THREE STATISTICAL PUZZLES

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

Preface

The puzzles in this note are meant to illustrate and elucidate two basic issues of causal inference in statistical analysis. The first puzzle, entitled "Dr. Pearson and the Hypothetical Counterfactual," concerns noncompliance. It was written to rebut the prevailing wisdom that randomized controlled experiments of the Fisherian type are the only reliable source of causal inference. My aim was (and is) to call attention to simple mathematical results [Balke & Pearl 1993] showing that significant, if not precise information can sometimes be obtained from indirect experiments, that is, experiments involving self-selection, noncompliance, or randomized encouragement.

The other two puzzles are structurally isomorphic and concern the problem of confounding bias in observational studies. "On Bottles and Drugs," written for a seminar I presented at UC Berkeley [November 1993], demonstrates the usefulness of covariates that are affected by the treatment. While textbooks warn us to avoid adjusting for such covariates, this puzzle shows that post-treatment covariates can be not only admissible but necessary for obtaining unbiased estimates of the treatments total effect. The third puzzle, "Dr. Simpson on Smoking," is extracted from [Pearl 1993] and [Pearl 1994] where a calculus is developed to facilitate the algebraic derivation of causal effect formulas.

The solutions to all three puzzles can be cast in closed form, and theories are now available (see references below) for systematic analysis of such problems, regardless of their size or complexity.

References

[Balke & Pearl 1993] Balke, A. and Pearl, J., "Nonparametric Bounds on Causal Effects from Partial Compliance Data," UCLA Computer Science Department, Technical Report R-199, September 1993. Submitted to *JASA*.

[Pearl 1993] Pearl, J., "Mediating Instrumental Variables," UCLA Computer Science Department, Technical Report R-210, December 1993.

[Pearl 1994] Pearl, J., "A Probabilistic Calculus of Actions," UCLA Computer Science Department, Technical Report R-212, January 1994. To appear in *Proceedings of UAI-94*.

Dr. Pearson and the Hypothetical Counterfactuals

The school board has decided to hire an expert statistician, Dr. Pearson, to examine the question of whether there is a real difference between the educational effectiveness of the two high schools in town, A and B. After in-depth discussions with all parties involved, Dr. Pearson has decided to resolve the issue by conducting a clean randomized experiment. He obtained the list of all the students about to enroll in next year's freshman classes and selected at random two groups, A and B, of 100 students each. He then went to the homes of all 200 students and persuaded their parents to enroll their children in school A or B, depending on the group the child happened to be in. Dr. Pearson then went back to his work at the university, after arranging to come back in the spring to analyze the test scores of the two groups.

When he returned, Dr. Pearson found out that something had gone terribly wrong immediately after he left town. Some kids found their assigned schools too boring, others complained about school being too hard, and still others were plain contrarians. Altogether, only 50% of those assigned to school A actually remained in school A after the first week of classes, and exactly the same had happened with those assigned to school B. Dr. Pearson was totally depressed. While it is true that the test scores showed a marked performance difference between the two groups – 50% of the students who went to school B, and none of those going school A managed to pass the state exam – he knew very well that the study was worthless by all statistical standards.

As he was preparing to quit town in disgust, his young assistant, Alex, asked to have one more look at the data. "What is there to look at?" snarled Dr. Pearson, "I can tell a classical case of non-compliance when I see one! Even when only 10% of the subjects switch groups, I have a hell of a time convincing my colleagues that the study is worth a dime. Can you imagine how they are going to react when I show them data with 50% cross over? They are going to say that exactly those students who would have performed better in school A have decided to switch over to school B and that those who switched from school B to school A were precisely the students who would not have performed well no matter where they went. You know how convoluted statisticians can be."

Surprisingly, after spending some time on the computer, Alex came back with a smile on his face: "Dr. Pearson, it wasn't a total waste after all. No matter how you slice it, school B is a clear winner over school A; barring sampling errors, it gives students a 50% greater chance of passing the state exam."

"What do you mean 'winner'? You talk like one of those computer freaks. In our profession, we do not talk about winners or losers, we talk about the data and let our clients jump to their own conclusions. Let's stick to the data, Alex."

"I am talking about the data," Alex replied. "What I mean is that if all the subjects we selected had remained in their assigned schools, the percentage of success in group B (school B) would have been exactly 50% higher than that of group A (school A)."

"You are a hopeless case, Alex. What's all this talk about 'had remained' and 'would have been'? Where the hell did you take your statistics 1 class? Next you are going to pretend the data tells you how many would pass the exam if we were to convince all our subjects to enroll in school B, right?"

"50%," said Alex.

"Here you are, a perfect oracle. Now how about if they enrolled in school A?" asked Dr. Pearson, somewhat amused.

"Hmm, they would all fail," said Alex, as he scanned the computer printout.

"This is too much. I give up," said Dr. Pearson. "All this talk about 'if we were' and 'if they would' is getting on my nerves. I am getting out of this place."

We solicit the readers' help in answering the following questions:

1. Could any model of population behavior be compatible with Alex's figures?

2. Could any data support Alex's confidence in his claims ?

3. What calculations are needed to gain such confidence?

On Bottles and Drugs

I went to a pharmacy to buy a certain drug, and I found that it was available in two different bottles: one priced at \$1, the other at \$10. I asked the druggist, "What's the difference?" and he told me, "The \$10 bottle is fresh, whereas the \$1 bottle one has been on the shelf for 3 years. But, you know, data shows that the percentage of recovery is much higher among those who bought the cheap stuff. Amazing isn't it?" I asked if the aged drug was ever tested. He said, "Yes, and this is even more amazing; 95% of the aged drug and only 5% of the fresh drug tend to lose the active ingredient, yet the percentage of recovery among those who got bad bottles, with none of the active ingredient, is still much higher than among those who got good bottles, with the active ingredient."

Before ordering a cheap bottle, it occurred to me to have a good look at the data. The data were, for each previous customer, the type of bottle purchased (aged or fresh), the concentration of the active ingredient in the bottle (high or low), and whether the customer recovered from the illness. The data perfectly confirmed the druggist's story. However, after making some additional calculations, I decided to buy the expensive bottle after all; even without testing its content, I could determine that a fresh bottle would offer the average patient a greater chance of recovery.

The druggist, who happened to be an amateur statistician, asked me if I was sure about my decision. "After all," he said, "this is merely non-randomized, non-experimental data, which we know cannot be trusted for decision making. Isn't it possible that some characteristic of the price-conscious customers (e.g., a different diet) makes them more likely to recover than the quality-minded customers?" But I replied that, based on two very reasonable assumptions, the data show clearly that the fresh drug is more effective.

The assumptions are:

1. Customers had no information about the chemical content (high or low) of the specific bottle of the drug that they were buying; their choices were influenced by price and shelf-age alone.

2. The effect of the drug on any given individual depends only on its chemical content, not on its shelf age (fresh or aged).

We solicit the readers' help in answering the following questions:

1. Is there a model of population that could support this scenario?

2. Could the effect of choosing the fresh versus the aged drug be determined from this type of data?

3. What calculations are needed to make the determination.

Dr. Simpson on Smoking

Consider the century-old debate on the relation between smoking and lung cancer. According to some, the tobacco industry has managed to stay anti-smoking legislation by promoting the theory that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype that also induces an inborn craving for nicotine.

To assess the degree to which cigarette smoking increases (or decreases) risk of lung cancer, a careful study was undertaken, in which the following factors were measured simultaneously on a randomly selected sample of 800,000 subjects:

1. amount of smoking,
2. amount of tar in the lungs, and
3. whether lung cancer has been found.

The data from this study are presented in the table below, where, for simplicity, all three variables are assumed to be binary. All numbers are given in thousands.

	TAR 400		**NO TAR 400**		**ALL SUBJECTS 800**	
	Smokers 380	Non-Smokers 20	Smokers 20	Non-Smokers 380	Smokers 400	Non-Smokers 400
Cancer	323 (85%)	1 (5%)	18 (90%)	38 (10%)	341 (85%)	39 (9.75%)
No Cancer	57	19	2	342	59	361

Two opposing interpretations have been offered for these data. The advocates of anti-smoking legislation argue that the table proves the harmful effect of smoking. They point to the fact that about 85% of the smokers have developed lung cancer, compared to only 9.75% of the non-smokers. Moreover, within each of two subgroups, tar and no tar, smokers show a much higher percentage of cancer than non-smokers.

However, the tobacco industry argues that the table tells an entirely different story – that smoking would actually decrease, not increase, one's risk of lung cancer. Their argument goes as follows: If you choose to smoke, then your chances of building up tar deposits are 95%, compared to 5% if you choose not to smoke (380/400 versus 20/400). To evaluate the effect of tar deposits, we look separately at two groups, smokers and non-smokers, as done in the table below. All numbers are given in thousands.

	SMOKERS 400		NON-SMOKERS 400		ALL SUBJECTS 800	
	Tar 380	No Tar 20	Tar 20	No Tar 380	Tar 400	No Tar 400
Cancer	323 (85%)	18 (90%)	1 (5%)	38 (10%)	324 (81%)	56 (19%)
No Cancer	57	2	19	342	76	344

We see that tar deposits have a protective effect in both groups; in smokers it lowers cancer rates from 90% to 85%, and in non-smokers it lowers cancer rates from 10% to 5%. Thus, regardless of whether I have a natural craving for nicotine, I should be seeking the protective effect of tar deposits, and smoking offers a very effective means of achieving them.

We now call upon Dr. Simpson to settle the dispute and answer our question: Should or shouldn't we smoke, Dr. Simpson?