USA Films: The Non-Transitivity and Simpson's Paradoxes

Introduction to Statistics and Computation with Data

Anshul Raj Singh, Anshuman Sharma, Gauransh Kapur, Ramdas Singh
Indian Statistical Institute, Bangalore

Contents

1 Introduction

In this report, we explore two prominent statistical paradoxes that reveal the subtleties and dangers of interpreting correlation and categorical data without sufficient scrutiny. Using data sampled from Leonard Maltin's 1996 Movie and Video Guide, we analyze how certain relationships that appear intuitive or obvious at first glance can, upon closer examination, behave in ways that defy our expectations.

Our analysis focuses on a random sample of 100 movies drawn from a population of approximately 19,000. Within this dataset, we investigate instances of the non-transitivity paradox in correlation, and demonstrate an example of Simpson's paradox. These paradoxes serve not only as theoretical curiosities but also as practical warnings for analysts working with observational data.

2 Dataset Overview

The dataset used consists of five primary variables:

- **Year**: The year in which the film was released.

- **Length**: The runtime of the film in minutes.
- **Cast**: The number of major cast members in the film.
- **Rating**: A critic rating score on a scale from 1 to 4.
- **Description**: The number of descriptive lines used for the film.

These variables were recorded for each of the 100 films in our sample. The sampling was performed through simple random sampling techniques, ensuring that movies across a range of years and genres were fairly represented. Extra care was taken to avoid over-representation of any particular genre.

3 The Correlational Paradox

Correlation is a measure that quantifies the degree to which two quantitative variables are linearly related. It is tempting to assume that if variable X is positively correlated with Y, and Y with Z, then X must be positively correlated with Z. This, however, is not necessarily true. The non-transitivity of correlation can produce surprising results.

Consider a case involving three variables: Length (runtime of the movie), Year (of release), and Rating (critic score). Length and Rating are positively correlated, meaning that longer movies tend to receive higher ratings. Year and Length are also positively correlated, indicating that newer movies tend to be longer. One might therefore expect that newer movies receive higher ratings. Paradoxically, the correlation between Year and Rating is negative.

To illustrate this, we examine a scatterplot matrix of the variables involved:

We now examine the following empirical results:

Variables	**Correlation (r)**	**p-value**
Length vs Rating	0.318	0.001
Year vs Length	0.509	0.000
Year vs Rating	-0.148	0.143

The p-values suggest that the correlations between Length and Rating, and between Year and Length, are statistically significant. The correlation between Year and Rating, however, is not. Still, the negative sign of the latter correlation contradicts the positive expectation derived from the other two.

This kind of paradox was studied rigorously by Langford et al., who showed that if the sum of the squares of the two known correlations exceeds 1, then the third correlation must also be positive. That is, if $\rho_{XY}^2 + \rho_{YZ}^2 > 1$, then it follows that $\rho_{XZ} > 0$. In our case, $0.509^2 + 0.318^2 = 0.360 < 1$, so the paradox is theoretically permitted.

To understand this formally, consider the correlation matrix:

$$R = \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix}.$$

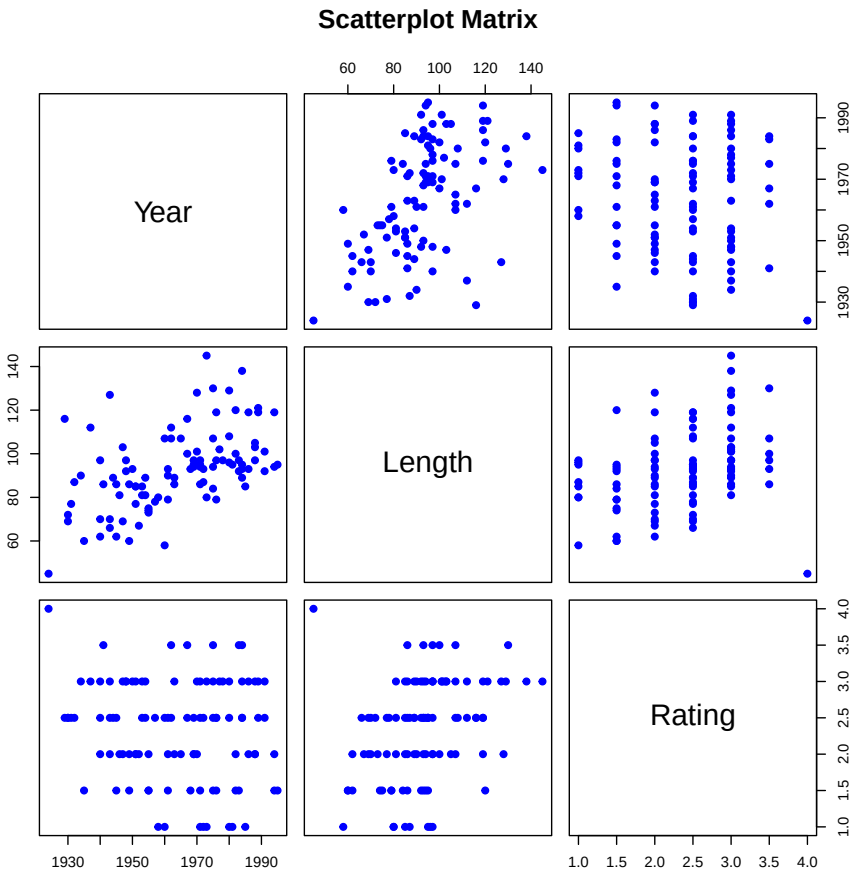

Figure 1: Scatterplot matrix of Rating, Length, and Year.

Since any covariance matrix (and therefore its normalized correlation matrix) must be positive semi-definite, its determinant must be non-negative. This condition leads to an inequality that constrains ρ_{XZ} in terms of ρ_{XY} and ρ_{YZ}.

Further analysis includes grouping the dataset by movie length. To aid in visual understanding, we include the following coded scatterplot of Rating against Year, colored by movie Length:

Movies were defined as *short* if their length was less than 90 minutes and *long* otherwise. When grouped in this manner, we find the following correlations between Rating and Year:

Group	**Correlation (r)**	**p-value**
Short films	-0.520	0.000
Long films	-0.280	0.033

The negative correlation is more pronounced within each category, reinforcing the paradox. A multiple regression model was also considered to understand the joint effect of Year and Length on Rating:

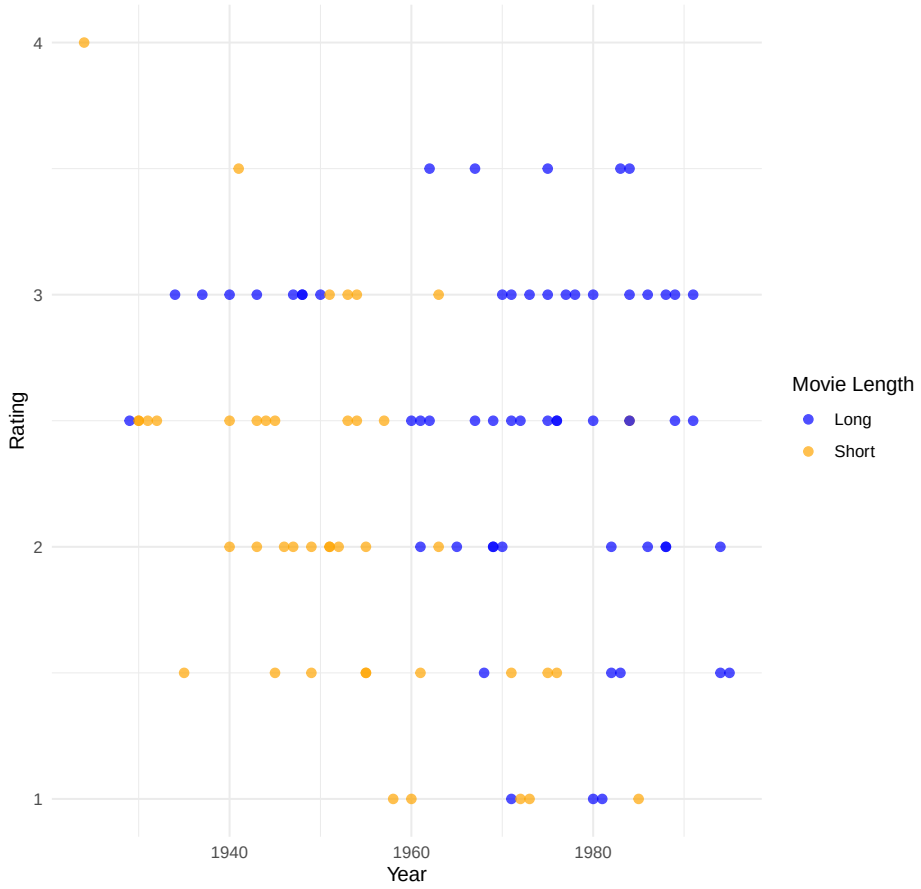

Figure 2: Scatterplot of Rating vs. Year, coded by movie Length.

$$\text{Rating} = 24.59 - 0.0119 \cdot \text{Year} + 0.0124 \cdot \text{Length},$$

which showed statistically significant coefficients. In contrast, the simple linear regression of Rating on Year:

$$\text{Rating} = 13.5 - 0.00570 \cdot \text{Year},$$

had a less significant slope. This shows how controlling for Length alters the apparent relationship between Year and Rating.

4 Simpson's Paradox

Simpson's paradox occurs when a trend observed in several groups reverses when the groups are combined. Classically, it arises with categorical data. To see this, consider a hiring scenario:

Group	Whites	Non-whites
Job 1	6/60 is 10%	5/45 is 11.11%
Job 2	30/40 is 75%	4/5 is 80%
Total	36/100 is 36%	9/50 is 18%

150 people applied for two jobs at a company. Of those 150, 100 were white and 50 were from other races. 60 of the whites applied for Job 1 of which only 6 were selected, and 30 were selected from the other 40 who applied for Job 2. Also, 45 of the non-whites applied for Job 1 of which only 5 were selected and out of the remaining 5, who applied for Job 2, 4 were selected. Although non-whites have a higher acceptance rate within each job, the aggregate acceptance rate is lower. This inversion exemplifies Simpson's paradox.

A more famous example comes from UC Berkeley's graduate admissions in the 1970s. Women had a lower overall admission rate than men. However, when admissions were analyzed by department, no bias was found. In fact, women tended to apply to more competitive departments.

Major	Men	% Admitted	Women	% Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

This discrepancy occurred because women were more likely to apply to departments with lower overall acceptance rates.

In the end, Simpson's paradox boils down to the behaviour of fractions and the type of numbers in the dataset; by the 'behaviour of fractions,' we mean that if

$$\frac{a}{b} < \frac{c}{d} \text{ and } \frac{e}{f} < \frac{g}{h}$$

then the following may occur:

$$\frac{a+e}{b+f} > \frac{c+g}{d+h}.$$

A more general statement statement works as $\frac{a_i}{b_i} > \frac{c_i}{d_i}$ for $1 \leq i \leq n$, but $\frac{\sum a_i}{\sum b_i} < \frac{\sum c_i}{\sum d_i}$. This is what occurs in the UC Berkeley's graduate admissions study.

5 Analysis of a Second Dataset

To validate the earlier observations on a different sample, we drew 100 movies from IMDb's top-rated films. We recorded the Year, Length, and Rating (out of 10) for each movie. A scatterplot matrix summarizing the relationships between these variables is shown below:

The correlations were:

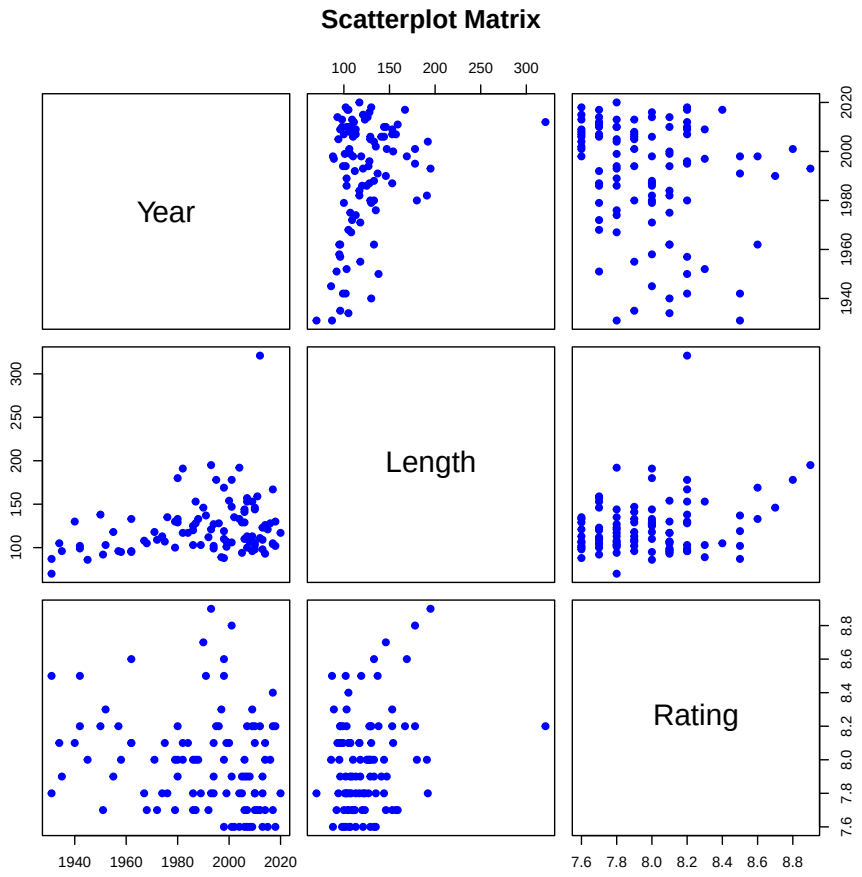

Figure 3: Scatterplot matrix for the IMDb top-rated movie dataset.

Variables	r	p-value
Length vs Rating	0.251	0.012
Year vs Length	0.275	0.006
Year vs Rating	-0.220	0.028

These again hint at non-transitive correlation. For the Simpson's paradox, we categorized:

- **Length**: short if under 105 minutes, long otherwise.
- **Year**: old if before 1998, new otherwise.
- **Rating**: good if at least 8, bad otherwise.

Group	Bad	Good	% Good
New, Short	12	20	62.5%
New, Long	24	15	38.5%
Old, Short	7	11	61.1%
Old, Long	7	4	36.4%
Total	50	50	50.0%

Although short films in both eras tend to have higher proportions of good ratings, aggregating across lengths reverses the trend.

6 Conclusion

The non-transitivity and Simpson's paradoxes serve as powerful illustrations of the pitfalls of naively interpreting statistical relationships. Even when individual relationships appear well-behaved, they may yield contradictory or unintuitive results when combined.

Length, in our study, acted as a confounding variable that masked the true negative relationship between Year and Rating. Similarly, categorization by length and year allowed us to construct an example of Simpson's paradox.

References

- Langford, E., Schwertman, N., Owens, M. (2001). *Is the Property of Being Positively Correlated Transitive?* The American Statistician.
- Leonard Maltin (1996). *Movie and Video Guide*. Penguin Books.
- Freedman, D., Pisani, R., Purves, R. (1998). *Statistics* (3rd ed.). W.W. Norton.
- Moore, T. L. (2006). *Journal of Statistics Education*, 14(1).