Paradoxes in Film Ratings

Thomas L. Moore
Grinnell College

Journal of Statistics Education Volume 14, Number 1 (2006),
ww2.amstat.org/publications/jse/v14n1/datasets.moore.html

Copyright © 2006 by Thomas L. Moore, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Controlling for a variable; Non-transitivity of positive correlation; Simpson’s paradox

I selected a simple random sample of 100 movies from the Movie and Video Guide (1996), by Leonard Maltin. My intent was to obtain some basic information on the population of roughly 19,000 movies through a small sample. In exploring the data, I discovered that it exhibited two paradoxes about a three-variable relationship: (1) A non-transitivity paradox for positive correlation, and (2) Simpson’s paradox. Giving concrete examples of these two paradoxes in an introductory course gives to students a sense of the nuances involved in describing associations in observational studies.

1. Introduction

For many years, I had asked my students to do projects of their own invention. One such assignment asked them to describe some dataset in a way that answered an interesting question or small set of questions. About 10 years ago, I decided to raise expectations on report writing, so I needed a project of my own to report on so that I could give my students a good model report to see prior to writing their own reports. Subsequently I have been able to use former student reports as models, but the dataset I chose for my project has turned into a dataset that I continue to use in my teaching because of the interesting and unexpected patterns I found in the data and because the context of the data is both interesting and understandable to my students.

The Movie and Video Guide by Leonard Maltin is an annual ratings guide to movies. While not all films ever made are in Maltin’s Guide, it does contain a very large number of movies covering the history of cinema. In this article, I discuss a dataset collected from the 1996 edition, which contained ratings on about 19,000 films.

I used Minitab to generate a simple random sample of 100 titles from the book. I recorded 5 variables on each movie sampled: The year the movie was released (Year), the running time of the movie in minutes (Length), the number of cast members listed (Cast), the rating that Maltin gave the movie on a rising scale of 1, 1.5, 2, ..., 4 (Rating), and the number of lines of description for the movie in the Guide (Description).
2. A correlational paradox in the data

In this and the next section, we investigate a three-way relationship between Length, Year, and Rating. Section 2 discusses the non-transitivity paradox for positive correlation: that if X, Y, and Z are quantitative it is possible for X and Y to be positively correlated and Y and Z to be positively correlated, but for X and Z to not be positively correlated. Section 3 discusses Simpson’s paradox. My primary use of the dataset has been to ask my students to explore these paradoxes through a guided classroom activity (e.g., a lab) or as a classroom example with discussion.

2.1 Correlational analysis of quantitative variables

The variables Length and Year are quantitative and we’ll also treat Rating as quantitative. One can look at the 3 bivariate relationships using scatterplots and computing correlations. Figure 1 gives the scatterplot matrix for the 3 variables. The correlations between the 3 variables are given in Table 1. There is a paradox at work in these relationships which we proceed now to investigate.

Figure 1. Scatterplot matrix of Rating, Length, and Year. Notice that longer movies tend to have higher ratings and more recent movies tend to be longer movies, but that Year and Rating appear to be uncorrelated or, perhaps, negatively correlated.

Table 1: Pairwise correlations between 3 quantitative variables, with P-values.
2.2 Investigating the paradox

Langford, Schwertman and Owens (2001) discuss what I am calling the non-transitivity paradox for positive correlation. The films data illustrate the paradox: Let X=Year, Y=Length, and Z=Rating. More recent movies tend to be longer movies, so that X and Y are positively correlated. Longer movies tend to get higher ratings, so that Y and Z are positively correlated. If more recent movies tend to be longer and longer movies tend to be rated higher, one might then assume that more recent movies would tend to get higher ratings. That this reasoning can fail, and fails in this instance, is the paradox. Indeed, the correlation between Year and Rating is negative. While Langford et al. (2001) did not discuss inference to a population, the P-values in Table 1 show us that the reasoning fails at the population level as well: We have statistically significant positive correlations between X and Y and between Y and Z, but not between X and Z. Langford, et al. (2001) prove that the non-transitivity paradox for positive correlation cannot occur when \((\rho_{XY})^2 + (\rho_{YZ})^2 > 1\), i.e., cannot occur when the two positive correlations are sufficiently close to 1. In this case, \((0.509)^2 + (0.318)^2 = 0.360 < 1\); the inequality does not hold.

Let’s explore the data further to see what is going on. Figure 2 shows a coded scatterplot of Rating against Year. We have defined a movie as short if its length is less than 90 minutes and as long if its length is 90 minutes or more. From the plot, we see that the longer movies tend to be more recent movies than the short movies, but within each length category there is a fairly clear negative relationship between Year and Rating: more recent movies tend to be rated lower and now the negative correlations are statistically significant. (See Table 2.) Length “masks” the negative relationship between Year and Rating—as Length increases Year tends to increase and the tendency of longer movies to get higher ratings negates the tendency of more recent movies to get lower ratings.

| Pair of Variables | R     | P-value |
|-------------------|-------|---------|
| Length vs. Rating | 0.318 | 0.001   |
| Year vs. Length   | 0.509 | 0.000   |
| Year vs. Rating   | -0.148| 0.143   |
Figure 2. Rating vs. Year, coded by Length. Movies less than 90 minutes are coded as short, while movies 90 minutes or longer are coded as long.

Table 2: Rating versus Year correlations, controlling for Length. The negative correlation between Rating and Year is more evident: within each Length category more recent movies get lower ratings.

| Pair of Variables                  | R       | P-value |
|------------------------------------|---------|---------|
| Rating vs. Year (short movies)     | -0.520  | 0.000   |
| Rating vs. Year (long movies)      | -0.280  | 0.033   |

In an elementary course, even at the descriptive statistics level, I like this example because it illustrates the perils of aggregating data. I have also used this example when introducing multiple regression in a more advanced course. The two-predictor model estimates the relationship between Rating (our response variable) and Year, controlling for Length:

\[
\text{Rating} = 24.6 - 0.0119 \text{ Year} + 0.0124 \text{ Length}
\]

| Predictor      | Coef   | SE Coef  | T       | P       |
|----------------|--------|----------|---------|---------|
| Constant       | 24.59  | 10.04    | 2.45    | 0.018   |
| Year           | -0.011856 | 0.005095 | -2.33   | 0.024   |
| Length         | 0.012407 | 0.006154 | 2.02    | 0.049   |
S = 0.6151 R-Sq = 14.2% R-Sq(adj) = 11.0%

Compare this to the simple linear regression Rating = 13.5 - 0.00570 Year, where the slope estimate of -0.00570 has the confirmatory non-significant P-value of 0.143.

The students can see how our regression output corroborates what we have learned through the coded scatterplots and correlations computed previously: there is a statistically significant, negative relationship between Rating and Year, controlling for Length.

3. Simpson’s paradox in the data

Simpson’s paradox refers to a reversal in the direction of an association between two variables X and Y. If the X versus Y association is one direction when variable Z is ignored, but reverses direction at each level of Z (i.e., the relationship reverses direction when controlling for Z), we say that Simpson’s paradox has occurred. Classically, the variables X, Y, and Z in this definition are categorical.

My favorite examples of Simpson’s paradox are summarized in Table 3. For example, in the Berkeley admissions data from Freedman, Pisani and Purves (1998), men applicants appear to have a higher rate of admission to graduate school than women, but when we control for the graduate program, men’s advantage disappears. Or in the Florida death sentence data from Witmer (1992), whites convicted of murder appear more likely to be given the death sentence, but when we control for the race of the victim, blacks are more likely to get the death sentence regardless of whether the victim is white or black. The reader can consult the references for the data and story for each example. The data for each example with an abbreviated description can be found at www.math.grinnell.edu/~mooret/reports/SimpsonExamples.pdf.

Table 3: Here is a summary of favorite examples of Simpson’s paradox. In each case, the direction of an X-versus-Y relationship is reversed when controlling for the Z variable. See the references for the complete data and the stories behind the data.

| Subject                | X                   | Y                        | Z                                | Reference                                      |
|------------------------|----------------------|--------------------------|----------------------------------|------------------------------------------------|
| Berkley Admissions Data| sex of applicant     | accept or reject         | grad program applied to          | Freedman, et al. 1998, pp 17-20.               |
| Airlines on-time data  | airline              | on-time or late          | airport location                 | Moore 2003, p 143.                             |
| Death sentence data    | race of convicted murderer | death sentence: yes or no | race of murder victim            | Witmer 1992, pp 110-112.                       |
| Comparing batting averages | person batting         | hit or out               | year of that at bat              | Friedlander 1992, p 845.                       |
| Prenatal care          | care status           | infant mortality         | clinic                           | Bishop, Fienberg and Holland 1975, pp 41-42. |
We can create a Simpson’s paradox from the films data as follows. As above, use 90 minutes to define two categories of movie length: short movies run less than 90 minutes and long movies run 90 minutes or longer. Then define two categories of movie based on Year: 1965 or prior are called ‘old’ and 1966 or later are called ‘new.’ Finally, define ‘bad’ movies as those with ratings at or below 2.5 and ‘good’ movies as those with ratings 3 or above. Based upon these definitions, we obtain a Simpson’s paradox, as Table 4 illustrates.

Table 4: The percentage of good movies is higher for new movies (32%) than for old (30%). But this comparison reverses itself when controlling for movie length (i.e., when disaggregating into Short or Long movies.)

|            | Short Movies | bad | good | good% | Long Movies | bad | good | good% | All Movies | bad | good | good% |
|------------|--------------|-----|------|-------|-------------|-----|------|-------|-----------|-----|------|-------|
| new        |              | 7   | 0    | 0.0%  | new         | 27  | 16   | 37.2% | new       | 34  | 16   | 32.0% |
| old        |              | 29  | 6    | 17.1% | old         | 6   | 9    | 60.0% | old       | 35  | 15   | 30.0% |
|            |              |     |      |       |             |     |      |       |           | 69  | 31   | 31.0% |

Not any choice of break points defining your categories will lead to an instance of Simpson’s paradox. Simpson’s paradox requires, by definition, an actual reversal in the relationship when controlling for the third variable, but I like to tell my students that the important point in studying Simpson’s paradox is not just that reversals can happen, but that with observational data relationships that look one way when aggregated can look quite different when disaggregated by a third variable. Calling this more general effect a “Simpson-like paradox,” I tell students that “Simpson’s paradox happens” and “Simpson-like paradoxes happen a lot.” Among famous paradoxes they have studied, Simpson’s may be one they encounter with some frequency in their later lives.

4. Other student investigations

Here are some other activities one can devise around the films dataset, listed by topic.

Sampling. How does one take a simple random sample of movies? This question provides lessons in confronting practical sampling issues in a simple, yet real setting. I sampled by having Minitab choose random (page number, item number) pairs. For example, the pair (1083, 3) would lead to the third movie listed on page 1083 of the Guide. To make the sample proper, one needs an upper bound on the number of items on a given page, which is admittedly a bit ad hoc. When the page selected contains fewer items than the item number selected, you ignore that random pair; so for a SRS of 100 you may need to select a few more than 100 random pairs. It takes some thought to convince oneself that all samples of 100 films have an equal probability of being selected under this scheme. The reason for selecting pairs is for convenience, as it would be prohibitive to number all 19,000 movies consecutively.

Identifying outliers. We can see one clear outlier in Figure 1: the movie with a **** rating that runs less
than 50 minutes. Identifying this outlier serves, at least symbolically, to make the point that outliers are often the most interesting cases in a dataset. The movie in question is “Sherlock, Jr.,” the 1924 Buster Keaton classic, which Maltin describes as a “sublime study of film and fantasy, which has undoubtedly influenced countless filmmakers.” But does Keaton’s classic influence our correlations? Minus the outlier, the correlation between Rating and Length rises from .318 to .408, but the outlier has no qualitative effect on the paradoxes described above.

EDA for a single variable. Of interest to me, and probably to any user of the Guide, would be the distribution of Rating. For example, I tended to assume that a rating of *** or better was a good movie and that ***1/2 or **** movies were rare. But one doesn’t know this until one looks. Figure 3 shows the distribution of Ratings. Only 31 of the 100 movies had ratings of *** or higher and only 7 had ratings of ***1/2 or ****.

| Rating | Count |
|--------|-------|
| 1.0    | 8     |
| 1.5    | 14    |
| 2.0    | 20    |
| 2.5    | 27    |
| 3.0    | 24    |
| 3.5    | 6     |
| 4.0    | 1     |
| N     | 100   |

Figure 3. The distribution of ratings for a SRS of 100 movies. We include both a dotplot and a frequency table. The average rating is about 2.5 (**1/2).

Confidence intervals. Given that we have a SRS from a population, one can ask students to compute confidence intervals for parameters of interest. For example, one could compute a confidence interval for the mean rating: the mean is 2.33 with a 95% confidence interval of 2.19 to 2.47. This assumes that we can treat Rating as a quantitative variable, an issue you can discuss in class as well. We might choose a confidence interval more relevant to the discussion above. For example, 31% of movies in the sample have ratings of 3 or above, with a 95% confidence interval of 22% to 41%. (This is the classical Wald interval; the “plus four” interval gives 31.7% with a confidence interval of 22.8% to 40.6%.)

Other relationships. One can also look at other bivariate relationships. For example both Cast and Description show statistically significant, positive correlations with Rating. There are plausible
explanations for these, which would make good class discussion or exercises.

5. Conclusion

The films data provides a good example of why one must be careful when aggregating observational data. Controlling for the Length of a movie, there is a clear negative relationship between a movie’s Year and its Rating. But if one ignores Length, the relationship between Rating and Year is weaker at the sample level and not statistically different from 0 at the population level. In this sense, Length masks the relationship between Rating and Year. Because films are part of the common experience of most students, this dataset provides a good addition to the teaching examples in the area of aggregation paradoxes.

6. Getting the Data

The file films.dat.txt contains the raw data. The file films.txt is a documentation file containing a brief description of the dataset.

Acknowledgements

I thank Roger Johnson and two anonymous reviewers for their great suggestions for improving my article.

References

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975), Discrete Multivariate Analysis: Theory and Practice, Cambridge, Massachusetts: The MIT Press.

Freedman, D., Pisani, R., and Purves, R. (1998), Statistics (3rd ed.), New York, NY: W.W. Norton and Company.

Friedlander, R. J. (1992), “Ol’ Abner Has Done it Again,” American Mathematical Monthly, 99(9), 845.

Langford, E., Schwertman, N., and Owens, M. (2001), “Is the Property of Being Positively Correlated Transitive?” The American Statistician, 55, 322-325.

Maltin, L. (1996), Leonard Maltin’s 1996 Movie and Video Guide, New York, NY: Penguin Books.

Moore, D. S. (2003), The Basic Practice of Statistics (3rd ed.), New York, NY: W.H. Freeman.

Witmer, J. A. (1992), Data Analysis: An Introduction, Prentice-Hall, Engelwood Cliffs, NJ.

Thomas L. Moore
