Confound Those Speculative Statistics

Milo Schield
Augsburg College, Minneapolis, MN

Abstract

Speculative statistics are model-based statistics. These include deaths attributable to being in a group (deaths linked to a distant cause). Such deaths are those due to primary smoke, second-hand smoke, obesity and radon. This paper reviews the epidemiological model used and introduces a graphical technique to present three big ideas: that a confounder can influence (1) an association in an observational study, (2) the speculative statistics generated by epidemiological models and (3) the statistical significance calculated in comparing these statistics. This paper argues that if students are to deal with the statistics of everyday life, they must appreciate these three big ideas. They must be aware that speculative statistics are often indistinguishable from actual statistics and are vulnerable to confounding. A separate statistical literacy course based on these ideas is recommended.

1. Speculative Statistics

Speculative statistics are arguably the most rapidly growing form of statistics in modern-day society. Speculative statistics are model-based statistics. Sophisticated examples include forecasts of weather and global warming. Another example is *green math* – the attempt to measure the carbon footprint or environmental impact of consumer products.

One type of speculative statistic involves group-based statistics. Common examples involve deaths, injuries or accidents *attributable to* being a member of a particular group. These speculative statistics are generally based on a simple epidemiological model. This epidemiological model uses a comparison of rates between two groups to calculate the percentage or number of cases *attributable* to being in the higher-rate group. The speculative statistics presented in this paper are all of this kind.

Take deaths. Deaths are tabulated in the U.S. Statistical Abstract where the proximate causes are verifiable by medical doctors and coroners. These verifiable deaths include those due to heart disease, cancer, accidents, Parkinson's, Alzheimer's, diabetes, HIV, etc. (Table 111, 2008 US Statistical Abstract).

Speculative deaths are actual deaths but they are speculatively tied to distant or remote "causes" such as smoking, obesity or global warming. These distant or remote "causes" presumably bring about the death through the proximate causes that are recorded on death certificates.

These group-based speculative statistics are indistinguishable from factual statistics (such as births and deaths) in their form. But they are different because they are arguable: they assert something that is unobservable – which is why they are created using models rather than counted or measured.

Speculative statistics are similar to survey statistics in that both involve some kind of inference. Speculative statistics differ from survey-based statistics in that survey-based statistics generalize from sample to population while speculative statistics imply causal connections based on group associations.

Speculative deaths – deaths where the remote causes are not verifiable -- are not recorded in the U. S. Statistical Abstract. Speculative deaths are calculated for remote causes. Some of these remote causes may have some causal efficacy (smoking) while others seem to lack any obvious causal efficacy (physical inactivity). No matter, this method works for any outcome involving two related groups: the exposure and control groups.

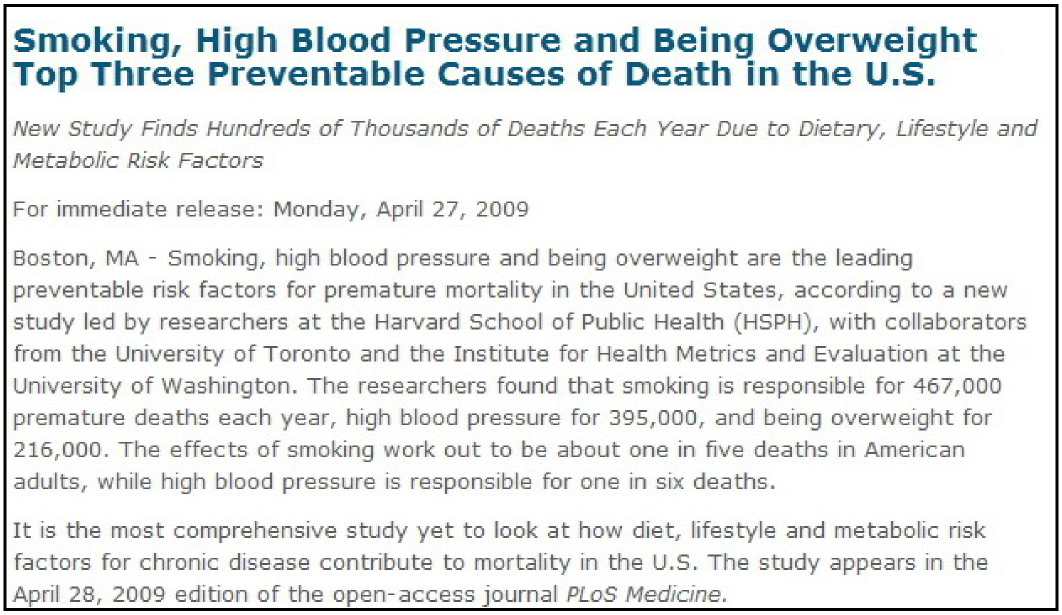
Counts of speculative deaths are available on the web by searching on “deaths attributed” or “premature deaths”. Sources of speculative deaths include smoking, obesity, physical inactivity, eating animal products, adverse drug reactions, COPD, avoidable medical mistakes, microbial agents, excessive alcohol use, gap in quality health care, pollution-related sickness, ESRD, second-hand smoke, flu, chemical exposure, genital NMSC, radon gas, illicit drug use, hypertension, heat waves, staph infections, MRSA (more than AIDS), drug overdose, pneumonia and influenza, aspirin, soot pollution and SIDS.

Speculative deaths are often described as premature deaths – although there is no way for a physician or coroner to know whether a particular person’s death was premature.

Danaei et al (2009) estimated premature deaths from these remote causes:

Table 1: Preventable Causes of Deaths in U.S.

Cause	**Number**		**Cause**	**Number**
Smoking	467,000		High LDL cholesterol	113,000
Hypertension (blood pressure)	395,000		High salt intake	102,000
Overweight-obesity	216,000		Low Omega-3 (seafood)	84,000
Inactivity, inadequate activity	191,000		High trans fatty acids	82,000
High blood sugar	191,000		Low fruits vegetables	58,000



Smoking, High Blood Pressure and Being Overweight Top Three Preventable Causes of Death in the U.S.

New Study Finds Hundreds of Thousands of Deaths Each Year Due to Dietary, Lifestyle and Metabolic Risk Factors

For immediate release: Monday, April 27, 2009

Boston, MA - Smoking, high blood pressure and being overweight are the leading preventable risk factors for premature mortality in the United States, according to a new study led by researchers at the Harvard School of Public Health (HSPH), with collaborators from the University of Toronto and the Institute for Health Metrics and Evaluation at the University of Washington. The researchers found that smoking is responsible for 467,000 premature deaths each year, high blood pressure for 395,000, and being overweight for 216,000. The effects of smoking work out to be about one in five deaths in American adults, while high blood pressure is responsible for one in six deaths.

It is the most comprehensive study yet to look at how diet, lifestyle and metabolic risk factors for chronic disease contribute to mortality in the U.S. The study appears in the April 28, 2009 edition of the open-access journal *PLoS Medicine*.

Figure 1 Harvard Press Release

Source: http://www.hsph.harvard.edu/news/press-releases/2009-releases/smoking-high-blood-pressure-overweight-preventable-causes-death-us.html

Students have no idea that these deaths due to smoking or obesity are speculative statistics. They have no idea that they can be influenced by a change in model assumptions such as by taking into account the influence of plausible confounder.

Speculative statistics abound. Speculative deaths can be calculated for almost any remote condition or event that an investigator deems dangerous. Speculative deaths have been calculated for lifestyle factors, personal decisions, being uninsured, new age fraud and global warming.

Speculative lives can be saved too. Speculative *lives saved* have sprouted from remote causes such as exercises, vitamins and the Los Angeles City's Living Wage Ordinance.

We must recognize that these speculative statistics may have great political significance.

For example, Mokdad et al (2004) projected 400,000 US deaths per year due to obesity and overweight. They predicted that deaths due to obesity would overtake deaths due to smoking in the next decade. Mokdad et al (2005) corrected this to 365,000 per year.

The Director of the Center for Disease Control (CDC), a co-author of these studies, used these numbers to argue – successfully – for an increase in the budget for the CDC.

But one year later, Flegal et al (2005) estimated the US deaths per year due to obesity and overweight as 26,000. Both groups used the same sources of data. How can these numbers change so rapidly? Welcome to the world of speculative statistics. These model-based statistics are based solely on group associations typically taken in observational studies. As such they are vulnerable to the influence of confounding.

The more remote the factor, the smaller the relative risk, the greater the vulnerability to the influence of confounding. Being overweight is very remote from causing death. And so it is readily influenced by taking into account factors that are much larger – and closer to causing death – such as age.

If students are to be properly prepared to deal with the statistics of everyday life, they must be aware that (1) speculative statistics are almost indistinguishable from actual statistics, (2) speculative statistics are vulnerable to the influence of confounding, and (3) confounding can influence statistical significance.

Statistical educators should consider ways to prepare their students for a world populated by an increasing number of speculative statistics. This paper recommends that statistical educators develop a separate course in statistical literacy that focuses entirely on these big ideas.

The rest of this paper provides details on the epidemiological model (section 2), applying this model to the same group at two different times (section 3) and applying this model to two groups (exposed and control) at the same time (section 4). Section 5 examines the influence of confounding on an association (section 5.1), on cases attributable (section 5.2) and on statistical significance (section 5.3). The pros and cons of incorporating this material in an introductory course are analyzed (section 6), the findings are summarized (section 7) and recommendations are made (section 8).

2. Epidemiological Model

Epidemiologists have a simple model for attributing a certain fraction of the cases in the exposed group to their membership in that group. From this they can easily calculate the number of cases in the exposed group that are attributable to being in the exposed group.

Attribution has a technical meaning in statistics that is different from that in journalism.

In journalism, attribution is all about quoting a source. The most common word is "said". Other words include "stated" (implies formality), "according to" (reserve this for written reports or stories), "charged" (reserve this for formal legal actions). Words like "added" or "mentioned" imply the statement was made in passing or incidentally. Words like "exclaimed" or "gushed" imply that the statement was made by a person in a certain emotional state. In journalism, *attributable* or *attributed to* is used when the person refuses to be quoted or to be named as the author of a quote.

Saying something is *attributable* is a way of saying it may be *associated with* or *due to* a related factor. This aspect of attributable is what is intended in the statistical usage.

The epidemiological model for attributing is extremely simple. The groups can be any two groups; the outcome can be anything that is countable. Consider two examples. The first compares the same rate for the same group at two different times. The second compares the same rate for two different groups at the same time.

3. Comparing Identical Rates for the Same Group at Different Times

The clearest example of this involves a comparison of two rates involving the same whole and part but taken at two different times. Consider these age-adjusted rates on US cancer deaths per 100,000: 216 in 1990 and 184 in 2005.[1,2]

> "Cancer society officials estimate that 650,000 deaths were avoided from 1990 to 2005 because of the decline in the death rate." [3]

The derivation is quite simple. For a given year (2005), calculate the difference in the death rates (216 – 184 = 32 per 100,000 population). Apply that difference to the 2005 US population (296.3 million). The product (94,816) gives the lives "saved" in 2005. Repeat this for the other years and total the number obtained.

As a short cut, we'd estimate the average to be half the maximum during those 16 years assuming a constant population. Multiplying 95,000 by eight gives an estimated 760,000 lives saved during those 16 years. This estimate would be closer to the 650,000 stated in the news story if the population were increasing.

Notice two things. (1) The mathematics is simple. It just involves taking a difference between two rates and applying the excess to the group being compared. (2) The

[1] 2008 US Statistical Abstract, Table 112 Age-Adjusted Deaths by Selected Causes.

[2] Cancer/Oncology Journal. July/August 2009.
http://www3.interscience.wiley.com/journal/121481132/home

[3] 5/27/2009 AP News story: US cancer death rate drops again in 2006 by Mike Stobbe, AP Medical Writer. http://news.yahoo.com/s/ap/20090527/ap_on_he_me/us_med_cancer_deaths/print

assertion is minimal. Saying the number of lives saved is “because of the decline in the death rate” is a no-brainer. The outcome is a deductive consequence of the premises.

So far this procedure doesn’t involve anything mathematically complex or philosophically disputable.

4. Comparing Rates for Complimentary Groups at the Same Time

Now consider an extension of this simple procedure. Epidemiologists often deal with cross-sectional studies: studies where all the data is counted or measured at the same moment or interval in time. One of the earliest applications involved the influence of smoking on deaths due to lung cancer. Consider the counts in Table 2.

Table 2: Lung-Cancer Deaths among Smokers

Counts of Deaths	**Cause of Death**		
	Other	**Lung Cancer**	**TOTAL**
Non-Smoker	784	16	800
Smoker	160	40	200
TOTAL	944	56	1000

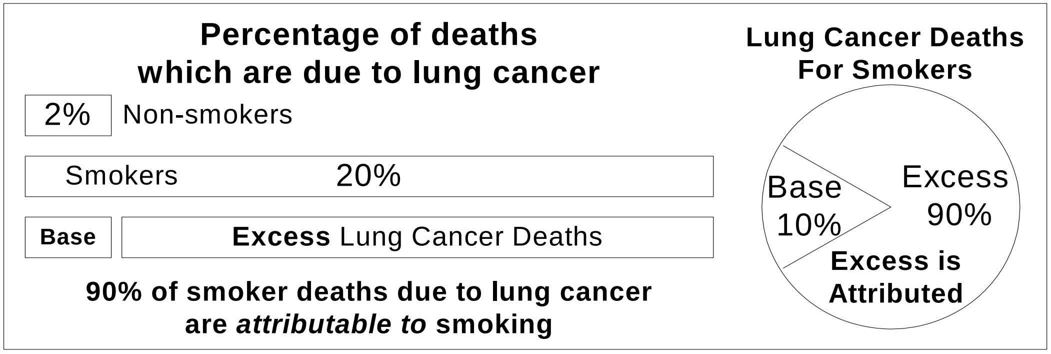
The cause of death listed on a death certificate identifies proximate causes (e.g., lung cancer) and does not speculate on distant causes (e.g., smoking).

Of the 40 lung cancer deaths among these smokers, what fraction is *attributable* to the smoking? Certainly not all of them. Non-smokers can die from lung cancer. We need the excess over what is expected if they did not smoke.

Lung cancer was the cause of death for 20% of smokers (40/200) and for 2% of non-smokers (16/800). The excess is 18%: 90% of 20%. So we say that 90% of these lung-cancer deaths among smokers are attributable to smoking. 90% = 100%(20% - 2%)/20%.

This procedure is illustrated in Figure 2.

Figure 2 Percentage of Deaths Attributable to Smoking



In any controlled study, any excess of an outcome found in the exposed group is always *attributable to* membership in that group, whether causal or not. In the following assume that the exposure group is the group having the higher rate of interest.

The **percentage of the exposure rate attributable to the exposure** is the excess between the exposure and control group rates as a percentage of the exposure rate.

% *attributable* to Exposure = 100% (Exposed_Rate – Control_Rate) / Exposed_Rate.

The word *attributable* may sound like a claim of causation but it is not. It seems easy to go from *attributable to* to *due to* to *because of*[4] to *caused by*.

The causal claim may in fact be true, but *attributable to* – a mathematical association – is seldom strong enough to validate that claim. '**Attributable to**' just means *associated with*. The strongest causal statements would be "*may be caused by* X." or "*is caused by X or something else.*" The second form is seldom used, because it so clearly indicates the limitation on this induction. The first form is commonly used.

5. Influence of Confounding on Speculative Statistics

Most speculative statistics are based on observational studies – and that is the basis for the problem. Associations obtained from observational studies are readily influenced by confounders – related factors that were not taken into account in the study design.

One way of taking into account the influence of a related factor involves multivariate regression. When the outcome and confounder are continuous this involves checking model diagnostics and validating the assumptions.

Schield (2006) demonstrated a simple graphical technique for a binary predictor and a binary confounder that bypasses the need to discuss the assumptions of linear regression and the underlying mathematics. This graphical technique involves weighted averages and uses a statistical principle from the 1960s called 'standardizing.' By using a binary predictor and a binary confounder, this model has no need for checking model diagnostics and validating assumptions.

The following exercise uses this technique to show the influence of a confounder on three things: (#1) the size of an association, (#2) the number of cases attributable to a related factor, and (#3) the statistical significance of a difference between two groups.

5.1. Influence of a Confounder on an Association

To see the influence of a confounder on an association, consider two hospitals: Rural and City. Patients in good condition can walk in; patients in poor condition are carried in. Suppose the death rates (hypothetical) are 2% and 7% for those in good and poor condition at the Rural hospital; 1% and 6% respectively at the City hospital. Suppose that 90% of City patients (30% of Rural patients) are in poor condition.

What are the average deaths rates at these two hospitals? 5.5% at City; 3.5% at Rural. This can be obtained algebraically as a weighted average: City (0.9*0.06 + 0.1*0.01), Rural (0.3*0.07 + 0.7*0.02). Figure 3 shows how this weighted average can be solved graphically for each hospital.

[4] There is a subtle difference between "*due to*" and "*because of*". "*Due to*" originated as an adjective; "*because of*" originated an adverb. Adjectives modify nouns and adjectives; adverbs usually modify verbs. Gibson would say, "His defeat was due to X" or "He was defeated because of X". Gibson would not say, "His defeat was because of X" or "He was defeated due to X." Malcolm Gibson's Wonderful World of Editing. See http://web.ku.edu/~edit/because.html

Figure 3 Raw Hospital Death Rates

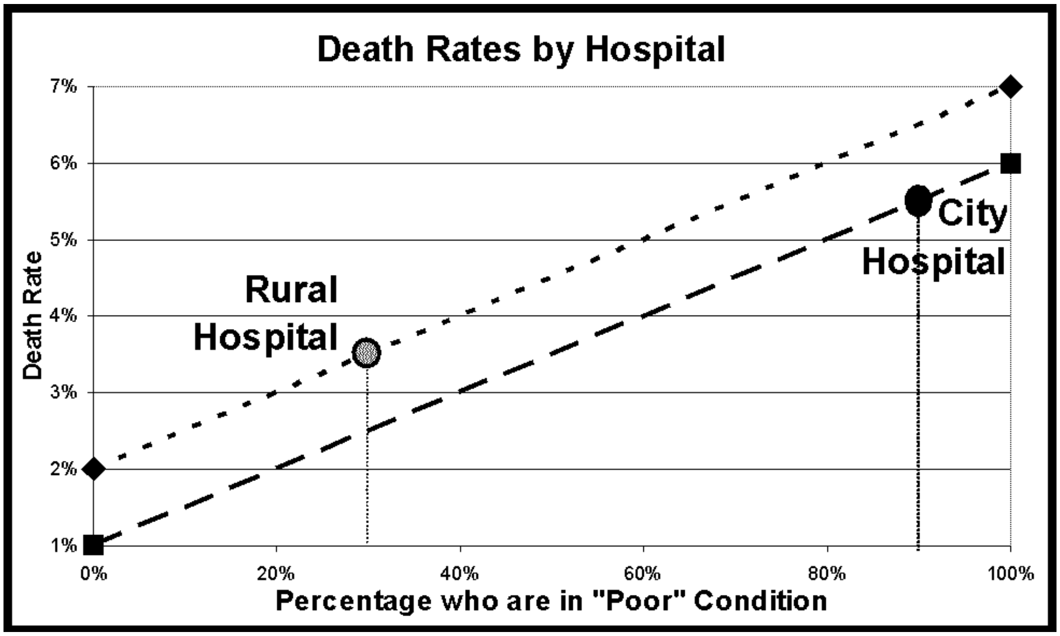
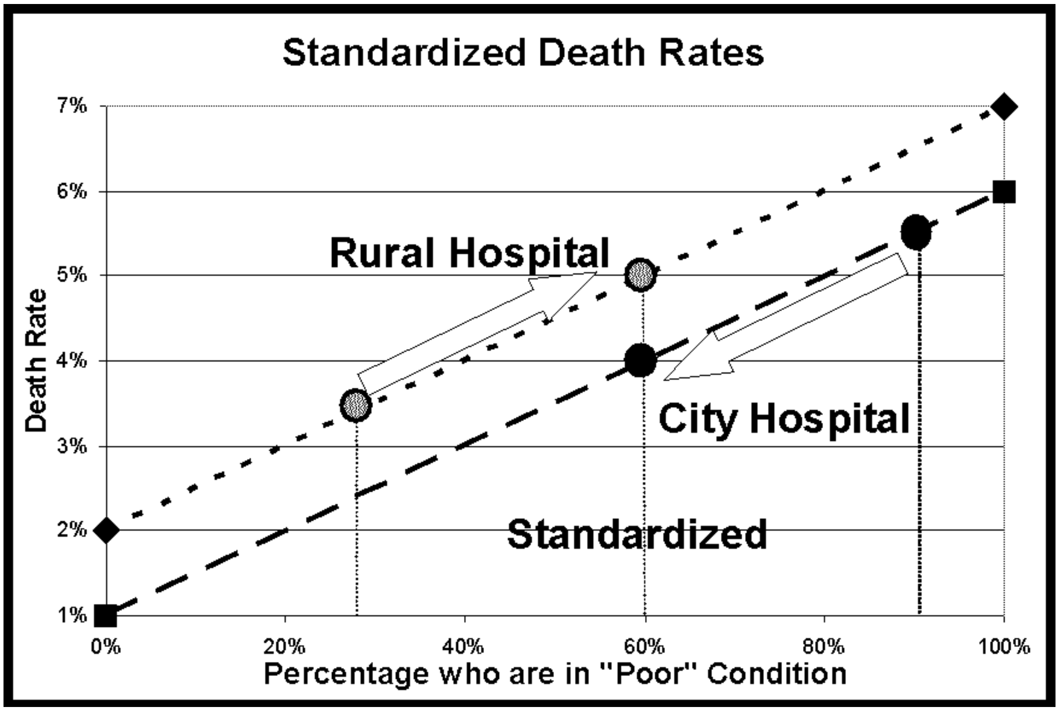


Figure 4 Standardized Hospital Death Rates



Which hospital has the higher death rate after taking into account the difference in patient mix? Figure 4 illustrates standardizing: taking into account a difference in mix. Suppose the combined hospitals have 60% of their patients in poor conditions. If we standardize on 60%, we see that Rural has a higher standardized deaths rate (5%) than City (4%). This reversal is an example of Simpson’s Paradox. This simple graphical technique allows students to work problems and to calculate the influence of a binary confounder on an association in a way that they can visualize so the result can make sense to them without using any difficult mathematics.

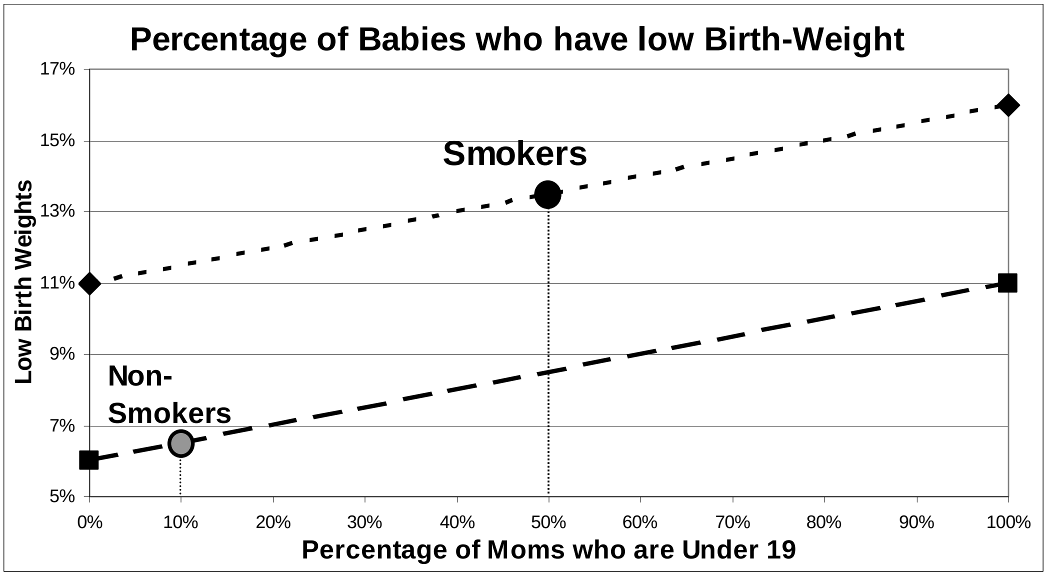
5.2. Influence of a Confounder on Speculative Statistics

Confounding can influence speculative statistics based on epidemiological models.

Suppose that among mothers who don't smoke the percentage of their babies who have low birth weights is 6% and 11% among older and younger mothers; 11% and 16% among those who do smoke. Suppose that those under 19 are 10% of non-smokers and 50% of the smokers.

Among non-smoking moms, what percentage of babies have low birth-weights? This problem in weighted averages can be solved algebraically. Among non-smoking moms: 0.10*0.11 + 0.90*0.06 = 0.065. Among smoking moms: 0.5*0.16 + 0.5*0.11 = 0.135. Or it can be solved graphically as shown in Figure 5. Either way, the percentage is 6.5% among non-smoking moms and 13.5% among moms who smoke.

Figure 5 Influence of Confounding on Percentages and Cases Attributed

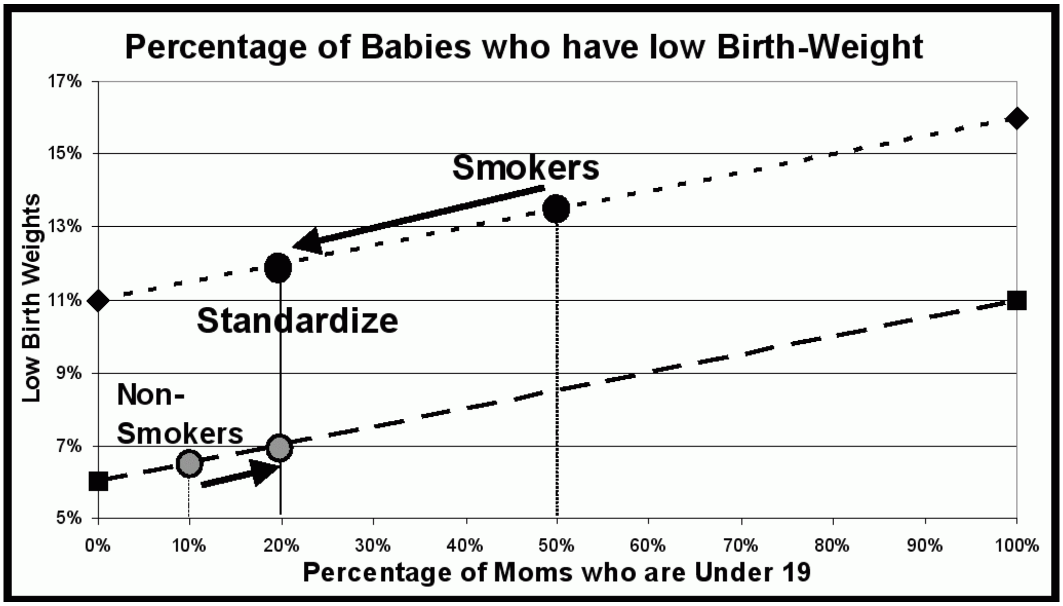


What percentage of low birth weight babies with moms who smoke are attributable to their mom smoking? Answer: 52%: (13.5% - 6.5%) / 13.5%. See prior discussion on excess risk.

How many babies having low birth-weights are attributable to their mother smoking? Assume there were 3.5 million births. Assume 25% of these mothers smoked. Of the 875,000 babies whose mothers smoked, 13.5% (118,125) have low birth weights. Of these 118,125 low birth-weight babies whose mothers smoked, 52% (61,250) are attributable to their mother smoking. Answer: 61,250.

After taking into account the influence of age, what are the standardized percentages of babies who have low birth weights? Assume that 20% of all moms are 19 or younger. Answer: The standardized percentage of babies who have low birth weight is 7.0% among non-smoking moms, 12.0% among moms who smoke. Algebraically: 0.2*0.11 + 0.8*0.6 = 7%; 0.2*0.16 + 0.8*0.11 = 12%. Figure 6 illustrates these standardized values graphically.

Figure 6 Influence of Confounding on Percentages and Cases Attributed

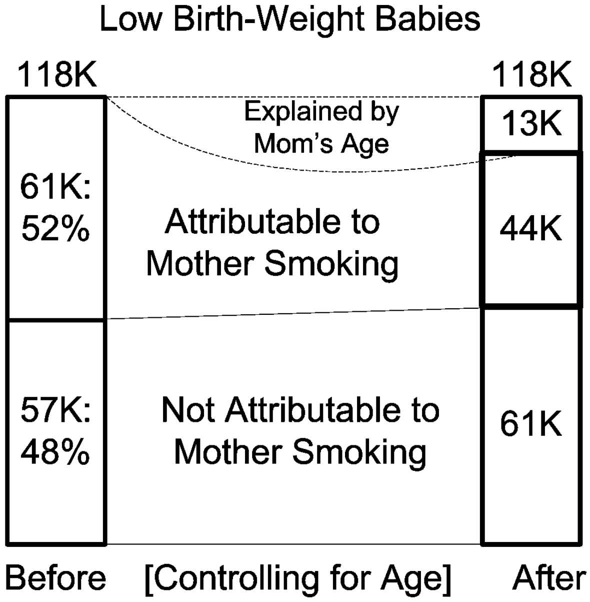


Using the standardized values what percentage of low birth weight babies whose moms smoke are attributable to their mom being a smoker? Answer: 42%: (12% - 7%) / 12%.

After taking into account the influence of age, how many babies having low birth-weights are attributable to their mother smoking? Assume 3.5 million. Assume 25% of these mothers smoked. Of the 875,000 babies whose mothers smoked, 12% (105,000) have low birth weights. Of these 105,000 low birth-weight babies whose mothers smoked, 42% (43,750) are attributable to their mother smoking: Answer: 43,750.

Compare the number of babies who have low birth weights that are attributable to their mother smoking before and after taking into account the influence of age. In both cases, there were 875,000 babies whose mothers smoked.

Figure 7 Low Birth-Weights Attributed to Smoking: Influence of Age



- Without taking into account age, 118,125 had low birth weights. Of these 61,250 (52%) were attributable to their mother smoking.
- After taking into account age, 105,000 had low birth weights. Of these 43,750 (42%) were attributable to their mother smoking.

Analyze the difference in these two cases. Taking into account age reduced the number of low-weight births attributable to smoking by almost 30% – from 61,250 to 43,750. Figure 7 illustrates these differences.

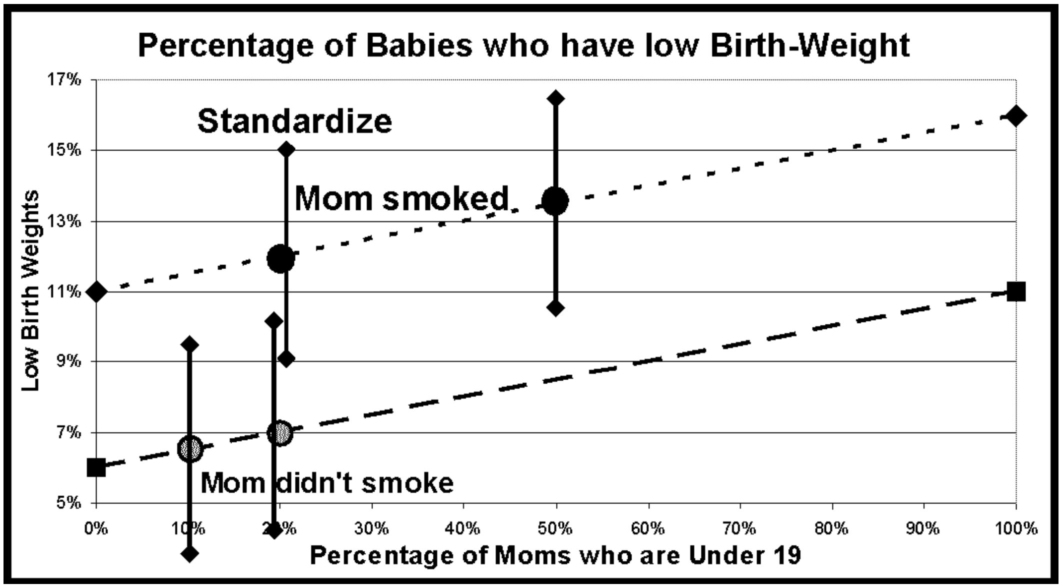
This figure preserves the original number of cases. Statistical educators can decide how best to make and explain these comparisons. Controlling for age decreased the number of low birth-weight births attributed to smoking from 61K to 44K – a reduction of almost 30%.

5.3. Influence of a Confounder on Speculative Statistics

Controlling for a confounder can influence whether a difference that is statistically significant becomes statistically insignificant – or vice versa.

Are these differences statistically significant if the 95% margin of error is three percentage points? Yes. Using the gap between confidence intervals as a simple – but conservative – test for statistical significance, the initial seven point gap (13.5% vs. 6.5%) is statistically significant. See Figure 8.

Figure 8 Influence of Confounding on Statistical Significance



Are these differences statistically significant after controlling for the influence of age? No. The standardized gap is five points: (12% vs. 7%). Using the simple overlapping confounder intervals test, this difference is not statistically significant.

6. Discussion

Statistical educators may have several concerns about this emphasis on speculative statistics and their vulnerability to confounding.

Statistical educators may worry that by focusing on how vulnerable observationally-based statistics are to the influence of confounders, they will generate students that distrust all statistics. Certainly college faculty who appreciate the value of critical thinking are not keen to teach students the values of nihlism. But, is the alternative acceptable? Can statistical educators justify their silence knowing that they promote their students mistaken idea that these speculative statistics – even when statistically-significant – are invulnerable to confounding?

A recent study found that children who were spanked had lower IQs than children who weren't – and the difference was statistically significant. Would we want our students to conclude this finding was strongly upheld because the IQ difference between these two groups was statistically significant? The issue isn't whether the difference is important; the issue is whether the difference is real or spurious. Only by seeing how a confounder might be responsible for some – if not all – of this difference, will students have a way to read and interpret these kinds of studies.

Statistical educators may worry that by using the lack of overlapping confidence intervals as the test for statistical significance, they will be promoting bad practice, since more sophisticated tests of significance are available. But as statisticians well know, there are "always" more sophisticated tests for statistical significance.

Moore (1998) focused on the needs of individuals in different roles to distinguish statistical literacy from statistical competence. "What is statistical literacy, what every educated person should know? What is statistical competence, roughly the content of a first course for those who must deal with data in their work?"

Moore (1998) thought that statistical literacy should involve two clusters of 'big ideas': 1) 'The omnipresence of variation, conclusions are uncertain, avoid inference from short-run irregularity, [and] avoid inference from coincidence.' 2) 'Beware the lurking variable, association is not causation, where did the data come from? [and] observation versus experiment.'

If statistical educators are going to uphold the claim that *association is not causation* as one of the core ideas of introductory statistics, they must be willing to point out the influence of confounding on associations and on the statistical significance based on those associations.

7. Summary

This paper presents a way of teaching *multivariate thinking* using a simple graphical model that allows students to work problems that illustrate what are arguably some of the top most important concepts in introductory statistics: the influence of confounding on an association (5.1), the influence of confounding on speculative statistics (5.2) and the influence of confounding on statistical significance (5.3)

With these new graphical tools, statistical educators now have a real opportunity. They can decide to include these three big ideas in the introductory statistics inference course or they can follow Moore's advice and field a separate course in statistical literacy. Either way they will be upholding Moore's call to focus more on the big ideas of statistics.

8. Recommendations

Statistical educators should support the need for a separate statistical literacy course that focuses on these big ideas and on the influence of Joel Best's social construction: the influence on counts, measures and comparisons of how groups are defined and how quantities are measured.

In his plenary address at USCOTS, Rossman (2007) noted, "You simply can't achieve these [GAISE statistical literacy] goals in one course if you also teach a long list of methods." He suggested that "Most students would be better served by a Stat 100 [statistical literacy] course than a Stat 101 [methods] course."

Acknowledgments

To the W. M. Keck Foundation for their grant "to support the development of statistical literacy as an interdisciplinary curriculum in the liberal arts." To the late Howard Eppler, a dairy farmer near Strawberry Point, Iowa, from whom I first heard the word *confound* used when he said to his cows, "Confound it Bessie, move over." This paper may be found by searching for *2009SchieldASA.pdf* at www.StatLit.org.

References

Danaei, Goodarz, Eric L. Ding, Dariush Mozaffarian, Ben Taylor, Jurgen Rehm, Christopher J.L. Murray and Majid Ezzati (2009). "The Preventable Causes of Death in the United States: Comparative Risk Assessment of Dietary, Lifestyle, and Metabolic Risk Factors." *PLoS Medicine*, April 28, 2009, Volume 6, Issue 4.

Flegal, Graubard, Williamson and Gail (2005). Excess Deaths Associated With Underweight, Overweight, and Obesity. JAMA, April 20, 2005; 293: 1861 - 1867.

Mokdad, Marks, Stroup and Gerberding (2004) Actual Causes of Death in the United States, 2000. JAMA, March 10, 2004; 291: 1238 - 1245.

Mokdad, Marks, Stroup, and Gerberding (2005). Correction: Actual Causes of Death in the United States, 2000. JAMA. 2005; 293(3):293-294.

Moore, D. (1998): Statistic Literacy and Statistical Competence in the 21st Century. Abstract for his talk at a 'Making Statistics More Effective in Schools of Business' (MSMESB) in Iowa City. www.StatLit.org/pdf/1998MooreMSMESB.pdf

Rossman, Allan (2007). Seven Challenges for the Undergraduate Statistics Curriculum in 2007. Slides at /www.statlit.org/PDF/2007RossmanUSCOTS6up.pdf. Handout at www.statlit.org/pdf/2007RossmanUSCOTS.pdf

Schield, Milo (2006). Presenting Confounding and Standardization Graphically. STATS Magazine, American Statistical Association. Fall 2006. pp. 14-18. Draft at www.StatLit.org/pdf/2006SchieldSTATS.pdf.