Jerry Johnson
University of Nevada
July, 2002

# Misleading Aggregates: Simpson's Paradox

**Suggested grade levels**: 12 and up due to subject matter and reading levels.

**Possible subject areas**: Social studies

**Math skills**: Arithmetic and percentages.

**Overview**: The verb *to aggregate* means "to collect or gather into a mass or whole." If you lump data together (aggregate it) you may get one picture, whereas if you break the data down into categories, you may get a very different picture - sometimes a seemingly contradictory one. The phenomenon is called "Simpson's Paradox." The examples below illustrate how this can happen and why one should be cautious before drawing conclusions from numbers.

---

## Student activities: Misleading Aggregates

**Example 1:** The local newspaper examined the town's two hospitals and found that over the last six months at Mercy Hospital 79% of the patients survived while at County Hospital 90% survived. The table below summarizes the findings.

| | Lived | Died | Total | % who lived |
|---|---|---|---|---|
| MERCY HOSPITAL | 790 | 210 | 1000 | 79.0% |
| COUNTY HOSPITAL | 900 | 100 | 1000 | 90.0% |

On closer investigation it was observed that the patients were categorized upon admission as being in fair (or better) condition or in poor (or worse) condition. When the survival rates were examined for these groups, the following tables emerged:

Patients admitted in fair condition or better:

| | Lived | Died | Total | % who lived |
|---|---|---|---|---|
| MERCY HOSPITAL | 580 | 10 | 590 | _____ |
| COUNTY HOSPITAL | 860 | 30 | 890 | _____ |

Patients admitted in poor condition or worse:

| | Lived | Died | Total | % who lived |
|---|---|---|---|---|
| MERCY HOSPITAL | 210 | 200 | 410 | _____ |
| COUNTY HOSPITAL | 40 | 70 | 110 | _____ |

**Exercises**:

1. Fill in the four blanks in the two tables above with the correct percentages.

2. Compare the percentages in the first table with those in the next two tables. Do you observe anything strange?

3. Which hospital would you choose, and why?

**Example 2**: In a recent hiring period, a hypothetical department store, U-Mart, hired 62% of the males who applied and 14% of the females. A lawsuit was contemplated since these numbers seemed to indicate that there was gender discrimination.

On closer examination, it was found that U-Mart's hiring was only for two of their departments: a hardware department and a ladies apparel department. The hardware department hired 60 out of 80 male applicants and 15 out of 20 female applicants. The ladies apparel department hired 2 out of 20 male applicants and 30 out of 300 female applicants.

| | Males Applied | Male Hired | Females Applied | Females Hired |
|---|---|---|---|---|
| Hardware | 80 | 60 | 20 | 15 |
| Ladies apparel | 20 | 2 | 300 | 30 |
| U-Mart total | 100 | 62 | 320 | 45 |

4. What percentage of male applicants for the hardware department was hired by U-Mart?

5. What percentage of female applicants for the hardware department was hired?

6. What percentage of male applicants for the ladies apparel department was hired?

7. What percentage of female applicants for the ladies apparel department was hired?

8. You are an attorney for a female plaintiff. How would you argue that there is gender discrimination?

9. You are an attorney for U-Mart. How would you argue that there is no gender discrimination?

10. How would you vote if you were the judge or you were on a jury?

**For the Teacher:**

Both of these examples illustrate something called "Simpson's Paradox." Aggregate percentages (obtained by lumping everything together) give one picture, but by examining the percentages in two separate categories one can be led to the opposite conclusion - a seeming paradox that leaves one wondering what to believe.

**Example 1:**

1. Patients admitted in fair condition or better:

| | Lived | Died | Total | % who lived |
|---|---|---|---|---|
| MERCY HOSPITAL | 580 | 10 | 590 | 98.3% |
| COUNTY HOSPITAL | 860 | 30 | 890 | 96.6% |

Patients admitted in poor condition or worse:

| | Lived | Died | Total | % who lived |
|---|---|---|---|---|
| MERCY HOSPITAL | 210 | 200 | 410 | 51.2% |
| COUNTY HOSPITAL | 40 | 70 | 110 | 36.4% |

2. This information seems paradoxical. How could it be that Mercy was ahead of County in both categories, but was behind overall? (Check the numbers - they add up.) This example shows that one should be aware that "aggregating" data may give a different picture than that obtained from categorizing it.

3. Even though County has a better overall record, one can see that Mercy is better in both categories. Breaking the data down into categories adds information that we did not have in the aggregate. It would seem that using this extra information may be the best thing to do, and thus Mercy may be the better choice - especially if your condition is poor. What this shows is that aggregated data may mask underlying factors. *Remark*: There are statistical tests that can be used to detect whether such differences are "statistically significant."

**Example 2:** This is another case where the percentages in the two separate departments give a different impression than the aggregate percentages.

4. What percentage of male applicants for the hardware department was hired by U-Mart? 75%

5. What percentage of female applicants for the hardware department was hired by U-Mart? 75%

6. What percentage of male applicants for the ladies apparel department was hired by U-Mart? 10%

7. What percentage of female applicants for the ladies apparel department was hired by U-Mart? 10%

8. An attorney for a female plaintiff might argue that U-Mart hired 62% of the males who applied (62 hired, 100 applied) and only 14% of the females (45 hired, 320 applied). In fact, more women applied than men and yet more men were hired! In addition, the numbers indicate that the store may have discriminated by subtly manipulating people to apply to departments based on gender. Your honor, these numbers indicate that there was clearly gender discrimination!

9. An attorney for U-Mart might argue that U-Mart's hardware department hired 75% of the males who applied and 75% of the females who applied. This is clearly equitable. Also, U-Mart's ladies apparel department hired 10% of the males who applied and 10% of the females who applied. This is clearly equitable as well. Your honor, these numbers are an example of Simpson's Paradox and indicate that there was clearly no gender discrimination at all!

10. One assumes a juror would vote for the defendant, U-Mart. The finer analysis of the data indicates that neither department is discriminating.

***Some Remarks:***

- Reference [1] below is a widely cited article that is somewhat similar to the "U-Mart" example. It discusses a real case of what appears to be gender bias, but suggests that there may be a different picture if the data are "disaggregated."

- Another example of what may seem "paradoxical" to some people is how a presidential candidate can win the popular vote and yet lose the election because he loses the electoral vote. This happened in the 2000 election as the table below shows. The phenomenon is different from Simpson's paradox but there's an underlying similarity in that breaking data up in different ways can lead to different outcomes.

| | Party | Popular vote | | Electoral vote | |
|---|---|---|---|---|---|
| George W. Bush | Republican | 50,459,624 | 47.87% | 271 | 50.40% |
| Albert Gore Jr. | Democrat | 51,003,238 | 48.38% | 266 | 49.40% |
| Ralph Nader | Green | 2,882,985 | 2.74% | 0 | 0% |

*The following two articles are real life examples of controversies caused in large part by the way in which numbers are interpreted. The first concerns class sizes at the University of Montana. The second concerns using statewide averages of test scores for measuring performance of schools.*

The Montana Kaimin is the student newspaper of the University of Montana in Missoula. The following article is from their web site http://www.kaimin.org/Apr01/4-18-01/index_4-19-01.html

**Math chair: Class size stats misleading**
**by Erik Olson**
**Montana Kaimin**

Class sizes haven't increased that much, according to the provost's office. However, at least one department chair said those numbers can be misleading, depending on how the statistics are interpreted.

Jim Hirstein, chair of the math department, said while the numbers for average class sizes that Provost Lois Muir released at last week's Faculty Senate are accurate, they don't show the whole picture for the average student taking Math 117.

According to the provost's report, the math department has an average of 35 students per class, an increase of seven from last spring. But, Hirstein said, that number doesn't show the significant leaps that occur in some lower-division classes.

"Here's a case where averages seem to be misleading," he said.

For an example, Hirstein said Math 117 is composed of one 240-student lecture section that meets three times a week and eight 30-student sections that meet once a week. According to the calculations used in Muir's report, the average was computed by considering those nine different times as each a separate class. That average for Math 117 comes to 53 students per class.

That number, Hirstein said, is the average from the faculty's point of view, and not the students'.

Students in Math 117 go to three class sessions per week with 240 students and one class session with 30 students. From the students' perspective, the average class size per week for Math 117 is 187 students.

Last semester, Math 117 was taught exclusively in small discussion sections. However, with the loss of adjunct professors and Muir's mandate to keep the same number of seats available for all classes, Math 117 ballooned to its current format.

Hirstein said numbers for Math 100 also increased from 30 students to 50, but upper-level courses remained steady or decreased in size, which balanced out the averages. "When you do it with averages, you're masking the changes," he said.

Muir stuck by the numbers she received from the Office of Institutional Research, saying the office did separate the numbers by specific courses. Larger courses such as Math 117 are balanced out by other smaller courses, she said.

"It tells you something about what kind of mixture of classes we have here," she said. "At a university, we have to have large courses, medium courses and small courses."

According to the statistics, class sizes campus wide increased from an average of 26 students per class last spring semester to 28 students this semester. In the College of Arts and Sciences, the size increased from 31 students last year to 35 this year.

---

**Warning: Why Average Isn't Average!**
**Simplistic Statistics Can Be Misleading**
**In Measuring for Accountability in Education**
(Reprinted by permission.)

Author: Douglas E. Hall Executive Director
New Hampshire Center for Public Policy Studies in Association with: Institute for Policy and Social Science Research, University of New Hampshire, June 1998
http://www.unh.edu/nhcpps/education/average.html

***Executive Summary*** Using statewide averages of test scores or other statistics as a benchmark for measuring performance of a school or a school district can be analytically misleading. The New Hampshire Center for Public Policy Studies raises a very strong caution in this regard. While the Center enthusiastically and unequivocally supports emphasizing results, we believe measuring, reporting and interpreting these results must be done carefully.

One of the proposed measures of achievement of individual schools and school districts includes the results of the state's assessment tests of 3rd, 6th, and 10th graders. As this paper demonstrates, it is possible for the average score of students in a school to be below the statewide average, yet the average for every sub-group of students in that same school to be above the statewide average for similar sub-groups. This can occur because the demographic composition of the student population of one school can vary considerably from the state average demographic composition for all schools.

The Center strongly cautions state policy-makers about moving too rapidly into an overly simplistic reporting, analysis, and use of student achievement measures. In order to avoid misleading analysis, we are recommending:

- Certain basic demographic information on each student needs to be collected as part of the statewide assessment program.
- The state should use demographic data, for example from the U.S. Census, to group schools and school districts serving roughly similar student populations so that meaningful analysis can occur.

We are not suggesting that overall standards should be modified to accommodate different student populations. On the contrary, whatever standards are agreed upon as indicative of an "adequate education" should hold for all students. Nor are we criticizing the state's standardized tests. It is simply the analysis and reporting of the results of the tests that needs to be more sophisticated if those results are to be used as a basis for effectively evaluating performance of schools, school districts, and the system as a whole. Our grave concern is that simplistic evaluations of progress toward educational goals will lead to a one-size-fits-all mentality and inappropriate solutions that could harm our public education system rather than improve it.

***Introduction*** Using statewide averages of test scores or other statistics as a benchmark for measuring performance of a school or a school district involves a major analytical fallacy. The Center raises a very strong caution in this regard. Equal funding does not and will not guarantee equal results. We all know this from practical experience. If it were not true, all students in the same class would achieve and perform at the same level. In New Hampshire, as we rework our financing system for public schools, we are also planning how to measure more accurately what that system actually accomplishes. This new focus on student results is a positive step forward that can lay a solid foundation for reform and improvement. The New Hampshire Center for Public Policy Studies enthusiastically supports emphasizing results, but believes measuring the results must be done carefully.

There are many ideas and proposals for measuring achievement of individual schools and school districts. Among the most often mentioned, are the results of the state's assessment tests of 3rd, 6th, and 10th graders. Other suggested measures include high school dropout rates, job success, and college entrance rates. Many citizens and some policy makers are already comparing the test results from their schools and school districts to state averages at each grade level. Their degree of concern or complacency is then based on those comparisons. Policy proposals are being made, or may soon be made, to reward schools and teachers whose pupils do well when compared to these statewide norms. And, on the other hand, some policy makers may propose that schools be selected for special attention and corrective action based on results that continue to fall below these averages. The Center's caution that simple statistics can be misleading when trying to achieve greater accountability in education can best be explained by an example.

***The Success of School X.*** For the sake of simplicity, assume that the statewide assessment test provides a score for each student between 0 and 100. Also assume that students are divided into two groups. Here are the results for each group, both state average, and in School X.

Table 1

| | Statewide Average | School X Average |
|---|---|---|
| Student Group A | 80 | 85 |
| Student Group B | 50 | 60 |

Based on these results, how well is School X doing?

The students in Group A are doing somewhat better that their counterparts throughout the state. Group B students are doing much better than the average of their counterparts. In fact, this might be one of the best performing schools among Group B students. Since this school has better than average results for both groups, we might validly conclude this school is doing pretty well.

Using real life categories, Group B might be all students whose family income is low enough to qualify them for the free and reduced-price lunch program. Group A would be those students whose family incomes are higher and who are ineligible for that program. Thus, School X is obtaining considerably higher achievement among its low-income students than the average around the state, and somewhat better than the average among students from higher-income families. The result is also more egalitarian than the state as a whole: the gap in average performance between the two groups is only 25 points compared to 30 points for the state. It would be fair to conclude that this school is above average and should be recognized and potentially rewarded for its relative success.

***The Struggle of School Y.*** In our second example, the test results are for all students without any differentiation.

Table 2

| | Statewide Average | School Y Average |
|---|---|---|
| All Students | 77 | 72.5 |

How well is school Y doing?

School Y's students are not performing up to the statewide average. A review of these figures by local school board members, legislators, or parents would likely cause concern. At some point, teachers and administrators might even be chastised, and it is possible that penalties of some sort might be imposed on schools or school districts.

***Caution: School X and School Y Are the Same School!*** This school obtains above average achievement for each of its student sub-populations, yet is below average overall. This is a compelling example of why "average isn't average" and how simple statistics can be misleading. Depending on which information is presented, Table 1 or Table 2, this school could be praised or chastised, rewarded or punished.

How is this possible? Simply because this school has a much larger percentage of its students that fall into Group B, the low-income group, than is the case on average for the state. The example uses 50% for the school and 10% for the state.

Under New Hampshire's current state educational assessment program, only aggregate data similar to the second table is available! Without more complete analysis and reporting, it is clearly possible that some members of the public will be led to unfairly evaluate this school as under-performing when, in fact, it is actually performing quite well.

***Factors of Importance*** There are many factors that might differentiate students into two or more groups such as Group A and Group B in the example. Most are not relevant to educational achievement. For example, if students were divided into two groups based on height, the achievement scores of the two groups would show little, if any, difference. However, there is a considerable body of research into factors that do make a difference. Among those are English-language ability, family income, parental education and presence of both parents in the home. In New Hampshire, the number of non-English speaking students is very small and not likely to impact overall achievement averages, except in a handful of school districts where the number is growing. Family income and parental education levels vary considerably by school district and by school in New Hampshire. The percentage of adults with a college degree varies from 5% in some districts to 73% in other districts. The percentage of students eligible for free/reduced-price school lunch, a family income measure, varies from 0% to 77%.

***Judging Achievement*** Before the state or local school boards begin to draw conclusions about how well schools are doing in reaching desired student achievement, it will be essential to factor in the underlying demographics of the students attending each school or school district. At the present time this is not being done. While some current policy proposals allude to this need, there is as yet no firm plan to do so.

There are two initial steps that the Center recommends be taken:

***First, certain basic demographic information on each student needs to be collected as part of the statewide assessment program.*** This is already done in other standardized tests such as the SAT and the National Assessment of Educational Progress. This information should then be used on a statistical basis to determine how well a school is performing with student sub-groups, as in School X above. This would provide the opportunity to evaluate a school's performance in a more complete and insightful manner than simply using overall averages as in School Y above.

***Second, the state should use demographic data, for example from the U.S. Census, to group schools and school districts serving roughly similar student populations***. Schools should then be identified from within each grouping that are performing well and asked to assist those in the same grouping that may not be reaching the average achievement for their underlying demographics. It would not be appropriate to assume that schools serving quite different student bodies could necessarily achieve similar results through

similar methods. (A similar principle is well understood in interscholastic athletics, where very small high schools are generally not asked to compete against the largest schools in major sports because larger student bodies are statistically more likely to have more talented athletes.)

***Maintaining High Standards*** We want to be very clear that nothing discussed here should be interpreted to mean that overall standards should be modified to accommodate different student populations. On the contrary, whatever standards are agreed upon as indicative of an "adequate education" should hold for all students. Expectations for our public educational system should be uniform.

However, no matter what measuring stick is used, reaching the desired outcomes will not be equally easy for all schools and school districts. The School X / School Y example was based on test results. The same statistical flaw occurs for all other measures when underlying differences among student sub-groups are not taken into account. We could have based the example on the high school dropout rate and made a similar point: the goal of achieving a high school dropout rate of 8% or less will be much harder to achieve in some schools than in others because of important underlying differences in the student populations. This must be recognized by everyone who participates in the public discussion of these issues.

We also want to emphasize that we are not criticizing the 3rd, 6th, and 10th grade tests created and conducted under the New Hampshire Educational Improvement and Assessment Program. The tests are an appropriate tool. It is simply the analysis and reporting of the results of the tests that needs to be more sophisticated if those results are to be used as a basis for effectively evaluating performance of schools, school districts, and the system as a whole. Our grave concern is that simplistic evaluations of progress toward educational goals will lead to a one-size-fits-all mentality and inappropriate solutions.

The Center strongly cautions state policy-makers about moving too rapidly into an overly simplistic reporting, analysis, and use of student achievement measures. As stated above, we unequivocally endorse focusing on pupil achievement and outcomes rather than budgets and inputs. But the processes, measures, and norms that are used to judge performance must be designed to illuminate real successes and failures, rather than hide them or obscure them as depicted in our example. Without sufficient care, we could harm our public education system rather than improve it.

---

***The following is a summary by Dr. Stacy Gordon of the article by Claudine Gay entitled "The Effect of Black Congressional Representation on Political Participation." American Political Science Review 95(3): 589-602. (Reference 3 below.)***

Theory on legislative representation has suggested that when African-American districts elect African-American legislators, it should lead to an increase in black voting

participation. The argument is that black constituents who are represented by black legislators feel better represented, feel as though the government may respond to their interests, and, therefore, will be more likely to participate in elections.

However, past research supports the argument that low aggregate turnout rates in districts represented by black congresspersons illustrate that black representation does not increase participation among African-Americans and may, in fact, decrease it. However, Gay (2001) notes that there is a potential "ecological fallacy" here. Just because a district is majority black, elects a black representative, and has low voter turnout, does not mean that *individual* blacks in that district are less likely to vote.

When Gay separated white voter turnout from black voter turnout in various districts, she found that *white* voter turnout was significantly lower in districts represented by an African-American than in similar districts (with similar challenger quality and similar electoral pressures) represented by a white legislator. White turnout in districts with a black representative was depressed by as little as 4.5 percentage points or as much as 18.3 percentage points (page 596). Conversely, black turnout in those same districts sometimes remained the same and sometimes, in specific types of circumstances, increased with black representation. In other words, it is a decrease in *white* turnout that leads to the aggregate decrease in voter turnout in majority black districts, not a decrease in black turnout.

Hypothetical Examples of Turnout in Districts with Black Representatives:

| District | Aggregate Voter Turnout In Dist With Black Rep.* | Black Turnout (% Black in District) | White turnout (% White in District) |
|---|---|---|---|
| #1 | - 2.0 | + 10.0 (60%) | -10.0 (40%) |
| #2 | - 4.0 | + 4.0 (65%) | -4.4 (35%) |
| #3 | -10.0 | 0.0 (70%) | -33.33 (30%) |
| #4 | - 5.0 | + 2.0 (55%) | -8.67 (45%) |

* When compared to similar districts with white representatives.

**References:**


1. *Sex Bias in Graduate Admissions: Data from Berkeley*, Bickel, P.J., Hammel, P.A., O'Connell, J.W, Science, (1975).

2. *Simpson's Paradox in Real Life*, Wagner,C.H., The American Statistician (1982)

3. Claudine Gay, "The Effect of Black Congressional Representation on Political Participation." American Political Science Review 95(3): 589-602.

4. www.amstat.org/publications/jse/secure/v7n3/datasets.morrell.cfm

5. http://www.unh.edu/nhcpps/education/average.html

6. http://cq-pan.cqu.edu.au/schools/smad/simpadox.html