关于若干著名统计佯谬的资料

2006/10/05 晚
李晓煦
香港中文大学教育学院
lixiaoxu@gmail.com

Wiki 上的 Simpson Paradox (公共版权网页，附 pdf 页)
https://secure.wikimedia.org/wikipedia/en/wiki/Simpson%27s_paradox

Simpson Paradox 的几何解释
http://www.cut-the-knot.org/Curriculum/Algebra/SimpsonParadox.shtml

Wiki 上的 Regression Towards the Mean(公共版权网页，附 pdf 页)
https://secure.wikimedia.org/wikipedia/en/wiki/Regression_toward_the_mean

Wiki 上关于这个词条的 discussion(公共版权网页)－－对于有兴趣深究的同学，这个页面非常值得一读，可以看到有信心在 wiki 上写词条的志愿者如何在这个著名的概念上犯错误。
https://secure.wikimedia.org/wikipedia/en/wiki/Talk:Regression_toward_the_mean

Wiki 上列出的 F. Galton 1886 年的原始文献下载地址
http://www.spss.com/research/wilkinson/Publications/galton.pdf

Wiki 上目前还没有 Lord’s Paradox 的词条，希望有英文好的同学在这个课后去编写。

一篇用同一组数据演示 Simpson Paradox、Regression Towards the Mean 和 Lord’s Paradox 的论文
http://www.statlit.org/PDF/2004Wainer_ThreeParadoxes.pdf

适合生成模拟数据研究多层分析的统计软件：

R
http://R-project.org/
http://cran.r-project.org/src/contrib/Descriptions/multilevel.html

SAS
http://support.sas.com/publishing/bbu/companion_site/57323.html “Example Code and Data”
http://www.gse.harvard.edu/~faculty/singer/ “DOWNLOADABLE PAPERS”

Simpson's paradox

From Wikipedia, the free encyclopedia

Simpson's paradox (or the **Yule-Simpson effect**) is a statistical paradox described by E. H. Simpson in 1951[1] and G. U. Yule in 1903, in which the successes of several groups seem to be reversed when the groups are combined. This seemingly impossible result is encountered surprisingly often in social science and medical statistics, and occurs when a weighting variable which is not relevant to the individual group assessment must be used in the combined assessment.

Contents

Explanation by example

To illustrate the paradox, suppose two people, Lisa and Bart, are let loose on Wikipedia. In the first week, Lisa improves 60 percent of the articles she edits while Bart improves 90 percent of the articles he edits. In the second week, Lisa improves just 10 percent of the articles she edits, while Bart improves 30 percent.

Both times, Bart improved a much higher percentage of articles than Lisa—yet when the two tests are combined, Lisa has improved a much higher percentage than Bart!

	Week 1	Week 2	Total
Lisa	60.0%	10.0%	55.5%
Bart	90.0%	30.0%	35.5%

This strange-looking result comes about because of the varying number of articles worked on by each person - information not presented in the initial presentation. In the first week, Lisa edits 100 articles, improving 60 of them, while Bart edits just 10 articles, improving all but one. In the second week, Lisa edits only 10 articles, improving one, while Bart edits 100 articles, improving 30. When two week's worth of work is combined, both edited the same number of articles, yet Lisa improved 55% of them (61 in total) while Bart improved only 35% of them (39 in total).

	Week 1	Week 2	Total
Lisa	60 / 100	1 / 10	61 / 110
Bart	9 / 10	30 / 100	39 / 110

To recap, introducing some notation that will be useful later:

- In the first week
  - $S_A(1) = 60\%$ — Lisa improved 60% of the many articles she edited.
  - $S_B(1) = 90\%$ — Bart had a 90% success rate during that time.

Success is associated with Bart.

- In the second week
  - $S_A(2) = 10\%$ — Lisa managed 10% in her busy life.
  - $S_B(2) = 30\%$ — Bart achieved a 30% success rate.

  Success is associated with Bart.

On both occasions Bart's edits were more successful than Lisa's. But if we combine the two sets, we see that Lisa and Bart both edited 110 articles, and:

- $S_A = \frac{61}{110}$ — Lisa improved 61 articles.
- $S_B = \frac{39}{110}$ — Bart improved only 39.
- $S_A > S_B$ — Success is now associated with Lisa.

Bart is better for each set but worse overall!

The arithmetical basis of the paradox is uncontroversial. If $S_B(1) > S_A(1)$ and $S_B(2) > S_A(2)$ we feel that S_B *must be greater* than S_A. However if *different* weights are used to form the overall score for each person then this feeling may be disappointed. Here the first test is weighted $\frac{100}{110}$ for Lisa and $\frac{10}{110}$ for Bart while the weights are reversed on the second test.

- $S_A = \frac{100}{110} S_A(1) + \frac{10}{110} S_A(2)$
- $S_B = \frac{10}{110} S_B(1) + \frac{100}{110} S_B(2)$

By more extreme reweighting A's overall score can be pushed up towards 60% and B's down towards 30%.

Who is more accomplished? Lisa and Bart's mutual friends think Lisa is better—her overall success rate is higher. But it is possible to retell the story so that it appears obvious that Bart is more diligent.

Real-world examples

The batting average paradox

The most common example of the paradox in America involves batting averages in baseball. It is possible — and in rare occasions it has actually happened — for one player to hit for a higher batting average than another player during the first half of the year, and to do so again during the second half, but to have a lower batting average for the entire year, as shown in this example:

```
            First Half     Second Half      Total season
Player A     4/10 (.400)   25/100 (.250)    29/110 (.264)
Player B   35/100 (.350)     2/10  (.200)    37/110 (.336)
```

Sports sabermetrician Bill James has called attention to this phenomenon.

A kidney stone treatment example

This is a real-life example from a medical study comparing the success rates of two treatments for kidney stones. [1] (http://bmj.bmjjournals.com/cgi/content/full/309/6967/1480)

The first table shows the overall success rates and numbers of treatments for both treatments.

success rates (successes/total)

Treatment A	Treatment B
78% (273/350)	83% (289/350)

This seems to show treatment B is more effective. If we include data about kidney stone size, however, the same set of treatments reveals a different answer.

Results accounting for stone size

small stones		large stones	
Treatment A	**Treatment B**	**Treatment A**	**Treatment B**
93% (81/87)	87% (234/270)	73% (192/263)	69% (55/80)

The information about stone size has reversed our conclusion about the effectiveness of each treatment. Now treatment A is seen to be more effective in both cases. In this example the lurking variable (or confounding variable) of stone size was not previously known to be important until its effects were included.

Which treatment is considered better is determined by an inequality between two ratios (successes/total). The reversal of the inequality between these ratios, which creates Simpson's paradox, happens because two effects occur together:

1. the lurking variable has a large effect on the ratios
2. the sizes of the groups which are combined when the lurking variable is ignored are very different

The Berkeley sex bias case

One of the best known real life examples of Simpson's paradox occurred when U. C. Berkeley was sued for bias against women applying to grad school. The admission figures showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.[2] However when examining the individual departments, it was found that no department was significantly biased against women; in fact, most departments had a small (and not very significant) bias against men.

The explanation turned out to be that women tended to apply to departments with low rates of admission, while men tended to apply to departments with high rates of admission.

2006 US school study

In July 2006, the United States Department of Education released a study[2] (http://www.nytimes.com/packages/pdf/national/20060715report.pdf) documenting student performances in reading and math in different school settings[3] (http://www.nytimes.com/2006/07/15/education/15report.html) . It reported that while the math and reading levels for students at grades 4 and 8 were uniformly higher in private/parochial schools than in public schools, repeating the comparisons on demographic subgroups showed much smaller differences which were nearly equally divided in direction.

"Lurking variable"

Simpson's paradox shows us an extreme example of the importance of including data about possible confounding variables when attempting to calculate correlations.

The "lurking variable" principle also works with the Electoral College, which determines the winner of U.S. presidential elections. For example, if Candidate A wins 35 of the states and Candidate B wins 15 of the states, the color-coded map will appear to be a landslide for Candidate A; but if Candidate A's states are less populated and Candidate B's states are more populated, it is still possible for Candidate B to win. The lurking variable is the differering number of electoral votes each state carries.

See also

- Low birth weight paradox, an example of Simpson's paradox in action.

References

1. **^** Simpson,E. H. (1951). "The Interpretation of Interaction in Contingency Tables". *Journal of the Royal Statistical Society, Ser. B* **13**: 238-241.
2. **^** Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science* **187**: 398-404.

External links

For a brief history of the origins of the paradox see the entries on Simpson's Paradox and Spurious Correlation in

- Earliest known uses of some of the words of mathematics: S (http://members.aol.com/jeff570/s.html)

Other links:

- Simpson's Paradox: An Anatomy by Judea Pearl (http://singapore.cs.ucla.edu/R264.pdf)
- Mediant Fractions (http://www.cut-the-knot.org/blue/Mediant.shtml) at cut-the-knot
- Simpson's Paradox (http://www.cut-the-knot.org/Curriculum/Algebra/SimpsonParadox.shtml) at cut-the-knot
- Stanford Encyclopedia of Philosophy entry (http://plato.stanford.edu/entries/paradox-simpson/)

Categories: Probability theory paradoxes | Statistical paradoxes

Regression toward the mean

From Wikipedia, the free encyclopedia

In statistics, **regression toward the mean** is a principle stating that of related measurements, and selecting those where the first measurement is either higher or lower than the average, the expected value of the second is closer to the mean than the observed value of the first. The degree of regression toward the mean becomes more extreme, other things being equal, as the distance of the first measurement from the average becomes larger.

Contents

Examples

Consider, for example, students who take a midterm and a final exam. Students who got an extremely high score on the midterm will probably get a good score on the final exam as well, but we expect their score to generally be closer to the average than their midterm score was. The reason: it is likely that some luck was involved in getting the exceptional midterm score, and this luck cannot be counted on for the final. Conversely, it is also true that among those who get exceptionally high final exam scores, the average midterm score will not have been as far above average as the final exam score, since some of the students will have obtained high scores on the final due to luck that they didn't have on the midterm.

Similarly, unusually low scores regress toward the mean. Thus, if students who obtain a very low midterm score are selected, their average on the final exam is expected to closer to the average for all students than their average midterm score.

It is a commonplace observation that matings of two championship athletes, or of two geniuses, usually results in a child who is above average but less talented than either of their parents.

History

The first regression line drawn on biological data was a plot of seed weights presented by Francis Galton at a Royal Institution lecture in 1877. Galton had seven sets of sweet pea seeds labelled K to Q and in each packet the seeds were of the same weight. He chose sweet peas on the advice of his cousin Charles Darwin and the botanist Joseph Dalton Hooker as sweet peas tend not to self fertilise and the seed weight varies little with humidity. He distributed these packets to a group of friends throughout Great Britain who planted them. At the end of the growing season the plants were uprooted and returned to Galton. The seeds were distributed because when Galton had tried this experiment himself in the Kew Gardens in 1874, the crop had failed.

He found that the weights of the offspring seeds were normally distributed, like their parents, and that if he plotted the mean diameter of the offspring seeds against the mean diameter of their parents he could draw a straight line through the points - the first regression line. He also found on this plot that the mean size of the offspring seeds tended to the overall mean size. He initially referred to the slope of this line as the "coefficient of reversion". Once he discovered that this effect was not a heritable property but the result of his manipulations of the data, he changed the name to the "coefficient of regression". This result was important because it appeared to conflict with the current thinking on evolution and natural selection. He went to do extensive work in quantitative genetics and in 1888 coined the term "co-relation" and used the now familiar symbol "r" for this value.

In additional work he investigated geniuses in various fields and noted that their children, while typically gifted, were almost invariably closer to the average than their exceptional parents. He later described the same effect more numerically by comparing fathers' heights to their sons' heights. Again, the heights of sons both of unusually tall fathers and of unusually short fathers was typically closer to the mean height than their fathers' heights.

Ubiquity

It is important to realize that regression toward the mean is unrelated to the progression of time: the *fathers* of exceptionally tall people also tend to be closer to the mean than their sons. The overall variability of height among fathers and sons is the same.

The original version of regression toward the mean suggests an identical trait with two correlated measurements with the same reliability. However, this character is not necessary, unless any pair of predicting and predicted variables had to be viewed with an identical potential trait. The necessary implicate presumption is that the standard deviations of the predicting and the predicted are the same to be comparable, or have been transformed or interpreted to be comparable.

One later version of regression toward the mean defines a predicting variable with measurement error which impairs the predicting coefficient. This interpretation is not necessary. For example, in the original case the measurement error of length could be ignored.

Mathematical derivation

Let X and Y be zero mean jointly Gaussian random variables with the *same* variance, and correlation coefficient r. The Cauchy-Schwartz inequality shows that $|r| <= 1$. From Gaussianity, the expected value of Y conditioned on the value of X is linear in X; more precisely, $E[Y|X]=rX$, hence the estimated value for Y is closer to the mean 0 than the observed value X since $|r| <= 1$. Similar results can be obtained for more general classes of distributions. For example, let (X,Y) be jointly normal as above, and define $W=AX$, $Z=AY$, where A is any absolutely integrable scalar random variable independent of X and Y. The variables W and Z have zero mean but are not Gaussian. Nevertheless, it is possible to prove that the linear regression property still holds: $E[Z|W]=rW$, and once again regression toward the mean is observed.

The example illustrates a general feature: regression toward the mean is more pronounced the less the two variables are correlated, i.e. the smaller $|r|$ is.

The phenomenon of regression toward the mean is related to Stein's example.

Regression fallacies

Misunderstandings of the principle (known as "**regression fallacies**") have repeatedly led to mistaken claims in the scientific literature.

An extreme example is Horace Secrist's 1933 book *The Triumph of Mediocrity in Business*, in which the statistics professor collected mountains of data to prove that the profit rates of competitive businesses tend towards the average over time. In fact, there is no such effect; the variability of profit rates is almost constant over time. Secrist had only described the common regression toward the mean. One exasperated reviewer likened the book to "proving the multiplication table by arranging elephants in rows and columns, and then doing the same for numerous other kinds of animals".

A different regression fallacy occurs in the following example. We want to test whether a certain stress-reducing drug increases reading skills of poor readers. Pupils are given a reading test. The lowest 10% scorers are then given the drug, and tested again, with a different test that also measures reading skill. We find that the average reading score of our group has improved significantly. This however does not show anything about the effectiveness of the drug: even without the drug, the principle of regression toward the mean would have predicted the same outcome.

The calculation and interpretation of "improvement scores" on standardized educational tests in Massachusetts probably provides another example of the regression fallacy. In 1999, schools were given improvement goals. For each school, the Department of Education tabulated the difference in the average score achieved by students in 1999 and in 2000. It was quickly noted that most of the worst-performing schools had met their goals, which the Department of Education took as confirmation of the soundness of their policies. However, it was also noted that many of the supposedly best schools in the Commonwealth, such as Brookline High School (with 18 National Merit Scholarship finalists) were declared to have failed. As in many cases involving statistics and public policy, the issue is debated, but "improvement scores" were not announced in subsequent years and the findings appear to be a case of regression to the mean.

In sports

Statistical analysts have long recognized the effect of regression to the mean in sports; they even have a special name for it: the "Sophomore Slump." For example, Carmelo Anthony of the NBA's Denver Nuggets had an outstanding rookie season in 2004. It was so outstanding, in fact, that he couldn't possibly be expected to repeat it: in 2005, Anthony's numbers had slightly dropped from his torrid rookie season. The reasons for the "sophomore slump" abound, as sports are all about adjustment and counter-adjustment, but luck-based excellence as a rookie is as good a reason as any. Of course, not just "sophomores" experience regression to the mean. Any athlete who posts a significant outlier, whether as a rookie (young players are universally not as good as those in their prime seasons), or particularly after their prime years (for most sports, the mid to late twenties), can be expected to perform more in line with their established standards of performance. The trick for sports executives, then, is to determine whether or not a player's play in the previous season was indeed an outlier, or if the player has established a new level of play. However, this is not easy. Melvin Mora of the Baltimore Orioles put up a season in 2003, at age 31, that was so far away from his performance in prior seasons that analysts assumed it had to be an outlier... but in 2004, Mora was even better. Mora, then, had truly established a new level of production, though he will likely regress to his more reasonable 2003 numbers in 2005. Conversely, Kurt Thomas of the New York Knicks significantly ramped up his production in 2001, at an age (29) when players typically start to play more poorly. Sure enough, in the following season Thomas was his old self again, having regressed to the mean of his established level of play. John Hollinger has an alternate name for the law of regression to the mean: the "fluke rule." Whatever you call it, though, regression to the mean is a fact of life, and also of sports. Regression to the mean in sports performance produced the "Sports Illustrated Jinx" superstition, in all probability. Athletes believe that being on the cover of Sports Illustrated jinxes their future performance, where this apparent jinx was an artifact of regression.

References

- J.M. Bland and D.G. Altman. "Statistic Notes: Regression towards the mean", *British Medical Journal* 308:1499, 1994. *(Article, including a diagram of Galton's original data, online at: [1] (http://bmj.bmjjournals.com/cgi/content/full/308/6942/1499))*
- Francis Galton. "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute*, 15:246-263 (1886). *(Facsimile at: [2] (http://www.mugu.com/galton/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf))*
- Stephen M. Stigler. *Statistics on the Table*, Harvard University Press, 1999. *(See Chapter 9.)*

External links

- A non-mathematical explanation of regression toward the mean. (http://davidmlane.com/hyperstat/B153351.html)

- A simulation of regression toward the mean. (http://onlinestatbook.com/stat_sim/reg_to_mean/index.html)
- Amanda Wachsmuth, Leland Wilkinson, Gerard E. Dallal. Galton's Bend: An Undiscovered Nonlinearity in Galton's Family Stature Regression Data and a Likely Explanation Based on Pearson and Lee's Stature Data (http://www.spss.com/research/wilkinson/Publications/galton.pdf) *(A modern look at Galton's analysis.)*

- Massachusetts standardized test scores, interpreted by a statistician as an example of regression: see discussion in sci.stat.edu (http://groups.google.com/groups?q=g:thl3845480903d&dq=&hl=en&lr=&ie=UTF-8&oe=UTF-8&safe=off and its continuation (http://groups.google.com/group/sci.stat.edu/tree/browse_frm/thread/c1086922ef405246/60bb52814483 .

Retrieved from "http://en.wikipedia.org/wiki/Regression_toward_the_mean"

Category: Statistics

- This page was last modified 00:39, 21 September 2006.
-