Marriages

"One third of the women students at Johns Hopkins University during its first year married faculty members."

Rent Increase

In order to justify a proposed rent increase, a landlord indicated that his maintenance expenses have risen by a greater percentage than his rental income over the past few years.

Skating

"Roller blading is the most dangerous skating form. Last year there were about 90,000 reported accidents involving roller blades, vs. 30,000 involving skate boards and less involving other forms of skating."

Accidents

A Minnesota study has shown that 61% of those involved in accidents have spent more than 10 years behind the wheel. Only 21% had 6-10 years of experience, and only 17% were between 1-5 years of experience. "Apparently drivers become more complacent about their driving as the years go by. As a consequence their records become worse," was the conclusion.

Family Size

A random sample of 30 children is selected at a school, and each child is asked the size of his/her family. The average of the resulting responses is taken as an estimate of mean family size for families represented in the school.

Polls

Before the parliamentary elections in Israel in 1988, the religious parties were predicted by polls to get about 10% of the votes. They ended up getting 15%.

Painting Costs

A painting contractor observed that, on the contracts he obtained, his actual costs tend to exceed his prior estimates. He began to worry that his cost-estimation procedures were faulty.

Smoking

A North Carolina study of smoking habits revealed that moderate smokers have a greater life expectancy than either heavy smokers or nonsmokers of the same age.

The obvious conclusion is that a little smoking is good for you.

Handwriting

In a study of schoolboys, a researcher found a positive correlation between hand size and quality of handwriting.

Education and Income

A recent report showed a strong relationship between the income of adults, and the quality of their childhood education (measured by such statistics as student-teacher ratio and teacher salaries).

Left-handed

A study of left-handed and right-handed people showed that the life expectancy of the left-handed ones was lower (on average).

How could being left-handed affect one's health or life expectancy?

Minimum Salary

The president of an institution proposed to set a minimum salary of $12,000 for employees in a particular category. An assistant found that there were 250 employees in this category, and that their average annual salary was $11,000. It was therefore concluded that the payroll increase would be about $250,000.

Teaching Good Students

A school administered a series of comprehensive standard examinations to a class of students at the ends of two consecutive years. It was found that the students who did best on the tests at the end of the first year tended not to maintain that level of performance on the second-year tests. It was therefore concluded that the second-year teacher failed to adequately stimulate the better students.

Salary Increase

A union claimed that the average monthly earnings of plant employees has fallen 8 percent in the past year. But management showed that every employee was making more than (s)he was the year before.

How come?

SAT Scores

From 1984 to 1985, the national mean SAT quantitative-skills score dropped (and a number of editorials lamented the decline in US educational standards). Yet the mean score for "whites" increased, as did the mean score for "non-whites".

Is this possible? Are we doing better or worse?

"On-Time" Airlines

In June 1991, counting all flights into 5 of the 30 busiest airports in the U.S., America West Airlines had an on-time percentage of 89.1 while Alaska Airlines was on-time only 86.7 percent of the time at these same airports.

Which airline should a passenger concerned about timeliness take?

Isthmus
7/31/98
p.6

ON THE TOWN

By BILL LUEDERS

MURDER SPREE: Rep. Scott Walker is right: Murder rates in Madison and Dane County are seeing a phenomenal increase. Walker, chair of the Assembly's Corrections Facility Committee, recently issued a press release noting that Madison's murder rate shot up 200% from 1997 to 1998, while Dane County saw a 300% increase.

Walker neglected to mention that the increase in Madison was from the uncommonly low 1996 total of one to 1997's total of three, which puts the city's murder rate per 100,000 residents at 1.5, well below the state average of 3.9. And Dane County's increase was from two to eight, for a 1997 rate of 2.0. Twenty Wisconsin counties had higher rates, led by Milwaukee County at 13.2.

Intelligence and Job Success

A study of Harvard graduates was conducted to determine the connection between intelligence and job success. In the data collected, there is no correlation between intelligence and career achievement.

Therefore, intelligence is not an important determinant of job success.

Summary of Main Points

Theme: Data Misrepresentations

Marriages:

It so happened, there were only three of them. Beware of small numbers!

Rent Increase:

But the maintenance expenses are only a fraction of the landlord's total cost. For instance, assume that taxation was fixed. Shouldn't the tenants have a share of this as well?

There is nothing very deep in this type of statistical abuse. It is simply a matter of representing correct data in a selective way. The selection and the focus may be misleading.

Theme: Sample and Population

Skating:

(i) First and foremost, we are not told how many people use these forms of skating, and how often. It may well be the case that the 90,000 figure is out of millions of users who use them daily, and the 30,000 is out of as many users who use them once a week.

(ii) Another issue has to do with "reported." You should always bear in mind that the way you receive information may be biased. For instance, it is possible that many of the skating-board users do not report injuries because they do not have health care coverage anyway.

(iii) Generally, whenever you look at "raw data," which were not collected in a controlled experiment, you should ask yourself what other factors may be influential here, what differences might there be between various groups (who were not randomly selected as would be the case in a controlled experiment) and so forth.

Accidents:

As in "Skating," we do not know what the population sizes are. There we were not told how many use which type of skates. Here we are not told what is the percentage of drivers with 10 years of experience (or more) in the whole population. It is very likely that most drivers are anyway in this group, and therefore it is more likely to see them involved in accidents – not because they are complacent, but because there simply are many of them.

To consider an extreme example, let's call a driver "experienced" if (s)he has been driving for more than one month. Then the vast majority of drivers are experienced. And even if a new driver is more likely to be involved in accidents than an experienced one, a randomly chosen accident is still more likely to find an experienced driver.

What we see here is an example of a phenomenon that the psychologists Daniel Kahneman and Amos Tversky call "ignoring base probabilities." They have shown that people tend to confound the probability of A given B with that of B given A. In our case, the fact that "an accident" (A) may have a high probability given "a new driver" (B), does not yet mean that "a new driver" has a high probability given "an accident."

Family Size:

Large families will typically have more children in the school than small families. Thus they get an "unfair" chance of being sampled. For instance, if there are five children in one family and one in another, and you ask all children, you'll get the answer "five" five different times. You average will then be 4.33 rather than 3.

The general principle we should bear in mind is that for a sample to be representative, each item in the population should have the same probability of being sampled. It is not always easy to design samples in such a way, but we should be aware of the pitfalls of non-representative samples.

Polls:

The poll was conducted over the phone. The statistical procedure for sampling among phone numbers may be perfect, but the phone customers population may not be identical to (or even representative of) the voters population. In this case, many ultra-orthodox families did not have phones at home, while they still voted.

At the end of this section you can find "A Primer on Polls." Note that the story we discuss here happened more than 50 years after the *Literary Digest's* famous poll in 1936.

Painting Costs:

This is related to the "Winner's Curse" phenomenon: suppose that the contractor's estimate of a project's cost is unbiased, that is, it is right "on average." Yet it has some variance. Now when our contractor overestimates a project's cost, he is less likely to get it, since some of his competitors are likely to make better offers than he makes. On the other hand, in those cases where he underestimated the true cost, he tends to win the contract more often. Conditioned on *having won the contract,* it is more likely that this is one of the "underestimate" cases than that it is an "overestimate" case.

In this case there is nothing wrong with the actual estimation procedure. However, the contractor may be better off by making offers which are not necessarily based on his estimation. Being aware of the "Winner's Curse," he should decide which offer to make as a matter of strategic choice, in the context of his competitors' behavior, rather than as a single-person decision making problem.

Theme: Correlation and Causation

Smoking:

The conclusion would seem warranted if a person choice of smoking level were independent of his/her health condition. But this is not the case: the "nonsmoker" group includes both elective non-smokers, and forced ones. That is, there are people who would smoke if their health condition allowed it. Some of them may even have smoked in the past, and had to quit because of the damages of smoking.

To make this example more extreme, one can argue that hospitals are detrimental to your health. This is hard to deny since most of the people in hospitals are sick, and most of the people outside them are healthy. This conclusion results from confounding cause and effect: the hospitalization, which is a *result* of sickness, is suggested as its *cause*. By a similar token, restaurants make you hungry, showers make you dirty and studying makes you ignorant.

Handwriting:

While this statement does not suggest a conclusion, many would believe that you need a large hand to write nicely. Maybe you should even start stretching your kids' hands to improve their handwriting.

And maybe not. Because correlation does not imply causation. We'll go back to this point in Section 4. In the meantime, let us just mention that it is very likely that age is a causal factor affecting both phenomena: older children tend to have larger hands, *and* to write better. So there may be no *causal* relationship between two factors which are correlated.

To test the existence of a causal relationship, we would like to "control for" age: to compare children in the same age group and see if the correlation survives. (We may include in our sample children of various age groups, but we should somehow "neutralize" the effect of age. Multiple regression is a tool which will allow us to do such tricks.)

OK, you may say – you controlled for age. What about other variables? How do we know there isn't something else hiding there? Good question. We don't. There is always a theoretical or not-so-theoretical possibility that we ignored some other factor. So what do we do?

Well, one way out is a controlled experiment, in which we randomly assign subjects to, say, two groups, and subject them to two different conditions. Assuming the samples are large, we can attribute the effect we measure to the factor we manipulated. But this is not entirely feasible here: we can't take children randomly, and make one group have larger hands. (Their parents may object to this idea.)

So it is very often the case that we are left with some uncertainty about the "true" cause of a certain effect. Furthermore, the history of science has many cases in which causal relationships were wrongly attributed to certain factors due to the unavailability of an unbiased sample. (See the "left-handed" story.)

Education and Income:

Well, there is certainly no evidence here of a causal relationship, as might be implicitly inferred from the report. For instance, it may be the case that your parents' income affects both your education quality and future income. Perhaps the level of effort that your parents put into selecting a school for you correlates with other factors which, in turn, contribute to future income. In short, explanations abound.

Left-Handed:

Well, it turns out that about 60 years ago and earlier people thought that there is something wrong about writing with one's left hand. Children who attempted to do that were spanked. The result is that among people over 60 there are relatively few who are left-handed.

Now the measurement of "life expectancy" is always a tricky business. But whichever way you do it, the fact that you almost don't find left-handed people over 60 is going to affect you results. If you only look at aggregate data, this is equivalent to a rare disease which attacks the left-handed population at the age of 60. (And if such a disease did exist, it certainly would and should have and effect on life expectancy measurement.)

Note that life-expectancy is a classical example in which it may be very problematic to have a controlled experiment: first, you have ethical problems. But even if you could ignore these, you'd typically have to wait quite some time to get any results.

Theme: Data Aggregation

Minimum Salary:

This calculation would be right only if there were no variance, i.e., if all employees were getting the same (average) salary of $11,000. But since there is variance in salaries, the payroll increase is likely to be *higher*. For instance, if half of the employees in this category were making $13,000 and half – $9,000, the increase would be $375,000. Intuitively, the fact that you don't have to raise a salary above $12,000 doesn't compensate for the need to raise those below $12,000. A $13,000 worker and a $9,000 worker are equivalent to two $12,000 workers when it comes to the average. They are not when "minimum" or "maximum" levels are set.

Teaching Good Students:

This is a well-known phenomenon, called "regression to the mean." The story is as follows: on any given test there is some aspect of luck, right? So those students who did well were partly good, partly lucky. (Any individual student could simply very good. Maybe (s)he even had a bad day, yet scored very high. But overall, we do observe some "noise" effect here as well as true quality.) Because of the "true quality" effect, we would expect the students who excelled last year, as a group, to be above average. But as a group, they will tend to be below their previous level, since not all of them will be equally lucky this year. (Note: the bias stems from our focusing on a group which contains more first-year-lucky-ones than the average.)

Correspondingly, you'd expect those students who did worst last year to be below average, but above their last year's level. On the other hand, some students who were about average last year are likely to replenish the "good" and the "bad" groups. That is – this year we are likely to have more or less the same number of "excellent" students as last year, but they do not have to be the *same* students as last year.

Salary Increase:

This is a little strange, isn't it? If everyone is making more, how can they all together make less? Well, the trick is in the definition of "everyone:" if, indeed, the same people are included in both populations, then this phenomenon cannot happen. But assume that some senior employees left the plant, and younger, lower-paid workers joined it during the past year. In this case the average can drop while any single worker who has been in the plant during both years had a salary increase.

Generally, there might be problems whenever we aggregate data for populations with different compositions. (See the "SAT scores" story.)

SAT Scores:

The problem is, again, the composition of the population. If we were guaranteed that the proportion of "whites" to "non-whites" is the same in the two populations under discussion (1984 students and 1985 students), an increase in each sub-population implies an increase in the overall average. But if the proportion of the sub-population changed, this need not be the case anymore. For instance, assume that a certain weak group, which has slightly improved, has become much larger relative to the others. In this case, despite the fact that *all* groups are doing better than before, the whole population still appears to be doing worse.

In this case, it seems plausible to argue that the education system is doing fine, since it shows improvement in each sub-group, and its evaluation should not take into account demographic changes. But there are purposes for which the "more relevant" data are the aggregate ones. For instance, assume that you consider the qualitative skills of the incoming class. You should take into account the fact that on the average, they are lower than last year's. True, we are not saying here anything about the school system, but we still have to deal with this fact.

If you wish to read more on this phenomenon, and see a numerical example, read the handout "On the Dangers of (Dis)aggregation." If you feel you understand the point, you are probably right.

On the Dangers of (Dis)aggregation

Suppose we are comparing a new medical treatment to an old one. Each was tried on 40 people, with the following results:

	Improved	Not Improved	% Improved
New	20	20	50
Old	24	16	60

So, it seems as if the new treatment is not as efficient as the old one. Tough luck!

However, a certain researcher was interested in the efficacy of the drug for the sub-populations of men and women. She analyzed the same samples by these categories. Her findings are summarized in the following tables:

Men Only	Improved	Not Improved	% Improved
New	12	18	40
Old	3	7	30

Women Only	Improved	Not Improved	% Improved
New	8	2	80
Old	21	9	70

She therefore concluded that, *whatever the patient's gender,* the new treatment is better than the old one.

What is going on here? This is quite puzzling, indeed, and this phenomenon was even dubbed a "paradox" ("Simpson's paradox").

A careful inspection shows the following:

a. Both treatments are more successful for women than they are for men;

b. The new treatment was tried on a population containing mostly men (75%); the old one was tried on a population containing mostly women (only 25% men);

c. Thus, the old treatment *appears* to be better simply because it was administered to a population with many women, who are anyway more likely to get better.

Comment: There are experiment design considerations which help avoid this problem whenever possible. Roughly, if you control the experiment, you would like to assign subjects randomly to the two types of treatments, to reduce the probability of uneven populations. However, in many cases you are confronted with data that does not result from a controlled experiment. In these cases you should be aware of the possibility that some additional factors are involved. For instance, if one is unaware of gender effects, one may be misled to assume that the old treatment is better than the new one. Unfortunately, if you do not control the experiment, there is no way to guarantee that some "hidden" factor is not actually responsible for the results.

Analysis

Simpson's paradox is easily explained by conditional probability analysis. Denote:

I: the event that the patient's condition is improved.

W: the event that the patient is a woman.

M: the even that the patient is a man ($M = W^c$).

Then, for each treatment separately:

$$P(I) = P(M)P(I|M) + P(W)P(I|W).$$

If the percentage of men and women in the two treatments is the same, then the paradox cannot occur: in this case, if $P(I|M)$ and $P(I|W)$ are both larger for the new treatment than they are for the old treatment, so will be the overall probability of improvement, $P(I)$. (And this is probably where our intuition comes from, i.e., we tend to implicitly believe that the proportion of women (and men) in both samples is identical.)

However, if $P(M)$ (and, thus, also $P(W)$) differ in the two cases, the aggregation may be misleading. In our example, for the old treatment we get

$$.6 = (.25)(.3) + (.75)(.7)$$

and for the new one

$$.5 = (.75)(.4) + (.25)(.8).$$

Thus, despite the fact that (.4) > (.3) and (.8) > (.7), the aggregation of (.4) and (.8) gives a lower value than that of (.3) and (.7). This is simply because the aggregation is "unfair:" in the new treatment case, most of the weight (.75) is put on the lower value (.4), while in the old treatment case most of the weight (.75) is put on the high value (.7).

<u>So What's the Lesson?</u>

Well, part of it we have already mentioned:

1. When two populations have different proportions of sub-populations, the aggregated conditional probabilities may appear quite different from the disaggregated ones. In our example, the conditional probability of improvement *given either gender* is larger for the new treatment than for the old one, but the aggregated probability of improvement is smaller for the new treatment.

2. Thus, one should be careful when analyzing data that may be a result of uneven aggregation. Similarly, one may be safer using controlled experiments when possible.

However, it is not always the case that the disaggregated data are "right" and the aggregated ones are "misleading." Consider, for instance, the case of affirmative action in a university with two schools. Using similar numbers (or even the same numbers), you can construct examples where both schools boast a higher percentage of, say, minority students, while the university as a whole has a lower percentage. (Do this as an exercise. Hint: you can take the same example, and replace "old treatment" and "new treatment" by "1980" and "1990", respectively; correspondingly, replace "women" and "men" by the two schools, and "improved" by "minority student.")

So we have a third conclusion:

3. Sometimes the aggregated data will be the "right" or more relevant ones, and then it is the *disaggregation* that may be misleading.

"On-Time" Airlines:

The airport-by-airport numbers were as follows:

	Alaska	Airlines	America	West Airlines
Destination	**% on time**	**# of arrivals**	**% on time**	**# of arrivals**
Los Angeles	88.9%	559	85.6%	811
Phoenix	94.8%	233	92.1%	5,255
San Diego	91.4%	232	85.5%	448
San Francisco	83.1%	605	71.3%	449
Seattle	85.8%	2,146	76.7%	262

So, at each of the five airports in the sample, Alaska Airlines had a better on-time percentage than America West! How could America West have a better overall percentage? The overall percentage is a weighted average of the individual percentages, where the weights are determined by the number of flights into each airport. It turns out that America West flies mostly into Phoenix, where good weather allows both airlines high on-time ratings, while Alaska Airlines flies mostly into Seattle where the weather is worse, making on-time ratings lower. The different weightings make the overall ratings meaningless from the point of view of a traveller going to a specific city.

If you wish to read more on this phenomenon, and see another numerical example, read the handout "On the Dangers of (Dis)aggregation." If you feel you understand the point, you are probably right.