it is expected that columns have equal length, so these have to be equalized by filling in the blanks with the missing data indicator NA (for Not Available). Thus, for the data under discussion King Kong (2005) has 3077 shots, so every other film has to be filled out with NAs to this length. Once this is done the file can be imported into R as a data frame with 134 columns of unequal length.

\[1\] Most simply done by highlighting the 134 columns and rows down to 3077 then using the ‘Replace’ command in Excel to replace blanks with NA.
columns. and 3077 rows. The file will be called SL.ALL below.

10.2.2 Data analysis

Extracting SL data for a single film for analysis, Harvey for example, which is the 33rd film (column) in the data set, named Harvey, can be done using

SL.Harvey <- na.omit(SL.ALL[, 33])

The SL.ALL[, 33] can be replaced with SL.ALL$Harvey, but the more general form is shown. The na.omit function strips the NAs out of the column. This is necessary because some operations you might want to perform don’t like them. It’s possible to tell other functions what to do with NAs when you invoke them, but simplest to remove them at the start.

Once extracted you can do things like mean(SL.Harvey) and median(SL.Harvey) to get the ASL and MSL or, better still, create a small function to do both jobs at once (Section 4.3). As a reminder, and to prepare for generalizing, the function was called filmstats, defined as

```r
filmstats <- function(film) {
  z <- film
  ASL <- mean(z)
  MSL <- median(z)
  list(ASL = ASL, median = MSL) # List of statistics
}
```

10.2.3 Dealing with several films at once

To obtain the ASL and median for all 134 films the following can be used.

```r
SL.AllStats <- function() {
  z <- SL.ALL/10 # convert SLs to seconds
  ASL <- apply(z, 2, function(x) mean(na.omit(x)))
  MSL <- apply(z, 2, function(x) median(na.omit(x)))
  stats <- as.data.frame(cbind(ASL, MSL))
  attr(stats, "row.names") <- names(z)
  stats
}
```

The apply function applies to each column of z the function defined by function(x). If operations are to be applied to rows just replace the argument 2 by 1. The cbind function combines the ASLs and MSLs into a table of two columns, converted into a data frame using the as.data.frame function. The attr function provides rows names that are those for the column headers for the films extracted by the names function; you can omit it; or use c(1:134) rather than names(z) for numeric row names. The table produced, stats, will have two columns with the names ASL and MSL, and whatever row names you choose to give it. You can now do things like

```r
test <- SL.AllStats()
plot(test$ASL, test$MSL)
```

to get the left-hand plot in Figure 10.1. It’s also possible to replace stats at the end of the function with plot(ASL, MSL) to get the same graph. Once set up you can, within reason, do what you like with the information provided by the function.

Suppose, for example, other information relating to the films has been entered into R including a variable Date containing the years the films were released. To incorporate this information into a plot

```r
ASL <- test$ASL
MSL <- test$MSL
Symbol <- ifelse(Date > 1970, 16, 15)
color <- ifelse(Date > 1970, "red", "blue")
plot(ASL, MSL, pch = Symbol, col = color)
lines(ASL, lm(MSL ~ ASL)$fitted)
title = "Date", bty = "n")
```
categorizes the films as 1970 or earlier, and later than 1970 (using the `ifelse` function), and produces the right-hand plot in the figure. In the code 16 and 15 define the plotting symbols (solid circles and solid squares) that the `pch` argument in the `plot` function specifies. The `ifelse` function is also used to define colors for the plotting symbols.

There is a reasonably good linear relationship between the MSL and ASL, particularly for the smaller ASLs that tend to characterize later films. The `lm` function fits a regression line through the data (of median against ASL) whose fitted values are extracted (``fitted``) and plotted using the `lines` function.

The plot suggests that, at least in terms for looking at broad patterns in sets of data, concerns that have been expressed about the effects of outliers on ASL calculations are largely misplaced, since the ASL and MSL tell a similar story (Baxter, 2012a). The pattern bears on the issue of the lognormality of SL distributions – detailed discussion of which is deferred to Chapter 12. Once a function like the above has been created it can be added to incrementally, as the need arises and as the next section illustrates.

### 10.3 Further basic code

Some of the code used in later analyses is provided below.

```r
SL.AllStats <- function() {
  library(moments)
  z <- SL.ALL/10
  date <- Date

  n <- apply(z, 2, function(x) length(na.omit(x)))
  ASL <- apply(z, 2, function(x) mean(na.omit(x)))
  MSL <- apply(z, 2, function(x) median(na.omit(x)))
  mu <- apply(z, 2, function(x) mean(log(na.omit(x))))
  sigma <- apply(z, 2, function(x) sd(log(na.omit(x))))
  Lmed <- apply(z, 2, function(x) median(log(na.omit(x))))
  GM <- apply(z, 2, function(x) exp(mean(log(na.omit(x)))))
  Lmedscale <- apply(z, 2, function(x) median(scale(log(na.omit(x)))))
  skew <- apply(z, 2, function(x) skewness(scale(log(na.omit(x)))))
}
```

Figure 10.1: *Plots of median SL against ASL for 134 films, 1935-2005, the right-hand plot being an enhanced version of that to the left.*
kurtosis <- apply(z, 2, function(x) kurtosis(scale(log(na.omit(x)))))

pairs(cbind(date, n, ASL, MSL))

stats <- as.data.frame(cbind(n, ASL, MSL, mu, sigma, Lmed, GM, Lmedscale, skew, kurtosis))
attr(stats, "row.names") <- names(z)
stats

Figure 10.2: A scatterplot matrix for date, sample size, ASL and MSL for 134 films.

Here z and date import data from previously created structures; n, ASL and MSL are based on the SL data; other calculations are based on log-transformation. The quantities mu and sigma are estimates of the parameters, \( \mu \) and \( \sigma \), that define the lognormal distribution (see the Appendix); Lmed is the median of the log-transformed data, and Lmedscale is the median after standardizing the log-transformed data (see Section 12.2.2 for an analysis using this). The skew and kurtosis statistics are defined for log-transformed data; they are discussed at length in Section 12.2.3. They are obtained using the skewness and kurtosis functions from the moments package. Obviously many other statistics could as readily be calculated (e.g., the skewness, kurtosis, range, IQR, MSL/ASL ratio etc. of the untransformed logged data) but enough has been shown for illustration.

A pairs plot (i.e. scatterplot matrix) has been created within the function (for illustration). Something like FirstAnalysis <- SL.AllStats() will both save the data collected in stats and produce Figure 10.2.

The reasonable linear relationship between the ASL and MSL has already been discussed, and both decline as \( n \) increases (in a non-linear fashion), and over time, \( n \) tending to increase with date. (Remember that the upper triangle of the plot is just the lower triangle with axes reversed.) None of this is startling, nor is it intended to be; the aim is to show how easily and rapidly data exploration – to identify obvious patterns or unusual features, for example – is possible in R before undertaking anything more ambitious that such exploratory analysis should inform.
10.4 Hypothesis tests

There is a point of view that statistical hypothesis testing has no serious role to play in cinemetric data analysis (Salt, 2012). The argument is that SL data for a film constitute a population. ‘Classical’ methods of statistical inference that involve testing hypotheses about populations on the basis of (random) samples are argued to be inappropriate because the data cannot be treated as a sample. This view is possibly Draconian and introduces its own problems, discussed in the context of testing for lognormality of SL distributions in Section 12.3.

For the moment it is sufficient to note that statistical hypothesis testing has been used in some cinemetric studies. The intention here is to illustrate how this can be done in R, and how tests can be applied to many films at once. The focus is on tests of log-transformed SL distributions discussed in detail in Chapter 12.

The Shapiro-Wilk (SW) test is one of the more widely used tests of normality. The Shapiro-Francia test is an approximate version of the SW test; a ‘competing’ test of normality is the Jarque-Bera (JB) test. The rationale for these tests is discussed in Section 12.3; here it suffices to know they have been used to test the hypothesis that the log-transformed SL data from a film are sampled from a normal distribution. Taking Fantasia (1940) for illustration, \texttt{shapiro.test(log(SL.Fantasia))} produces

\begin{verbatim}
  Shapiro-Wilk normality test
  data: log(SL.Fantasia)
  W = 0.9963, p-value = 0.1093
\end{verbatim}

The ‘business part’, p-value = 0.1093, tells us that the null hypothesis of normality cannot be rejected at the 5% (or even 10%) level of significance. That is, we’re pretty happy to accept that the logged data can be treated as normal, and hence that the original SL data can be regarded as approximately lognormal. To get the p-value directly use \texttt{shapiro.test(log(SL.Fantasia))}$p.value$. The command

\begin{verbatim}
SWall.p <- apply(SL.ALL/10, 2, function(x) shapiro.test(na.omit(log(x)))$p.value)
\end{verbatim}

computes the p-values for all the films in the sample. To get the values of the test statistic itself replace $p.value$ in the above with $statistic$.

Statistics for the SF test can be produced in exactly the same way after loading the \texttt{nortest} package and replacing \texttt{shapiro.test} with \texttt{sf.test}. The JB test is available in several packages, the function name needed to execute it differing from package to package. If the \texttt{tseries} package is used replace \texttt{shapiro.test} with \texttt{jarque.bera.test}.
Chapter 11

Median or mean? (and more)

11.1 Introduction

The question *Median or Mean?* was chosen as the first subject for debate in the *On Statistics* discussion on the *Cinemetrics* website. Aspects of the debate have been touched on in Section 4.5 and elsewhere in these notes. The present chapter discusses the issues involved with an update on some of the ideas put forward and with additional case studies.

Salt (1974) proposed the ASL as a useful measure of ‘style’ in film – the title of the paper *Statistical Style Analysis of Motion Pictures* emphasizes that it is quantitative measures of ‘style’ that are of interest, and in this chapter only those based on SLs will be of concern. The ASL has the merit that it only requires knowledge of the length of a film and number of shots in it to calculate; the MSL requires the length of each shot to be measured and is thus more time-consuming to obtain. The consequence of this is that the ASL but not the MSL was used routinely for the purposes of statistical style analysis, until fairly recently.

I’m not aware that anyone who finds the ASL a useful measure of style has doubted that the MSL is also useful, but the opposite is not the case. This is most manifest in the writing of Redfern (2012d) who concludes that the ASL ‘is an obviously inappropriate statistic of film style’. Elsewhere on his website Redfern has claimed that the use of the ASL ‘as a statistic of film style leads to fundamentally flawed inferences about film style’ and been very critical of the work of scholars who have so used it. Should Redfern be correct the issue is not, therefore, a trivial one, given the number of analyses based on the ASL without the benefit of access to the MSL.\(^2\)

The case against the ASL is that it ‘does not locate the center of a shot length distribution and is not resistant to the influence of outliers’ (Redfern, 2012d). These are two separate issues that I shall refer to as the ‘argument from skewness’ and the ‘argument from robustness’. The latter is dealt with in Section 11.3. The argument from skewness – that the ASL doesn’t locate the center of a shot length distribution – is taken up here, and elaborated on in both theoretical and practical terms in Sections 11.2 and 11.4.

I take one aspect of statistical style analysis to be the computation of statistics descriptive, in some way, of the properties of an SL distribution. This includes the ASL, MSL and other statistics discussed in Section 4.6. I take it almost as axiomatic that the purpose of computing such statistics is for comparison, and that the merits of a statistic should be judged by their effectiveness for the comparative purposes to which they are put.

It will be argued below that when the ASL and MSL are used for the same comparative purposes either (a) they lead to the same conclusions, in which case it doesn’t matter which is used, or (b) they lead to different conclusions for reasons which, if explored, reveal more about an

---

1[^note1]: http://nickredfern.wordpress.com/category/statistics/page/4/ – *Some notes on cinemetrics IV.*

2[^note2]: My own view (Baxter, 2012a), elaborated on in this chapter, is that it is best to work with the ASL and MSL together. In the absence of the latter the question is whether or not the ASL really does lead to ‘flawed’ inferences, or whether the one statistic can be used as a surrogate for the other. If so the claims about the superiority of the MSL cannot be upheld.

113
SL distribution than either statistic does when used separately (often because neither statistic is an appropriate summary measure).

Crudely put, the argument from skewness is that SL distributions are generally skewed; that the ASL does not identify a ‘typical’ value because more SLs lie below than above it; that the MSL is ‘typical’ because half the SLs lie above and half below it; and that therefore the MSL but not the ASL is an appropriate measure of ‘style’. This would seem to equate ‘style’ with ‘typicality’ and there is usually an appeal to what might be called the ‘common perception’ of what is typical. The ASL and MSL, as well as the mode or modal class, are just different measures of location which will give different values if a distribution is not symmetrical. Divorced from the notion that the central value is the appropriate measure of ‘style’ – leading by definition to the MSL – any of these statistics can be proposed as a measure of ‘style’\(^3\). What matters is whether they do the job for which they are intended (i.e. are they useful).

Whether or not a statistic is useful depends on the shape of the distribution. Redfern’s (2012d) statement that the ‘median has a clearly defined meaning that is easy to understand, and fulfils precisely the role film scholars erroneously believe the mean performs’ is problematic. The ASL (mean) is also clearly defined; film scholars may or may not erroneously think the ASL is the center of a distribution in the way that the MSL is\(^4\), but it does not follow logically from this that comparisons effected with the ASL don’t perform the task they think it does.

The argument for the MSL seems to take it as axiomatic that the center of a distribution is what we should be trying to measure and that the median is the center. Thus, the argument is that the ASL is ‘wrong’ because it is not the median (i.e. center) and the median is ‘right’ because it is the median. I think this might be a tautology.

The argument from robustness – the potential effect outliers have on ASL but not MSL calculations – has more force and is open to empirical investigation. It is not obvious that outliers must be a ‘bad thing’. Redfern (2012d) contrasts *Lights of New York* (1928) with *The Scarlet Empress* (1934), which have similar means and different medians, to argue that outliers distort the ASL and once discounted lead, via the MSL, to the conclusion that *Lights of New York* is cut more quickly. Salt (2012) contrasts *Lights of New York* with *The New World* (2005) which have similar medians and different means, the ASL for *Lights of New York* being the larger. This is because the film has far more ‘outliers’ than *The New World* which Salt attributes to ‘technical constraints on shooting and editing synchronized sound’ operating in the early days of sound. Relying solely on the MSL would miss this important aspect of ‘style’. It can be argued (Baxter, 2012b) that the reason why the phenomena that Redfern and Salt observe arise stems not from outliers but from the unusually structured tail of *Lights of New York*. This renders both the MSL and ASL unsuitable as summary statistics, but attention is drawn to the structure by the joint use of the ASL and MSL. The lesson here, reiterated below, is that it is much more fruitful to use the ASL and MSL in tandem than to prescribe one as the more correct statistic.

### 11.2 Some theory

To set later discussion in context it is helpful to provide a little statistical ‘theory’. For the purposes of exposition an ideal model will be assumed, in which SLs follow a lognormal distribution exactly\(^5\).

An idealized lognormal distribution is completely determined by two parameters, \(\mu\) and \(\sigma\), which are the mean and standard deviation of the log-transformed SLs. These two numbers are necessary to describe the distribution and sufficient to describe it exactly. Figure 11.1 is based on two idealized films with parameters identical to two rather different real films that look to have distributions fairly close to lognormal. The film associated with the red lines mimics *Brief*.

---

\(^3\)If you insist on ‘typicality’ as a criterion and allow appeals to ‘common perception’ many would opt for the mode rather than the median.

\(^4\)The assertion is rather sweeping; it is unclear what the evidence is for it.

\(^5\)The extent to which this may be approximately the case for a majority of films is the subject of Chapter 12. Assuming the ideal means that the ‘complexity’ of summarizing an SL distribution – ‘statistical style’ if you wish – in even the *simplest* situation is made transparent. Any departure from the ideal increases complexity.
Encounter (1945) with \((\mu, \sigma) = (2.28, 0.89)\); the blue lines mimic Wedding Crashers (2005) with \((\mu, \sigma) = (0.92, 0.73)\). The ASL and MSL for the two films are \((14.6, 9.2)\) and \((3.3, 2.4)\).

![Idealized lognormal SL distributions and their logged values for two films with \((\mu, \sigma)\) given by \((2.28, 0.89)\) and \((0.92, 0.73)\) – the red and blue lines respectively.](image)

Whether the ASL, MSL or graphs are used it is fairly obvious that the second film is cut much more rapidly than the first, but conclusions are not always that simple. Barry Lyndon (1975), another film reasonably well-approximated by a lognormal distribution, has ASL and MSL values of \((14.0, 9.8)\), so in comparison to Brief Encounter conclusions about whether it is cut faster (the ASL) or slower (the MSL) depend on which statistic you choose to use – if you must use only one. For Barry Lyndon \((\mu, \sigma) = (2.31, 0.80)\). The value \(\mu = 2.31\) is greater than the 2.28 for Brief Encounter and if these are compared leads to exactly the same conclusions as a comparison of the MSLs. This is inevitable because of the mathematical relationship between the MSL (\(\Omega\)) and \(\mu\), because \(\Omega = \exp(\mu)\), so the film with the larger MSL will have the larger \(\mu\).

The expression for the ASL, \(\mu_L\), in terms of the parameters that define the distribution, is more complicated, and given by

\[
\text{ASL} = \mu_L = \exp(\mu + \sigma^2/2) = \Omega \exp(\sigma^2/2).
\]

The ratio of the ASL for two films can be expressed as \(\text{ASL}_1/\text{ASL}_2\) where the subscript identifies the film. If logarithms of this ratio are taken the result can be expressed as

\[
\log \text{ASL}_1 - \log \text{ASL}_2 = (\log \Omega_1 - \log \Omega_2) + 0.5(\sigma^2_1 - \sigma^2_2) = (\mu_1 - \mu_2) + 0.5(\sigma^2_1 - \sigma^2_2).
\]

It is instructive to look at some special cases. If \(\sigma_1 = \sigma_2\) then the ASL and MSL ratios are the same – that is, if you want to interpret them in terms of relative speed of cutting – they tell the same story.

If the MSL ratio is 1 then \((\mu_1 - \mu_2) = 0\). Whether the ASL ratio is greater or less than 1 depends on whether \(\sigma_1\) is bigger or smaller than \(\sigma_2\). A similar argument applies if the ASL ratio is 1 and differs from the MSL ratio. The situation is more complicated if the ASL and MSL ratios both differ from 1, since the stories that emerge depend on the \(\mu_i\) as well as the \(\sigma_i\) but the same idea holds. That is, even if SLs follow an exact theoretical distribution use of the ASL or MSL in isolation to compare different films can lead to different conclusions about relative cutting speeds, depending on the values of \((\mu_i, \sigma_i)\).

The real message here is that attempts in the literature to draw conclusions about the relative rapidity of cutting based on the ASL or MSL which assert that where the ASL and MSL lead to different conclusions results are ‘contradictory’, and which imply that one is ‘right’ and the other
'wrong', are misguided. Redfern (2012d), for example, observes that *Lights of New York* (1928) and *The Scarlet Empress* (1934) have similar ASLs (implying that they are cut equally quickly), but that *Lights of New York* has the smaller MSL implying quicker cutting. It is stated that these conclusions are ‘contradictory and cannot be true at the same time since they purport to describe the same thing’, but the statement is wrong.

It is wrong because it appears to be based on the assumption that ‘rapidity’ of cutting can be measured in some absolute sense, for which some statistics (the MSL in this case) are more appropriate than others. In the context of film, rapidity is relative to the measure chosen to represent it and different measures may be equally appropriate. To put this another way, the ASL and MSL do not ‘purport to describe the same thing’ – they are different ways of trying to summarize a ‘thing’ that does not have a uniquely ‘correct’ definition.

The problem is that the lognormal distribution can’t be summarized using a single statistic, and arguments that focus entirely on the relative merits of the ASL and MSL sometimes ignore this. The ASL and MSL are both simply different measures of location for the lognormal distribution, which is skewed and for which no single measure is appropriate. Both measures have been interpreted in terms of cutting ‘speed’, but this interpretation is imposed by the analyst and, in comparing two films, there is no logical reason for expecting them to lead to similar conclusions, even in the ideal case.

It is a lot easier to make judgments about location and scale (or shape) on a log-scale, as the right-hand panel of Figure 11.1 suggests. One film clearly has a much smaller value of $\mu$ than the other and also a smaller $\sigma$. The pair $(\mu, \sigma)$ completely describes the shape of the log-transformed data and also the untransformed data in the left-hand panel.

A ‘problem’ is that for mathematical purposes it is often convenient to work with logarithms but this is, in a sense, an abstraction. Films unravel in real time and, I suspect, most people are more comfortable dealing with quantitative descriptors that reflect some aspect of that reality. You need (at least) two descriptors that involve both the parameters $(\mu, \sigma)$ and have a choice. Baxter (2012a) noted that the pair (ASL, MSL) was one possible choice and that, in looking for broad patterns in data, it was more fruitful to use both together, rather than in isolation. This is pursued in Section 11.4. Before this some of the realities of practice, when SL distributions are very far from the ideal assumed in this section, need to be addressed.

### 11.3 Outliers and robustness

#### 11.3.1 General remarks

As just suggested, I suspect that most people who are interested in the subject are happier (or more accustomed to) dealing with graphical displays like those to the left of Figure 11.1 or, more usually, their empirical equivalent in the form of histograms of SL data. I also suspect that most people (myself included) have far less ‘intuition’ when it comes to dealing with skewed distributions such as the lognormal than with symmetric distributions such as the normal.

It is fairly obvious from the left-hand panel of the figure that the SLs for one film are generally smaller than those for the other and there is no argument about it being cut much faster, however one wishes to define ‘fast’. The films which these distributions mimic were chosen to be very different and such judgments are much harder to make if the distributions have ASLs or MSLs that are closer. In making finer judgments, and making a case for the MSL, the argument from robustness carries potentially more weight and is more frequently put. The essence of the

---

6It has already been noted that the MSL has been promoted in favour of the ASL because it is a more ‘appropriate’ measure of what a typical SL is. It is not obvious that it should be equated with ‘quickness’ of cutting in the way Redfern does. You can argue that an ‘obvious’ measure of ‘quickness’ is the cutting-rate – given two films of equal duration, $S$, the more quickly cut film is that with the greater number of shots, $n$, so the film with the larger value of $n/S$ has the faster cutting-rate. This is just the inverse of the ASL, smaller ASLs implying faster cutting. Arguments of these kind, in favor of either the MSL or ASL, ignore how the SL distribution may look and how SLs are distributed within a film.

7Everybody knows this, but the fact and its implications are sometimes forgotten in debate.
argument is that SL distributions are prone to outliers and that this distorts the ASL, making it an ‘unreliable’ or ‘misleading’ measure of location. That is, the ASL is not a robust statistic compared to the MSL which is unaffected by outliers.

What is to be understood by ‘robust’ in the context of the ASL and MSL was discussed in Section 4.6 but bears repeating here. If there are \( n \) shots in a film only one of these needs to be changed by an arbitrarily large amount to change the ASL by as much as you wish. The ASL thus has what is called a breakdown point of \( 1/n \). As \( n \) gets very large this tends to zero, which is as bad as you can get. It means that, in theory, the ASL can be very sensitive to changes in a single SL or, in other words, is an unstable measure of location. The MSL, by contrast, requires change in 50% of the data (0.5) to be modified, which is as good as you can get. This theoretical argument about the potential instability, or lack of robustness, of the ASL is – as far as I can tell – the main serious plank in the argument for using the MSL in preference to the ASL.

The theoretical argument can’t be gainsaid, but does it have much practical merit so far as the analysis of SL distributions is concerned? Among the points made in (Baxter 2012a,b) were that (a) the prevalence of outliers in SL data has been grossly exaggerated; (b) their effect on ASL calculations is usually asserted rather than demonstrated and, where outliers exist, they are often of limited importance in terms of their effect; (c) outliers, or what are designated as such, can have a substantial effect and lead to different conclusions from the MSL, but it is then more interesting to ask why this is happening rather than just asserting the superiority of one statistic over the other. Point (b) is pursued in a little more detail here.

Theoretical arguments about the effect of arbitrarily large changes in an SL on the ASL ignore the reality of film – you don’t have SLs that approach infinity. There are films like Russian Ark (2002) lasting 99 minutes that consist of a single shot, but most films are not like that. In the sample of 134 films from the database of Cutting et al. (2010) used elsewhere in these notes the maximum of 145,319 shots, at just over 6 minutes, is one of 361.7 seconds in Those Magnificent Men in their Flying Machines (1965). There are three shots longer than 4 minutes, 17 (0.01%) greater than 3 minutes; and 60 (0.041%) greater than 2 minutes. If we deal with (extreme) practicalities and add a 6-minute shot to Wedding Crashers, with \( n = 2194 \), the ASL increases from 3.3 to 3.5 seconds. This is not ‘arbitrarily large’; whether it matters would depend on the point of any comparison. If an extra 2-minute shot rather than 6-minutes is added there is no change (to one decimal place).

Obviously had a film with fewer shots been selected there would be a greater difference. Such films tend to be both earlier and have larger ASLs to begin with so the proportionate change will be equally undramatic. The film with the smallest number of shots in the database The Seven Year Itch (1955), with \( n = 233 \), has an ASL of 26.2. Add a 6-minute shot to this and it changes to 27.6; add a 2-minute shot and it changes to 26.6. It is the film with the largest ASL in the sample so the changes don’t affect this.

### 11.3.2 Outliers and Brief Encounter

Before worrying about the effect of outliers they need to be detected. There are formal statistical tests for doing this, but they mostly depend on assuming that a majority of data conform to an underlying probability model. If the lognormal distribution was generally applicable to SL data one approach might be to use tests based on a normal probability model applied to log-transformed SLs, but a consensus about the applicability of this model is lacking (Chapter 12). Consequently, more informal graphical methods are usually used. Differing views are aired in Baxter (2012a,b), Redfern (2012d) and Salt (2012). This section summarizes some of the discussion there with an update.

Brief Encounter (1945) will be used for illustration; nobody I know of has claimed this is much affected by outliers so it is ideal for illustrating what is problematic about methods that have been used in the cinemetric literature for detecting them. Several plots based on the film are shown in Figure 11.2. The upper-left plot for the SLs is of the kind often presented. It is difficult to make visual assessments about outliers from this kind of plot unless they are very obvious. For
a sample from a lognormal or other long-tailed skew distribution a long and ‘bitty’ tail is to be expected and it is easy to be misled into thinking that ‘bitty’ parts are outliers.

There are perhaps two observations in the tail that stand out as possible outliers. If the data are log-transformed, as in the histogram to the upper-right, there is nothing untoward to be seen. Since *Brief Encounter* is one of few films that comfortably passes stringent tests of lognormality (Chapter 12) it might reasonably be concluded there are no outliers of consequence.

There are three boxplots for the data shown in the lower part of the graph; the (mis)use of boxplots for detecting outliers has already been discussed in detail in Section 5.3.3, so discussion here can be brief. The left boxplot in the lower part of Figure 11.2 is the default produced by R. The ‘whiskers’ extend to a distance of $(1.5 \times \text{IQR})$ from the box defined by the quartiles $Q_1$ and $Q_3$, points beyond this being plotted separately. These have been used to identify outliers in SL data, typically only the upper tail being of interest. Redfern (2012d) equates outliers with points which are greater than $Q_3 + (1.5 \times \text{IQR})$, all those shown separately on the boxplot, and extreme outliers with points greater than $Q_3 + (3 \times \text{IQR})$.

With these definitions *Brief Encounter* has 28 outliers and 6 extreme outliers. This assessment is obviously excessive. All the supposed ‘outliers’ do is reflect what is to be expected from a long-tailed skew distribution. This misinterpretation in terms of outliers has contributed to an exaggeration of the prevalence of outliers in the cinematic literature. As noted in Section 5.3.3 it is possible to construct boxplots ‘adjusted’ for skewness. This is shown in the central boxplot and suggests, more sensibly, two possible outliers. The final boxplot to the right, using log-transformed data, suggests no real problems.

Taking all the evidence into account it can be concluded that ASL calculations for *Brief
Encounter are not seriously affected by outliers. If we take the pessimistic view that the two most extreme cases should be treated as such omitting them reduces the ASL from 14.6 to 14.0. There are three films, in the reference sample of 134 used here, with ASLs between 14.0 and 14.6, whose ASL ranking relative to Brief Encounter would change.

The effect of indisputable outliers on calculations can always be investigated. Given large \( n \) a single extreme outlier will have little effect; a shot of three minutes in a film with \( n = 500 \) will change the ASL by a little under 0.4 seconds. Only 16/134 films from the reference sample have a shot of three minutes or more, mostly in films with \( n > 500 \) and \( ASL > 10 \). The practical issue would be whether a change in the ASL of 0.3–0.4s in films with ASLs of this magnitude matters much for comparative purposes. It is easy enough to construct hypothetical examples with multiple outliers having a ‘large’ effect; where such instances arise in practice their designation as ‘outliers’ might be questioned as might the appropriateness of the ASL and MSL as summary statistics.

11.4 Case studies

11.4.1 ASL and MSL trends over time

In Figure 4.1 it was noted that there was a strong linear relationship between the MSL and ASL for 134 films from the database of Cutting et al. (2010) (product-moment and Spearman’s rank-correlations of 0.95 and 0.96). It might be expected that in looking for broad patterns in a large body of data it would not matter which was used. This is illustrated with reference to trends in the ASL and MSL over time in Figure 11.3 using these data.

![Figure 11.3: For 134 Hollywood films 1935-2005, at five-year intervals, plots of the mean of the yearly ASLs and median MSLs against date.](image)

It can be seen that the pattern of medians of the MSLs mimics that of the mean of the ASLs for each year, and this is what might be expected from the high correlation between the MSL and ASL. Arguments to the effect that the ASL is an ‘inappropriate’ measure of film style were cited in Section 11.1 and, if accepted, would lead to the conclusion that the ‘picture’ presented by the line for the MSLs is the more ‘correct’. In fact the ASL and MSL tell the same story of a fairly steady decline in SLs since the 1950s, with less simply summarized variation before then.
11.4.2 ASLs, MSLs and the silent-sound transition

The theme of the previous sections, that the ASL and MSL can be equally useful with the additional emphasis that they are more useful studied together, is pursued in this section. Some attention is paid to the MSL/ASL ratio, bearing in mind Salt’s (2011) comment that he considered it a ‘remarkable fact’ that in a sample of 1520 films 82% had an MSL/ASL ratio in the range 0.5–0.7. Figure 4.1 was based on films from the period 1935–2005. Here we go back in time to the period 1920–1933 covering the silent-sound transition.

The data are those used in Redfern’s (2012f) paper on the impact of sound technology on SLs, augmented with information on the ASLs. The data consists of two samples of 54 silent films from the period 1920–1928 and 106 films from the early period of sound, 1929–1933. A plot of the ASL against MSL is shown in Figure 11.4.

![Figure 11.4: A plot of ASL against MSL for 160 films from the period 1920–1933. The lines are robust regression fits of ASL against MSL for silent and sound films separately.](image)

Ignoring, for a moment, the silent-sound distinction and the fitted lines it can be seen that there is a fairly good correlation between the two statistics. The product-moment and Spearman’s rank-correlation coefficient are 0.91 and 0.91. Eight fairly obvious outliers are highlighted\(^8\) – all sound films – and if these are omitted the correlations change to 0.97 and 0.95. The outliers are the only ones where the MSL/ASL ratio is less than 0.5. Forty-three ratios are greater than 0.7 so 68% lie in the range 0.5–0.7, a number that rises to 85% if the range is extended to 0.75. Silent films account for 34% of the total sample, but 72% of those with an MSL/ASL ratio larger than 0.7. This observation is pursued shortly.

\(^8\)These are, to my eye, visually fairly obvious but the identifications can be justified more ‘objectively’ by analyzing the residuals from a regression analysis. Details are not presented – it makes no difference whether ordinary or robust regression is used.
Redfern (2012f) bases analysis on the MSL accompanied by a robust measure of dispersion $Q_n$ (Section 4.5). It is fairly obvious from Figure 11.4 that both the MSL and ASL tend to increase with the advent of sound. Redfern confirms the increase in the MSL is highly significant using the Mann-Whitney U test and estimates the increase as 2.0s with a 95% confidence interval of (1.6, 2.4). It is further noted that the variation in the MSL for sound films is somewhat greater than for silent films.

It would be possible to conduct a similar analysis for ASLs only, but is more interesting to observe that, if the ASL is plotted against the MSL, there appears to be some difference in the pattern for silent and sound films. This can be confirmed by fitting separate regression lines through the two sets of points, as in Figure 11.4, but the difference can be more directly appreciated by examination of the MSL/ASL ratio$^9$.

![KDEs of the MSL/ASL ratio for 54 silent films (1920–1928) and 106 sound films (1929–1933).](image)

Figure 11.5: KDEs of the MSL/ASL ratio for 54 silent films (1920–1928) and 106 sound films (1929–1933).

Figure 11.5 shows KDEs for the MSL/ASL ratio for silent and sound films, omitting eight outliers from the latter. It has already been noted that the smallest values tend to be associated with sound films and the largest with silent films and the figure confirms the general pattern. A formal statistical test is not really needed to confirm the difference. If one is wanted, the distributions are sufficiently normal to justify the use of a two-sample $t$-test, which confirms that the difference in mean MSL/ASL ratios is very highly significant with a 95% confidence interval of (0.061, 0.010).

One possible reason for the difference may be the effect of titles of various kinds in silent films, which are included in the ASL and MSL calculations. The measurements for the silent films in the Cinemetrics data base are a mixture of simple and advanced mode, so information on the lengths of title shots is only available for some films. If they tend to be on the shorter side their removal would increase the MSL and ASL but would be likely to have more effect on the ASL, bringing

$^9$The regression lines are robust fits based on the lmrob function in the robustbase package. In effect this means that the more obvious outliers have little or no effect on the fitted lines. One or two silent films, not outlying in the context of the total data set, are suggested as possible outliers relative to other silent films because of their larger ASLs. The regressions are of the ASL against MSL and would change if these roles were reversed, but the general picture remains the same.
the MSL/ASL ratios more in line with those for sound films. For example, *The Kid* (1921) has 441 shots and an MSL/ASL ratio of 0.66. The majority of shots (382) are action, with the ASL and MSL for the 25 title shots being about half of the values for action shots. For the film as a whole, ASL = 6.7s and MSL = 4.4s. Omitting the titles increases the ASL to 6.9s and the MSL to 4.5s and reduces the MSL/ASL ratio to 0.64. This is as predicted but, if repeated exactly across all silent films, would still leave a difference between them and the sound films; more analysis (and data) than is possible here would be needed to pursue this further.

Another reason why MSL/ASL ratios might be smaller for early sound films has been discussed by Salt (2012) who noted that ‘Like many films made at the very beginning of the use of synchronized sound, *The Lights of New York* is a mixture of scenes done in long takes shot with direct sound, and action scenes that were shot wild and post-synchronized. The use of normal fast cutting in the latter was of course unconstrained by technical factors’. The need for long takes with direct sound is attributed to ‘technical constraints on shooting and editing synchronized sound’ operative in the early days of sound. The statistical effect of this, relative to silent films, would be to increase the ASL, affecting the MSL less, and reduce MSL/ASL ratios.

*The Lights of New York* is a 1928 film. Salt’s use of the term ‘very beginning’ raises the possibility that any effect on the MSL/ASL ratio may be date dependent. Figure 11.6 shows boxplots of MSL/ASL for the 54 silent films and for sound films for each of the years 1929–1933 with (20, 26, 21, 23, 16) films (outliers are included).

<table>
<thead>
<tr>
<th>Year</th>
<th>MSL/ASL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent</td>
<td>0.3</td>
</tr>
<tr>
<td>1929</td>
<td>0.4</td>
</tr>
<tr>
<td>1930</td>
<td>0.5</td>
</tr>
<tr>
<td>1931</td>
<td>0.6</td>
</tr>
<tr>
<td>1932</td>
<td>0.7</td>
</tr>
<tr>
<td>1933</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Figure 11.6: *Boxplots of the MSL/ASL ratio for silent films and sound films from 1929-1933.*

That the ratio tends to be larger for silent films is clear; at least half of the silent films have ratios greater than three-quarters of the sound films for any of the years. There is no consistent pattern for the sound years; 1929 tends to have lower values than other years, except that it is matched by 1932. It may be worth mentioning, though details are not presented, that if analyses similar to those above are carried out for the films from 1935–2005 in the Cutting et al. (2010) database patterns comparable to those for the sound films from 1929–1933 are obtained. The chief difference is a longer left-tail of ratios less than 0.5, associated with 13 films (of 134) in total, from the period 1935–1975. Only six films have a ratio greater than 0.75 and these are spread across the years.

A thesis of Redfern (2012f) is that the advent of sound can be associated with a significant
increase in SLs, but as measured by the ‘typical’ MSL is much less dramatic than suggested by a ‘typical’ ASL\textsuperscript{10}. Specifically he states that his ‘results support the conclusions of earlier studies that the shift from silent to sound cinema led to an overall increase in shot lengths but the size of this effect is shown to be much smaller than that described by studies using the mean shot length’ (Redfern, 2012f, p.86). The view taken here is that it does not really make sense to try and pin a single exact number on the size of the effect. It is relative to the measure used to define it, and both the MSL and ASL tell the same story that the size increased.

Given the strong correlation between the MSL and ASL it should be generally the case that any patterns in the data associated with the one will be identified by the other. If the two statistics are examined together other features than a pure size effect – the change in ‘typical’ MSL/ASL ratios – that invite explanation are revealed that analysis of the ASL or MSL in isolation will not detect.

Joint analysis of the ASL and MSL statistics also identifies detail, in the shape of individual films that depart noticeably from the general pattern, that invites explanation. Of the eight outliers in Figure 11.4 by far the most extreme is \textit{Rain} (1932)\textsuperscript{11}. Charles O’Brien’s analysis of this in the Cinemetrics database lists the film as having 308 shots, with ASL = 18.0s, MSL = 5.1s and MSL/ASL = 0.28. There are 138 ‘action’ shots, 169 ‘dialog’, and 5 ‘singing’.

Some of the dialog shots are long; four are over two minutes with one close to seven minutes. The ASL drops to 14.9 if these are omitted and MSL/ASL rises to 0.34. This is still extreme so the unusual nature of the film cannot be attributed solely to ‘outliers’\textsuperscript{12}.

There are few ‘singing’ shots so they will be ignored in what follows. The MSL and ASL for the action shots are 2.5s and 5.0s with a ratio of 0.5; for dialog shots the figures are 11.8, 28.9 and 0.41 changing to 10.8, 23.3 and 0.44 if shots longer than 2 minutes are omitted. Figure 11.7 is based on log-transformed SLs and shows a KDE using all the data with separate KDEs for dialog and action shots superimposed.

The KDE using all the logged data is very ‘lumpy’, in the sense discussed in Baxter (2012a). There it is suggested that neither the ASL or MSL would be an appropriate summary statistics for films with this kind of distribution. The major mode to the left is associated with the peak for action shots; the central ‘shoulder’ and minor mode to the right are associated with modes in the bimodal distribution for dialog shots. Similar analyses can be undertaken for \textit{Applause} (1929), \textit{Movie Crazy} (1932) and \textit{Trouble in Paradise} (1932) which are the three other films with an MSL/ASL ratio of less than 0.5 for which shot type information that distinguishes between ‘action’ and ‘dialog’ is available.

None of these three films, or \textit{Rain}, are especially suited to summary using either the ASL or MSL because of the marked difference is distributions of the action and dialog shots, which induces very ‘lumpy’ patterns in the (logged-)SL distributions. The dialog shots are typically located at much larger SL values (whatever measure of location you choose to use), and for three of the four films have distributions that are multi-modal or close to it. For three of these four films (\textit{Applause} is the exception) the MSL and ASL values for action shots only would place them comfortably in the cluster of points associated with silent films in Figure 11.4.

It is legitimate to wonder if the proportion of sound films with characteristics like those discussed is sufficient to render attempts to summarize the effects of the introduction of sound in a single statistic (whether based on the ASL or MSL) potentially misleading. If action shots in early sound films have similar MSL and ASL characteristics to those displayed by silent films then for sound films largely dominated by action and dialog the rather prosaic conclusion might be reached that the difference in SL lengths – which all seem to agree exists – arises because sound films have more (audible) dialog than silent films!

\textsuperscript{10}I’m deliberately not defining ‘typical’ to avoid discussing whether a typical MSL or ASL should be measured by the median or mean!

\textsuperscript{11}The residual from the robust regression is 10.14 compared to 6.10 and 6.08 for the next most extreme, \textit{Applause} (1929) and \textit{The Front Page} (1931).

\textsuperscript{12}If \texttt{adjbox} is used to produce a boxplot adjusted for skewness (Section 11.3.2) only two ‘outliers’ are suggested.
11.4.3 ASLs and MSLs in D.W. Griffith’s one-reelers

Salt (2011) suggests ‘it is probable that about half of all films, with an ASL [Average Shot Length] less than 20 seconds, conform well to the Lognormal distribution’. In the same article he also suggests that there are interesting empirical regularities in the relationships between statistics descriptive of SL distributions, among them the MSL and ASL discussed above. In the context of the MSL/ASL ratio he comments that that the fit to the lognormal distribution begins to break down at an ASL of 15.

We need not insist on lognormality, but can abstract from Salt’s observations and interpret them to mean that there are fairly stable relationships between statistics descriptive of SL distributions that break down at higher ASLs, leaving the definition of ‘higher’ a little vague. Analyses in the previous two sections broadly confirm the stability of the MSL/ASL relationship, with some subtleties of variation and inevitable exceptions to the general rule. To see how far this can be pushed we go even further back in time to the 1909–1913 period when D.W. Griffith was at Biograph and explore a sample of 62 of his one-reelers from that period.

These have been the subject of a study by Baxter (2013d), where the focus was on internal cutting patterns. The films are mostly of about 1000 feet in length, with none less that 920 feet, so some control is exercised over their duration. Over the period 1909 to 1913 Griffith’s cutting-rate, however measured, tended to increase. Give the relative constancy of duration of the films this means that the number of shots in the films, \( n \), also tended to increase. In the sample \( n \) varies from 22 to 135. All 14 of the 1909\textsuperscript{13} films in the sample have \( n \) less than 50, a property that characterizes five of the seven earliest 1910 films in the sample, which come from the first half of that year, and three late films from that year. The value of \( n \) for 1910 films is quite variable; for

---

\textsuperscript{13}Dates are dates of production as given in Henderson (1970).
later films only *The Indian Brothers* with \( n = 55 \), has a value less than 69. That is, with this exception there is a difference in \( n \) of 23 between the maximum for 1909 and the minimum for 1911–1913, and differences are generally somewhat greater. This is reflected in the ASL and MSL values.

![Figure 11.8: A plot of ASL against MSL for 62 of D.W. Griffith’s Biograph one-reelers.](image)

Figure 11.8 shows a plot of ASL against MSL for these films labelled by production date. Sectioning the graph at an ASL of just over 15s and an MSL of 10s neatly separates all the 1909 films from all the 1911–1913 films. Films in the south-west quadrant exhibit something of the linear relationship found between the MSL and ASL for the samples of later films studied in previous sections; there is only a weak relationship between the MSL and ASL for films outside this quadrant (all but 1 with ASLs greater than 15).

All but one of the films in the south-west quadrant have \( n = 69 \) or more, the exception having \( n = 55 \). Only three films with \( n > 55 \), all 1910, lie outside the quadrant, two lying very close to it and which fit without problems into the pattern in the quadrant. The pattern that emerges here is of a fairly volatile relationship between the ASL and MSL in the early films, going into 1910, associated with ASLs greater than 15s and with fewer than 50 shots. This evolves into a more stable pattern during the 1910s that ‘crystallizes’ in 1911 when the number of shots regularly hits 70 or more. A closer look is now taken in Figure 11.9 at the pattern for those films with at least 55 shots.

The product-moment and Spearman’s rank-correlation between the ASL and MSL are 0.89 and 0.91, slightly lower than for the bodies of films examined in previous sections. The 1910 film in the top-right, *The Usurer* is an outlier, but nothing else stands out. Interestingly, there is the suggestion of subsets of films with different MSL/ASL patterns. Specifically there is a cluster of 11 films in the lower left with smaller ASL and MSL values that sit within the range occupied by the silent films from 1920-1928 from Figure 11.4. The robust regression fit for this latter group
of films is superimposed on Figure 11.9 and fits this group of 11 well but not, with two or three exceptions at larger values, other films on the graph. Most other films are better described, though not so tightly, by the robust regression fit for early sound films.

Other than the 1910 film (A Child’s Stratagem, with \(n = 86\)) in the group of 11 the remaining films have \(n\) greater than or equal to 92, and are 10/13 films with this number of shots. It is tempting to suggest that in the latter half of his Biograph career Griffith, in those films where he used about 90 shots or more, evolved a style – as far as the MSL/ASL ratio was concerned – that anticipated the pattern that emerged in later silent film when features were the norm. This suggestion is based on a small sample of mainly 1911–1913 films and obviously needs much further investigation, using both more of Griffith’s films from the period and films from other directors from the period 1910–1919.

Why other 1910–1913 films with \(n\) mostly in the range 70-90 follow an MSL/ASL pattern more akin to sound films (both from the 1929-1933 period and later) is hard to explain. It may be some kind of ‘generic’ pattern but that is not an explanation. The reasons that have been suggested for a possible change in pattern for early sound films – that they include dialog and dialog shots were much longer than action shots which remained much the same as in the silent era – does not apply to the Biograph films.

---

14 The three films not in the group plot just above it at the lower MSLs. These three are The Mender of Nets \((n = 108)\), The Adventures of Billy \((n = 108)\) and The Lonedale Operator \((n = 107)\). Films in the group of 10 are The Battle, Fighting Blood, Swords and Hearts, Billy’s Stratagem, Friends, The Girl and Her Trust, The Lesser Evil, The Unseen Enemy, Death’s Marathon and A Beast at Bay.
Chapter 12
Lognormality of SLs – a case study

12.1 The lognormality issue

In the beginning were these words

“Looking at the frequency distributions of shot length, a considerable similarity of overall shape is apparent. This is a surprise . . . The profiles of the distributions approximate in nearly all cases to that of the Poisson distribution . . .” Salt (1974, p.15)

with subsequent research replacing the Poisson by the lognormal as the preferred model

“The generality of the Lognormal distribution for shot lengths in movies is illustrated by some examples from films made between 1916 and 1998. . . . For ASLs above 20 seconds, the fitting of the actual shot length distributions to the lognormal distribution is markedly less good, and an example of this is Carmen Jones (1954), which has an ASL of 46 seconds [and for which] the Gamma distribution proves a slightly better fit . . . but is still not really good.” Salt (2006, pp.391-394)

‘generality’ later being qualified to mean

“it is probable that about half of all films, with an ASL [Average Shot Length] less than 20 seconds, conform well to the Lognormal distribution. It is worth noting that of the remainder, a substantial proportion just miss out on being Lognormal.” Salt (2011)

and with dissenting voices to be heard.

“while the lognormal distribution is an adequate model for a handful of films, this is not the case for the vast majority of films . . . there is no justification for the claim that the lognormal distribution is an appropriate parametric model for the shot length distributions of Hollywood films.” Redfern (2012a)

The two views enunciated here could be described as polarised. It is interesting to ask why. Both protagonists base their positions on extensive empirical analysis of essentially similar material. Both, it will be argued, use statistical methods to support their position that have many similarities.

Redfern’s (2012a) paper appeared in the journal Literary and Linguistic Computing and provoked responses from Baxter (2013a) and DeLong (2013). These papers form the basis of much of this chapter. All three authors base analyses on 134 of the 150 films in the database of Cutting et al., excluding those with SLs recorded as zero or negative. The supplementary material for DeLong (2013) usefully provides graphical summaries for all 134 films – histograms of the raw and log-transformed SL data, and probability plots of the log-transformed SLs.
I take Salt’s (1974) initial position to have been that many SL distributions possess what I shall call distributional regularity in the sense that they have a similarity that goes beyond the purely qualitative (singly-peaked, with long right tails), which can be expressed mathematically in terms of a common model that fits a majority\(^1\) of SL distributions at least approximately.

The term distributional regularity is used deliberately. The lognormal is what I’ve seen described as the ‘ruling’ model; however, Salt’s (1974) first thoughts were of the Poisson distribution; he has explored the Gamma distribution (Salt, 2006, p.394); and others have considered the Weibull/Erlang and Pareto distributions (e.g., Taskiran and Delp, 2002; Vasconselos and Lippmann, 2000). That is to say, the lognormal distribution does not enjoy any kind of monopoly, even if it is the ‘ruling’ cinemetrics model.

It might therefore be a good idea to decouple the idea of distributional regularity from that of lognormality, though the latter is explored in some detail below. The idea of distributional regularity is an important one; if it can be sustained all the benefits to be derived from successful pattern recognition accrue (Chapter 1). The idea of lognormality, and beyond, is explored in a fairly leisurely fashion below, partly to illustrate how useful R can be for exploratory data analysis at various levels. The exploration is, however, directed at examining the merits of what might be called an ‘hypothesis’ about SL distributions. Nobody believes that all SL distributions are lognormal, and some films are obviously not. Some examples were provided in Section 6.4.5 (and Figures 6.8 and 6.10) and this chapter provides a more thorough discussion of issues raised there.

12.2 Exploratory analyses

12.2.1 Graphical inspection

Section 6.4.5 illustrated the use of KDEs for comparing SL distributions after log-transformation, in order to make judgements about lognormality. The idea is that if SLs have a lognormal distribution then they are normal after log-transformation, and it is much easier to make visual assessments about normality than lognormality. It is possible to superimpose fitted lognormal or normal models on the graphs to aid interpretation, as is done in the supplement of DeLong (2013) who emulates Redfern (2012a) in the additional use of probability plots (Section 6.4.3).

Figures 12.1 and 12.2 are similar to DeLong’s graphs for Harvey (1950) and Brief Encounter (1945), perhaps the least and most lognormal of the films in the sample. It is fairly easy to see, from histograms of the untransformed SL data – with which Salt (2006) tends to work – why the idea that SL distributions were frequently lognormal arose. Typically they are skewed with a single peak and long tail, all characteristics of the lognormal and other distributions that have been proposed as a model. It is, however, quite difficult to make visual assessments about lognormality. The concentration of the mass at the lower SLs makes it difficult to assess departures from lognormality in the visually less dominant right portion of the plot. It is easier to assess departures from a normal model, and if log-transformed SLs are examined evidence of non-normality implies that the untransformed data are not lognormal. This is evidently the case for Harvey.

To my eye – and these judgments are subjective – about a quarter of the log-transformed SL distributions presented by DeLong don’t look particularly normal because of what I have elsewhere called their ‘lumpiness’ (Baxter, 2012a) or their skewness. Judgments about this kind of thing can be affected by the bin-widths and anchor points used for the histograms and the use of KDEs is an alternative (Sections 6.4.5, 6.4.5)\(^2\).

One problem with histograms (and KDEs) is that they may not be very good at picking up aberrant tail behavior, for which probability plots are better (Section 6.4.5). If the data are normally distributed the plot should follow a straight line, as Brief Encounter does and Harvey

\(^1\)Meaning anything more than 50%, to be clear.
\(^2\)The histograms in Figures 12.1 and 12.2 differ in some respects from those of DeLong (2013). A probability rather than frequency scale has been used and he truncates his histograms for the untransformed SLs at 50 seconds. The interval used also differs – illustrating, in fact, the effect, that this can have.
Figure 12.1: Histograms (on a probability scale) of the raw and log-transformed SLs for Harvey (1950), and a normal probability plot of the log-transformed SLs. Fitted lognormal and normal curves are shown for the raw and log-transformed SLs respectively.

Figure 12.2: Histograms (on a probability scale) of the raw and log-transformed SLs for Brief Encounter (1945), and a normal probability plot of the log-transformed SLs. Fitted lognormal and normal curves are shown for the raw and log-transformed SLs respectively.

does not. Apart from non-normal tail behavior the general shape of the plot can indicate how a distribution departs from normality, though this can be difficult to interpret in isolation from other methods. If the issue is whether or not the data are normal, hypothesis tests based on the probability plot are available; roughly, they are based on statistics that measure the departure of the plot from ‘perfect straightness’. Such tests are an important plank in Redfern’s (2012a) argument that SL distributions are typically not lognormal, and it is his interpretation of these that both Baxter (2013a) and DeLong (2013) dispute.

Though he does not quite express it this way DeLong makes the interesting suggestion that a more appropriate model for many films may be a lognormal distribution truncated at about 0.5 seconds. The idea here is that there is a ‘kind of perceptual floor that limits how many frames can be in a shot’ and 0.5 seconds is proposed for this floor. If a sample is from a log-normal distribution the expectation is that a normal probability plot of log-transformed data will show evidence of non-normality in the form of upward curvature in the left-tail of the plot. The thinking is that rejection of lognormality in the tests used in Redfern (2012a) may often be occurring for this reason so lognormality with truncation remains a reasonable model.

12.2.2 Comparing estimates of the median

Another plank in Redfern’s (2012a) argument was the observation that there was a mismatch between two estimates of the median of a lognormal distribution that could be expected to be similar if lognormality applied (see the Appendix). If a set of data is exactly lognormal then the median (MSL) and the geometric mean (GM) of the data are identical. The GM is $\exp(\mu)$ where
\( \mu \) is one of two parameters that define the lognormal distribution, the mean of the data after log-transformation.

If real data follow lognormality approximately the MSL and GM should be similar. The left-hand plot of Figure 12.3 is of the GM against the MSL, with a reference line (MSL = GM) about which points should cluster if lognormality applies generally\(^3\). They don’t. The graph can be viewed as compelling evidence against assertions about the prevalence of lognormality\(^4\).

![Graph showing GM against MSL with reference line](image)

**Figure 12.3:** Plots based on different estimates of the median of SL distributions (untransformed and log-transformed) bearing on the question of log-normality.

Before rushing to judgment though, some caveats need to be entered. One, previously noted, is that some of the films have SL distributions for which neither the MSL nor ASL are especially suitable summary statistics. Another is that where the difference between the MSL and GM is small the possibility that the lognormal is an adequate approximation to the distribution is not precluded. Redfern’s (2012a) film-by-film demonstration that only 8–10 films are acceptably lognormal, using tests akin to the Shapiro-Wilk test, is problematic for reasons discussed below.

As evidence against assertions about the generality of lognormality – interpreted as ‘strict’ rather than ‘approximate’ – what is ‘compelling’ about the plot is the pattern of departure from lognormality, rather than what is happening with individual films, specifically that the GM is mostly larger than the MSL. This can be viewed informatively, in another way, on a log-scale, as now discussed.

\(^3\)This can be looked at in two ways. The SL data can be thought of as a sample from a population whose median is of interest. The observed sample median and GM are both estimates of the population median. If the data are lognormal, and for large enough samples, they should give similar estimates. Non-statisticians might think the sample median is ‘obviously’ the best estimator to use, but statistical theory does not support this.

If the idea of SL data as a sample is eschewed, on the grounds that such data constitute a population (Salt, 2012), then equality in the ideal case between the MSL and GM is a mathematical fact. If they differ at all systematically with real data an explanation is needed. That the data aren’t generally lognormal is one, while leaving ‘approximate’ lognormality open as an option.

\(^4\)The left-hand plot in Figure 12.3 is Figure 1 in Baxter (2013a). Delong’s (2013) Figure 6 is similar but with a reference line designed to emphasize the strong correlation between the MSL and GM rather than the fairly ‘systematic’ difference between them. Redfern (2012a) notes that GM > MSL in all but five cases, without pursuing this in any detail, and bases some of his analysis on what is called a ‘consistent ratio’ \( \max(\text{GM, MSL})/\min(\text{GM, MSL}) \) which discards information about which is the larger of the two statistics.
12.2.3 Skewness and kurtosis I

Skewness and kurtosis will be dealt with in two stages. In this section they are defined and ways in which they might be used in an exploratory fashion presented. Their further exploitation using hypothesis testing ideas is evaluated in the next section.

If a distribution is lognormal then it is skew and the ASL is larger than the MSL. After log-transformation, if the data are exactly normal, the distribution will be symmetrical and the mean and median of the transformed data should be the same. If the transformed data are subsequently standardized to have a mean of zero then the median after standardization would also be zero. The right-hand plot of Figure 12.3 shows that with six exceptions the median of the standardized log-transformed data is less than zero. This shows, in a different way, information similar to that in the left-hand plot.

One of the purposes of log-transformation, if the data are lognormal, is to remove the skewness. The plot suggests that in general it is failing to do this; skewness of the kind that should be removed remains after transformation. This kind of ‘failure’ is illustrated well for A Night at the Opera in Figure 5.2. The possibility remains that for many films the departure from symmetry after log-transformation may be weak enough not to compromise assertions about approximate lognormality, but the fact remains that there is strong evidence to suggest reasonably systematic departures from the posited ideal.

This can be pursued further, and more directly, by looking at the skewness and kurtosis of the log-transformed data. Some ‘technical’ discussion is needed here. Assume that we are dealing with reasonably smooth distributions having a single peak with the normal as the ideal model of a singly-peaked symmetric distribution whose spread is neither too wide or narrow\(^5\). Departures from symmetry relative to the normal can be defined and measured in different ways, dealt with in reliable introductory statistics texts. In the skew function from the moments package, that used here, skewness is measured as

\[
S = \frac{\sum(x - \bar{x})^3/n}{\sum\{(x - \bar{x})^2/n\}^{3/2}}
\]

which should be close to zero for normal data (e.g., Thode, 2005, pp.45-46). For untransformed SL data this can be expected to be strongly positive; 3.23 and 5.83 for Brief Encounter and A Night at the Opera respectively, for example, which doesn’t tell us very much. After log-transformation the values for the two films are 0.06 and 0.77. There is little dispute that Brief Encounter is well-approximated by a lognormal distribution so the log-transformed data are reasonably normal, and 0.06 counts as ‘close to zero’; 0.77 for A Night at the Opera does not, what it means in terms of the distribution is best appreciated by looking at Figure 5.2.

Skewness is a fairly simple idea; kurtosis less so (Balanda and MacGillivray, 1988; DeCarlo, 1997). Balanda and MacGillivray (1988) note that in some ways it is a ‘vague’ concept, often ‘operationally defined’, in sample terms, as

\[
b_2 = \frac{\sum(x - \bar{x})^4/n}{\sum\{(x - \bar{x})^2/n\}^2}
\]

which is expected to be 3 for an exactly normal distribution. It can be defined in terms of what is sometimes called excess kurtosis as \(K = (b_2 - 3)\), which should be zero for a normal distribution.

Distributions with negative kurtosis, \(K < 0\) are sometimes referred to as platykurtic, and with negative kurtosis, \(K > 0\), as leptokurtic. These terms are, respectively, also sometimes equated with distributions that are squatter/shorter-tailed and higher/longer-tailed relative to the normal\(^6\). There can, however, be problems in equating positive and negative kurtosis with shape, so here \(K\) and \(S\) will be used mainly as indicators of degrees of departure from normality.

---

5Not ‘technical’ terms.

6It’s easy to forget this. A popular mnemonic of longstanding is to think in terms of the shapes of duck-billed platypuses and kangaroos that were supposedly known for their ‘lepping’. According to the dictionaries lying around ‘lep’ is obsolete Scottish dialect for ‘leap’ (OED) or Irish dialect (Chambers). You probably don’t want to know this.
Figure 12.4 is a plot of the skewness against kurtosis for log-transformed data for our sample of 134 films. If lognormality applies, reasonably generally and approximately, then the log-transformed data should cluster around the (0,0) point, since the transformed data should be approximately normal with \((S, K)\) close to these values. They don’t; in particular, with three exceptions all the distributions exhibit positive skew. The issue of whether they are close enough to (0,0) to be treated as approximately normal remains, and is discussed in the next section.

The figure is really just another way of displaying the systematic departure from what is expected under lognormality, manifest in Figure 12.3, but there is additional information. The kurtosis values spread to either side of zero, but the labeling shows that the later films (defined a little arbitrarily as 1975 and later) tend to have positive kurtosis and earlier ones tend to have negative.

Additionally, films that I’ve deemed to be ‘unusual’ are highlighted. This identification was based on inspection of KDEs of log-transformed data where the distribution was sufficiently ‘lumpy’ to cast doubt on the ASL, median, skewness and kurtosis as sensible summary statistics (see Baxter, 2012a, for the general idea). Some of the judgments (e.g., Harvey, Pursuit to Algiers) are fairly indisputable; others might be questioned (see the discussion of Baxter, 2012b) (the reader is spared the details). The important point to note is that, even allowing for potential differences of opinion about how obviously ‘unusual’ such films are, with one exception they belong to the earlier group, and with one exception have negative values of kurtosis.

Two things are, arguably, happening here. One is that the values of skewness are displaying a pattern counter to the idea that a majority of SL distributions can be interpreted as ‘approximately’ lognormal. The other is that, with reference to kurtosis, there is a pattern of sorts in the difference between ‘earlier’ and ‘later’ films, different from the much-remarked on decline in ASLs and not logically entailed by this decline.
12.2.4 Skewness and kurtosis II

The problem with a lot of this, and any attempt to assess claims about the generality of ‘approximate’ log-normality, lies in the usually ill- or un-defined nature of what exactly is to be understood by the term ‘approximate’. Here we examine how an hypothesis testing approach might be exploited using skewness and kurtosis statistics. This provides both a prelude to a fuller discussion of Redfern (2012a), and also illustrates some well-known but sometimes neglected pitfalls involved in using such an approach.

Theory concerning the use of $S$ and $K$ to test the (null) hypothesis that a set of data is drawn from a normal distribution is summarized in Chapter 3 of Thode (2002). Only separate tests of $S$ and $K$ are considered in this section; joint tests are touched on in the next section.

The results of an hypothesis test are affected by the sample size used; other things being equal, the larger a sample size, $n$, is the more likely a null hypothesis is to be rejected. The statistic $S$ can be used as a basis for tests of the null hypothesis that population skewness $\beta_1 = 0$. Since $\beta_1 = 0$ is expected of a normal distribution, rejection of the null hypothesis amounts to rejection of the normality hypothesis; failure to reject is consistent with the idea of normality but does not guarantee it. Similar comments apply to $K$, defined to be zero if the data are perfectly normal.

How large $S$ and $K$ need to be to reject the null hypothesis of normality depends both on the sample size and the level of significance (decision rule) an analyst chooses to use. For small samples the theory is quite messy, but for large samples simplification is possible. For $S$ and using a 5% decision rule you’d reject normality if $S$ lies outside the range $(−1.96\sqrt{6/n}, 1.96\sqrt{6/n})$. For a 1% rule substitute 2.58 for 1.96; for $K$ substitute 24 for 6 in either case.

Since the number of shots in a film varies – about 10% in our sample have 500 or less, with the median a bit over 1000 – broad guidelines will be used in the plots to come assuming sample sizes of 500 and 1000. Thus, do not reject the null hypothesis of normality, with the sample sizes and significance levels indicated if films lie within the limits provided; the regions so defined are color-coded in Figure 12.5.

For skewness

1. $n = 1000$; significance = 5%; limits (-0.15, .15); color = pink
2. $n = 500$; significance = 1%; limits (-0.28, .28); color = light-green

For kurtosis

3. $n = 1000$; significance = 5%; limits (-0.30, .30); color = pink
4. $n = 500$; significance = 1%; limits (-0.57, .57); color = light-green

Films in the pink zone are acceptably lognormal as judged separately by both $S$ and $K$ at the 5% level of significance, for films with 1000 or fewer shots. Conditionally on the sample size, the use of a 5% rule for deciding not to reject the hypothesis of normality is quite strict and also arbitrary (even if explicit); a 1% rule is more forgiving in that you need larger values of $S$ and $K$ to reject normality. The light-green shown is at the generous end in using a 1% rule assuming samples of 500.

The main points to be made are that conclusions about the generality of lognormality are dependent on both the level of significance chosen and the sample sizes involved. Ignoring the dictates of statistical rigor and using the guidelines given above simply as ‘rules-of-thumb’ you can be strict, stay in the pink zone, and conclude that relatively few films can be regarded as having lognormal $SL$ distributions (about 7%). This is more-or-less what Redfern (2012a) does using more rigorous hypothesis testing procedures (Section 12.3). With an eye on, and sympathy for, the rather undefined notion of ‘approximate’ one could take a generous view and work in the light-green zone, concluding that about a quarter of the films are approximately lognormal. This is still somewhat less than the ‘majority’ claimed by Salt, and the issue will be revisited.

These large-sample results are adequate for $S$ with most of the sample sizes used here; they are less satisfactory for $K$, but good enough for the broad-brush approach being adopted.
Figure 12.5: A plot of skewness against kurtosis for 134 1935-2005 Hollywood films. See the text for an explanation of the shading.

It needs to be mentioned that for large enough sample sizes any null hypothesis will be rejected, however trivial the departure of what is observed from what is being tested. Plenty of statisticians would regard tests of normality with the sample sizes available here as largely a waste of time (some having said as much in print). This is different from Salt’s (2012) objection to using hypothesis tests on the grounds that a population rather than sample is involved, but if the ‘sample’ is large its properties will be very close to that of the population, so the argument can be viewed as an objection to using hypothesis tests with large samples.

12.3 Omnibus tests of (log)normality

Redfern (2012a) robustly opposes Salt’s view that a majority of films, which need not be much more than 50%, have approximately lognormal distributions. I want to explore Redfern’s methodology mainly to prepare the ground for my own later explorations but readers can also view this as an illustration of how statistical tests can be rapidly and repeatedly applied to a large body of films using R, even if the specific application is of less interest.

The skewness and kurtosis statistics, $S$ and $K$, can be used for purely descriptive purposes, or as the basis for testing normality of the log-transformed SLs. In this latter role they are only designed to detect specific kinds of departure from normality, and if they fail to do so this does not imply the data are normal.

Omnibus tests don’t focus on specific kinds of departure from normality: the ideal is that any kind of departure will be detected, subsequent more focused analysis being needed if non-normality is detected where its specific nature is then of concern. There are numerous omnibus tests buried in the statistical literature. Most of them are, one suspects, never much used even by their inventors, being mathematically too complex for routine application with no known advantages of
performance over more established methodology.

The Shapiro-Wilk (SW) test is one of the most popular omnibus tests, and is usually the one against which any competing new test is compared. It is widely viewed, with supporting evidence from computer-intensive studies, as being as good as or better than most other tests at detecting the kinds of non-normality one is mostly interested in.\(^8\)

It is difficult to explain simply in mathematical terms but, allowing a bit of licence, essentially it measures the ‘correlation’ between the observed data and values they should have if the data are normal. Perfect normality results in an SW test statistic of 1; values significantly less than 1 indicate non-normality. Application in R is simple enough (with the reminder that the proper thing to do in practice is give the film a thorough graphical inspection before attempting anything fancy like a statistical test). Thus `shapiro.test(log(SL.Exodus))` produces

```
Shapiro-Wilk normality test

data:  log(SL.Exodus)
W = 0.9947, p-value = 0.05506
```

which has a p-value of 0.05506. If using a 5% decision rule the equates to a p-value of 0.05, and as our observed value of 0.05506 is greater than this we can accept that the log-transformed data for Exodus are acceptably normal and hence that the film can be deemed to have an approximate lognormal distribution.

Exodus has not been chosen by accident. It has 547 shots. Construct a new film, call it BigExodus with twice as many shots that simply repeats each shot twice, that has exactly the same SL distribution. Then `shapiro.test(log(SL.BigExodus))` produces a p-value of 0.0006 which, with its very small value and even a 1% rule leads to the emphatic rejection of normality. All that’s changed is the sample size. This is a concrete illustration of the sample-size effect, where conclusions about log-normality, arrived at via hypothesis testing, are affected by the number of shots in a film.

Redfern (2012a) uses the Shapiro-Francia (SF) test, an approximation to the SW test, developed at a time when computing power wasn’t up to handling the calculations needed for the SW test as easily as can be done now.\(^9\) It gives a p-value of 0.12 for Exodus somewhat larger than that for the SW test, but both lead to the same decision. With minor differences at the margins both the SW and SF tests arrive at similar conclusions about which films are acceptably lognormal using a 5% decision rule. He notes the sample-size effect can lead to rejection of the null hypothesis ‘in the presence of trivial deviations from the theoretical model’ and supplements it with the Jarque-Bera (JB) test, used ‘as an additional test of normality suitable for large samples’.

The JB test is based on the statistic

\[
JB = \frac{n}{6} \left( S^2 + \frac{K^2}{4} \right) = z_S^2 + z_K^2
\]

and asymptotically (i.e. for ‘large’ samples), under normality, is distributed as \(\chi^2_2\). The statistics \(z_S = S/\sqrt{6/n}\) and \(z_K = K/\sqrt{24/n}\), are asymptotically \(N(0,1)\) under the null hypothesis of normality, and are used separately for asymptotic tests of normality based on skewness and kurtosis (Section 12.2.4)

The JB test, as used in Redfern (2012a), is only ‘suitable’ in the sense that the standard theory that justifies the use of the significance estimates in most of the several R implementations is only valid if the sample is large enough (which they mostly are). It does not ‘correct’ for sample size effects, as the explicit dependence on \(n\) shows. Increase \(n\) and JB increases without any corresponding change in the critical chi-squared value used to assess its significance. Apply the

\(^8\)When a new method is developed its ‘merit’ is usually demonstrated by trying to find some specific form of departure from normality where it seems to perform as well as the SW test. For some very specific kinds of non-normality, such as symmetric distributions with longer tails than the normal, tests like the Jarque-Bera (JB) test and variants of it may do better. The JB test is also popular, but depends only on \(S\) and \(K\) so can be poor at detecting non-normality that is not a function of skewness and kurtosis – a fact not advertized by its enthusiasts.

\(^9\)The SF test is implemented by `sf.test` in the `nortest` package.
JB test to *Exodus* and *BigExodus* and *p*-values of 0.10 and 0.0094, leading to different conclusions about lognormality, are obtained.

This mistaken use of the JB test as an *independent* check on the results of the SF test, unaffected by sample size, vitiates much of Redfern’s discussion of the significance testing results, to which more emphasis is given than the more informal (but more compelling) evidence against lognormality provided by Figure 12.3. This reduces the strength of the critique of Salt’s claims about the prevalence of lognormality. It is more constructive to examine what their approaches have in common, rather than their differences.

Stripped of technical detail and some minor issues, both assess the claims of the lognormal using a measure of the correlation between the observed data and what is to be expected under lognormality. They differ in how a reasonable ‘approximation’ to lognormality is determined. Redfern uses an ‘objective’ hypothesis testing approach to assess whether his correlations show that the lognormal is an ‘appropriate’ model; Salt’s interpretation of his correlations is more ‘subjective’ and typically would accept lower values of the correlation than Redfern as indicative of a reasonable ‘approximation’. The ‘objective’ approach is compromised by the sample-size-effect problem; the subjectivity of the ‘subjective’ approach is its main problem since there is scope for anyone to disagree about what constitutes an acceptable approximation.

![Shapiro–Wilk test statistic](image)

**Figure 12.6:** *A histogram of the Shapiro-Wilk 134 Hollywood films 1935-2005.*

In Section 10.4 R code was give for generating the Shapiro-Wilk statistic for each of a large body of films. Applying this to all 134 films and plotting the results as a histogram gives Figure 12.6. A cut-off point could be determined to the right of the graph and films to the right of this can be deemed ‘approximately’ lognormal. The only real way in which the approaches being described differ is where you choose to place the cut-off point.\(^{10}\)

\(^{10}\)This simplifies things to get the idea across. The ‘objective’ approach involves a different line for each film
It is possible to rearrange the JB statistic to get an ‘index’ of ‘conformity to normality’

\[ I = \frac{6 \text{JB}}{n} = S^2 + \frac{K^2}{4} \]

for the log-transformed data that might be used as the basis for subjective interpretation. The median of \( I \) for the 134 films in Redfern’s sample is 0.21. If you assert that about 50% of films conform to the lognormal this can be viewed as using a rule-of-thumb with 0.21 as the criterion on which ‘conformity’ is based. Viewed from the perspective of a strict hypothesis testing approach, at the 1% level of significance the critical value of the chi-squared reference distribution is 9.21. Solving for \( n \) gives \( n = 263 \). That is, a ‘subjective’ approach based on using index \( I \) as a rule-of-thumb could be viewed from the hypothesis testing perspective as behaving as if all films have about 260 shots.

### 12.4 On distributional regularity – an alternative approach

My own view is that it is more fruitful to investigate distributional regularity allowing this to be more complex than lognormality. The left-hand plot of Figure 12.3 is compelling ‘holistic’ evidence that lognormality does not apply generally; but this and the right-hand plot suggest that there is some form of regularity to the departure from lognormality that may be worth trying to ‘capture’.

This ‘regularity’ is manifest in the fact that most distributions remain skewed (in the same direction) after log-transformation. The log-transformation is intended to remove skewness and is failing to do so; one thought is to apply a second log-transformation to cure this, but this can’t be done if there are negative values on the log-scale.

One possibility is to apply the Box-Cox transformation rather than log-transformation to the raw SLs. This is discussed in the Appendix and includes the log-transformation as a special case. After such a transformation 26% of the transformed data for the 134 films is acceptably normal at the 5% level, and 45% at the 1% level. This compares with about 7% for a log-transformation. The Yeo-Johnson (YJ) transformation (see the Appendix) applied to the log-transformed data produces even better results, and is that described here. As can be seen from it looks rather complex, the complexity being needed to cope with the fact that log-transformed data may be negative. It may help to think of it as attempting the same thing as a double log-transformation would, were this possible.

Code for applying the YJ transformation to the log-transformed data – double-transformation – and obtaining the skewness and kurtosis of the doubly-transformed data, is given in Section 12.5. A plot of skewness against kurtosis, akin to the earlier Figure 12.5 for log-transformed data, is shown in Figure 12.7.

It is evident that the double-transformation does a remarkably good job of producing values of \( S \) compatible with the normal distribution, so that the more obvious departures from normality are associated with the larger values of \( K \). It is also obvious that many more films have values of \( S \) and \( K \) compatible with normality than was the case after a single log-transformation. If a Shapiro-Wilk test is applied normality is not rejected for 40% of the films using a 5% level of significance, and 59% using a 1% level.

These results are not immune from sample-size effects but the critique that was applied to Redfern’s (2012a) interpretation of his results does not apply here. In the former case results were largely ‘significant’, but given the large sample sizes disentangling substantive significance from the sample-size effect is problematic. In our analysis the widespread non-significance occurs despite the large sample sizes, so that the evidence that double-transformation is achieving normality is strong.

and the SF and JB tests are used; the ‘subjective’ approach is more akin to using a single cut-off point but, at the detailed level, a different correlation measure. Neither author applies the criteria suggested quite as rigidly as the description has it; for example judgment is exercised in the few cases where the SF and JB tests lead to different conclusions.
It can be objected that the 40% achieved at the 5% level is not a ‘majority’, but some of the significant results, superficially indicative of non-normality, are attributable to large samples. Baxter (2013a) describes a small simulation experiment, coupled with graphical analysis, that investigated this. The simulation was based on taking repeated randomly selected half-samples and testing these for normality at the 5% level after double-transformation. Details are not provided here, but it was concluded that about half of the later films with large numbers of shots, initially classified as ‘non-normal’, could be reclassified. This raises the proportion of films that can be regarded as distributionally regular to well over 50% even if the less generous 5% decision rule is used.

Once distributional regularity for a majority of films is established it becomes of interest so see why and how films depart from the ‘norm’ – it is, perhaps, the main reason for trying to establish a ‘norm’ in the first place. This is touched on only briefly in Baxter (2013a); it was suggested that, once one strays from distributional regularity, earlier films (with 1935 being the earliest in the sample) are differently irregular from later ones and show considerable variety. There is possibly a pattern to the form of irregularity shown by later films along the lines that, compared to what is expected from normality after double-transformation, there are rather more shorter SLs than to be expected, with this often showing in the left-tail of the distribution. This was left as an issue to be pursued.

12.5 R code

Code for generating the skewness and kurtosis for log-transformed SL data, among other statistics, was given in Section 10.3. The Yeo-Johnson transformation is applied after an initial log-transformation. It works in the same kind of way as the Box-Cox transformation, allowing for the
fact that some of the log-transformed data may be zero or negative, depending on an exponent \( \lambda \). To produce the data needed for further analysis it is necessary to calculate the exponent, \( \lambda \), designed to make the data as nearly normal as possible. Once this is done the data need to be transformed using it. The following code does this for the YJ transformation, and obtains the skewness and kurtosis of the transformed SL data for each film. Other than knowing it provides the data for Figure 12.7 most readers may want to ignore this.

The \texttt{lambda} command obtain \( \lambda \) for a film, after log-transformation. The \texttt{car} package is needed for the \texttt{powerTransform} function. The \texttt{skewness} and \texttt{kurtosis} functions are from the \texttt{moments} package, with \texttt{yjPower} from \texttt{car} doing the necessary data transformation using \( \lambda \). For an equivalent Box-Cox analysis replace all occurrences of \texttt{yj} with \texttt{bc}, and remove \texttt{log} in the two places it occurs.

```r
library(car)
library(moments)

z <- SL.ALL/10

lambda <- function(x) powerTransform(log(x), family = "yjPower")$lambda
skewYJ <- function(x) skewness(yjPower(log(na.omit(x)), lambda(x)))
kurtosisYJ <- function(x) kurtosis(yjPower(log(na.omit(x)), lambda(x))) - 3

skew <- apply(z, 2, skewYJ)
kurtosis <- apply(z, 2, kurtosisYJ)
```

The following code produces Figure 12.7; \texttt{kurtosis} and \texttt{skew} are the statistics for the YJ-transform generated by the function defined above. Replace these with the values for log-transformed data to get Figure 12.5. \texttt{Date} is a previously created variable with the dates of the films. The main new feature here is the use of the \texttt{polygon} function to generate the shaded areas in the plots.

```r
plot(kurtosis, skew, xlab = "kurtosis", ylab = "skew", type = "n")
polygon(c(-.57, -.57, .57, .57), c(-.28, .28, .28, -.28), border = NA, col = "lightgreen")
polygon(c(-.3, -.3, .3, .3), c(-.15, .15, .15, -.15), border = NA, col = "pink")

color <- ifelse(Date < 1975, "black", "red")
Symbol <- ifelse(Date < 1975, 16, 15)
points(kurtosis, skew, pch = Symbol, col = color, cex = 1.1)
abline(h = 0, lwd = 2)
abline(v = 0, lwd = 2)

col = c("black", "red"), bty = "n", title = "Date")
```
Chapter 13

Multivariate methods – basics

13.1 Introduction

It is only fairly recently that standard methods of applied multivariate statistical data analysis have begun to be applied to cinemetric data analysis. As treated here multivariate methods can be thought of as exploratory methods for detecting patterns in \( n \times p \) tables of numbers. Correspondence analysis (Section 9.2) is such a technique, typically applied to tables of counts\(^1\).

Baxter (2012c) and Redfern (2012c, 2013c) have exploited correspondence analysis to investigate pattern in data tables derived from film-related studies. In Baxter (2012a) (and Chapter 9) the application is to counts in tables of shot-scale type by film; in Redfern (2012c) to a sample of individuals cross-tabulated by age/gender categories and genre preferences; in Redfern (2013c) to counts for different categories of four variables (shot-scale, camera movement, camera angle, shot-type) within each of four passages in Rashomon (1950). Earlier work by Murtagh et al. (2009) used correspondence analysis and hierarchical clustering to investigate narrative patterns in film scripts.

This chapter reviews methods for tables of data based on a set of continuous variables, which in our case will be shot lengths (SLs), measured for a sample of \( n \) films. Two of the most common are principal component analysis (PCA) and cluster analysis. Baxter (2013c,d) has used these to explore cutting patterns in samples of D.W Griffith’s Biograph one-reelers from 1909–1913. Redfern (2013b) has illustrated how cluster analysis might be used to explore time-series patterns in SL distributions within films, using for illustration a sample of 20 Hollywood films from 1940–2000 from the database compiled by Cutting et al. (2010).

Good statistical texts on multivariate analysis, of which there are many, cover the three main techniques used in these notes. Greenacre (2007), Jolliffe (2002) and Everitt et al. (2011) provide accounts of correspondence analysis, PCA and cluster analysis respectively.

The emphasis in this chapter is on conceptual and computational aspects of PCA and cluster analysis as applied to SL data. The ideas are exploited in the case study of Chapter 14. An immediate problem to be faced is that films consist of different numbers of shots. PCA and cluster analysis are applied to rectangular tables of data where \( p \), the number of columns, needs to be the same for each row (film). The SL data needs to be processed in some way so that this condition is achieved. This is discussed before further illustration.

\(^1\)It can, and has been, applied to tables of non-negative numbers other than counts, but is most often presented in the context of counted data.
13.2 Data processing

13.2.1 Brisson’s partitioning method

An idea of Keith Brisson’s, posted on the Cinemetrics website, is described here\(^2\). A film is partitioned into a fixed number of \( p \) intervals of equal duration; ‘fractional’ counts of shots contained within each interval are obtained; and ‘localized’ ASLs are computed for each interval by dividing the interval duration by the fractional count. Call these partition SLs. The partition SL ‘profile’ of two or more films can be compared across films, ignoring the fact that interval widths are different for different films.

If \( p = 100 \) is used, the index of an interval can be read as the elapsed duration (in percentage terms) of the film; if, for example, \( p = 200 \), divide the index number by 2 to get the same interpretation. Choice of \( p \) may be an issue, and it is easy enough to experiment with different choices. The original purpose for which Brisson’s method was devised was to allow the averaging of profiles across films and this is possible without any further data manipulation, though smoothing the data either for individual films or for the average is a possibility to be discussed shortly.

In slightly more formal terms what the Brisson methodology does is to transform the ‘raw’ SL data for film \( i \) to a set of partition SLs, \((x_{i1}, x_{i2}, \ldots, x_{ip})\), which can be stacked-up to get an \( n \times p \) table (or data matrix), \( X \), that can be manipulated using standard statistical methods for whatever purpose the investigator has in mind. Some R code for effecting this processing is given in Section 13.6.

Figure 13.1 shows the SL pattern for D.W. Griffith’s A Corner in Wheat (1909) before and after applying the Brisson partitioning process with 100 intervals. In the former case the time axis is scaled to lie between 1 and 100, with plotting positions defined by the cut-points after this rescaling. By construction, plotting positions in the latter case are the integers in the range \([1,100]\) and can be read as the proportion of a film’s duration. It is clear that there is little difference between them, any difference being most evident in intervals in the second plot where a cut occurs.

![Figure 13.1: The SL distribution of A Corner in Wheat before and after applying the Brisson method.](image-url)

The advantage of using the partitioning method is two-fold. Figure 13.1 shows examples of what Yuri Tsivian, in his introduction Looking for Lookalikes? the third topic on the On Statistics pages of Cinemetrics, calls bi-axial graphs. They were used to compare how similar a small sample

of D.W. Griffith’s Biograph films were. By partitioning into an equal number of intervals, films are comparable in terms of the number of partitioned SLs involved, and multivariate methods such as PCA and cluster analysis can be applied directly to them to investigate the question.

This is not without problems, as other contributors to the discussion note. Film SL data patterns are highly variable and it makes sense to simplify the problem of comparison by smoothing the SL data to identify broad patterns in what is otherwise ‘noisy’ data. This raises its own issues, among them the fact that any pattern detected is conditional on the level of smoothing used. Smoothing can also be done in more than one way, as discussed in Section 13.3. Although smoothing does not require equal numbers of shots for different films for it to be applied, it is easier to implement, for the purposes of multivariate analysis, if there are an equal number of data points to begin with. The partitioning method achieves this.

13.3 Smoothing

13.3.1 General issues

One approach to smoothing SL data is to apply KDEs to the cut-points (Redfern, 2013a). Section 7.7 compares this approach with the loess smoothing of SLs against cut-point. They produce similar results in the sense that the results of either method are sensitive to the level of smoothing, and the one can be matched against the level chosen for the other.

Section 8.3.2 discussed how it was possible to exploit the `density` function in R to extract estimates of the KDE at a fixed number of \( p \) equally spaced points. If this is done for \( n \) films this generates an \( n \times p \) data matrix suitable as input to multivariate methods for comparing the SL structure of films. Redfern (2013b) illustrates the idea, in his cluster analysis of SL time-series. R code showing how to generate the data for a film is given in Section 8.3.3.

Much the same idea can be applied, in principle, using loess smoothing. If the raw rather than partitioned SL data are used it is necessary to first fit the model and then predict values at the desired number of equally spaced points. For comparative purposes this can cause problems if the shots at either end are unduly long, since algorithmic problems can arise in predicting outside the observed range. This problem disappears if partitioned SLs are used since, by construction, there are an equal number of equally spaced points to begin with and the fitted values at these points can be used for subsequent multivariate analysis.

There are a number of issues in application that will be addressed as they arise in the following sections – the choice of level of smoothing is an obvious and important one (Sections 13.3.3, 14.4). A general issue with smoothing methods is that estimates are less reliable at the boundaries of a film; roughly this is because smooths depend on ‘averaging’ over points in a region, and you run out of points near the boundaries. KDEs are not especially satisfactory in this respect, and other density estimation methods such as local polynomial fitting (the `locpoly` function in R) have been suggested as preferable if end-effects are a concern (Venables and Ripley, 2002, p.132).

The main aim in the rest of this and the next chapter is to illustrate some of the statistical ideas and issues involved in applying multivariate methods to SL data. For this purpose partitioned SL data will be used. Where smoothing is involved, unless otherwise stated, loess smooths with a span of 1/2 and otherwise the defaults in the `loess` function of R are used.

13.3.2 Scaling

In applying multivariate methods generally, and specifically if the aim is to effect numerical comparisons between SL distributions, the way the data are scaled is important. Figure 13.2 illustrates the kind of considerations that are involved.

The upper plots show smooths of the SL data for *A Corner in Wheat* (1909) and *The Lonedale Operator* (1911). Other than the first 20% of the film or so they look not too dissimilar as these
things go\textsuperscript{3}. If, however, the smooths are compared on the same plot, as in the bottom-left, then they are well-separated. Over most of the range of the film there is a consistent but quite large difference between them which translates into a large ‘distance’ between them when measured more formally (Section 13.4).

The reason for the large difference is that the films have rather different ASLs, that for the earlier \textit{A Corner in Wheat}, 25.7s, being more than two and a half times the 9.5s for the later film. \textit{The Painted Lady} (1912), with an ASL of 9.6s, is shown in the lower plots for comparison. It has a markedly different shape from the other two films, but appears generally closer to \textit{The Lonedale Operator} (which has a similar ASL) than is \textit{A Corner in Wheat} over much of the range. The consequence for analysis, if smoothed SLs are compared in this form, is that grouping will be based more on differences in the ASL (‘size’ effects) than differences in temporal patterns (‘shape’ effects). This may not matter if size differences are the focus of interest, but if, as in what follows, it is shape that is of interest then size effects need to be removed.

![Graphs showing smoothed SL distributions for A Corner in Wheat (1909), The Lonedale Operator (1911), and The Painted Lady (1912).](image)

Figure 13.2: \textit{The smoothed SL distributions of A Corner in Wheat (1909), The Lonedale Operator (1911) and The Painted Lady (1912), illustrating the effects of scaling. See the text for detail.}

This can be done in different ways. For example, the SLs could be scaled in some way before further analysis, division by the maximum SL for a film being a possibility. Another alternative

\textsuperscript{3}Remembering that it is the smooths that are being compared, so this judgment is conditional on the degree of smoothing being used.
would be to standardize the smoothed estimates for a film to have zero mean and unit variance. In the lower right-hand plot in Figure 13.2 the smoothed estimates from the loess fit have been scaled to lie in the region \([0, 1]\). Redfern (2013b) treats his KDE estimates in a similar way. The similarity in shape of *A Corner in Wheat* and *The Lonedale Operator*, and their difference from *The Painted Lady*, is more strongly emphasized than in the lower left-hand figure. More importantly, for later usage, is the fact that these evident shape differences correspond, after the scaling, much more closely to the distances between pairs of films than is evident in the lower-right.

### 13.3.3 Degrees of smoothing

The effect of varying the degree of smoothing is explored more fully in Section 14.4. The aim here is to illustrate the ‘problem’. Figure 13.3 shows smooths for spans of 1/2, 1/5 and 1/10 for each of *A Corner in Wheat* and *The Lonedale Operator*.

In a 1926 essay in *Liberty* magazine, *The Pace in Movies*, Griffith wrote the following \(^4\).

The pace must be quickened from beginning to end. This is not, however, a steady ascent. The action must quicken to a height in a minor climax, then slow down and build again to the next climax which should be faster than the first, retard again, and build to the third, which should be faster than the second, and on to the major climax, where the pace should be fastest.

This was written some time after his major achievements as a director of full-length feature films. It is of interest to ask to what extent did his earlier one-reelers actually follow this kind of pattern, and what level of smoothing is needed to make this clear?

![Figure 13.3: Smoothed SL distributions of A Corner in Wheat (1909) and The Lonedale Operator (1911) for different degrees of smoothing.](image)

The implication is that a smoothed pattern ought to show distinct troughs, getting deeper as the film progresses. We can ignore minor bumps that disturb the ideal pattern. The smooth for *A Corner in Wheat* does rather well with a span of 1/5; with a span of 1/10 the general picture is similar but bumpier and the individual shots are being tracked fairly closely, probably because of their rather small number (32). Arguably the 1/2 span over-smooths the data but the pattern is what Griffith’s prescription leads one to expect, with the minor climaxes smoothed out \(^5\). None of these smooths are ‘wrong’ though if the data are under-smoothed, as with the span of 1/10, it rather negates the purpose of smoothing in the first place. Baxter (2013d) consistently used

---

\(^4\) There is some doubt that Griffith actually wrote this but, if not, it is assumed that he was aware of and did not dispute the content.

\(^5\) Since writing this the question of whether or not the pattern is a function of the placement of intertitles in the film has been raised. It isn’t.
spans of 1/2 as an aid to identifying groups of films with broadly similar cutting patterns, but noted that this over-simplified the structure in some cases for which less smoothing might have been appropriate.

For *The Lonedale Operator* the patterns are not quite as neat, but are there. A span of something like 1/10 is needed to get a four-trough pattern, and the troughs are not as nicely differentiated in terms of depth, or as evenly spaced, as with *A Corner in Wheat*. The smooth with 1/2 is reasonably in accord with the prescription; the difference from *A Corner in Wheat* at the start is attributable to four relatively short shots among the first five. In terms of attempts to compare film structure via smoothing methods this raises questions about choosing the degree of smoothing that are deferred to the next chapter.

### 13.4 Distance measures

However it is generated, analysis is conducted on a data matrix $\mathbf{X}$ with $n$ rows corresponding to films, each with $p$ partitioned or ‘adjusted’ SLs. A fundamental idea is that the (dis)similarity between a pair of films can be measured by the distance between their profiles. Two films, identical with respect to the set of numbers defining their SL profile, have a distance of zero between them, but this will get larger as their profiles diverge.

Mathematically, ‘distance’ can be defined in many ways. Most commonly, and for practical purposes, *Euclidean distance* is used, and is the default in many statistical software packages that include methods of multivariate analysis. Euclidean distance will be used in the rest of this and the next chapter without much further discussion.

Many multivariate methods, including PCA and cluster analysis, are sensitive to the scaling used. The default in some software packages (though not R) is to standardize the columns of a data matrix to zero mean and unit variance. Some form of standardization is necessary if variables are measured in different units, and it is often regarded as the ‘safe option’. If numbers are in the same units but differ by orders of magnitude then the variables in columns associated with the larger numbers will predictably dominate an analysis and standardization in these circumstances is also common. As has been discussed in Section 13.3.2 SL data are affected by this kind of problem but it is the rows of the data matrix (corresponding to films) that are affected because of the variation in ASLs. This will be dealt with by scaling smoothed (partition) SLs to [0, 1], though other treatments are possible.

---

6Tsivian, in his introduction to the third *On Statistics* discussion presents a graph of the average cutting pattern for 61 of Griffith’s Biographs that is very similar to that obtained for a *A Corner in Wheat* with a span of 1/2; Baxter (2013d) shows a similar (smoothed) plot of the average for a very similar sample of 62 Biographs. Appendix 2 of the latter paper shows smooths with a span of 1/2 for all 62 films. It is perhaps interesting to note that *A Corner in Wheat* bears what is possibly the closest resemblance to the ‘Griffith average’, and that a lot of films do not exhibit the patterning – at this level of smoothing – that Griffith’s prescription might lead us to expect.

7Imagine you have a cat sitting in a garden that has a tree in it. The cat walks to the tree, climbs up it (vertically), and gets stuck. You rescue it by placing a ladder at the point where the cat was sitting that reaches just to where it is stuck; it hops on and climbs down. If the cat is sitting at a distance, $x$, from the tree and climbs to a height of $y$ it travels a distance ($x + y$) to get stuck and a distance of $\sqrt{x^2 + y^2}$ to escape. These are different but (mathematically) legitimate ways of measuring distance; the latter distance, the escape route by the ladder, is Euclidean distance. Most people are comfortable with this idea in two or three dimensions; we inhabit a three-dimensional world and make judgments about (Euclidean) distances on a daily basis without thinking about it. Mathematically a point in three-dimensional space can be defined by three ordered coordinates, $(a, b, c)$ and the distance between two such points can be expressed mathematically. In multivariate analysis more than three coordinates are needed to define a point, $(a, b, c, d, \ldots )$. In the present context $a$, $b$, $c$ etc. correspond to the first, second, third etc. (partition) SLs. We can’t visualize the distance between two films characterized in this way in quite the same way we can visualize the cat, or measure it with some physical device as we might measure the length of the ladder, but we can measure it mathematically. As with the cat you can do the measurement in more than one way, but if you stick with Euclidean distance you are working with an idea you should be comfortable with (even if you don’t know it).

8Some statisticians would choose to work with log-transformed data in these situations.
13.5 Principal component (PCA) and cluster analysis

13.5.1 Principal component analysis

Principal component analysis is actually a fairly simple mathematical technique if you are familiar with matrix algebra, but non-mathematicians lacking such familiarity can be put off by the appearance and regard it, in consequence, as a ‘difficult’ method. It isn’t. Carrying out a PCA is (almost) as easy as calculating an ASL and what it does can be explained in ordinary language.

There are, admittedly, pitfalls to be aware of, but then this is true of the ASL. If SL data for a single film is coded as $x$ then the ASL is obtained as $\text{mean}(x)$. Given a suitably constructed data table of scaled partition SLs for a body of $n$ films, $X$, a basic principal component (PC) plot, of the kind to be found in the next chapter, can be obtained from

$$\text{PC} \leftarrow \text{prcomp}(X) \cdot x; \text{plot}(\text{PC[,1]}, \text{PC[,2]})$$

which produces a two-dimensional ‘map’ of $n$ points.

The points correspond to films, and the idea is that films that are close to each other on the map as measured by the distance between them will have an SL structure that is more similar than either is to some other film at a much greater distance. The configuration may suggest clusters of similarly structured films that one would expect to be different, in identifiable ways, between clusters.

The catch – or pitfall – to be aware of is that the map distance is only a two-dimensional approximation to the real distances between films, and can mislead. In the context of SL data as treated here what PCA does is to define new variables, principal components (PCs), that describe different aspects of the ‘shape’ of the film. There are as many PCs as the smaller of $n$ and $p$, and they are constructed so that the first PC is the most ‘important’, the second the next most ‘important’ and so on. It’s possible to construct a measure of the total ‘variation’ between film profiles. The hope is that most of this variation is attributable to (or ‘explained by’) the first few PCs, variation explained being equated with ‘importance’. There are different rules-of-thumb for defining what is meant by ‘most’ and ‘few’: the simplest require that ‘few’ is the number of PCs that ‘explain’ some arbitrary percentage of the total variation, such as 70% or 80%.

Common practice is to present a plot of the first two PCs, and this will often be useful even if the proportion of variation explained is somewhat smaller than these numbers. It is easy enough to examine plots using PCs other than the first two and this is illustrated in Section 14.3. The need to choose some level of smoothing complicates matters. The less smoothing used the more ‘wiggly’ the curves that describe a film become. Increased ‘wigginess’ causes the distance between any pair of films to increase; this inflates the total variation in the data, with the consequence that the amount of variation explained by a fixed number of PCs decreases as the smoothing decreases. This really just reflects the fact that if some level of simplification – smoothing – is not imposed on a film then all films look different to a greater or lesser degree. The consequences of this are explored in more detail in the case study in the next chapter.

13.5.2 Cluster analysis

Implementing a cluster analysis is as straightforward and the idea as simple. The following code

$$\text{plclust(hclust(dist(X), method = "average"))}$$

will carry out (an average-link) cluster analysis and plot the results in a form of a tree-diagram or dendrogram. This consists of branches and leaves (films) and the usual idea is to try and cut the

---

9There are statisticians who regret the ease with which PCA and other multivariate methods can now be applied. Attempting to stem this is reminiscent of the efforts of King Canute. Ease of application should be welcomed, but allied to an appreciation of limitations of a technique, which doesn’t usually require a deep understanding of the underlying mathematics to be aware of.

10The ASL can be sensitive to SLs of excessive duration and is pretty useless in the perhaps rare situation where the SL distribution is multi-modal.
branches of the tree in such a way as to cluster leaves corresponding to similar cases (as measured by distance) that differ from those in other branches.

The idea is simple, but also deceptive. The **method = "average"** argument specifies the choice of clustering algorithm (average-linkage in this case). There are many choices of algorithm available which at some point have seemed to someone to be a ‘sensible’ means of grouping data; most of these have no solid theoretical grounding or, if they do, embody assumptions that are usually unrealistic. Different methods can give rise to different results and can give rise to uninterpretable trees, or trees with an obvious interpretation that may be wrong. That is, the simpler clustering algorithms are designed to produce clusters and may suggest them even if the pattern is random.

For these kind of reasons not all statisticians are especially keen on the simpler methods favored by practitioners\(^{11}\). Everitt *et al.* (2011) provides a sympathetic account that can be referred to for a fuller discussion of the simpler methods.

Very briefly, hierarchical clustering methods – which all those used in Chapter 14 are – begin by treating each case as a single cluster and successively merge cases until all are united in a single cluster. All methods will begin by merging the two most similar cases but diverge as soon as clusters are formed because of the different ways similarity between a pair of clusters is measured. Of the common linkage methods, for single- and complete-linkage the similarity depends only on two cases, one from each cluster, so that group structure is ignored. Single-linkage tends to suffer from an effect called **chaining** which produces a dendrogram looking much like a staircase that is difficult to interpret.

By contrast average-linkage and Ward’s method attempt to account for group structure by measuring similarity in terms of the difference between ‘average’ properties of two clusters, differing in the way ‘average’ is conceived. Some users like Ward’s method because it produces dendrograms that can be clearly interpreted, but this can be illusory since it can do this even if the data are random. Average-linkage can be thought of as a method that tries to avoid the under-interpretability of single-linkage and over-interpretability of Ward’s method.

The ‘popularity’ of different methods in the past has almost certainly been swayed by the defaults in the statistical software available at the time. Ward’s method was the default in **CLUSTAN**, one of the earliest specialist packages for cluster analysis, and average-linkage is the default in some of the more widely used general-purpose packages. In application areas I’m familiar with (chiefly archaeology) where cluster analysis has been widely used, where I’ve seen informed opinion based on extensive experience expressed, average-linkage has sometimes been favored on pragmatic rather than theoretical grounds. My own practice is to compare results from different methods in conjunction with PCA as a method of data visualization. This is at the ‘quick-and-simple’ end of the analytical spectrum but can often produce satisfactory results.

The message here is to use cluster analysis by all means, but don’t rely on a single method, and invest effort in validating conclusions about the reality of any clusters that are suggested by the methodology. As with PCA, which is a useful tool for checking results from a cluster analysis, ways of doing this are explored in the context of real data in the next chapter.

### 13.6 R code for Brisson’s partition method

Let the SLs in a film be \( y_i \), where \( i = (1, 2, \ldots, S_i) \), which define cut-points, \( t_i \). The number of shots, \( S_i \), varies from film to film. Assume a partition of \( B \) equally spaced intervals is required, so that interval boundaries are given by \((0, b_1, b_2, \ldots, b_B)\). The ‘trick’ is to combine the cut-points and boundary intervals, treating the latter as ‘augmented’ cut-points, and compute fractional counts of the form \( d/y \) adding the result to interval \( j \), where \( d \) is the difference between adjacent

---

\(^{11}\)The tone was set by Cormack’s (1971) seminal paper *A Review of Classification*, which began with the words ‘The availability of computer packages of classification techniques has led to the waste of more valuable scientific time than any other "statistical" innovation’. Model-based methods of clustering, which some prefer, are typically too complex for routine use and/or not much suited to data sets with many variables and/or need assumptions one doesn’t usually believe in to ‘work’. The widespread use of simple and accessible methods, whatever their statistical pedigree, is understandable.
augmented cut-points. The code below cycles through the successive \((S_i + B)\) augmented cut-points, computing \(d\) on each cycle. An indicator variable, 1 for a true cut-point and 0 for a boundary point, determines when to move on to the next value of \(y_k\) as a divisor and the interval \(j\) to which the fractional count is added.

In the code 100 intervals \((B = 100)\) are used for illustration. The data have been set up initially as a table in an Excel file then imported into R as described in Section 10.2. That is, the film that is an argument to the function is the name of of a column of the table, and may contain NAs that need to be omitted using na.omit. The variable \(y\) scales SLs to lie in the interval \((0, 100)\) and is perturbed by a small amount so that the scaled cut-points derived from it, \(t\), will not be coincident with the boundary points, \(b\).

```r
BRISSON.method <- function(film, B = 100) {
  ## scale data so the cut-points lie in the interval(0,100)
  y <- as.numeric(na.omit(film)); S <- sum(y)
  y <- 100 * y/S - .00000001
  t <- cumsum(y); t <- 100 * t/max(t)
  ny <- length(y) ## number of shots

  ## do the same for the interval boundaries
  b <- 100* c(1:B)/B
  T <- c(t, b)
  ind <- c(rep(1, ny), rep(0, B)) ## indicator variable
  ind <- ind[order(T)] ## 1 for a cut-point, 0 for a boundary
  T <- sort(T)
  T0 <- c(0,T) ## add the 0 for lowest boundary
  dT <- diff(T0) ## compute differences

  ## initialise variables; I will hold fractional counts
  I <- rep(0,B)
  j <- 1
  k <- 1

  ## compute fractional counts; assign to the correct interval
  for(i in seq(1,length(T),1)) {
    I[j] <- I[j] + dT[i]/y[k]
    j <- ifelse(ind[i] == 0, (j + 1), j)
    k <- ifelse(ind[i] == 1, (k + 1), k)
  }
  I[B] <- ny - sum(I[1:(B - 1)])

  SL <- S/(I * B) ## scale so results compare with original SLs
  SL
}
```

**BRISSON.method**

Suppose \(film\) is one column in a table of a body of films, filmbody. The following can be used to generate the \(n \times p\) table of partitioned SLs for the full data set.

```r
brisson <- function(x) {f <- na.omit(x); BRISSON.method(f)[1:100]}
X <- t(apply(filmbody, 2, brisson))
```

Given \(X\) the following code will generate the average of the partitioned SLs, a loess smooth of the average with a span of 1/2, and the smoothed values scaled to lie in the interval \([0,1]\).
order <- c(1:100)  # the number of partition intervals
average <- apply(X, 2, mean)
average.smooth <- loess(average ~ order, span = 1/2)$fitted
# scale to [0,1]
Min <- min(average.smooth); Max <- max(average.smooth);
scaled.smooth <- (average.smooth - Min)/(Max - Min)
Chapter 14

Structure in Griffith’s Biographs

14.1 Introduction

The second of the *On Statistics* discussions, *What Do Lines Tell?*, included contributions that examined how different methods of time-series analysis might be applied to the descriptive presentation and comparison of temporal structure within films (see, also Chapter 7). Some of the papers in the third discussion, *Looking for Lookalikes?*, took this further by examining how multivariate methods might be used to identify films that had temporally similar SL structures. Some of the basic methodology has been outlined in the previous chapter. My second contribution to the discussion was an exploration of how structurally similar films might be identified in a sample of 62 of D.W. Griffith’s Biograph one-reelers from the years 1909–1913, also the subject of some of the other contributions.

This case study pursues some of the issues raised in these discussions, along with further methodological illustration. A sample of 25 films from the years 1911–1912, listed in Table 14.1, will be used.

<table>
<thead>
<tr>
<th>Index</th>
<th>Title</th>
<th>Date</th>
<th>Shots</th>
<th>ASL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>His Daughter</td>
<td>1911</td>
<td>88</td>
<td>10.1</td>
</tr>
<tr>
<td>2</td>
<td>Lonesdale Operator, The</td>
<td>1911</td>
<td>107</td>
<td>9.5</td>
</tr>
<tr>
<td>3</td>
<td>Indian Brothers, The</td>
<td>1911</td>
<td>55</td>
<td>15.0</td>
</tr>
<tr>
<td>4</td>
<td>Fighting Blood</td>
<td>1911</td>
<td>101</td>
<td>6.4</td>
</tr>
<tr>
<td>5</td>
<td>Last Drop of Water, The</td>
<td>1911</td>
<td>72</td>
<td>10.8</td>
</tr>
<tr>
<td>6</td>
<td>Country Cupid, A</td>
<td>1911</td>
<td>69</td>
<td>12.9</td>
</tr>
<tr>
<td>7</td>
<td>Swords and Hearts</td>
<td>1911</td>
<td>114</td>
<td>8.5</td>
</tr>
<tr>
<td>8</td>
<td>Adventures of Billy, The</td>
<td>1911</td>
<td>106</td>
<td>8.3</td>
</tr>
<tr>
<td>9</td>
<td>Battle, The</td>
<td>1911</td>
<td>113</td>
<td>8.6</td>
</tr>
<tr>
<td>10</td>
<td>Miser’s Heart, The</td>
<td>1911</td>
<td>82</td>
<td>11.7</td>
</tr>
<tr>
<td>11</td>
<td>For His Son</td>
<td>1911</td>
<td>72</td>
<td>12.2</td>
</tr>
<tr>
<td>12</td>
<td>Billy’s Stratagem</td>
<td>1911</td>
<td>111</td>
<td>7.4</td>
</tr>
<tr>
<td>13</td>
<td>Sunbeam, The</td>
<td>1911</td>
<td>83</td>
<td>10.6</td>
</tr>
<tr>
<td>14</td>
<td>Mender of Nets, The</td>
<td>1912</td>
<td>108</td>
<td>6.9</td>
</tr>
<tr>
<td>15</td>
<td>Girl and Her Trust, The</td>
<td>1912</td>
<td>135</td>
<td>6.6</td>
</tr>
<tr>
<td>16</td>
<td>Female of the Species,The</td>
<td>1912</td>
<td>73</td>
<td>11.3</td>
</tr>
<tr>
<td>17</td>
<td>One is Business, the Other Crime</td>
<td>1912</td>
<td>77</td>
<td>11.7</td>
</tr>
<tr>
<td>18</td>
<td>Lesser Evil, The</td>
<td>1912</td>
<td>105</td>
<td>7.4</td>
</tr>
<tr>
<td>19</td>
<td>Beast at Bay, A</td>
<td>1912</td>
<td>105</td>
<td>6.3</td>
</tr>
<tr>
<td>20</td>
<td>Unseen Enemy, An</td>
<td>1912</td>
<td>130</td>
<td>6.9</td>
</tr>
<tr>
<td>21</td>
<td>Friends</td>
<td>1912</td>
<td>92</td>
<td>8.1</td>
</tr>
<tr>
<td>22</td>
<td>Painted Lady, The</td>
<td>1912</td>
<td>77</td>
<td>9.6</td>
</tr>
<tr>
<td>23</td>
<td>Musketeers of Pig Alley, The</td>
<td>1912</td>
<td>88</td>
<td>11.2</td>
</tr>
<tr>
<td>24</td>
<td>New York Hat, The</td>
<td>1912</td>
<td>76</td>
<td>12.7</td>
</tr>
<tr>
<td>25</td>
<td>Burglar’s Dilemma, The</td>
<td>1912</td>
<td>86</td>
<td>10.6</td>
</tr>
</tbody>
</table>

Table 14.1: *D.W. Griffith Biograph films from the years 1912–1913*. The ordering of the films, the ‘Index’ used in labeling some plots, is chronological by production date, as given in Henderson (1970). Films 1–5 and 14–19 were produced in California.

The original analysis was based, among other things, on PCA and cluster analyses of smoothed
cumulative frequency distributions (CFDs) of the films. Here similar analyses will be investigated using partition SLs allied to loess smoothing. The shift in emphasis, and the use of a smaller data set, occurs for a number of reasons, methodological and substantive. The original CFD analysis was judged on how successfully (or not) it identified groups of films with reasonably similar looking SL smooths of the kind illustrated in Figure 13.2. Basing analysis directly on such smooths is an obvious avenue to explore\(^1\). The smaller data set is used because it makes it easier to explore and present aspects of the use of PCA and cluster analysis absent from earlier discussions.

The more substantive reason for just using the 1911–1912 results is that I have some doubts about using some of the earlier films with comparatively few shots for this kind of analysis. These doubts were raised in my earlier analysis, particularly with respect to films from 1909, and have since intensified, partly as a consequence of analyses similar to those reported below. One-readers are of comparable length so those with fewer shots will tend to have larger ASLs, and in many cases contain individual shots of unusual duration. These can be differently placed in a film and the consequence is that with other than quite high levels of smoothing the earliest films tend to look different from each other and from later films. All the 1909 films previously analysed had fewer than 50 shots, mostly less than 40, and the same is true for several of the 1910 films, particularly those in the sample from the first half of that year. All the films in the sample used here have at least 55 shots, all but one with 69 or more.

### 14.2 Initial cluster analyses

Until further notice all analyses are based on loess smooths of partition SLs with a span of 1/2, otherwise using the defaults in the \texttt{loess} function in R. We proceed from implementation to interpretation then explore the effects of varying the level of smoothing. Figure 14.1 shows a dendrogram obtained using the average-linkage method with Euclidean distance as the (dis)similarity measure and with leaves labeled by the index of the film given in Table 14.1.

The dendrogram is simply obtained using the code

```r
h <- hclust(dist(X), method = "average")
plot(h, labels = c(1:25), sub = " ", xlab = " ", main = "Average-Linkage (Span = 1/2)"
```

where \(X\) is the matrix of partition SLs, generated as described in Section 13.6; the \texttt{labels} argument is an instruction, in this case, to label leaves with the index of the film; and the \texttt{sub = " "} and \texttt{xlab = " "} arguments get rid of the default labeling otherwise provided.

Use \texttt{?hclust} to see what is available for the \texttt{method} argument. Replacing \texttt{average} with \texttt{single} or \texttt{ward} will produce a single-linkage and Ward’s method analysis, and any of these methods can be abbreviated to their first letter. Single-linkage is illustrated in Figure 14.3, and Ward’s method is that used for illustration by Redfern (2013b). In the applied literature average-linkage and Ward’s method are perhaps the most widely used, along with complete-linkage (\texttt{method = "complete"}) which is the default in \texttt{hclust}. The additional argument \texttt{hang = -1} will cause the leaves to extend to the baseline ‘Height’ of 0\(^2\).

More exotic ways of customising dendrograms exist, some of which will be illustrated, but however it’s done the first task (usually) is to make a decision about how many clusters there are in the data. A lot of rules have been proposed, some quite complex and, one suspects, not much used. Most simply what is often done is to try and cut the tree at a fixed level to try, ideally, to isolate branches that seem distinct from each other and which contain tightly clustered leaves within the branches. The tree in Figure 14.1 has been cut at a height of 3.5 to isolate five clusters of – from left to right – (2, 3, 12, 5, 3) leaves. If you are a ‘lumper’ you might initially cut at a height of 4 to get a (2, 15, 8) clustering; a ‘splitter’ might be inclined to split the larger cluster of 12 at a height of about 2.6 into smaller clusters of (3, 5, 4).

\(^1\)That I did not do this originally was because of unease about the sensitivity of results to the degree of smoothing used, based on experimental work with other data sets. The nettle is grasped here.

\(^2\)Mentioned because this is often the default in other software, and to be seen in many published applications.
Figure 14.1: A dendrogram from an average-linkage cluster analysis of partition SLs for 25 D.W Griffith’s Biograph films from 1911–12, from Table 14.1. Clustering is based on loess smooths of the SL data using a span of 1/2.

The chief merit of this approach is its simplicity, and you have to start somewhere, but it can have its problems. It is not obligatory, as just implied with the split of the larger cluster above, to cut all branches at the same level. In fact it often makes more visual sense to cut at different levels. What seems visually proper can, however, be misleading. This can be especially problematic with Ward’s method which, more so than other methods, can suggest quite clearly distinct clusters even when the data are random. Sometimes there will be more than one plausible way of cutting a tree and sometimes none at all.

This raises, in a fairly natural way, the question of how do you judge how ‘good’ a proposed clustering is, or adjudicate between competing clusterings? The question does not admit a simple answer unless grouping is so obvious that you don’t really need clustering methodology to detect it. Considerable intellectual effort has been expended on devising methods to address this issue, much of it too complex to be applied routinely by the non-specialist. Only some fairly simple and easily applied ideas are discussed here.

As a prelude to this an enhanced version of Figure 14.1 is shown in Figure 14.2, where the index of the film is replaced by the name of the film and the leaves are color-coded according to the cluster to which they are initially assigned (from the left, turquoise, magenta, red, blue and...}

---

3It can also be pretty bad at indentifying unusual data, which other methods are better at picking out. Users often like Ward’s method because of its apparent clarity, but appearances can be deceptive.
4Single-linkage is prone to this latter problem and is one reason why practitioners often avoid the method despite the ‘good’ theoretical properties claimed for it in some of the statistical literature.
This is the sort of thing produced for publication purposes after the more routine exploratory analyses using less ‘exciting’ graphs such as Figure 14.1. The color-coding is not strictly necessary here since, with the cut indicated, the proposed clusters are clear enough, but the coloring is useful for assessing how stable clusters are when the data are subjected to other forms of processing. For example, a common recommendation is to subject a data set to more than one method of clustering to assess how similar results are in terms of the clusterings they suggest. This requires an initial ‘starting position’ and Ward’s method can be useful for this because it can suggest clear clusters that are easy to identify (even if they turn out to be illusory and in need of refinement). Here, though, the five-cluster average linkage results will be taken as a starting point and used for labeling the single-linkage output in Figure 14.3.

I’d not normally use this for choice, if looking for clusters rather than outliers, because the output can be hard to interpret. That it can be difficult to use as the basis for identifying clustering is because the ‘staircase effect’, noticeable between The Miser’s Heart and The Female of the Species, often characterizes the entire dendrogram. An obvious level at which to cut it may not exist. Here, though, the ‘blue’ cluster suggested by average-linkage, separates out nicely, and the ‘red’ cluster, with the Burglar’s Dilemma as an interloper, also separates out reasonably well.

The higher up a cluster forms (and a ‘cluster’ can be a single case) the more isolated, or distinct, the cluster is. Single-linkage can be quite good at suggesting such cases, and in the present instance the suggestion is that, to the left, films between The Lonedale Operator and
Single–Linkage (Span = 1/2)

Figure 14.3: An enhanced dendrogram from a single-linkage cluster analysis of partition SLs for 25 of D.W. Griffith’s Biograph films from 1911–12. Clustering is based on loess smooths of the SL data using a span of 1/2.

(perhaps) The Miser’s Heart can be treated as reasonably distinct clusters or isolated cases. It will be seen shortly, using PCA that this is a reasonable interpretation. The dendrogram from a Ward’s method clustering (not shown) is consistent with the average-linkage clustering if branches are cut at different levels.

14.3 Principal component analysis (PCA)

In order to get some idea of how reasonable the proposed clustering is, plots based on PCA can be used to visualize the data. A plot of the first two PCs is shown in Figure 14.4, with labeling as in Table 14.1 and coloring as in Figure 14.2. The idea is that if a suggested cluster is split up in some way on the plot then it cannot be regarded as homogeneous. Thus film 1, His Daughter, clearly separates out from the other two films in the green cluster to which it was provisionally assigned. The principle operating here is that if films are well-separated on a PC plot this reflects the fact that they are genuinely different, whatever the cluster analysis might suggest. This applies to plots based on components other than the first two, if they contribute noticeably to variation in the data.

If points plot coherently for a cluster, as is largely the case for the red cluster, this is consistent with the notion that the cluster is a coherent one, but does not ‘prove’ the case. It is possible that plots using other than the first two components will split it up. His Daughter provides a nice
Figure 14.4: A plot of the first two principal components based on the scaled loess smooths (with a span of 1/2 of partition SLs for the films in Table 14.1. Colors correspond to the clustering suggested by the average-linkage analysis in Figure 14.2.

Illustration of the idea; although not placed in the red group by the cluster analysis it appears to plot with it in Figure 14.2. The first two PCs only explain 53% of the variation in the data, and the third component is needed to explain 71% of the variation. A quick way of investigating pattern among components other than the first two PCs is to use a scatterplot matrix, and a labeled plot of this kind for the first three PCs is shown in Figure 14.5. The upper-triangle of plots is just a reflection of those in the lower-triangle, which are those that would normally be inspected.

It can be seen that plots involving PC2 separate out the small cyan and magenta clusters, and those involving PC3 separate out the green cluster, so that a plot of PC2 against PC3 separates out all three clusters. This last plot does not separate out the red and blue groups, but they are separated on PC1. Taken together these plots establish that the five groups suggested by the cluster analysis can be regarded as distinct.

It does not establish their homogeneity. We have already seen that His Daughter separates out on PC1 from the other two in the green cluster, so should perhaps be treated as an outlier. On some plots some of the clusters are quite dispersed, suggesting that even if the grouping is sensible they may not be tightly defined. The other possibility, with the smaller clusters, is that the group is formed more because of the way they differ from everything else than because of intra-cluster homogeneity.

The analysis can be taken further in various ways. For individual clusters one possibility is to return to the original SL plots with the smooths of a span of 1/2 superimposed. This is shown for
Given the level of smoothing, what separates these films from many others in the sample is the peaking at round about the 80–85% mark. What distinguishes *His Daughter* from the other two films in the group is the ‘acceleration’ over much of the first half of the film, a characteristic of many of the films in the red cluster that it plots with in Figure 14.4.

Although this kind of pairwise comparison is possible for all the films and clusters it is tedious and partly misses out on one of the points of the exercise, which is to see if it is possible to group films in a way that allows a fairly simple summary of the differences between them. One way this might be done, which respects the possibility that films may be inappropriately assigned to a cluster, is illustrated in Figure 14.7 which is a side-by-side comparison of the average smoothed pattern for films within the ‘red’ and ‘blue’ groups, superimposed on a background of smooths of the individual films in the groups. A span of 1/2 has been used, and everything has been scaled to the interval [0, 1].

The difference between the average profiles for the two groups is clear. Films in the red group tend to have relatively fast cutting in their central portion (30–80%) being slower either side of this (with considerable variation between them at the start). Ignoring the first 5% or so of the film the general pattern is slow-fast-slow, [SFS]. Films in the blue group, by contrast, have an average...
profile that is [SFSFSF] with the earliest ‘fast’ section exhibiting the quicker cutting.

What is also evident is the variation of individual films about the average profile and differences between films within a group. This is particularly so for the blue cluster. That it is a distinct group can be seen from the single-linkage analysis in Figure 14.3 where, however, it can also be seen that some of the films cluster at a high level. This suggests that the cluster is being identified as such more because of large differences in their cutting pattern compared to other films rather than because of great similarity between films within the cluster. The red cluster looks more tightly defined though even here there are films (e.g., the one with an obvious peak just past the 60% mark) that deviate noticeably from the average pattern.

### 14.4 Varying the level of smoothing

The foregoing examples are intended to illustrate that interpreting the results from a cluster analysis or PCA is not straightforward if structure in the data is not ‘obvious’. There is the additional complication that the data on which analysis is based, as treated here, depend on the level of smoothing used. Figure 14.8 shows what happens if an average-link cluster analysis is applied to data smoothed using a span of 1/4 rather than 1/2. The labeling of the groups is that suggested by the analysis with a span of 1/2 in Figure 14.2.

It is immediately apparent that, as reflected in the dendrogram, the same grouping as previ-
ously suggested is not being reproduced. This is also apparent from the plot of the first two PCs of the smoothed data shown in Figure 14.9. There are, however, also strong similarities between the different clusterings. In the cluster analysis, with three exceptions (*The Last Drop of Water, Fighting Blood* and *The Lonedale Operator*), the red and green groups separate out from the other three groups. The last of these films, number 2 in Figure 14.2, sits among films from the red cluster in the cluster analysis, but can be seen to be something of an outlier on the PC plot.

![Average-linkage (Span = 1/4)](image)

**Figure 14.8**: A dendrogram from an average-linkage cluster analysis of partition SLs for 25 D.W Griffith’s Biograph films from 1911–12. Clustering is based on loess smooths of the SL data using a span of 1/4.

The PCA continues to separate out the rather dispersed blue cluster from the others, while films in the original turquoise, magenta and green clusters no longer group together. Some films do remain tightly clustered in all the analyses. Films 8, 18 and 20 (*The Adventure’s of Billy, The Lesser Evil and An Unseen Enemy*) form a cluster of three in all the analyses, joining with *The Girl and Her Trust* in the analysis with a span of 1/4. The four films plot closely in the PC plots for both analyses. In the analyses with a span of 1/2 *Fighting Blood*, film 4, plots closely with these four films but less so with a span of 1/2. From the original red cluster films 9 and 19, *The Battle and A Beast at Bay*, also form a tight pair that plot in the same region.

With the reduction in smoothing, films 3, 4, 5 (*The Indian Brothers, Fighting Blood, The Last Drop of Water*) separate out on the PC plot from others in the red cluster. It is tempting to interpret these as a sub-cluster of the orginal group, but the temptation should be resisted. Inspection of the dendrogram in Figure 14.8 show that they cluster separately. They remain separate from the blue cluster and from other outliers on the lower extremity of Figure 14.9 but, at this level of smoothing, the conclusion would seem to be that they differ from each other, and
from other films in the original red cluster. This is not surprising – as the level of smoothing is reduced individual characteristics of films that cause them to differ in structure will increasingly be emphasized and will increase the distance between them. What is perhaps more surprising is the way some films continue to group together as the level of smoothing is reduced.

Figure 14.9: A plot of the first two principal components based on the scaled loess smooths (with a span of 1/4) of partition SLs for the films in Table 14.1, from which the numbering is taken. Colors correspond to the clustering suggested by the average-linkage analysis in Figure 14.2.

To get some idea of what is going on Figures 14.10 and 14.11 are presented. Three films, Fighting Blood (4), The Girl and Her Trust (15) and An Unseen Enemy (20) have been selected. As already noted these cluster and plot closely in the analyses with a span of 1/2, but Fighting Blood becomes ‘detached’ as the smoothing is reduced to a span of 1/4. In Figure 14.10, for all three films, the light-blue lines were obtained from smooths for all spans between 0.2 and 0.8 at intervals of 0.01. The smooths for spans of 1/2 and 1/4 are highlighted. The strong ‘band’ in each plot is accounted for by smooths with the smaller values and is well-represented by the smooth with a span of 1/2. This gives rise to a ‘dominant’ trend that is broadly U-shaped that characterizes most of those films grouped together in the red cluster at this level of smoothing. The similarity, or otherwise, of the smooths is not immediately apparent because of variation in the magnitude of SLs. This was adjusted for in the multivariate analyses by scaling the fitted smooths to lie between [0, 1] and this is done in Figure 14.11. In the left-hand plot, where the three films are highlighted, Fighting Blood and The Girl and Her Trust are closer to each other than to An Unseen Enemy, which is why they cluster together in Figure 14.2.

As the smoothing is decreased the smooths become more ‘wiggly’ with the wiggles being differently placed for different films, causing the distance between films to increase. The extremities
SLs which allows a data matrix to be constructed for a body of films where there are an equal number of observations per film.

Tsivian’s (2013) observation that all films have different SL structures but some are more different than others was one of the stimuli for the investigation conducted in this chapter. The question is, how do you establish that films are differently different, and how do you characterize ‘different’? The starting point here was to use Brisson’s methodology to convert a set of SLs to ‘partition’ SLs which allows a data matrix to be constructed for a body of films where there are an equal number of observations per film.

Brisson’s methodology was originally used to generate a curve representing the averaged pattern across a body of D.W. Griffith’s Biograph films, of which the data used here is a subset. A problem is that few individual films have a cutting pattern that closely resembles the average...
(Baxter, 2013d), and it seems better to try and disaggregate the body of films into subsets having different average structures before computing averages. This was the approach adopted in Baxter (2013d) and the present and previous chapters explore and develop ideas introduced there and in Redfern (2013b).

The central idea is to apply standard methods of multivariate analysis to rectangular tables of ‘adjusted’ SLs to identify groups with common cutting patterns. This can be based on either the original data or partition SLs. In the former case some form of smoothing seems necessary so the smoothed curve can be sampled at an equal number of points to generate the required data matrix. Brisson’s partitioning methodology produces a table of data that might be processed without further smoothing, but it is advantageous to do so and raises explicitly the question of how much smoothing to apply.

This is not easily answered, but a tentative suggestion is offered here. What has been referred to above as a ‘dominant’ trend in a set of SL results can be identified using a relatively high degree of smoothing, and for loess smooths a span of 1/2 (with non-robust localized quadratic fitting) seems to work reasonably well. This smooths out detail that does differentiate between films, but also opens up the prospect of grouping films with broadly similar cutting patterns over their length. Without any smoothing all films are different and this will become increasingly apparent as the degree of smoothing is reduced.

What may happen, though, is that films become increasingly differentiated from those with which they share the same basic pattern but maintain a considerable distance from films that have a rather different dominant pattern. This thought is complicated by the fact that dominant patterns form a continuum which any analysis of this kind divides in an unavoidably arbitrary way. As the level of smoothing is varied films may ‘shift their allegiance’ to groups close in the continuum to that which they were originally assigned to. The idea, however, is that they are much less likely to ‘shift’ to a group whose dominant pattern is not close to that to which they were originally assigned.

More exploration is needed to see if this kind of idea ‘works’. If it does it provides some basis for making assertions of the kind that film A is ‘more’ or ‘less’ different from film B than C and, with all the obvious caveats, making some attempt at classification.

As far as the Griffith Biographs go, the sample used here is too small to admit much generalization, as are many of the ‘clusters’ suggested by analysis. My current thinking is that early films with relatively few shots – less than 50 say – do not admit easy classification. Some evidence for this, from another angle, is provided by the analysis of MSL/ASL ratios in Section 11.4.3. Analyses similar to those described in this section that include all 62 films used in Baxter (2013d) tend to place the 1909 films and 1910 films with fewer than 50 shots around the periphery of PC plots, distinct from most of the 1911–1912 films. That is, the search for common statistical structure is possibly best confined to later films with a reasonable number of shots – say 70 or more – where there is greater freedom for the expression of structure.

The other thing that now seems fairly obvious is that it is statistically possible to define a ‘Griffith average’ but a lot of films have a structure that bears little relation to this. The average for the 62 films used in Baxter (2013d) has something like a reverse-J pattern with (ignoring beginnings) the suggestion of accelerated cutting up to about the 70–75% mark and a slowing down thereafter. The suggestion here and in Baxter (2013d) is that a more prevalent ‘dominant’ pattern is the U-curve shown by the films in Figure 14.10 where there is – within the film – relatively fast cutting in a central portion that starts before the 50% mark. There are lots of films which don’t have this pattern but it seems to be the most common one in the available sample. That the ‘Griffith average’ is not like this is because a lot of films have some sort of ‘climax’ somewhere in the 70–75% region (by no means all and it can be at a distance from this). This averages out as a climax in this region; otherwise, dominant patterns that have this feature differ in the positioning of other more quickly-cut passages and these cancel each other out in the overall averaging process.

That is, attempting to identify pattern in structure using statistical methodology is complicated. The present case study is a continuation of earlier work that has explored methodological possibilities. Ultimately I think judgment of the ‘value’ of the outcome is not a statistical issue.
In a sense all films are unique but they get classified anyway – into ‘genres’ for example. Classification by cutting patterns depends, at least as attempted here, on the degree of smoothing that is considered acceptable and this, I think, needs to be judged from a ‘filmic’ rather than statistical perspective.
References


Cleveland, W.S. 1979: Robust locally weighted regression and smoothing scatterplots. Journal of


Faraway J.J. 2006: Extending the Linear Model with R. Boca Raton: Chapman and Hall/CRC.


Bibliography

This is a first attempt to compile a list of cinemetrics publications. There are sufficiently few of these around at present for the attempt to seem feasible, though I am conscious that it is currently confined to English language publications. A few papers that involve the statistical analysis of film-related material that cannot be classified as cinemetrics have been included. It will be obvious from looking at the bibliography that much of what has been published is on the web and that a fairly small number of scholars have generated much of the content. There is some overlap between entries in this list and those in the references, but not everything below is noted in the text.


SPIE Conference on Storage and Retrieval for Media Databases 2002


**APPENDIX:**  The Lognormal distribution

**Introduction**

The lognormal distribution has been proposed as a model for shot length (SL) distributions by Barry Salt, the chapter *The Numbers Speak* in Salt (2006) being the fullest exposition (see, also, Salt, 2011). There has been some debate about how generally applicable this model is, discussed in Chapter 12. What follows collects together the mathematical details concerning the lognormal and normal distributions, interpretation of parameters in them and their estimation. Some attention is paid to notational distinctions, not always made in the literature, which can lead to confusion if neglected. An attempt is made to explain what logarithmic transformation does; the idea is not what many people would regard as ‘natural’ but it is a useful one that anyone seriously interested in cinemetric data analysis should try to become comfortable with.

**The lognormal and normal distributions**

**Definitions**

The mathematical form of the lognormal probability density function is

\[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2\sigma^2}(\ln x - \mu)^2\right] \]

where \( x \) is strictly positive. The distribution is skew, bounded below by zero and tails off to a density of zero to the right as \( x \) approaches infinity. It is completely defined by the two parameters \( \mu \) and \( \sigma \).

Other ways of writing (parameterizing) the equation can be used such as

\[ f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2\sigma^2}(\ln x/\Omega)^2\right] \]

where \( \ln \Omega = \mu \). This is the form used by Salt (2006: p.391).

Let \( X \) be a random variable with a lognormal distribution and write

\[ X \sim \text{L}(\mu, \sigma^2) \]
to symbolize this. The logarithm of a lognormal variable, \( Y = \ln X \), has a normal distribution and we write

\[ Y \sim N(\mu, \sigma^2) \]

where the probability density function of \( Y \) is

\[ f(y) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma^2} (y - \mu)^2 \right] \]

which is also completely specified by \( \mu \) and \( \sigma \) and is singly-peaked and symmetric about \( \mu \).

**Parameter interpretation**

It is simplest to start with the normal distribution, for which the *mean* is \( \mu \) and *standard deviation* \( \sigma \). The *variance*, \( \sigma^2 \), is the square of the standard deviation. The *median* and *mode* of the normal distribution are the same as the mean, \( \mu \), but this is not generally true for non-symmetric distributions.

The mode, median, and mean are *measures of location*, the mode being the point at which the maximum value of the density occurs, the median being the point that has 50% of the distribution either side of it, and the mean being the centre of gravity of the data. The median can be referred to, though this is less common, as the 50th *percentile*. In common usage the undefined term ‘average’ is used sloppily without any clear indication of what is intended. Sometimes it is implicit that the (arithmetic) mean is intended and sometimes the median. The cinematics literature is not immune to this.

The standard deviation is a *measure of dispersion* with the useful interpretation, for the normal distribution, that just over two-thirds of the distribution lies within one standard deviation of the mean, and about 95% lies within two standard deviations. Again, this is not generally true.

The mean and standard deviation are often used in conjunction to summarize the location and dispersion of a set of data, whether normal or not. The median is the 50th percentile of the data; percentiles for other values can be defined; for example the 25th percentile has 25% of the distribution to its left and 75% to the right, and is sometimes
called the first \textit{quartile}. The 75\% percentile, the third quartile, can be similarly defined. The difference between them, the interquartile range (IQR), is often used as a measure of dispersion in conjunction with the median as a measure of location. The \textit{coefficient of variation} is a standardised (scale-free) measure of dispersion, defined as the standard deviation divided by the mean, so that for the normal distribution 

\[
CV = \frac{\sigma}{\mu}.
\]

For the lognormal distribution there is not a simple exact equality between population parameters and readily understood characteristics of interest descriptive of the population. The mean is greater than the median which is greater than the mode median and mode. In terms of the parameters of the distribution the mean is 

\[
\mu_L = \exp(\mu + \sigma^2/2) = \Omega \exp(\sigma^2/2)
\]

the median is 

\[
\Omega = \exp(\mu)
\]

and the mode is 

\[
\exp(\mu - \sigma^2).
\]

The mean, \(\mu_L\), is the ‘average shot length’, what is often denoted by ASL being its estimate. Note that if lognormality holds exactly there is a perfect linear relationship between the mean and median. The median, \(\Omega\), is also the \textit{geometric mean}, to which we return in the next section. The variance of the lognormal distribution and CV are 

\[
\sigma^2_L = [\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2).
\]

\[
CV_L = \sqrt{\exp(\sigma^2) - 1}
\]

The median and CV are functions, respectively of \(\mu\) and \(\sigma\) only. Unlike the normal distribution the parameter \(\sigma\), does not equate with dispersion in any simple way. In the context of the lognormal it is often referred to as a \textit{shape} parameter, but the shape of the lognormal, as many would comprehend it, depends on both \(\sigma\) and \(\mu\). The lognormal can be described mathematically in terms of two numbers, \(\mu\) and \(\sigma\),
which don’t have the ‘direct’ interpretation that they have for the normal.

It can be emphasized that even in the ideal case of the lognormal two quantities are needed to describe the distribution fully. They should involve both $\mu$ and $\sigma$ and this pair $(\mu, \sigma)$ is one obvious possibility. Baxter (2012a) suggests $(\mu_L, \Omega)$; Salt (2011) suggests $(\mu_L, \sigma_L)$. There is no mathematical basis for preferring any of these, which in practice must be estimated, so whatever is most convenient to calculate, or most usefully interpretable in context, are reasonable grounds for choice, if one has to be made.

Models and estimates

Models

The lognormal and normal distributions are idealized models for data, of value practically because many different kind of real-life data approximate the distributions sufficiently well that the models, and their properties, can be used as the basis for very concise descriptions of the data aiding, in some circumstances, an understanding of the processes that generate the data.

Models are (almost) never correct. This is obvious for the theoretical normal and lognormal distribution, which disappear into infinity in both directions for the normal and to the right for the lognormal. Real data can’t do this. In practice the most one can hope for is that over the relevant range of a distribution observed data ‘approximately’ match that which would be predicted by the model under consideration.

Descriptive statistics – the ASL, MSL, standard deviation, IQR, whatever – are attempts to usefully quantify aspects of data that provide, one hopes, useful ‘filmic’ insights into it. Such efforts can be conceived of as either of merit in their own right or, additionally, as estimating characteristics of a distribution of which the SL data are a sample. In this latter case it is not necessary to assume any particular distribution that underlies the data; if one does make such an assumption and wants to check it model-fitting, is reached.
Estimation and notation

Differentiating between characteristics \textit{estimated} from a set of data and the analogous characteristics or parameters of the hypothesized and idealized underlying model is important. The latter are unknown and estimates, which are calculated and known quantities, are typically meant to be usefully informative about what is unknown. It is important to maintain a notational distinction between the two, which can be confusing when done properly, but is even more so when not. As an example take the average shot length (ASL) which can be written variously, as

\[
\text{ASL} = \hat{\mu}_L = \bar{x} = \frac{\sum x}{n}
\]

where the last term just means to add the individual shot lengths (the \(x\)) and divide by the number of shots, \(n\). The circumflex, or ‘hat’ notation is applied to distinguish estimated sample quantities from their unknown population counterparts. Conventionally, the latter are indicated using letters from the Greek alphabet. Thus \(\hat{\mu}_L\) estimates the unknown population mean \(\mu_L\), the subscript, \(L\), distinguishing this from the population parameter \(\mu\).

Other notation is in common use and \(\bar{x}\) is standard notation for the mean of a set of numbers denoted by \(x\). Terminology, alas, can also be confusing. Salt, in coining ASL, was well aware that ‘arithmetic mean’ was the strictly correct terminology but opted for ‘average’ on the grounds that it was what his intended audience could live with (Salt, 2012). The qualifying adjective, ‘arithmetic’ is often dispensed with but it can be useful to retain, as other kinds of mean can be defined mathematically, and appear in cinemetric studies.

If similar calculations are applied to \textit{logarithmically transformed} data

\[
\hat{\mu} = \bar{y} = \frac{\sum y}{n}
\]

which provides a simple and direct way of estimating \(\mu\), which may be used for fitting the lognormal model. Note that \(y = \ln x\) is used in the calculations. The parameter \(\sigma\), which happens to be the standard
deviation of the normal distribution, is most simply estimated from
\[ \hat{\sigma}^2 = s^2 = \frac{\sum (y - \bar{y})^2}{n - 1} \]
where \( s \) as an estimate of standard deviation is also in common use\(^5\). Replacing \( y \) with \( x \) in the above provides an estimate of \( \sigma_L^2 \).

The median of the lognormal, \( \Omega \), can be estimated in various ways. Thinking of this as a population characteristic, the obvious estimate of it is the sample median, \( \hat{\Omega}_M \), where the subscript \( M \) is introduced to distinguish it from other possibilities. One such, as noted in passing above, is to use the geometric mean of the data, that is \( \exp(\hat{\mu}) \) where \( \exp() \) is the exponential function and \( \hat{\mu} \) is the arithmetic mean of the logged data. The geometric mean of a set of \( n \) numbers is defined as the \( n \)th root of the product of the numbers so
\[ \hat{\Omega}_G = (\prod x)^{1/n} \]
giving
\[ \ln \hat{\Omega}_G = \ln [(\prod x)^{1/n}] = \frac{\sum \ln x}{n} = \frac{\sum y}{n} = \hat{\mu} \]
and back-transforming to
\[ \hat{\Omega}_G = \exp(\hat{\mu}) \].

The mathematics here is for the record and can be ignored, so long as it’s appreciated that more than one estimate of the median is available. Parkin and Robinson (1993) study four possibilities, including the two mentioned here which are the simplest. Their conclusions are not simply summarized and neither of the two estimators considered here are unequivocally better than the other. The comparative use of the statistics is illustrated in Chapter 12.

**Why logarithms?**

The use of logarithms is unavoidable in discussing the lognormal distribution. What is involved is a transformation from one set of numbers to another, \( X \rightarrow Y \) say, or
\[ y = f(x) = \ln x. \]

\(^5\)In some statistics texts the divisor \( n \) rather than \((n - 1)\) is used. The reasons for this, where it is knowingly and correctly done, are not of concern here, since \( n \) is usually so large the difference has no effect on calculations. Similar comments apply when, as sometimes happens, \( \hat{\sigma}^2 \) and \( s^2 \) are defined with different divisors.
Back transformation, \( x = \exp(y) \), is possible so you can always get from one set of numbers to another. The transformation is *monotonic* meaning that it retains the same ordering of numbers. In terms of application to SLs a log-transformation does several things. SLs are bounded below by zero and distributions are skewed, often with a long tail. For some statistical purposes the lower bound of zero is an inconvenience and the log-transformation frees the data from this constraint so that negative numbers are possible and the data are (theoretically) unbounded. The values of SLs, measured in seconds, differ by orders of magnitude, meaning that a shot of 1 second is ten times longer than a 1 deci-second (0.1 second) shot; one of 10 seconds is 10 times greater than a 1 second shot and so on. This can be inconvenient for both statistical analysis and graphical display because results can be dominated, not necessarily beneficially, by the larger SLs.

The effect is removed by logarithms which transform SLs to the same order of magnitude so that (using logarithms to base 10) the logarithms of 0.1, 1, 10, 100, become -1, 0, 1, 2. (Logarithms to other bases can be used and the numbers will differ, but the difference between them will be constant\(^6\)) What this means, in graphical terms, is that equal weight is given to short and long SLs so that, for example, SL data that are lognormally distributed are symmetrically and normally distributed after log-transformation. Departures from log-normality on the original scale will appear as departures from normality on the log-scale, and it is easier to make visual judgments about such departures with a normal distribution as the reference. For making such assessments, including formal tests of log-normality, log-transformation is not only a possibility, but also the ‘natural’ thing to do. Several examples of the comparison of SL distributions on a log-scale are provided in the text.

Log-transformation also has the effect of downweighting the influence of outliers if they exist, without sacrificing information about the fact that some SLs are noticeably bigger than others – information that is ‘lost’ if rank-transformations are used, for example. In this sense

---

\(^6\)Logarithms can be to different bases; natural logarithms, to base \(e\), have been used in the text unless otherwise stated. These are sometimes written as \(\log_e\) in contrast to \(\log_{10}\) for base 10 logarithms. In \(R\) the \(\log\) function – that mostly used – provides natural logarithms. Using \(\log_{10}\) will provide logarithms to base 10.
using a log-transformation has some of the advantages claimed for ‘robust’ methodologies, without the sacrifice of information inherent in the latter usage.

The price paid for the often considerable advantages of using log-transformation is the discomfort that can be induced induced by the abstraction involved – a distancing from the ‘reality’ of the raw data. A shot is composed of a continuous sequence of frames, broken at either end by a cut or some other recognisable form of transition. If you’re looking for it you can recognise a shot when you see one and get some ‘feel’ for its length; you don’t think or ‘feel’ in terms of the logarithms of SLs.

What’s often sensible is to take a deep breath, do an analysis using logarithms then, if this causes discomfort, translate back to the original scale and see if the conclusions drawn from the log-based analysis convince you.

Other normalizing transformations

A logarithmic transformation of the data is often useful for getting a better ‘look’ at some aspects of the data. Beyond this there is a specific interest in using it to investigate the claim that a majority of films have SL distributions that approximate lognormality. This involves transforming the original data, $X$ to $Y = \ln X$ and investigating normality of $Y$.

Where this is not successful it is worth asking whether the data can, nevertheless, be transformed to normality using other than a logarithmic transformation. At the cost of what is only a slight additional mathematical complexity, this is readily explored. A rather simplistic way of thinking about the lognormal transformation is that it ‘squeezes’ in the long right tail, allowing more freedom to the left tail, in an attempt to get the data to ‘behave’ normally.

It sometimes works, but not always – the data may be recalcitrant, but we can look to see if a bit of an extra squeeze might remedy matters. The thought could also occur that you might have tried squeezing differently in the first place. This is, more-or-less, what’s attempted with the two transformations to be described.
The *Box-Cox* (BC) transformation applies differential ‘force’ from the word go; mathematically the amount of force needed to get the data to behave as well as possible (it may still end up behaving badly) is determined by an extra parameter, $\lambda$, that is where the additional complication comes in. This value is close to zero if the data are acceptably lognormal, and tends to become more negative as more force needs to be exercised to achieve normality (this observation is based on SL data I’ve looked at and does not apply generally to other kinds of data).

The other approach investigated initially applies equal force to the data, via a gentle log-transformation, but then applies an additional and differential squeeze to recalcitrant data sets. What motivates this idea is that initial log-transformation is intended to remove skewness, but often leaves the data still looking skew. A second log-transformation might seem in order but mathematical complications arises with this. Specifically, the first log-transformation may produce negative values, and a second log-transformation cannot be applied to these. To get round the problems here the use of the *Yeo-Johnson* (YJ) transformation was explored. It looks (and is) more complicated than the BC transformation, but depends on only a single extra parameter.

The BC and YJ transformations are outlined mathematically below. Their application, particularly of the latter, is explored in the Chapter 12. What’s involved is a double-transformation or double-‘squeeze’ of the data, with uniform squeezing at the first stage and differential squeezing at the second stage.

**Box-Cox (BC) transformations**

Power transformations of the form $X^\lambda$ have been used in statistical analysis for some time; the slightly more ‘complex’ version studied by Box and Cox (1964) (see Sakia, 1992, for a review of early work)

$$X^\lambda = \frac{X^\lambda - 1}{\lambda}$$

is that used here. In applications a transformation, $\lambda$, is sought that makes the data as nearly normal as possible. It is defined for $X > 0$,
in such a way as to ensure continuity, with a value of $\lambda = 0$ equivalent to a log-transformation. That is, the log-transformation that can be used to investigate the lognormality of SL distributions is a special case. If it is concluded that, for any particular film, lognormality does not hold it is possible to see if some other value will induce normality. Theory and advice on practical application is readily enough found in the literature; for practical analysis a variety of functions in R can be found and the `powerTransform` and `bcPower` functions from the `car` library are those used in the text.

**Yeo-Johnson (YJ) transformations**

The Yeo-Johnson transformation looks somewhat nastier, though at heart it is the same kind of thing as the BC transformation modified to deal with negative values. Here it is applied after an initial log-transformation, $Y = \ln X$. This may induce negative values, which is why the extra complexity of the YJ transformation is needed. It is defined as

$$
\begin{align*}
(Y^\lambda + 1) - 1 &/ \lambda & \text{if } \lambda \neq 0; y \geq 0 \\
\ln (Y + 1) & & \text{if } \lambda = 0; y \geq 0 \\
-[(Y + 1)^{2-\lambda} - 1] / (2 - \lambda) & & \text{if } \lambda \neq 2; y < 0 \\
-\ln (-Y + 1) & & \text{if } \lambda = 2; y < 0
\end{align*}
$$

and can be implemented using the `powerTransform` function from the `car` package in R. The meaning of $\lambda$ can be hard to interpret; the transformation can be viewed as a generalization of the BC transformation that amounts to applying different power transformations to negative and positive $Y$, using $(|Y| + 1)$ or $(Y + 1)$, with powers $(2 - \lambda)$ or $\lambda$, the additional complexities ensuring continuity.
Index

act structure in films, 11, 78, 87, 93

ASL
compared to MSL, 27
may be uninteresting, 46
national comparisons, 6–7
original use of, 31
variation over time, 5–6, 28
average shot length, see ASL

boxplots, 41–45, 49
adjusted for skewness, 44
construction, 41
definition of unusual values, 41
limitations of, 42
of MSL/ASL ratio, 122
outlier detection, 118–119
outliers
misinterpretation, 43

Cinemetrics website, x, 2
On Statistics discussions, x
Looking for Lookalikes?, 2, 141, 150
Median or Mean?, 1, 27, 113
What Do Lines Tell?, 150
database, 10, 22, 24, 108, 123
measurement of SLs, 32
graphs, 25
measurement software
graphs, 79

Salt/Cutting debate, 12
cluster analysis, 140, 146–147
applications, 151–154
average-linkage, 146, 151
completelinkage, 151
dendrogram, 146
dendrograms, 151
hierarchical clustering, 147
quality of clustering, 152
selecting number of clusters, 151
single-linkage, 151, 153
Ward’s method, 151, 153

comparison
ASL vs. MSL, 113–126
D.W. Griffith’s Biograph films, 124–126
silent-sound transition, 120–123
temporal variation, 119
controlling for shot numbers, 86–89
Hitchcock and Ophuls
use of shot-scales, 102
Hitchcock shot-scale data
silent vs. sound, 96
ASL vs. MSL, 110
of cluster analyses, 153
of estimates of median, 129–130
of KDEs after log-transformation, 128
correspondence analysis, 140
of film scripts, 140
of genre preferences, 140
of shot-scale data, 15–20, 94–103, 140
of shot-type in Rashomon, 140
seriation, 20
supplementary points, 102–103
types of map, 100
data transformation
Box-Cox, 137, 180
Box-Cox
dichotomization, 82
double-transformation, 137
log-transformation, 37, 50
and orders of magnitude, 145
effect of, 75
effect on smoothing, 66
logarithmic
normalization, 13
ranked data, see robust statistics
scaling to remove ‘size’ effects, 142–144
standardization, 51
confused with normalization, 13
Yeo-Johnson, 137–139, 181
descriptive statistics, 24–34
$Q_n$ (dispersion), 30, 121
$S_n$ (dispersion), 30
arithmetic mean
definition, 31
average shot length, see ASL
coefficient of variation (CV)
definition, 33
confusing the median and mean, 33
five-number summary, 34, 41
geometric mean
definition, 33
estimate of median in lognormal, 129
interquartile range (IQR), 25, 27, 41
definition, 33
kurtosis, 111
maximum, 32
measure of dispersion, 32
measure of location, 32
measures of dispersion, 30
measures of location
as measures of film style, 114
median
comparing estimates of, 129–130
definition, 32
median absolute deviation (MAD), 30
median shot length, see MSL, see MSL
minimum, 32
mode, 8
and boxplots, 42
definition, 32
in KDE, 123
modal class, 32
multi-modality, 32
percentile, 25
definition, 33
quartile, 25
definition, 33
range
definition, 32
skew, 111
skewness, see skewness
standard deviation, 27
definition, 33
variance
definition, 33
directors
Griffith, D.W., 124–126, 140
The Pace in Movies, 144
Hitchcock, Alfred, 16, 94–97
Lang, Fritz, 5, 13
May, Joe, 16
Ophuls, Max, 16, 102–103
von Sternberg, Josef, 10
distance measures, see multivariate analysis
distributional regularity, 128, 137–138
(, 128
films
Bambi (1942), 86
Ronde, La (1950), 103
A Night at the Opera (1935), 55
Adventures of Billy, The (1911), 126, 150
Applause (1929), 123
Aristocats (1970), 52, 56
Barry Lyndon (1975), 56, 115
Battle, The (1911), 126, 150
Beast at Bay, A (1912), 126, 150
Beyond a Reasonable Doubt (1956), 13, 16
Billy’s Stratagem (1911), 126
Birds, The (1963), 94
Blackmail (Silent version)(1929), 94, 95
Blackmail (Sound version)(1929), 94, 95, 106
Blue Angel, The (1930), 10
Brief Encounter (1945), 8, 9, 56, 108, 115, 117–119, 128, 131
Burglar’s Dilemma, The (1912), 150, 153
Captain Blood (1935), 56
Carmen Jones (1954), 127
Cast Away (2000), 29
Caught (1949), 102
Champagne (1928), 94
Child’s Stratagem, A (1910), 126
Cloak and Dagger (1946), 13
Coal Miner’s Daughter (1980), 29
Corner in Wheat, A (1909), 141–145
Country Cupid, A (1911), 150
De Mayerling à Sarajevo (1940), 102
Death’s Marathon (1913), 126
Detour (1945), 28
Dial M for Murder (1953), 94
Dial M for Murder (1954), 17
Downhill (1927), 94
Dr. Mabuse der Spieler (1) (1922), 13, 16
Dr. Mabuse der Spieler (2) (1922), 13
Easy Virtue (1928), 46–50
Easy Virtue (1928), 94
Exile, The (1948), 102, 103
Exodus (1960), 28, 52, 56, 135
Fantasia, 112
Fantasia (1940), 56
Farmer’s Wife, The (1928), 94
Female of the Species, The (1912), 153
Female of the Species, The (1912), 150
Fighting Blood (1911), 126, 150
Five Easy Pieces (1970), 29
For His Son (1911), 150
Foreign Correspondent (1940), 94
Friday the 13th (1980), 92
Friends (1912), 126, 150
Front Page, The (1931), 123
Fury (1936), 13
Girl and Her Trust, The (1912), 126, 150
Great Dictator, The (1940), 56
Halloween (1978), 85–89, 92
Hangmen Also Die (1943), 13, 15
Harvey (1950), 8, 9, 28, 42, 56, 109, 128, 132
His Daughter (1911), 150, 154–156
House on Sorority Row, The (1983), 90–92
Human Desire (1954), 13
I Confess (1952), 94
Indian Brothers, The (1911), 125, 150
Indische Grabmal, Das (1921), 16
Indische Grabmal, Das (1959), 13, 15
Intolerance (1916), 76, 79, 84
Jamaica Inn (1939), 94, 101
Jaws (1975), 29
Juno and the Paycock (1930), 94, 101
King Kong (2005), 11, 108
Komodie om Geld (1935), 102
Kriemhild’s Rache (1924), 13, 15
La Ronde (1950), 17
Lachende Erben (1932), 17, 102
Lady Vanishes, The (1939), 94, 95
Last Drop of Water, The (1911), 150
Le Plaisir (1953), 17
Lesser Evil, The (1912), 126, 150
Letter from an Unknown Woman (1948), 102
Liebelei (1932), 17, 102
Lifeboat (1944), 94
Lights of New York (1928), 22, 26, 43, 79–84, 88, 114, 116, 122
Lodger, The (1927), 94
Lola Montès (1955), 102
Lonelydale Operator, The (1911),
126, 142–145, 150, 153
*Müde Tod, Der* (1921), 13
*Madagascar* (2005), 56
*Madame de . . .* (1953), 17, 102
*Man Who Knew Too Much, The* (1935), 94
*Manxman, The* (1928), 94
*Mender of Nets, The* (1912), 126, 150
*Metropolis* (1926), 13
*Ministry of Fear, The* (1945), 13
*Miser’s Heart, The* (1911), 150, 153, 154
*Movie Crazy* (1932), 123
*Murder* (1931), 94
*Musketeers of Pig Alley, The* (1912), 150
*M* (1931), 13, 15, 16
*New World, The* (2005), 114
*New York Hat, The* (1912), 150
*Night at the Opera, A* (1935), 21, 24, 35–42, 45, 52, 55, 69, 131
*Night of the Hunter* (1955), 56
*North by North-West* (1959), 94
*Notorious* (1947), 94
*Number Seventeen* (1932), 94
*One is Business, the Other Crime* (1912), 150
*Painted Lady, The* (1912), 143–144, 150
*Paradine Case, The* (1947), 94
*Pinocchio* (1940), 90–92
*Plaisir, Le* (1952), 102
*Pleasure Garden, The* (1925), 94
*Pursuit to Algiers* (1945), 56, 132
*Rain* (1932), 123
*Rashomon* (1950), 140
*Rear Window* (1954), 94
*Rebecca* (1940), 94
*Reckless Moment, The* (1949), 102
*Rich and Strange* (1932), 94, 102
*Ride Lonesome* (1959), 58, 62–64
*Ring, The* (1928), 94
*Rocky Horror Picture Show, The* (1975), 29
*Ronde, La* (1950), 102
*Russian Ark* (2002), 117
*Sabotage* (1937), 94
*Saboteur* (1942), 94
*Sans Lendemain* (1939), 102
*Scarlet Empress, The* (1934), 114, 116
*Secret Agent, The* (1936), 94
*Secret Beyond the Door, The* (1948), 13
*Seven Year Itch, The* (1955), 28, 117
*Shadow of a Doubt* (1943), 94
*Siegfried* (1924), 13, 15
*Signora di Tutti, La* (1934), 102
*Skin Game, The* (1931), 46–50, 94
Slumber Party Massacre, The (1982), 92
Spinnen, Die (1) (1919), 13, 16
Spinnen, Die (2) (1920), 13
Spione (1928), 13, 16
Strangers on a Train (1951), 94
Sunbeam, The (1911), 150
Suspicion (1942), 94
Swords and Hearts (1911), 126, 150
Tendre Ennemie, La (1936), 102
Testament des Dr. Mabuse, Das (1933), 13, 15, 16
The Scarlet Empress (1934), 43
The Trouble with Harry (1954), 17
Thirty-Nine Steps, The (1935), 94
Those Magnificent Men in Their Flying Machines (1965), 29
Those Magnificent Men in their Flying Machines (1965), 117
Tiger von Eschnapur, Der (1959), 13, 15
To Catch a Thief (1955), 17, 94
Top Hat (1935), 64–74, 92
Trouble With Harry, The (1956), 94
Trouble in Paradise (1932), 123
Unseen Enemy, An (1912), 150
Unseen Enemy, The (1912), 126
Verkaufte Braut, Die (1932), 102
Verliebte Firma, Die (1931), 102
Vertigo (1958), 94
Walk the Line (2005), 54–55
Waltzes from Vienna (1934), 94, 101
Wedding Crashers (2005), 115, 117
Werther (1938), 102
While the City Sleeps (1956), 13, 16
Woman in the Window, The (1945), 13
Wrong Man, The (1957), 94
Yoshiwara (1937), 102
You Only Live Once (1937), 13, 16
You and Me (1938), 13
Young and Innocent (1938), 94

graphical methods
bar chart, 13
shot-scale data, 95
bi-axial graphs, 141
boxplots, see boxplots
color-coding plots, 79–82
convex hull, 15–17
cumulative frequency diagram (CFD), 7, 49–50
histogram, 6, 8, 9, 13
too few bins, 13
histograms, see histograms
kernel density estimate (KDE), see KDE, 74
loess smooth, see smoothing
moving average, see time-series
order structure matrices, see time-series
probability plots, 52–56
and tail behavior, 128
scatterplot matrix, 27, 111
of principal components, 155
violin plot, 41, 43, 49

histograms, 35–37, 46
anchor point, 36
bin-width, 35
comparison of, 128
confusion with bar charts, 13
construction, 36–37
limitations of, 38
too few bins, 54
unwieldy for comparison, 46

hypothesis test
of normality, 56
Shapiro-Francia test of normality, 57

hypothesis tests, 112
p-values, 112
Jarque-Bera normality test, 135
as basis for ‘index’ of normality, 137
misinterpretation of, 136
Shapiro-Francia normality test, 135
Shapiro-Wilk normality test, 135
idea behind, 135
in autoregression models, 11
Jarque-Bera test of normality, 112, 135
Mann-Whitney, 121
Mann-Whitney test, 70–71, 77
of kurtosis, 133–134
of skewness, 133–134

omnibus tests of normality, 134–137
populations and samples, 31
sample-size effect, 135
Shapiro-Francia test of normality, 112
Shapiro-Wilk test of normality, 112
two-sample t-test, 121
using shotscale data, 104–107

KDE, 38–41, 46–49
advantage over histogram, 56
and color-coding plots, 87
bandwidth, 38
bandwidth choice, 38
bandwidth selection, 88
comparing ASL distributions, 7
comparison after log-transformation, 128
construction, 38
for assessing normality, 55
for SL comparisons, 52–57
in scatterplot matrices, 28
modes, 123
of cut-point data, 73–74
of MSL/ASL ratio, 121, 123
of SL data, 9
kernel density estimation, see KDE
kurtosis, 131–134
definition, 131
excess kurtosis, 131
hypothesis tests of, 133–134
platykurtic and leptokurtic, 131
testing significance, 133
loess smoothing, see smoothing
lognormal distribution
and SLs, 8
lognormality
and outliers, 44
comparison of SL distributions with, 51–52
of SL distributions, 28, 36, 127–139
MSL
compared to ASL, 27
may be uninteresting, 46
multivariate analysis, 140–149
cluster analysis, see cluster analysis
correspondence analysis, see correspondence analysis
distance measures, 145
Euclidean, 145
principal component analysis, see principal component analysis (PCA)
order structure matrices, see time-series
outliers
in lognormal distribution, 118
outliers, 26, 43–45
and robustness, 116–119
detection of, 117–119
effect on ASL, 43, 113
exaggerated, 117
effect on histogram, 35
effect on plot scale, 29
films with unusual MSL/ASL ratios, 28
in bivariate plots, 30
in boxplots, 42
in lognormal data, 44
in PCA, 158
polynomial regression, see smoothing
principal component analysis (PCA), 140, 146
applications, 154–157
used to check cluster analysis, 154–160
probability distributions
chi-square, 137
Erlang, 128
Gamma, 127
lognormal, see lognormality, see lognormal distribution, 111, 114–116, 127, 172–179
truncated, 129
normal, 9, 37, 51–52
Pareto, 128
Poisson, 127
Weibull, 128
probability plots
(see) graphical methods, 52
R data entry
multiple films, 108–109
single films, 21
R functions, 22, 25–26
Qn, 31
Sn, 31
abline, 52, 99
adjbox, 44, 123
apply, 109
as.data.frame, 109
attr, 109
axis, 50, 100
boxplot, 42
cbind, 27, 109
chull, 102
cumsum, 61, 63
density, 38, 87, 142
dnorm, 54
ecdf, 50
edit, 25
fivenum, 34, 41
for, 81
graphics.off, 81
gray, 83
hclust, 151
heat.colors, 85
hist, 35
ifelse, 110
image, 73
jarque.bera.test, 112
kurtosis, 111
legend, 47, 50
library, 22
lines, 47, 100, 110
list, 26
lm, 61, 110
locpoly, 142
loess, 142, 151
  degree argument, 66
  family argument, 66
  span argument, 66
mean, 26, 109
median, 26, 109
na.omit, 109, 148
names, 109
nls, 11
pairs, 27, 111
par, 35
  new (overlay plots), 56
plot, 47, 80
  col (color) argument, 47
  lty (line type) argument, 47
  lwd (line width) argument, 47
  pch (plotting symbol) argument, 56
  xaxt, yaxt arguments, 50, 54
  xlim, ylim arguments, 47
points, 100
polygon, 139
poly, 61
powerTransform, 139
predict, 107
qqline, 52
quantile, 25
rank, 71
read.table, 21, 108
rep, 99
rollmean, 62
rollmedian, 69
round, 26
rug, 74
scale, 55
scatterplotMatrix, 28
sf.test, 112
shapiro.test, 112
skewness, 111
sort, 26
sqrt, 71
summary, 25
text, 100
vioplot, 43
win.graph, 81
yjpowen, 139
tt heat.colors, 89

R packages
  car, 28
  robustbase, 31

R packages, 22
  FactoMineR, 96
  MASS, 96
  ade4, 96
  anacor, 96
  car, 139
  ca, 95
  library, 22
  moments, 111, 131, 139
  nortest, 112
  robustbase, 44
  tseries, 112
  vegan, 96
  vioplot, 43
  zoo, 62, 69
  MASS, 22, 99

regression analysis
  linear regression
    of ASL, MSL, 121
  polynomial regression, see smoothing
  robust regression, see robust statistics

robust statistics
  and non-parametric statistics, 68
  and outliers, 116–117
  ASL, lack of robustness, 43
  breakdown point, 34, 117
  graphical methods, 67–73
  IQR, 34
  Mann-Whitney test, see hypothesis tests
  measures of dispersion, 30–31
  median, 34
  of skewness, 45
  order structure matrices, 86
  ranked data, 69–71
    and moving average plots, 77
    in shot-scale analysis, 104–105
  Mann-Whitney test, 70–71, 77
  moving average of, 69
  of cut-points, 75
  robust regression, 120, 121, 123
  Spearman’s rank-correlation, 119, 125

Salt, Barry
  database of, 6, 13, 16, 95
  shot length (SL)
    partition SLs, 141
  shot lengths, see SL
  shot-scale analysis, 5, 13–20, 94–107
    definition of types, 13
    skewness, 131–134
    and ASL, 113
    definition, 131
    hypothesis tests of, 133–134
    in boxplots, 41
    of SL distributions, 8
    robust measures, 45
    testing significance, 133

SL
  distributions, 8–10
  log-transformation, 9
  log-transformation of, 39
slasher films
  compared to Pinocchio, 90
slasher films, 85–89
smoothing, 142–145
  applications
    ASLs over time, 6
    of autoregression indices, 12
    degree of, 144–145
loess, 63, 64–67
  as non-parametric regression, 67
  varying degree of, 157–160
moving average, see time-series
of SLs, 10
polynomial regression, 59–61, 76–77
  as parametric model, 67
trendline, 58
problems at boundaries, 142
software
  Excel, 1, 21, 80, 108
  S-Plus, 23
time-series, 58–78
  analysis of SLs, 11–13
  and cluster analysis, 140
autoregression, 11
kernel density estimate (KDE), 73–74
loess smooth, see smoothing
moving average, 58, 62–64
  double-smoothing, 58, 63–64
moving median, 68–69
  evaluation of, 75
order structure matrices, 71–73, 77, 85

alternative to, 82–84
robustness, 86
partial autocorrelation function (PACF), 11