# ASSOCIATION-CAUSATION PROBLEMS IN NEWS STORIES

Milo Schield
Augsburg College, United States of America
schield@augsburg.edu

*Statistical educators strongly emphasize the importance of distinguishing association and causation. Yet this semantic distinction is often obscured due to an inappropriate choice of words. This paper investigates related problems–inaccuracies, omissions and ambiguities–in number-based news stories. It investigates association-related problems involving large numbers, confusion of the inverse, missing context, times-less and times-more comparisons, incomplete comparisons, slope comparisons and confusing "frequently" with "likely." It investigates causation-related problems involving causation words and action verbs. These problems may create reader confusion and misunderstanding. Data is needed on how readers understand the presentation of association and causation in the media. Statistical educators, journalism faculty and quantitative journalists should join together in analyzing these problems in number-based news stories.*

## INTRODUCTION

Some news stories involve errors, omissions or ambiguities in number-based associations.

### *Confusing Big Numbers*

According to Smith (2010) in a three year period there were 23 mix-ups of the words *million* and *billion* in the LA Times and 38 mix-ups in the New York Times.

*Analysis*: This is a big mix-up! Data is needed on how readers understand these. In a convenience survey, 200 Augsburg college students were asked "How big is a billion?" They said: 1,000 million (59%), 100 million (18%), a million million (10%), 10 million (7%) and "Don't know" (6%).

### *Confusion of the Inverse*

Sometimes the part and whole (numerator and denominator) in part-whole percentages are reversed: a fallacy of the transposed conditional. Consider these cases:

- *Racial Imbalance Persists at Elite Public Schools* by Javier C. Hernandez. New York Times 11/08/2008. "at Stuyvesant...2% of blacks, 3% of Hispanics, 24% of whites and 72% of Asians were accepted." He should have said "among those accepted at Stuyvesant...2% are blacks, 3% are Hispanics, 24% are whites and 72% are Asians."
- *Study says too much candy could lead to prison.* AP 9/30/2009. "Of the children who ate candies or chocolates daily at age 10, 69 percent were later arrested for a violent offense by the age of 34." The truth: "69% of respondents who were violent by the age of 34 years reported that they ate confectionary nearly every day during childhood."

*Analysis*: In the first story, the reporter got things backward by using *race* as the whole and *accepted* as the part. In the second story, the reporter got things backward by using *eating candy* as the whole and *arrested for a violent crime* as the part. One clue of a possible inversion in the first story was the fact that the percentages added up to approximately 100%. A clue of the inversion in the second story was the high success rate (69%) in predicting a rare event 20 years in advance.

### *Missing Context*

In some stories, a comparison of two ratios (e.g., "cancer rate doubles") is presented without giving the size of the ratios being compared. This is not confusing when readers have some idea of their sizes (e.g., interest rates or unemployment rates). This omission can be misleading when readers have no idea of the sizes (e.g., cancer rates). Consider this story:

- *Soft Drinks Could Boost Pancreatic Cancer Risk.* HealthDay News (2/8/2010). Lead sentence: "People who down two or more soft drinks a week may have double the risk of developing deadly pancreatic cancer, compared to non-soda drinkers." Quote from the story: "But the overall number of people developing the malignancy remains low, with the U.S. National Cancer Institute (NCI) estimating 42,470 new cases last year."

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php 

*Analysis*: "Double the risk" is an attention-getter, but if the underlying rate is low, a doubled risk may not be a big problem. The incidence rate of pancreatic cancer is 14 per 100,000 per year: 42,000 divided by the US population. This rate is less than half the 2008 US accidental death rate: 38 per 100,000. Woloshin et al. (2008) identified this lack of context as a major problem.

*Confusing association with causation*

Confusing million and billion may be of far less consequence than confusing association and causation. Although the AP Stylebook (Christian et al., 2009) has little to say on this matter, using statistical associations as evidence for causal connections is central to many–if not most–number-based news stories. To see this, consider the following headlines all based on the same journal article: Wilper et al. (2009):

- 45,000 deaths *attributable to* uninsurance (PNHP 9/17/09)
- 45,000 American deaths *associated with* lack of insurance (CNN 9/18/09)
- Study: Uninsured Americans Have 40 Percent *Higher* Death Risk (Ivanhoe, 9/18/09)
- Study: 45,000 Uninsured *Die* a Year (CBS News, 9/17/09)
- No health coverage *tied to* 45,000 deaths a year (Rueters MSNBC 9/17/09)
- Lack of insurance *linked to* 45,000 deaths (White Coat News, 9/17/09)
- Study *links* 45,000 U.S. deaths to lack of insurance (Reuters, 9/17/09)
- Study: 45,000 U.S. Deaths *From* Lack of Insurance (MoneyNews 9/17/09)
- One American dies every 12 minutes *due to* no health insurance (blog DR 9/17/09)
- 45,000 Americans die … *because of* lack of health insurance (blog MyDD 9/17/09)
- Lack of Health Insurance *Kills* 45,000 a Year (Health Insurance com, Inst.)
- Lack of Health Insurance *cause* 44789 deaths in United States every year (blog)
- Lack of insurance *to blame for* almost 45,000 deaths: Study (HealthDay 9/17/09).

*Analysis*: Notice the movement from pure association (*attributable to*, *associated with*) to connectors (*tied to*, *linked to*) to quasi-causal keywords (*due to*, *because of*), to action verbs (*kills*), to causality (*cause*) and moral judgment (*to blame*). But the original study just claimed association: "Lack of health insurance is associated with … 44,789 deaths per year." Wilper et al. (2009).

Each of these issues is central to forming or describing an association or to distinguishing association from causation. Statistical educators and journalism faculty need to review these issues. In the million-billion mix-up, journalists might be encouraged to use rates. Instead of saying "the US debt is $12 trillion," say that it is "$113,000 per taxpayer." Instead of saying "$108 trillion in US unfunded liabilities," say "$349,000 per taxpayer." In the percent inversion cases, journalists may need special training to spot these errors. In the association-causation confusion, journalistic standards may be needed on what wording to use and what wording should be avoided.

## APPROACH

This paper investigates how association and causation are described–properly or improperly–in number-based news stories. There are many good books on numbers in the media: Gigerenzer (1986), Meyer (1991), Mauro (1992), Paulos (1995), Cohn (1999), Wickham (1999), Cohen (2001), Best (2004), Miller (2004), and Woloshin et al. (2008). Few–if any–focus on the confusion in forming comparisons and in distinguishing association and causation.

These findings are based on two analyses of number-based stories in the everyday media. Raymond and Schield (2008) involved a judgment analysis of 256 stories and a computer content-analysis of 899 stories. Schield and Raymond (2009) did a computer content-analysis of headlines in 2,000 stories. In this paper, topics are introduced, news stories presented and problems analyzed.

## ASSOCIATION

The simplest form of association is a two-number comparison. Comparisons create context. Comparisons are common in number-based news stories. The computer-match of words in number-based news stories found these prevalences: *more* (90%), *more than* (47%), *less than* (42%), *er than* (32%), *more likely* (25%), *less likely* (11%), *times more* (7%), *percent more* (4%), *times as* (2%) and *times less* (1%).

An arithmetic comparison is one way to describe a quantitative association involving two groups. These comparisons can involve arithmetic differences (A is X more/less than B), ratios (A is X times [as much as] B), percentage differences (A is X% more/less than B) or times differences

(A is X times more/less than B). With this background, consider the problems in using "times less" and "times more" in number-based news stories.

A. *Times-Less Comparisons*

Raymond and Schield (2008) found the phrase "times less" in 1% of number-based news stories. Consider these examples:

- Circumcised men were two to three times less likely to contract HIV. AFP 5/28/2009. *Botswana circumcision drive will prevent HIV infections.*
- Those [patients] who had end-of-life talks were three times less likely to spend their final week in intensive care, four times less likely to be on breathing machines, and six times less likely to be resuscitated. AP 6/15/2008. *Most cancer doctors avoid saying it's the end.*
- The new study shows that only 3% of 16- to 29-year-old non-Christians express favorable views of evangelicals. This means that today's young non-Christians are eight times less likely to experience positive associations toward evangelicals than were non-Christians of the Boomer generation (25%). Barna 9/24/2007. *A New Generation Expresses its Skepticism and Frustration with Christianity*.

*Analysis*: In each case "times less" involves a ratio comparing two positive numbers. This ratio usage states, "*small#* is <large-to-small ratio> times less than *large#*." It converts fractions into whole numbers. It converts "80% less than ten" (20% of ten) into "two is five times less than ten."

This ratio usage is quite different from a percent-change usage. Schield (1999) and Wickham (1999). If *small#* is P percent less than *large#*, a percent-change usage means "*small#* is 100*P times less than *large#*." Two is 80% (0.8 times) less than ten. If the number before "times less" is more than one, the result is negative. Two times (200%) less than ten is -10.

The ratio usage of "times less" seems much more common than the percent-change usage. We can spot a ratio compare when the number in a *times-less* comparison is greater than one and the numbers being compared are positive. Of the 130 instances of "times less" in the WordBanks (2010) database, 117 were preceded by a number–each greater than one., e.g., "The infant mortality rate in all parts of Cuba is about three to four times lower than in Philadelphia."

In our convenience survey of 200 Augsburg students, 26% said they did not know how to interpret *two times less than six*: 34% picked three, 21% picked negative six, 12% picked zero, 6% picked negative 12 and 1% picked 12. When asked about the *times-less* usage in the AP story, the online editor of the AP Stylebook said, "Not in stylebook." [Personal communication]

B. *Times-More Comparisons*

As a ratio, ten is certainly five times [as much as] two. As a percent change, ten is 400% more than two. Ten is described as "five times more than two" when it is ratio based, but as "four times more than two" when it is percent-change based. Raymond and Schield (2008) found "times more" in 7% of number-based news stories. Consider these stories:

- *Swimming pool chlorine raises asthma, allergy risks: study*. "teenagers who spent more than 100 hours swimming in chlorinated pools were up to six times more at risk of having asthma [12%] than other teens [2%]." Reuters 9/14/2009.
- *Dementia Risk Higher for NFL Players*. "retired players … are 19 times more likely to struggle with memory problems [1.9%] than similarly aged men who never played professional football [0.1%]. HealthDay 9/30/2009.

*Analysis*: Based on the numbers being compared in these stories, each of these *times-more* comparisons is ratio-based. Unfortunately, there is no easy way to tell whether a *times-more* comparison is ratio based or percent-change based. But this ambiguity in a *times-more* comparison (a difference of one between two positive numbers) is much smaller than this ambiguity in a *times-less* comparison (a difference between positive and negative numbers).

Note that *times-more* and *times-less* comparisons are closely related in the ratio-based usage. Saying "six is three times more than two" supports saying "two is three times less than six."

In our convenience survey of 200 Augsburg students, 19% said they did not know how to interpret *two times more than three*: 36% picked nine, 34% picked six, 11% picked 12 and 1% picked three. Of those surveyed, 87% had completed or were taking a college math course.

Merriam-Webster's Dictionary of English Usage (1994) cites historical precedent for the ratio-based usage. Volokh (2009) and Freeman (2007) uphold a ratio-based usage. Meyer (1991), Wickham (1999), Schield (1999) and Perlman (2009) support a percent-change based usage.

Data is needed to see if mathematically-sophisticated readers interpret this comparison differently from others. Perhaps journalists should avoid using any "times more/less" comparison.

C. *Incomplete Comparisons with multiple groups*

Incomplete comparisons are common in headlines because they involve fewer words – they are punchier–than complete comparisons. There is little problem in determining the full comparison when the subjects involve a single group as in the first headline. Assuming that both groups are people, the comparison group is readily determined by negation: "than shorter people."

- *Taller People Earn More Money*. Reuters 8/25/2006.
- *Tall Women Earn More Money*. [Author-based revision of the preceding title]
- *Obese women less likely to be screened for cancer.* Reuters 5/8/2007.
- *Study: Young girls more likely to be fat*. AP 5/9/2007.
- *Severely Obese Women More Likely to Skip Cancer Scans.* HealthDay 5/17/2007.

*Analysis*: But problems arise when the subjects are members of several groups as in the second "headline." A complete negation (*short men*) seems inappropriate. But which partial negation: *short women* or *tall men*? The subsequent headlines all involve multi-group comparisons. These ambiguities may be untangled with rules such as "negate the adjective", "negate the noun" or "negate the item closest to (or further from) the comparison." But are there any such rules and do readers understand them? Data is needed on how readers interpret this ambiguity.

D. *Confusing "Frequently" with "Likely."*

*Frequently* and *likely* are similar but their difference may signal a confusion of the inverse. "The group *most likely* to experience an outcome" may differ from "the group found *most frequently* among those experiencing that same outcome".

- *1995 Honda Civic: Most Frequently Stolen Car*. State Farm Insurance. 7/9/2008.
- *New car study lists most likely to be stolen – '96 Honda Civic*. Mountain Times 8/27/2009
- *Study: Cadillac Escalade most likely stolen.* AP 6/7/2006.

*Analysis:* How can two different cars both be the US car most *likely stolen* in the same year? *Likely stolen* can mean *likely to be stolen* or *likely to be found among those stolen*. The two uses give the same results for equal size groups; they can give different results for different size groups.

Suppose that all cars were produced in equal numbers. In this case, the most stolen car – the car stolen most often, the car most likely to be found among stolen cars – would also be the car [that is] most likely to be stolen. Here the distinction between two kinds of *likely* has no difference.

Now consider reality. Among stolen cars in the US, the Civic is most prevalent – most likely to be found. But the car with the highest theft rate – the car most likely to be stolen – is the Cadillac Escalade. The car most likely among stolen cars is not the car most likely to be stolen.

So what occurs most frequently–what is most likely to be found–is sometimes–but not always–what is most likely to occur. This distinction–a confusion of the inverse–is quite subtle. In a survey, 200 Augsburg students were asked if these statements mean the same thing: (a) *This car is most frequently stolen*. (b) *This car is most likely to be stolen*. 89% said *No* (they have different meanings); 7.5% said *Yes* (they have the same meaning) and 3.5% said they didn't know.

E. *Slope Comparisons*

An association between two quantities is often described by a slope: comparing the change in one quantity with the change in the other. A good example is the slope of a stairway (rise/run): for each additional unit of rise, the run increases by 1.5 units.

Slope comparisons can be *within subjects* (comparing things measured at different times for the same subject) or *between groups* (comparing things measured at the same or different times for different groups), e.g., "As a child learns more words, their reading speed increases" versus "Among adults, as weight increases, height increases." A *between-group comparison* involves a change in focus, e.g., Adults who weigh more tend to be taller. The *within-subject comparison* may give more support for a causal interpretation by reducing confounder influence.

Since these two slope comparisons give different support for causation, it is important for journalists to know and communicate which kind of comparison is involved in a given association. This is not easy. Consider some examples in number-based news stories:

- *Alzheimer's memory loss faster among well-educated.* Reuters 10/22/2007. "the rate of decline unfolded 4 percent more quickly for each additional year of education."
- *Each Daily Soda Increases Obesity Risk 60%*. The Lancet, 2001; 357:505-508. "For every can or glass of sugar-sweetened beverage a child drank [a day] …, a child's … chance of becoming obese increased 60%."
- *Study Finds Link Between Television Viewing And Attention Problems In Children*. Science Daily 4/6/2004. "each hour of television watched per day at ages 1-3 increases the risk of attention problems, such as ADHD, by almost 10 percent at age 7."

*Analysis*: If each of these outcomes is considered to be unrepeatable, then each of these slope associations involves a *between-group change-in-focus comparison*. "Children who watched a given amount of TV per day at ages 1-3 had a 10% greater prevalence of attention problems at age 7 than those children who watched an hour less per day." Readers may ignore this subtlety and treat a *between-group* comparison as strong evidence of a temporal or causal connection within a person. They may expect the result if they made the associated change. Confused readers may overestimate the strength of between-group associations in supporting causation.

## ASSOCIATION VERSUS CAUSATION

Schield and Raymond (2009) did a computer-match on causation-association words in 2,000 headlines for number-based news stories. As a percentage of headlines, they found:

- Association words (2%): Words that just state that two things go together: verbs (*associate, relate, correlate, predict* or *attributed*) or nouns (*association, link, factor* or *relation*).
- Causation words (8%). Words that claim the presence of one factor changes the presence, amount, duration or frequency of a condition or event. Causal words (2%) include *cause, effect* and *result in*. Sufficient words (6%) include *prevent, stop, end, kill, cure* and *ban*.
- Between words (67%): Words that imply association for some readers and causation for others. These include action verbs (43%), comparative keywords (10%), connector keywords (5%), temporal keywords (3%), quasi-causal keywords (3%), superlative keywords (3%), comparison adverbs (0.8%) and action nouns (0.1%).

For more detail on these terms, see Schield and Raymond (2009). This framework sets a context for examining headlines that use causal words and action verbs.

### A. *Ambiguity in Interpreting Causal Words*

Causal words such as *cause, effect* and *result* are used in about 2% of the headlines for number-based news stories:

- *Obesity causes later onset of puberty in boys*. HealthDay News 2/2/2010.
- *Junk food causes a third of heart attacks*. Reuters 10/20/2008.
- *Obesity growing to be top cancer cause*. AP 2/15/2008.

*Analysis*: If these events are considered to be unrepeatable then same-subject experiments are impossible. Since these studies were observational–not clinical trials–the evidence for causation is circumstantial at best. Using unqualified causal words to describe the results of any observational study gives observational studies the same status as controlled experiments. This ambiguity may create confusion among readers about the quality of the science. Journalists might be told to avoid using causal words when describing associations based on observational studies.

### B. *Ambiguity in Using Action Words*

In almost half (43%) of number-based headlines, journalists describe an association by using action words such as *ups, cuts, raises* and *boosts.* In a *within-subject comparison*, this action reflects a change within a given subject: "Eating less cuts weight." When a *within-subject comparison* is impossible, the action reflects a *change in focus comparison* between different groups: "Eating less cuts risk of diabetes." Consider these headlines:

- *Study: Estratest doubles breast cancer risk.* AP 7/25/2006.
- *Gene increases depression risk: study.* AFP 3/2/2006.
- *Weddings boost mood: study.* Reuters 8/13/2006.
- *Expanding waist worsens kids' sleep apnea*. Reuters 1/29/2010.

*Analysis*: If developing breast cancer is considered a one-time event, then the first association must involve a change in focus. While being depressed can be a continuing state, the inability to change one's genes means that second headline must also involve a change in focus. Only the last two could involve a *within-subject before-after comparison*. If so, these associations would give stronger support for a causal connection than the first two. Yet there is no indication of this difference in the wording. By using action words to describe both kinds of comparisons, journalists may inadvertently contribute to public confusion about the value of science.

## CONCLUSION

Research is needed to see whether journalists or readers understand the differences between association and causation and how readers interpret various devices for indicating this difference. Are connection verbs (*linked*, *tied* or *related*) interpreted differently than action verbs (e.g., *cuts*)?

The topics mentioned in this paper are just a beginning. There are many other topics in describing association and causation that need attention. All sources of ambiguity and error involving association and causation should be reviewed to see which ones are most likely to create confusion in everyday usage. Statistical educators, journalism faculty and quantitative journalists should join together in analyzing statistical illiteracy in the everyday media so they can give better guidance to journalists and to the readers of numbers in the news.

## REFERENCES

Christian, D., Jacobsen, S., & Minthorn, D. (2009). *2009 AP Stylebook*. The Associated Press.

Best, J. (2004). *More Damned Lies and Statistics.* University of California Press.

Cohen, S. (2001). *Numbers in the Newsroom: Using Math and Statistics in the News*. IRE, Inc.

Cohn, V. (1999). *News and Numbers: A Guide to Report Statistical Claims*. Blackwell Publishing.

Freeman, J. (2007). Is 'three times less' three times worse? New York Times Opinion. Online: www.nytimes.com/2007/10/28/opinion/28iht-edfreeman.1.8081659.html?_r=2.

Gigerenzer, G. (1986). *Calculated Risks*. Simon & Schuster.

Mauro, J. (1992). *Statistical Deception at Work.* Erlbaum Associates, Inc.

*Merriam-Webster's Dictionary of English Usage* (1994). Online: http://books.google.com/.

Meyer, P. (1991). *The New Precision Journalism*. 4th ed. (2002). Rowman & Littlefield.

Miller, J. (2004). *The Chicago Guide to Writing About Numbers*. University of Chicago Press.

Paulos, J. A. (1995). *A Mathematician Reads the Newspaper*. 14th ed. Anchor.

Perlman, Merril (2009). Times Up: Is "three times more" the same as "three times as many"? *Colombia Journalism Review*. Online: www.cjr.org/language_corner/times_up.php

Raymond, R. and M. Schield (2008). Numbers in the News: A Survey, *2008 American Statistical Association Proceedings of the Section on Statistical Education.* [CD-ROM]. Online: www.StatLit.org/pdf/2008RaymondSchieldASA.pdf.

Schield, M. (1999). Common Errors in Forming Arithmetic Comparisons, 1999 APDU *Of Significance* (pp. 47-48). Online: www.StatLit.org/pdf/1999SchieldAPDU2.pdf.

Schield, M., & Raymond, R. (2009). Distinguishing Association from Causation in Media Headlines. *2009 ASA Proceedings of the Section on Statistical Education.* [CD-ROM] (pp. 4371-4385). Online: www.StatLit.org/pdf/2009SchieldRaymondASA.pdf.

Smith, D. (2010). But who's counting? LA Times. Online: www.statlit.org/pdf/2010SmithLATimes.pdf.

Volokh, E. (2009). Times Less Than. Online: http://volokh.com/posts/1253897118.shtml.

Wilper, A., Woolhandler, S., Lasser, K., McCormick, D., Bor, D. & Himmelstein, D. (2009). *Health Insurance and Mortality in US Adults. American Journal of Public Health* (Sept. 17) Online: http://ajph.aphapublications.org/cgi/content/abstract/AJPH.2008.157685v1.

Wickham, K. W. (1999). *Math Tools for Journalists.* Marion Street Press, Second edition 2003.

Woloshin et al. (2008). *Know Your Chances: Understanding Health Statistics*. U. California Press.

WordBanks Database (2010). Harper Collins. Online: http://wordbanks.harpercollins.co.uk