

Epidemiological Models and Spotty Statistics

Schild, Milo

Augsburg College (W. M. Keck Statistical Literacy Project)

2211 Riverside Drive

Minneapolis, MN 55454 USA

schild@augsborg.edu

Abstract

Epidemiological models generate statistics that look like ordinary counts and percents of real things. But these statistics are anything but ordinary counts and percents. These statistics are generated by models – they are speculative statistics. And like all model-generated numbers, they depend critically on the assumptions involved – the choice of what to take into account. This paper reviews how these statistics are generated and shows how sensitive they are when taking into account confounders. These speculative statistics are described as “spotty statistics” for the same reason we might say a worker’s resume with big gaps between jobs is a “spotty work record.” We are not saying that anything is necessarily false. We are saying that what is not shown may be extremely relevant to the situation at hand. In the case of spotty statistics, their sensitivity to assumptions or confounders means they should definitely be handled with care.

Introduction

The focus of this paper is the presence of model-based statistics in the everyday media. Here are some examples involving linear models:

- “the rate of [Alzheimer’s] decline unfolded 4 percent more quickly for each additional year of education.” Reuters 10/22/2007.
- “For every can or glass of sugar-sweetened beverage a child drank [a day] ..., a child’s ... chance of becoming obese increased 60%.” The Lancet, 2001; 357:505-508.
- “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.” Science Daily 4/6/2004.

Statistics generated by epidemiological models appear even more frequently. Here are some examples.

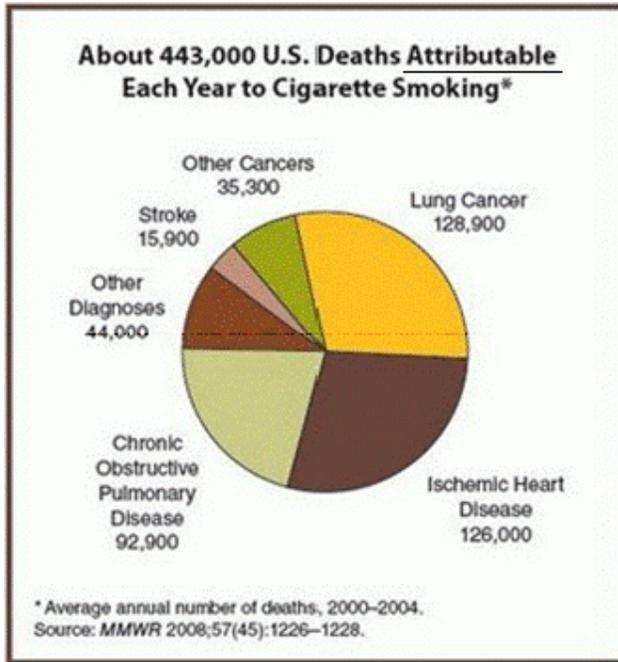
Figure 1: Smoking-Related Deaths



A casual observer might think that these five-million deaths represent real deaths – deaths that are certifiable by a coroner. All deaths are coroner certified. So what is speculative about this statistic? It is the unstated claim that of all the actual deaths, 5 million of them are caused by cigarette smoking. That is the speculative claim, the model-based claim.

Figure 2 illustrates more of these speculative statistics.

Figure 2: Smoking-Related Deaths



Again, a coroner has certified various causes of death as shown on each slice of this pie chart. But a coroner never certifies such deaths are caused by cigarette smoking. These, too, are model-based statistics. So how are they generated?

The arithmetic involved is elementary. Take any situation in which you have individual subjects with a binary outcome (say died or survived) and where these subjects can be formed into two subgroups (say smokers and non-smokers). Simply obtain the death rate for each group. Call the death rate among those exposed to cigarette smoking the RateExposed. Call the death rate among those in the control group (the non-smokers) the RateControl. With just those two rates, use Equation 1 to generate the exact percentage of the deaths in the exposed group that are attributable to being a member of that group:

Equation 1: Percentage *Attributed* to Exposure = 100% * (RateExposed – RateControl)/RateExposed

To understand this formula better, consider this “toy” data involving cancer deaths. Note that 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). See Schield (2009).

Table 1: Lung-Cancer Deaths among Smokers

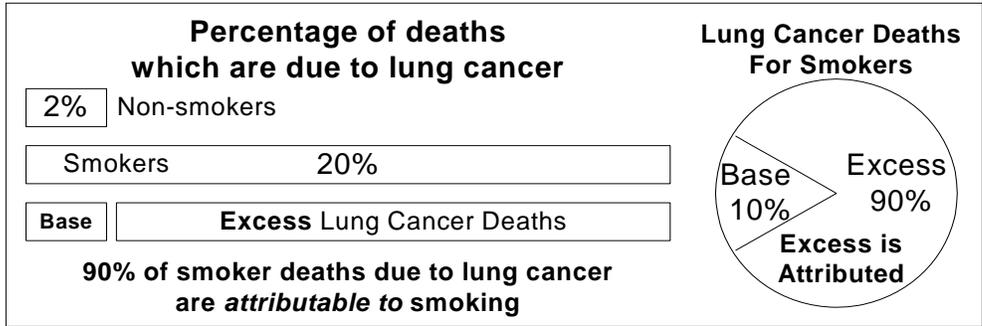
	Cause of Death		TOTAL
	Other	Lung Cancer	
Non-Smoker	784	16	800
Smoker	160	40	200
TOTAL	944	56	1000

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 the excess – the 90% of the lung-cancer deaths among smokers – are attributable to smoking where 90% = 100%*(20% - 2%)/20%.

This procedure is illustrated in Figure 3.

Figure 3: Percentage of Lung Cancer 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. One does not even need the two percentages – just their ratio: the relative risk. So if lung-cancer deaths are 10 times as prevalent for smokers as for non-smokers, then **90%** of smoker lung-cancer deaths are *attributable* to smoking.

Now suppose the *coroner-certified* lung-cancer deaths among smokers number **100,000**. The lung-cancer deaths among smokers *attributable* to smoking number 90,000: **90% of 100,000**.

This simple arithmetic allows one to compute statistics on almost any subject at any level. Consider deaths attributable to “diesel pollution.” As long as the death rate of those living closer to highways or truck stops is greater than that of those living further away, we can calculate the percentage of deaths attributable to “diesel pollution” among the former and the associated number of deaths.

Figure 4: Diesel-Related Deaths

Study blames diesel for deaths
 By Jon Brodtkin / Daily News Staff
 Wednesday, February 23, 2005

Diesel pollution is responsible for more deaths than drunk drivers and homicides, according to a new study that estimates how many premature deaths, asthma attacks and heart attacks are caused by diesel pollution in every U.S. county.

Nationwide, diesel pollution causes 21,000 premature deaths each year, including 475 in Massachusetts and 81 in Middlesex County, robbing those who die of an average of 14 years of their lives,

Isn't this incredible? The researchers who did this study know exactly how many premature deaths diesel pollution is responsible for nationwide (21,000), in Massachusetts (475) and in Middlesex county (81). A coroner counted the deaths, but the fraction of those attributable to diesel pollution is totally model-based.

Death counts seem reliable – especially when we say they are coroner-certified. But these counts of diesel-related deaths are not coroner-certified. These are statistical deaths. Yes, a real person actually died, but the cause of death is speculative. Asking whether these deaths are coroner-certified is an ambiguous question. The deaths are real. But of these real deaths, the number that is attributable to diesel pollution” is totally speculative.

Other Examples of Deaths Involving Speculative Statistics.

This simple arithmetic procedure can be used to calculate the number of deaths attributed to a wide variety of related factors as shown in Table 2. See Danaei et al (2009).

Table 2: 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

www.emaxhealth.com/2/24/30740/smoking-high-blood-pressure-obesity-top-preventable-death-causes.html

Examples of Speculative Statistics on the Web.

Speculative statistics don't generally have a unique grammatical identifier. The simplest case involves deaths. Here are web statistics on hits for various phrases involving death.

Table 3: Google Matches with various Search Phrases

Matches	WEB SEARCH	Matches	WEB SEARCH
7 M	<i>deaths associated with</i>	591 K	deaths attributed
87 K	deaths connected with	246 K	deaths attributable
6 M	deaths linked to	65 K	attributable deaths
39 M	deaths due to	384 K	unnecessary deaths
5 M	deaths because of	363 K	excess deaths
28 M	<i>deaths caused by</i>	1.2 M	premature deaths

These statistics were obtained on April, 2011 using an "exact phrase" match.

On the left side, notice the clear distinction between "deaths associated with" at the top and "deaths caused by" at the bottom. Next to "deaths associated with" we have "deaths connected with", "deaths linked to". These are generally seen as synonyms for "associated with". Next to "death caused by" we have "deaths because of." While *cause* and *because* may differ when dealing with people, they certainly share causation as a common factor. The largest group in between "associated with" and "caused by" is "deaths due to". *Due to* is a nicely ambiguous phrase. Here is where the deaths on the right side might appear.

On the right side, the first three phrases clearly indicate their epidemiological origin by the presence of "attributed" or "attributable." The last three are much more common, however. Reflect for a moment. When is a death *unnecessary*, *excess* or *premature*? With the exception of executing criminals, all deaths could be regarded as *unnecessary*, *excess* or *premature*. So what do these phrases really mean?

Unnecessary, *excess* and *premature* all indicate model-based statistics as compared with some other group having a lower rate. These phrases are code for speculative statistics.

Epidemiological statistics are easily generated

To this point, all the examples have involved deaths. But the speculative statistics are easily generated for all sorts of outcomes. All we need are two subgroups having different rates for a common outcome (e.g., two complementary wholes having a common part). Here are some examples:

Percentage attributable: percentage of college graduates attributable to having educated parents, percentage of high-school dropouts attributable to having a single-parent, and percentage of prisoners

attributable to having low IQ.

Cases attributable: Premature babies attributable to mom being a smoker, auto accidents attributable to having DWI conviction, and deaths attributable to living in US vs. Mexico

Here are some examples from the everyday media.

Figure 5: Junk-Food Related Deaths [Emphasis added]

The screenshot shows a Yahoo! News article from Reuters. The headline is "Junk food causes a third of heart attacks". The sub-headline reads: "WASHINGTON (Reuters) – Diets heavy in fried foods, salty snacks and meat account for about 35 percent of heart attacks globally, researchers reported on Monday." The article includes a photo of french fries and text stating: "Their study of 52 countries showed that people who ate a 'Western' diet based on meat, eggs and junk food were more likely to have heart attacks, while those who ate more fruits and vegetables had a lower risk."

Notice the move from “accounts for” in the text to “causes” in the title. These deaths are speculative; no coroner could determine that a heart-attack death was caused by junk food. Consider another example:

Figure 6: Smoking-Related Deaths [Emphasis added]

The screenshot shows a USA Today article from HealthDay. The headline is "Coffee, sex, smog can all trigger heart attack, study finds". The sub-headline reads: "A major analysis of data on potential triggers for heart attacks finds that many of the substances and activities Americans indulge in every day — coffee, alcohol, sex, even breathing — can all help spur an attack." The article includes a photo of a latte and text stating: "Because so many people are exposed to dirty air, air pollution while stuck in traffic topped the list of potential heart attack triggers, with the researchers pegging 7.4% of heart attacks to roadway smog. But coffee was also linked to 5% of attacks, booze to another 5%, and pot smoking to just under 1%, the European researchers found."

Again notice the move from “linked to” at the bottom to “triggers” in the earlier paragraphs and the title. Notice also the modifier *can*. Using this modal helps “pull the punch” in ways that are not clear. The *can* could refer to differences in individuals, or differences in circumstances. The *can* could also refer to the basic question of whether any of these have any real effect in anyone. This latter form is of course

the big question; the earlier forms are matters of degree in comparison.

A Harvard study (Figure 7) identified the top-three preventable causes of death based on an epidemiological model.

Figure 7: Top Three Preventable Causes of Death in the US

**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*.

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

Epidemiological statistics encourage seductive grammar

Consider these titles of news stories from Schield (2010c):

- 45,000 deaths *attributable to* uninsurance
- 45,000 deaths *associated with* lack of insurance
- Lack of insurance linked to 45,000 deaths
- 45,000 die ... *because of* lack of health insurance
- Lack of Health Insurance *Kills* 45,000 a Year
- Lack of Health Insurance *cause* 44789 deaths
- Lack of insurance *to blame* for almost 45,000 deaths

All of these statistics are based on the same research report which reported the results using phrases shown in the first three lines. But in the headlines after the first three, you can see how news reporters and headline writers added some punch to the story by using phrases like “because of”, “kills” and “cause”. The final extension is to add a moral element with the phrase “to blame for”. Yet, none of these four extensions is justified given the original statement of association.

Epidemiological statistics are sensitive statistics

Statistics generated by epidemiological models are extremely sensitive to the choice of assumptions. Here is one example of their sensitivity. In 2004, the US Center for Disease Control (CDC) concluded that there were 400,000 deaths attributable to obesity in the US. They concluded by saying “*Obesity might soon pass smoking as the country's leading cause of preventable death.*” A year later, the number was revised based on additional data and a small change in methodology. The same researchers now claimed there were 27,000 **deaths attributable to obesity**. This revision transformed obesity **from fearsome killer to pitiable also-ran**, ranked in 7th place.

So what changed? Was it the addition of one more year of data on the 25 years previously collected? Or perhaps they took something into account that is closely associated with death rates. Something such as age ...

To better illustrate this sensitivity, suppose that the rates before taking a related factor into account are identified as R_1 ; those afterward are R_2 . In both cases we have two groups: one with the higher rate R_H and one with the lower rate, R_L . The relative risk (RR) is defined as $RR = R_H/R_L$. Note that the percentage attributable is determined entirely by the relative risk:

$$\text{Equation 2: Percentage Attributed to Exposure} = 100\% * (R_H - R_L)/R_H = 1 - R_L/R_H = (1 - 1/RR)$$

As the relative risk increases, the percentage attributed to the related factor increases.

Now consider the influence of a confounder on the relative risk. Suppose that taking into account the influence of a related factor increases both the initial rates by the same amount (dR) in percentage points.

$$\text{Equation 3: } R_{2H} = R_{1H} + dR; \quad R_{2L} = R_{1L} + dR; \quad R_{2H} / R_{2L} = [R_{1H} + dR] / [R_{1L} + dR]$$

To better see the influence of controlling for a related factor, consider the ratio of these relative risks: after controlling for the factor versus before:

$$\text{Equation 4: } [R_{2H} / R_{2L}] / [R_{1H} / R_{1L}] = [1 + dR / R_{1H}] / [1 + dR / R_{1L}]$$

When dR is positive, this ratio decreases since $R_{1H} > R_{1L}$. Thus, if taking into account age on the association between obesity and death tends to increase the death rates for both obese and non-obese, we can see how this might decrease the relative risk and thus decrease the number of deaths attributable to obesity.

Now suppose that taking into account the influence of a related factor increases the lower rate by a certain percentage of the spread between high and low while decreasing the higher rate by the same amount.

$$\text{Equation 5: } R_{2H} = R_{1H} - dR; \quad R_{2L} = R_{1L} + dR; \quad R_{2H} / R_{2L} = [R_{1H} - dR] / [R_{1L} + dR]$$

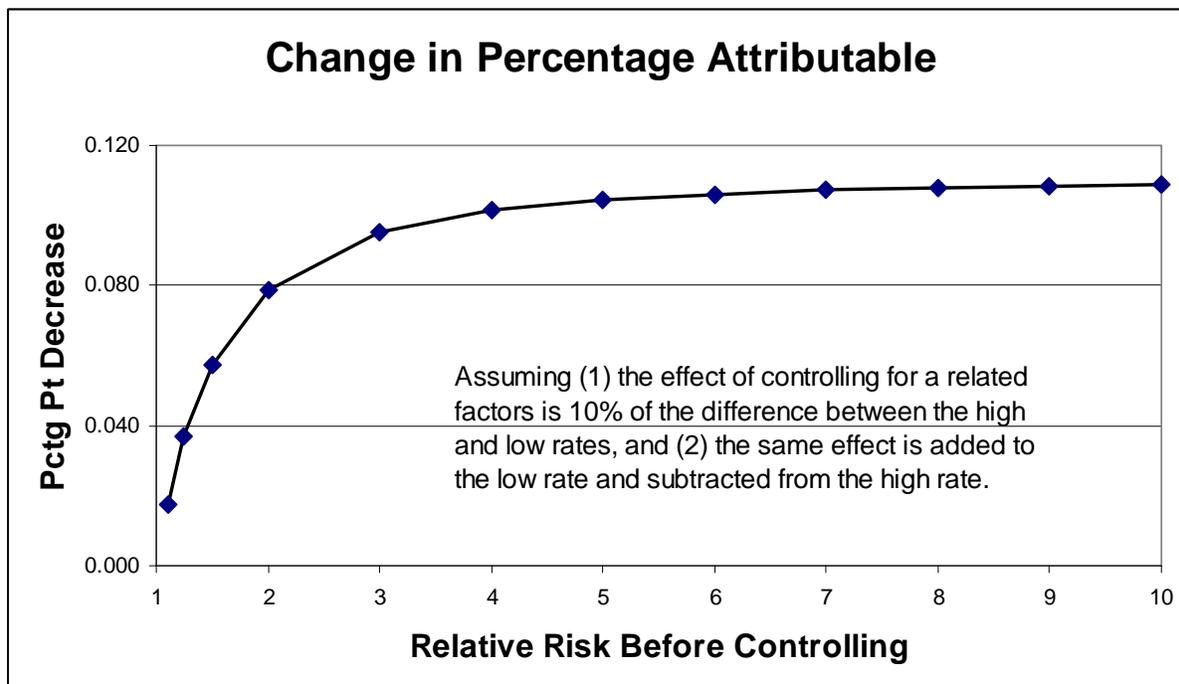
To see why this example is relevant consider the association between obesity and annual incidence of death. Assume that age is positively associated with obesity. Presumably those that are obese are above-average in age, while those that are non-obese are below average in age. If we take into account age, we need to change the composition of the obese group to have a lower average age which means a lower death rate. And we need to change the composition of the non-obese group to have a higher average age which means a higher death rate. In this particular case, we are assuming the two changes have equal magnitudes.

Now we can examine the sensitivity of the percentage attributable to a given change in rates. Suppose that the effect of confounding involves 10% of the point spread between high and low. Subtracting this from the high and adding it to the low gives results that depend solely on the initial relative risk.

For relative risks above 2 with a confounder that is positively associated with the predictor and outcome of interest, a change involving a fixed percentage of the point difference between the high and low rates that is subtracted from the higher and added to the lower will decrease the percentage attributable by at least that amount. So a change involving 10% of the point difference will decrease the percentage attributable by at least 10%; a change involving 30% of the point difference will decrease the percentage attributable by at least 30%.

Figure 8 illustrates the exact results of a 10% change:

Figure 8: Change in Percentage Attributable



Note that a 50% change may involve just a single percentage point – or even a small fraction of a percentage point -- when the outcomes are rare and the rates are low. The rarer the outcome, the easier it is for a confounder to influence the change in the percentage attributable substantially.

The main point is that a small change in assumptions can generate a big change in the statistics. This sensitivity to assumptions is true of all model-based statistics. But the sensitivity to the assumptions of what to take into account appears to be much bigger in these epidemiological studies than in other kinds of models. Perhaps because the results of the epidemiological studies involve differences or ratios of rates. If the goal were simply to model the size of a single rate under various conditions, the influence of errors on the result wouldn't be nearly as big as when modeling a difference or ratio of rates.

Epidemiological statistics are spotty statistics

We need a name that clearly identifies the nature of these statistics. Calling them “model-based” is true, but this phrase doesn't capture what essential about being based on an epidemiological model. Calling them “epidemiological statistics” is more fundamental, but identifying the source doesn't give any indication of how trustworthy they are. Calling them “speculative statistics” is better. It indicates they are model based and it indicates these statistics are disputable. See Schield (2009).

My wife, Cynthia and I were talking about this while we were having wine at dinner in Malahide, Ireland at the 2011 IASE conference. She thought all of these phrases had too many syllables. She argued that if consumers of statistics are to have any hope of recalling the phrase it should be short: one or two syllables. After several tries, she coined the phrase “spotty statistics” based on her work in Human Resources with spotty work records.

I used this phrase in my talk at ISI. I noticed at least a half-dozen to a dozen of the 100 in attendance nodding their heads “yes” at the usage. I asked several colleagues about this phrase afterwards. They liked the fact that the word *spotty* didn't claim the numbers were false and were generally supportive.

Regardless of the name, these epidemiologically-based statistics are common – but hidden – in the news. They have no unique grammar or keywords. They look plausible – coroners might count them. We treat counts as facts. Journalists and politicians don't question them. And, more importantly, we don't question them. Spotty statistics are hard to detect and they leave no clue as to their model-based

assumptions.

Statistical education is reluctant to focus on the process of obtaining statistics

In spring 2011, I surveyed a group of statistical educators in Minnesota on whether statistical education should focus more on context (what is – and is not – taken into account) or assembly (the process by which statistics are defined). As to whether statistical education should focus more on context, 6 agreed, 3 were neutral and 2 disagreed (no one strongly agreed; no one strongly disagreed). As to whether statistical education should focus more on assembly, 3 agreed, 3 were neutral, 3 disagreed and 2 strongly disagreed (no one strongly agreed).

Notice the reluctance among statistical educators to increase the focus on context. One statistical educator said that increasing the focus on how sensitive statistics are to context might decrease the respect for our discipline. Others said it would decrease the time available for teaching hard topics such as statistical inference.

Note the even-stronger reluctance among statistical educators to increase the focus on assembly: a key step in the process in which statistics are collected. Why this reluctance? One explanation is historical.

“because dealing with statistics as numbers seems to be part of nearly everyone's work, we are seen as special problem solvers instead of as prime mover” “Scientists who want to have their work be objective and true separate from the political context want to follow what once was the motto of the Royal Statistical Society, whose English translation is ‘Let others thresh it out,’ and that is why their necktie has as its emblem a sheaf of wheat. The idea was that statisticians would get the proper data and then others would decide what to do in the light of the information. I believe that the British later figured out that when others thresh it out the results may not be so satisfactory, because people who do not know what the numbers mean and do not mean may not be able to use them as well as those who do.” Mosteller (1988).

I agree with Mosteller that statisticians and statistical educators are not subject-matter experts. They should not presume to know which measures, comparisons and displays are best. But they are not mere bookkeepers either. They can see that choices have consequences. They can point out the consequences of choices and note the sensitivity of conclusions to these choices.

Statistical education should focus more on spotty statistics

Statistical educators should try to introduce these spotty statistics in the traditional statistical inference course. As more students come to college having studied descriptive statistics in school, statistical educators may be able to spend less time on descriptive statistics leaving more time for special topics such as confounding and spotty statistics. By remaining silent on these “spotty” statistics, a most common type of model-based statistics in the everyday media, traditional introductory statistics may be seen by outsiders as an archaic subject stuck in the 20th century fields of Fisher’s agricultural station at Rothamsted.

If there is no room for this topic in the introductory statistical inference course or no desire to include it, then it should be included in separate statistical literacy courses. **Statistical literacy** is the ability to read and interpret summary statistics in the everyday media: in graphs, tables, statements, surveys and studies. See Schield (2010b).

Statistical educators have been resolutely silent on whether to support a separate course in statistical literacy. It is as though statistical educators are afraid of leaving the safety of mathematics and having to really deal with the influence of context on real problems. As John Myles White (2010) said in his blog titled “Three-quarter truth”: “no idea has stifled the growth of statistical literacy as much as the endless repetition of the words correlation is not causation. This phrase seems to be primarily used to suppress intellectual inquiry”

Events may simply overtake what statistical educators at the national level prescribe. As more

teachers try to prepare their students to be able to read news stories involving statistics or to deal with current topics involving statistics, they end up creating their own course in statistical literacy. With 19% of US four-year colleges offering a course they label as *statistical literacy*, statistical education is changing. See Schield (2010a).

By including spotty statistics, statistical educators will make statistical literacy even more relevant to the new forms of statistics being generated in increasing numbers.

REFERENCES

Schild, M. (2009). Confound Those Speculative Statistics. *2009 American Statistical Association Proceedings of the Section on Statistical Education*. [CD-ROM] P. 4255-4266. See www.StatLit.org/pdf/2009SchildASA.pdf

Schild, Milo (2010a). Quantitative Graduation Requirements at US Four-Year Colleges. MAA JMM. See www.StatLit.org/pdf/2010SchildJMM.pdf

Schild, M. (2010b). "Assessing Statistical Literacy: TAKE CARE" in *Assessment Methods in Statistical Education: An International Perspective*. Edited by P. Bidgood, N. Hunt and F. Joliffe. Wiley Publishers, Ch. 11, p. 133-152. Excerpts at www.statlit.org/pdf/2010SchildExcerptsAssessingStatisticalLiteracy.pdf

Schild, Milo (2010c) Association-Causation Problems in News Stories. International Conference on Teaching Statistics (ICOTS-8) in Ljubljana, Slovenia. See www.StatLit.org/pdf/2010SchildICOTS.pdf

White, John Myles (2010). "Three-quarter truths: Correlation Is Not Causation" blog. www.johnmyleswhite.com/notebook/2010/10/01/three-quarter-truths-correlation-is-not-causation/