

## CHAPTER 14

# Cheating

Copyright 2008 William M. Briggs, wmbiggs.com

This Chapter is in the spirit of and is dedicated to Darrell Huff who, in 1954, published *How to Lie With Statistics*, a wonderful book that guided generations of statistical cheaters. That book is still in print. Most of an issue of *Statistical Science* in 2005 contained homages from well-known authors on how to lie in areas which Huff had not touched on. I try not to cover the same ground as Huff or the *Stat. Sci.* authors and have angled my tips especially for those who use statistics in their academic papers. J. Michael Steele *Statistical Science* 2005, Vol. 20, No. 3, 205209

### 1. Statistics on the loose

Here is a case study to show you how easy it is to cheat with statistics.

I saw a commercial for Glad ForceFlex trash bags<sup>1</sup>, in which they said, in bold, animated letters, that “7 out of 10 *consumers*<sup>2</sup> preferred” ForceFlex (then in small small print) “over the other leading brand.” So what is the probability that a “consumer” would prefer a Glad bag? You’ll be forgiven if you said 0.7. That is exactly what the advertiser wants you to think. But it is wrong, wrong, wrong. Why? Let’s parse the phrase they used and see how you can learn to cheat from it.

The first notable comment is “over the other leading brand.” This heavily implies, but of course does not absolutely prove, that Glad commissioned a market research firm to survey “consumers” about what trash bag they preferred. The best way to do this is to ask people, “What trash bag do you prefer?” But evidently, this

---

<sup>1</sup>Viewed on Channel 11, WPIX, 19 July 2007, at 6:56 pm.

<sup>2</sup>This is one of the most idiotic terms invented by businessmen. “Hey, I just saw a *consumer* walking down the street!”

is not what happened. Here, the “consumer” was given a choice, “Would you rather have Glad? Or *this other particular* brand?” Here, we have no idea what that brand was, nor what was meant by “*leading* brand.” Do you suppose it’s possible that the advertiser gave in to temptation and chose, for his comparison bag, an inferior one? One that, in his opinion, is obviously substandard to Glad (but maybe cheaper)? It certainly *is* possible. So we already suspect that the 0.7 guess is off. But we’re not finished yet.

In tiny type at the bottom of the screen, we find these words: “Versus the other leading brand’s Tall Kitchen Drawstring trash bag” and “Among those with a preference.” So now we know that the “other leading brand” was not just some other bag, but a very specifically chosen one, just as we suspected. But how about that other bit? The phrase “Among those with a preference” should have your system announce *Red Alert!* Because it tells us that there were some people who just didn’t give a damn about trash bags, or, at least, the two trash bags presented to them. How many people? We have no idea. But we might suspect it’s a lot. Which means that the original guess of 0.7 for the implied, but false, question “What proportion of people prefer Glad”, is way off, and certainly far too large.

The commercial had wanted you to believe that the background premise was  $E =$  “7 out of 10 people prefer ForceFlex” therefore there would be a 70% chance that *you* would prefer the bag. Closer inspection showed that the evidence  $E$  was very different, such that that we can’t adequately identify the exact premises. Lesson 1 in how to cheat is obvious: conceal contrary evidence in small print, or somehow obfuscate it.

Incidentally, it is also reasonable to infer that the real evidence is such that the probability you would prefer a bag is less than 0.7 based on the premise that if the advertiser did have better evidence in his favor, he certainly would have used it. He did not, ergo, etc.

The moral of the story is: always be suspicious of other people’s statistics, especially when somebody is trying to sell you something.

## 2. Conditioning

A typical academic study is one, say, that gathers two groups of college kids, maybe about 50 in each set, and has them do some task or asks them to rate something. Another common type of study gathers data from a small area, say a neighborhood in a city, where the sample size may be as high as a few hundred, and asks sociological and economic questions of the people that live there. A medical experiment might try two treatments in two groups of a hundred or so people. When the data from these studies are in, the results are compiled and papers are published. Claims are made in these papers. The college kids paper will say that people act one way and not another; the city paper will say that poor people have less money; and the medical paper will claim treatment A is better than treatment B.

We already know that if all these researchers wanted to do was to say something about their datasets, then they do not need statistics or probability models. They can look at their data and say, yes, more people got better under treatment A than under treatment B. They would be finished. Evidently, the creators of these studies do not want to make statements only about the past data; they want to imply their findings are more widely applicable.

By far the majority of these kinds of studies, published in academic journals, concern humans. As of this writing, there are over 6.6 billion humans alive, about 100 billion are dead, and God only knows how many more are yet to live. Incidentally, whatever you do, do *not* mention these facts in your results (unless, of course, you happen to be writing about demography), it will weaken your argument.

Are the results from the college kids study applicable to all humans? All those that lived in the past, those that will live in the future, even those that live now but not in the town in which the college lie? Those who are in their 50s?, 80s? who are less than 10? Poorer people and those with enough money to “get a degree<sup>3</sup>”? Kids at other universities? Let’s be clear: the researchers will gather data on their 100 kids, create a probability model, and

---

<sup>3</sup>Kids go to college to “get a degree” nowadays, and not usually for anything else. Well, maybe socialization. These are rational choices given the way things are.

since they have read this book, they will not just make a statement about the parameters, but calculate the probability distribution of future observables. The only problem is, about whom do we apply this probability distribution?

Before we answer that, think about the medical trial, which was conducted at a hospital in a city on the East Coast of the United States of America. The physicians also use their data to create a probability distribution of future patients. But who exactly are these patients? People who live in other cities on the east coast?, anywhere in the USA? Canada, too? Or only cities of a certain size? Or do the future patients merely have to “look like” the patients in the old data; that is, be of the same ages, sex ratio, weights, economic condition, have eaten the same things in their lifetimes, traveled to the same places, engaged in the same activities, and so on. Would it have applied to the people who used to be alive, and to people not yet born, indefinitely into the future?

Nobody knows the answers to these questions, which is highly in your favor, especially if you have just completed a study using data “at hand”, that is, that was easy for you to collect. You certainly want to imply that your results are as broadly applicable as possible because this makes you more of an expert than somebody who merely claims to know the habits of a small group of college kids in the year 2008 only, in city C and who are unmarried, between 19 and 22 years old, and whose parents are upper middle class, etc. Openly stressing these limitations might be noble and correct, but it will not get you far. State your results in terms of all people. For example, say “People choose option A over B which gives weight to our theory of psychology.” Do *not* say, “College kids in our freshman psychology class, who might not be anything like the rest of the population, carried out an experiment for us—and surely they took this task seriously—and...”

Same thing in the medical trial. Emphasize your small p-value, spend more time talking about how the two groups of patients (those that received treatment A and those that got B) were not different than one another. Tell how there were roughly equal numbers of men and women in both treatments, and the same with age, weight, etc. This is an excellent strategy because it is useful information: if the two groups did differ, then your results

may be biased. Well, this is a wonderful distraction because it allows you to ignore or downplay the discussion of how your results might only be useful for a small subset of patients.

In short, be sloppy describing the nature of your sample; or, rather, say as much about your sample as you like, but say little or nothing about whom you expect your results are applicable. Certainly imply that all humanity falls under your results. With any luck, a reporter will find your paper and help you along this road by summarizing your results, leaving out all hint of limitation with his headline, “Kumquats reduce risk of toenail cancer.”

### 3. Randomization

In classical statistics, all analyzed data must possess the mysterious quality of *randomness*. This is in part reinforced by the mistake of calling data, say, “normal”, as in WBC is “normal”, when what really should be said is that “our knowledge of WBC is quantified by a normal distribution”. A lot more words, but a lot more correct and surely less apt to be misleading. Modern statistics will take data as they come. This is not to say that we should ignore any data’s provenance. This obviously becomes part of our background evidence E, as we just discussed. Data gathered under suspicious or irregular circumstances should rightly not be fully trusted.

People are savvy about randomness. They know that non-randomized trials aren’t as trustworthy as randomized ones. For example, in the medical trial, a (computerized) coin is flipped as a patient walks in the door; if it is heads, he gets treatment A, else B. Wait a minute. Randomized? What does that mean?

Data need to be “random”<sup>4</sup> to justify use of the classical theory. Modern statistics recognizes that this concept is not needed (Jaynes calls the old belief a “mind projection fallacy”). Anyway, in classical theory, and specifically in “randomized” trials, some mechanism is invoked that takes the decision out of human hands

---

<sup>4</sup>Does the piece of data itself contain this “randomness”? Do some pieces have more “randomness” than others? Can we extract it, put in a jar so to speak? Randomness, like the Jabberwock, is elusive.

about how to allocate the groups. A set of “sealed envelopes” containing “random” numbers generated by a computer says which patient goes to which group.

You must understand that computerized coin flips, even the results from real coin flips, are not “random”. The output from a “random number generator” on a computer is nothing but a deterministic sequence of numbers: if you know the starting point, you know (I mean *know*) every number in order the computer will show you. A real coin flip is constrained by the same laws of motion that were responsible for dropping the apple on Newton’s head. If we knew the initial conditions of the flip (weight, air viscosity, amount of spin), we could predict exactly what the result would be (Jaynes, *The Other Guy*). These events—computerized numbers, actual coin flips—appear “random” because we turn a blind eye to the initial conditions and to the equations that govern the outcome. We want the outcomes to be unpredictable—they are *not* unpredictable, we just act as if they are. These acts work as “randomizers” because, even if we turned our attention to the initial conditions and equations of motion, we would never have enough time to solve them before the outcome is realized.

It is solely because you cannot *trust* human beings that “random” trials are necessary. People will lie to others and to themselves, they will cheat when able (using guides like this), they will maneuver, shade, and finagle, they will engage in intrigue, they will contrive and conspire, in short, they will use every method under the sun to “help” the results work out the way they want them to. The reason nobody trusts the results of a study touted by a homeopath is not because his method of treatment is ludicrous, it is because he has not conducted a “randomized” trial. Nobody will believe that he did not pick and choose this patients so that he could get the results he wanted.

Nobody will trust real doctors or researchers either when they report results from an “observational” or non “randomized” study, not because these people would purposely lie to us, but they might lie to themselves. They might have picked data that confirmed their suspicions and not sought out data that was contrary to them. Out of sheer humanity, a physician might have let a sicker patient receive the new drug rather than the placebo, and so bias the results. Or a caring researcher might seek out data that proves

some injustice befell a select group, but not look for data that shows this same injustice is common to most groups, or that it has nothing to do with the groups as he categorized them, but does have to do with some other feature that was ignored.

Because people are so wily is the same reason nobody would trust a referee just picking one of the teams to receive the kickoff at a football game. He is forced to flip a coin to remove the suspicion that he favors one team over the other.

You can't really cheat when it comes to studies like these. Best you can do is to not mention you failed to "randomize", but you surely will be caught. Your only option will be to opt for a lesser journal. Even better would be to issue a press release, thus bypassing any criticism.

#### 4. Surveys Polls, & Questionnaires

"Ninety-eight percent of Americans like to read about opinion polls. This result is accurate to within plus or minus four points." There are (at least) two things wrong with that statement; by the time you finish this section you should be able to find both. If you are on the ball, you should be able to find the most glaring error immediately.

A survey or poll nearly always consists of a set of fixed questions together with pre-determined responses asked of small samples of people (kind of like multiple choice exams). The results from surveys and polls are obviously statistical. Why? Because the results are never intended to be just about the sample of people polled: they are meant to apply to larger groups of humans; usually all Americans, or even all of humanity.

You can work miracles with surveys and polls. Consider asking these two questions, "Would you support a law that requires the rich to pay their fair share?" and "Would you like your income tax to go up 15%?" Both are meant to show how much support there is for a new bill to be passed. A Congressman and newspaper who want the new tax will commission a survey wherein respondents are asked question 1; the other side of the aisle will counter with a survey that asks question 2. Both will announce that "Americans support my position!" You will *never—never*—see the wording of the questions. You will only hear the inferences made from them.

The true beauty of surveys and polls is that they are infinitely flexible. You can prove support for any point of view by creating one. All it requires is two things: clever question writing, and wild extrapolation. There is a professional class of people called pollsters or market researchers whose entire careers are devoted to the art of imaginative question writing. It isn't too hard to do yourself, but if you have doubts, it's easy to find these firms on the internet. Tell them what you hope to prove and they will provide the questions for a fee. It goes without saying that sometimes those who commission studies want to discover the truth of a matter; for example, a business wants to know whether a new product will sell. But even then, the firms that do the surveys know well the adage of bearing bad news.

Obviously, the firms cannot reach all Americans, or all "Europeans", or whomever. So they call up a few people on the phone (land lines, not cell phones usually), or head to the mall with clipboard in hand. This small sample (who was at home at the right time or who passed by Auntie Annie's Pretzels) is invariably said to represent fairly the entire population. Don't worry about this; nobody ever questions your sampling method, or if they do, it's easy enough to overwhelm them with technicalities (this is done by citing published works that suffer the same problems you do).

It is also imperative to ignore the abovementioned fact that people lie. They lie like dogs and often, particularly when presented with the question, "How much money do you make?" Credit card companies which ask this question on applications know that nearly everybody claims to make over \$100 thousand. Even if the question isn't so bold as "Have you recently committed tax fraud?" and is simple as "What is your age?", people will lie. They sometimes lie because of the pure pleasure of it, or they lie to help you out by giving you answers they think you want to hear. Or maybe they answer wrongly because they didn't understand what you asked. These facts, known to everybody, should decrease the certainty in all survey and poll results. Strangely, however, they never do.

In some fields, such as medicine and psychology, a *survey* goes by the glorified name of *instrument*. You do not just ask people a bunch of questions, as you do on a survey, you *administer* an *instrument*, which certainly sounds impressive. This is a very

technical subject, but I will try to summarize adequately the main problems with a typical example. I mention no names not wanting to hurt anybody's feelings.

Two groups, one fat the other thin (measured by body mass index), are administered an instrument intended to measure their "depressive status", that is, whether they are depressed. The instrument consists of several questions such as "I feel sad" to which the respondents rate on a scale from 1 to 5, higher numbers indicating stronger agreement. A score is created from by, more or less, adding up the ratings across all questions. If the person has a score higher than some cut off, they are said to be depressed. Average scores for both fat and thin people are computed and a classical test is performed which, of course, gives us a p-value, which we can imagine is 0.05, and is therefore publishable. A paper is written announcing "Thin people suffer more depression than fat ones." What is the probability that this statement (call it S) is true? We already know that is has nothing to do with the p-value. But ignore that problem and let's think about the data itself.

First is that the questionnaire, the instrument I mean, is said to measure depression. Does it measure it exactly? Nobody makes this claim for any instrument (that is measures the thing it purports to; unless that thing is a real, physical entity, like weight). Psychiatrists and psychologists will not always agree whether a given patient is in fact depressed. So here are two sources of error: (1) the instrument does not and cannot measure depression exactly, and (2) depression itself is hard to define. These two sources of error have to be incorporated into our probability of S. Error (1) is usually large, (2) is less, but is not negligible.

It is the case that people, if given the instrument twice, will not answer in the same way. Their internal state might have changed between the times between administration or they might just answer differently because they do not think too much or cannot recall their previous answers (what's the difference between a "4" and a "5" on the question "I feel sad"?). This is two more sources of error, (3) the internal states of people might fluctuate too rapidly to be of use, and (4) inconsistency in answers. Error (3) is probably small or even negligible, but error (4) is not and it is well

known not to be. These further make the probability of S less certain.

Any more? Well, people lie, they either don't want people to know the truth or they angle it towards what people want to hear; that's (5). The size of this error is usually unknown, but, ever hopeful, people assume it is near zero. Not only might not people be able to distinguish between "4" and "5" on the scale, they might not know what you are asking. For example, one instrument asks something like "I feel blue", which surely depends on cultural information not possessed by all. Confusion about the questions is error source (6). This source is generally acknowledged.

So how to cheat? Well, same way everybody else does. Just do not mention the sources of error. Ignore them. This allows you to claim your results are far more certain than they truly are. People will see your small p-value and assume S is very probable, even true. Don't sweat it, either; nobody will ever tag you because nobody wants to give up on these paper-generating questionnaires. If a referee for your paper questions your validated (which usually means you gave your instrument in at least two samples and got similar answers) instrument because of the errors mentioned above, it means he would have to question his own.

You can easily create your own instrument, but it's far easier to use a well-established one on a new source of data. The vast number of previously-published studies that use that instrument give weight to the idea that this is a reasonable thing to do.

## 5. Publishable p-values

Most journals, say in medicine or those serving fields ending with "ology", are slaves to p-values. Papers have a difficult, if not impossible, time getting published unless authors can demonstrate for their study a p-value that is publishable, that is, that is less than 0.05. Sometimes, the data are not cooperative and the p-value that you get from using a common statistic is too large to see the light of print. This is bad news, because if you are an academic, you must publish papers else you can't get grants, and if you don't get grants, then you do not bring money into your university, and if you don't bring money into your university, you

do not get tenure, and if you do not get tenure, then you are out the door and you feel shame.

So small p-values are important. I of course advise against using classical statistics methods, but if you are forced to (and some journal editors insist on it), then all is not lost if an initial large p-value is found. In fact, I would go so far to say that if you cannot find a publishable p-value in any situation, then you are not trying hard enough. There are several ways to lower your p-value.

The most well known is *to increase your sample size*. This one is a lock. Let's take a look at the t-test statistic from Chapter 10 to see why.

$$t(x) = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

There is a mathematical phrase that begins "without loss of generality" which I now invoke by letting, for ease of notation,  $n_A = n_B = n$  and  $s_A^2 = s_B^2 = s^2$ , so that  $t(x)$  becomes

$$t(x) = \sqrt{n} \frac{(\bar{x}_A - \bar{x}_B)}{s}$$

Remember that we want a large statistic, a large  $t$ , the larger the better, because larger  $ts$  mean smaller p-values. Do you see the trick? A larger  $n$  means a larger  $t$ ! All you have to do is to increase your sample size and just wait for the small p-values to start rolling in. This trick *always* works in any classical situation, even when the difference  $\bar{x}_A - \bar{x}_B$  is too small to be of interest to anybody. This is why having a small p-value is called attaining *statistical* significance and not practical or useful significance.

Incidentally, this trick also works in Bayesian statistics in the sense that the posterior distribution of  $\mu_A - \mu_B$  will have most probability above or below zero. But it fails miserably in modern observable statistics because a trivial difference in  $\mu_A - \mu_B$  won't make a tinker's dam worth of difference in the probability distribution of future observables.

The next trick, if you cannot increase your sample size, is to *change your statistic*. This comes from the useful loophole in classical theory that there is no rule which specifies which statistic you must use in any situation. Thus, though some creativity and willingness to spend time with your statistical software, you can

create small p-values where other people see only despair. This isn't so easy to do in **R** because you have to know the names of the alternate statistics, but it's cake in **SAS**, which usually prints out dozens of statistics in standard cases, which is one reason **SAS** is worth its exorbitant price. Look around at the advertising brochures of statistical software and you will see that the openly boast of the large number of tests on offer.

For example, for use in “testing differences between proportions”, just off the top of my head I can think of the  $z$  statistic, the proportions test with and without correction for continuity (two or three to choose from here),  $\chi^2$  test, Fisher's exact test, McNemar's test, logistic regression. There are dozens more and teams of academic statisticians constantly add to the pile. Don't believe it? Here's a small table of these tests for the TSD/Sex data from Chapter 11.

Test	p-value
Prop test	0.78
Fisher's	0.70
Logistic Reg.	0.52
$\chi^2$	0.50
$z$ test	0.49
McNemar's	0.24

Because I was only able to get to 0.24 just means I didn't try hard enough. Which is *the* correct p-value? They *all* are; that's the beauty of this trick. Not one of these p-values is more “right” than any other one. Each is valid. If all you know is classical statistics, let this knowledge sink in. It should prove to you that p-values are not what you probably thought they were.

For “testing differences between means”, there is the t-test (a couple of versions of this, actually), Wilcox test (also called Mann-Whitney), sign tests, Spearman correlation tests, Kendall's  $\tau$ , Kruskal-Wallis test, Kolmogorov-Smirnov test, permutation test, Friedman two-way analysis of variance—I'm running out of breath—and many more. Here's some of those tests for the advertising data:

Test	p-value
Spearman	0.87
Perm.	0.20
t-test	0.19
Wilcox	0.14
Kol.-Smi.	0.08

Nearly there!

Please remember that in this example, like the previous one, the *data is the same*; the only thing that changes is that classical statistical test.

The key to this deceit is to never admit what you did. When it comes time to write up your result boldly and authoritatively state, “We used Johnston’s (Johnston, 1983) frammilax test for differences in means.” Tossing in a citation always cows potential critics; tossing in two or more guarantees editorial acquiescence. Do not tell the reader that you went through a dozen tests to find the lowest p-value. Act as if “Johnston’s test” was what you had in mind all along.

This technique is unavailable in Bayesian or observable statistics. True, you can change your default prior distribution on the parameters or even change the model (see below), but editors in most fields are still suspicious of modern methods and tend to be conservative and will likely insist on a well-known default. There will be more room for creativity in, say, ten years when modern methods become familiar.

Our last option, if you cannot lower your p-value any other way, is to change what is accepted as publishable. So, instead of a p-value of 0.05, use 0.10 and just state that this is the level you consider as statistically significant. I haven’t seen any other number besides 0.10, however, so if your p-value is larger than this the best you can do is to claim that your results are “suggestive” or “in the expected direction.” Don’t scoff, because this sometimes works.

You can really only get away with this in secondary and tertiary journals (which luckily are increasing in number) or in certain fields where the standard of evidence is low, or when your finding is one which people want to be true. This worked for second-hand smoking studies, for example, and currently works for anything negatively associated with global warming.

## 6. Expand Your Data

Here is ancient wisdom:

*Seek and ye shall find.*

Nowhere does this better apply than in data analysis. Sometimes, despite all your efforts, you can not find a way to produce a publishable p-value with a given set of data. You tried all the tricks above, you can't increase the sample size and all the classical tests under the sun bring no joy. What to do? Increase your data! No, I don't mean increase the sample size, but increase the data on which you are making tests.

Everybody is constrained by time at least, but by budget usually, which puts a cramp on the sure-fire method of increasing sample size to get a small p-value. Well, friends, I am here to tell you, leave your "significant" p-value set at 0.05, leave your sample size as it is. You can still find a significant result with the method of *multiple testing*. This one requires a little more planning because you have to think of it before you start collecting data. For example, in the TSD example, don't just collect the fact that there were men and women; also observe the age, the weight, the race, day of the week, hour of the day, whether the person wore jeans, or a hat, stop and ask the people their income, their political party, their views on this and on that, whether the day was sunny, whether it was raining, the traffic density, and any other thing you can imagine. The only trick is to record as many different things as possible. Five is too few, fifteen is better, a hundred or more is practically a guarantee. I once was the statistician on a study that collected over 5000 items per person! I promise this is true. It was a medical study, wherein everything in a patient's chart was recorded, not once, but five to six times over a period of time, plus the individual questions from several "instruments", some homemade some "validated" (which means more than one person in print used it).

Your main interest is still whether or not there is a difference between men and women, which we have already seen only gets our p-value (after trying several tests) to a non-publishable 0.24. The next thing to try is *sub-group analysis*. See if there is a difference between men and women on just the sunny days or the

cloudy, or when the traffic density was high or when it was low, or whether it was a weekday or weekend, and on every other cut of the other variables. Race is always popular. One of these is bound to give you a publishable p-value.

Statisticians are on to this one, so be careful in how you describe your results. Whatever you do, do *not* say you tried every possible combination. You will be busted. Some statistician will immediately point out that you should have used so-and-so's method of correcting for multiple testing (the result of which is to inflate all your p-values). So be daring and just state, "Our results indicate that poorer Latino men and women wear TSDs at different rates" and nobody will ever question you. Try to angle your writing towards the idea that this subgroup was your main interest all along.

It can still happen that, even after exhaustive efforts, you still cannot find a difference between men and women in any of the subgroups. This kind of thing is rare, and its more likely you will have got bored of looking than there isn't a statistically significant result lurking somewhere. Can you guess what to do next? Right! Abandon the quest to find differences between men and women and simply find a difference between some other group; sunny and cloudy days, or whatever. If you have collected enough data, you simply cannot go wrong.

## 7. Models

Suppose you ran a classical regression (the `glm` model) and found that some of the coefficients of interest did not have a small enough p-value. You can try the tricks above, but you could also scan through the data itself to make sure that nothing is causing problems.

It happens that sometimes in your data an exceptionally large or small value appears. Statisticians call these *outliers*. I don't mean bad data values that arise from, say, bad typing, or by transposition, transcription, or some other honest mistaken entry. You'll find those when you look through the data and remove them anyway. No—what I mean are large or small values that are real, that were really measured, but that stick out and which

cause your model to go astray. This happens a lot with medical and economic data where the use of the normal distribution to quantify uncertainty is ubiquitous. Very large and small values show up all the time. What to do? Smack the label *outlier* on them, and then shun them, by which I mean, toss them out. Recompute your model after this and you will usually find improvement. I have seen this done on my presence on more than one occasion.

What's an outlier? A piece of data that does not fit your expectations. With surgical precision, then, you can cut out any offending variable so that, in the end, your data will act just like you wanted it to. Your chosen model will now fit. Of course, you will have learned nothing new, you will merely have reinforced your preconceptions, but that is always a comfort, isn't it?

Gottfried Leibniz, co-discover of calculus, said this

Let us suppose for example that some one jots down a quantity of points upon a sheet of paper helter skelter, as of those who exercise the ridiculous art of Geomancy; now I say that it is possible to find a geometrical line whose concept shall be uniform and constant, that is, in accordance with a certain formula, and which line at the same time shall pass through all of those points...?

What the old boy is saying is that it is always possible, given *any* set of data, to find a model that fits those data to any level of exactness, even perfectly. The implication, of course, is that if you can't find a model that fits your data well, then you aren't trying.

In linear models, it's easy to find good fits. You have  $n$  data points. One variable you want to predict, the remaining variables help you predict it. There are  $p$  of these. If you read the section above, you know you want  $p$  to be a big number. A well known trick is to let  $p$  get close to  $n$  in size. If  $p = n$  then, with a regression model, you will meet Leibniz's criterion exactly, such that you will have found a line that goes through each data point perfectly. Now, chances are that if you do this the p-values on the coefficients will likely not be publishable, so you have to change strategy. Do not tout p-values, trumpet your model fit. I earlier

skipped over the measure  $R^2$  (which you can get from running `lm` instead of `glm`, which gives you AIC instead; I skipped it because it doesn't take the uncertainty in the guesses into account, which here works in your favor). The highest and best this can be is 1 and the lowest and worst is 0. If  $p = n$  your  $R^2$  will equal 1, no matter what set of data you have. Obviously, you cannot report an  $R^2 = 1$ . This is like a psychic reading your mind exactly. People would be suspicious that a fast one is being pulled.

Take a few variables out of your model and report a modest, say,  $R^2 = 0.6$ , which sounds low, but believe me, it is not. Some fields would celebrate a value this high (these fields, which shall remain nameless, routinely see  $R^2$ s in the 0.1 to 0.2 range). You can try this with regression models, and it's an OK trick, but if you do people usually get curious about the p-values on the coefficients, which is an annoyance because we know these won't pass inspection. To get around this, skip regression and move on to what are called *latent variable* models. These go by names like *path* and *factor* analysis.

The way these work is that you have an observable  $y$  and a bunch of observable  $x$ s which are used to help explain  $y$ . So far, this is the introduction to regression, which isn't that exciting. Now here's the beautiful part. What you do is to pretend that there are a series of hidden, unobservable or *latent* variables  $\alpha_1, \alpha_2, \dots, \alpha_q$  that lie between  $x$  and  $y$ . Since  $x$  has  $p$  different variables, and there is no limitation on how many hidden variables that can lie between each  $x_j$  and  $y$ , and how many different paths can be between each of the  $\alpha_i$  (yes! they can be connected too!), you have an inexhaustible supply of models. Ready for the best part? These are usually used just to find something like high  $R^2$ , the influence of the observable coefficients (hence, their p-values) are deemphasized. You can report on the p-values of the unobservable latent variables instead!

Besides the usually academic specialty suspects, these kinds of models find favor in marketing.

## 8. Sleight of hand

I obviously cannot teach you every possible way to turn leaden data into gold (peer-reviewed) published papers. Statistics is too

big a field and the number of methods is huge and ever growing. The techniques I have given are the easiest and most reliable and you can nearly always get away with them as long as you are careful about your language.

As I mentioned above, boldness is imperative. Simply write your results as if the findings were what you were looking for and expected all along. If you have to use some obscure classical test or model, be sure to include at least two references that show that other papers (doesn't have to be in your field) have used them. People, especially journal editors, dislike novelty. Reassure them that what you are doing *everybody* does.

Vagueness is ever useful. Do not confess all the steps you had to go through to get your desired finding. Let people think you are as honest as they are.

At the worst, if all else fails, then at least claim that your results are *suggestive* or *in the direction* one expects if your beloved theory were true.

## 9. Homework

- (1) Find one use of statistics in an advertisement. Print is best, because you can just clip it out or photocopy and hand it in. If it's television or advertising, try to tape it and then copy down *exactly* what was said and done, and *exactly* where and under what circumstances you heard or saw it (what channel, time, webpage, etc. etc.).
- (2) Crack open a journal in one of the so-called softer sciences. Find an example of a paper that might have—just *might* have, I emphasize (don't accuse anybody unless you're sure you can get away with it)—used some of the creative techniques of statistical analysis mentioned in this chapter.
- (3) Multiple testing.
- (4) Try Leibniz's Geomancy example. Take out a piece of paper and draw a cross on it right down the center. This will be our standard mathematical axes. Get a pen, close your eyes, and stab at the paper at least a dozen times. Open your eyes (if I didn't specify this step, some clown would insist that he couldn't perform the remaining steps) and stare at the points. Try to find a curve that passes through most of the points.

- (5) You are going to do a survey. The main goal is to... Collect the person's sex, age, college status (i.e. freshman, etc.), birth-country (USA or not), *Ask at least five people.*