Statistics 999
Lecture 2
Simpson's Paradox
Robert L. Wardrop

First, thanks for the comments to and compliments for Lecture 1. Some of you made very accurate comments; thus, below I will be repeating some of what you said.

Denise Tardiff, with whom I shared a great two-hour-per-day class back in 1965–66, wins the award for the most humorous—I was going to type *funniest*, but that might have been misconstrued—response. There is no prize—cash or otherwise—for Denise because I don't want to encourage undue silliness, although humor is not to be discouraged.

I am determined to keep these lectures short, so let's get to it.

The medical doctor/researcher who gave the lecture on vasectomies was quite certain and forceful in his statement that it makes no sense to believe that a little clip and tie in a remote part of a man's body could possibly have a positive influence on any, let alone all, measures of health. Thus, he continued, the vasectomy study is a great example of the influence of *background* or *lurking* variables. (Full disclosure: I am far from an expert on the origins of terminology in my field, so I want to mention that I was introduced to the more colorful term *lurking* by my creative and brilliant friend and colleague Tom Leonard who read Lecture 1 of this series and, in case he continues to read these lectures, I am hoping that these compliments, while accurate and sincere, will help him decide to be gentle when he notes errors that I undoubtedly will make!)

The broad idea is that we have a collection of people we want to study. Primary focus is on the **response** we will obtain from each person. Typically in serious research there are many responses, but in this lecture I will consider a single response. The response could be a number—how long somebody lives, for example—or a dichotomy—a yes/no situation, such as *Is a disease present?*. To keep this manageable I will focus on dichotomous responses although the results I present have direct analogies for a numerical response.

The researcher decides to divide the collection of people into two groups, in the current study the collection are men and the two groups correspond to whether or not a man has had a vasectomy. Note it is possible to divide the collection into many groups, but again for convenience, I will focus on two groups. We have two dichotomies floating around here so don't confuse them! There are two groups being compared and the response of interest is a dichotomy, yes or no.

This brings us to the major point, a point which, in my experience, many people overlook.

The groups the researcher chooses to compare are not necessarily meaningful with regards to the response. And certainly they are not unique. For example, we could divide men into groups based on any number of features: age, race, marital status, socio-economic status, occupation and so on.

The vasectomy researcher examined the entirety of his data and found that, compared to men without vasectomies, the sterilized men were wealthier, had more access to better health care and so on. Thus, to be totally clear about this, the vasectomy group might be healthier because of better medical care and not because of the surgery.

I now want to give you some numbers to illustrate these ideas. The numbers are totally hypothetical. In the next lecture I will give you some real data.

Table 1: Hypothetical Observational Data.

	Released?			
Sex	Yes	No	Total	$\hat{p}$
Female	60	40	100	0.60
Male	40	60	100	0.40
Total	100	100	200	

Years ago I worked as an expert witness in several cases of workplace discrimination. As a result of this work, I was invited to make a very brief presentation at a continuing education workshop for State of Wisconsin administrative judges. (In Wisconsin, the norm was (is?) to have workplace discrimination cases settled administratively rather than by a jury of citizens.) Below I am going to show you what I presented in my 10 minutes.

These are totally and extremely hypothetical data. A company with 200 employees decides it must reduce its work force by one-half. Table 1 reveals the relationship between sex and outcome. To be precise, let me note that the response of interest is *Was the employee released?*, the collection were employees of the company and the groups are women and men.

The table shows that the proportion of women who were released was 20 percentage points larger than the proportion of men who were released. I analyzed data like these for lawyers—I worked only for plaintiffs—and was careful not to say something like, "Good news for our female client, these data show that the company discriminated against women." Two big reasons not to say this:

1. **Legal:** *Discrimination* is a legal term and lawyers do not want statisticians to make statements about law.

2. **Scientific:** See below.

Instead, I would say, "Show these data to the defense and see what they say." A good defense team would look for a lurking variable as I demonstrate now.

Now, if I worked for the defense I would think, "What are some valid reasons for releasing an employee? Several answers come to mind. The job classification might be relevant; the years of seniority of the worker; and, of course, quality of work performance, although this last one is more difficult to measure. For the sake of this example, let's suppose that the defense settles on the lurking variable of job classification and again, for simplicity only, let's assume that there are exactly two job classification for the 200 employees.

My first possibility is shown in Table 2; it shows that bringing job into the analysis might have no effect whatsoever. The proportions in each sex and each job match exactly what we had in Table 1. Henceforth we will refer to our original table as the collapsed table and tables such as the two in Table 2 as the component tables. Before we continue note that while these are hypothetical data, they must be **consistent** in a way I will now describe. When we break down the data in a collapsed table into two component tables, data are neither created nor destroyed. In the collapsed

Table 2: Hypothetical Observational Data with Background Factor: **Case 1**.

	Job A					**Job B**			
	Released?					Released?			
Sex	Yes	No	Total	$\hat{p}$	Sex	Yes	No	Total	$\hat{p}$
Female	30	20	50	0.60	Female	30	20	50	0.60
Male	20	30	50	0.40	Male	20	30	50	0.40
Total	50	50	100			50	50	100	

Table 3: Hypothetical Observational Data with Background Factor: **Case 2**.

	Job A					**Job B**			
	Released?					Released?			
Sex	Yes	No	Total	$\hat{p}$	Sex	Yes	No	Total	$\hat{p}$
Female	30	10	40	0.75	Female	30	30	60	0.50
Male	30	30	60	0.50	Male	10	30	40	0.25
Total	60	40	100			40	60	100	

table there are 100 females and 100 males. In the component tables the 50 women in Job A added to the 50 women in Job B yield the same 100 women. And so on for the men.

Also, there are 30 women terminated in Job A and 30 in Job B giving the totsl of 60 in the collapsed table. And so on.

My next possibility is in Table 3. In this Case 2 we find that job does matter and it matters in the sense that women are doing even worse in both jobs than they are doing in the collapsed table: the difference is up from 20 to 25 percentage points.

Our next possibility, Case 3 in Table 4 shows that if we incorporate job into the description, the difference between the experiences of the sexes can disappear.

Finally, Case 4 in Table 5 shows that if we incorporate job into the description, the difference between the experiences of the sexes can **be reversed**! This reversal is called Simpson's Paradox.

When I made the above presentation at the aforementioned workshop the participants got really

Table 4: Hypothetical Observational Data with Background Factor: **Case 3**.

	Job A					**Job B**			
	Released?					Released?			
Sex	Yes	No	Total	$\hat{p}$	Sex	Yes	No	Total	$\hat{p}$
Female	60	15	75	0.80	Female	0	25	25	0.00
Male	40	10	50	0.80	Male	0	50	50	0.00
Total	100	25	125			0	75	75	

Table 5: Hypothetical Observational Data with Background Factor: **Case 4: Simpson's Paradox**.

	Job A					**Job B**			
	Released?					Released?			
Sex	Yes	No	Total	$\hat{p}$	Sex	Yes	No	Total	$\hat{p}$
Female	56	24	80	0.70	Female	4	16	20	0.20
Male	16	4	20	0.80	Male	24	56	80	0.30
Total	72	28	100			28	72	100	

excited and wanted to know which of the four cases were true in my study. I reminded them that this was all hypothetical.

To this response they asked, "Well, which case is most likely?" My reply was to the effect, "I have no idea. I don't know anything about workplaces. All I wanted to show you were a few possibilities." I went on to explain that whenever you see a *collapsed table* you should be aware that the *message in the data could change with further analysis*. This might be frustrating, but I am a big believer that truth trumps convenience.

My last comment. It seems to me that if one takes a general education course in Statistics then it is imperative that one be exposed to this topic; i.e., how easily a conclusion can change. Sadly, I have seen only two books—other than my two—that even mention Simpson's Paradox. If this isn't bad enough, the two other examples are bad.

In particular, both books I have seen show the collapsed table and say something snarky like, "Denise looks at these data and says this isn't fair." They then show the table I have in Table 4 and state, categorically, "This shows that Denise is wrong!"

Not so fast. The component tables I show above are themselves collapsed tables awaiting some additional division into more component tables. The whole topic reminds me of Matruska Dolls (Russian Nesting Dolls).

In short, it is bad science to think that a collapsed table is the truth. It is worse science to search for component tables that you agree with and then proclaim, "Here is the real truth!"

More to come in Lecture 3 on this topic. (I don't want to wear you out.) Thanks for your attention.