Contents

Chapter 1

Drawing Inferences from Data: Common Fallacies and Pitfalls

1.1 Introduction

Here are some remarks about statistics that I have been audience to:

- *Statistics has never solved anything for me.*
- *Did you know you can use statistics to prove anything you want?*
- *In my entire life I have never needed to use a significance test.*

What are the reasons for these attitudes? Should we try to convert these unbelievers?

There is no doubt that there are numerous examples where statistics is used wrongly. Sometimes this is unintentional, and sometimes it is not. Other times, we see an analysis and come to a conclusion or make a decision. Later, it seems this decision was wrong. This could have been due to an incorrect analysis, or due to our own misunderstanding of the analysis.

In this session, through the presentation of a few case studies and examples, I shall try to highlight some things to watch out for, whether we are consuming or presenting an analysis. Instead of specific rules to apply in different situations, you might conclude that what you need most is **patience, logic and an open mind**.

1.2 Summary of Pitfalls

Before going through the examples and case studies, here is a summary of common mistakes in statistical analyses. When we go through the examples, watch out for how the investigators avoided (or committed) the following errors in their analyses. The rough structure of this breakdown, and some of the examples noted here, come from [8] and [9].

1.2.1 Sources of Bias

Statistics allows us to make inferences about a large group (*a population*) based on observations of a smaller subset of that group (*a sample*). It also allows us to make comparisons between populations, and decide if observed differences between them are **real**, and not simply due to random variation.

When we collect data to make such decisions, we must be careful about the following:

- The sample must be **representative** of the target population. One way to achieve this is through **randomisation**.
- If we wish to apply statistical tests, the variables we measure must conform to certain **assumptions** which underlie the statistical procedures.
  - Several statistical tests assume that our observations are **independent** of one another. This is the hardest assumption to check, and *so we usually don't think too much about it*!
- We must have a valid **control** group in order to make comparisons. Without this, the primary goal of statistics, making comparisons, is nullified.

1.2.2 Errors in Methodology

Statistical methodology are varied, but the inductive logic behind testing is common to all models. Here we highlight some common mistakes made when conducting hypothesis tests, and conclusions from them.

- If we wish to perform a statistical hypothesis test, we need to collect data. How much data to collect? Too little, and we might not detect a significant difference (insufficient power) and too much, and we will find a difference that is statistically significant but not practically so. We should consider the **power** of our test before we go ahead.
- In consulting sessions, I am routinely asked questions similar to this one:

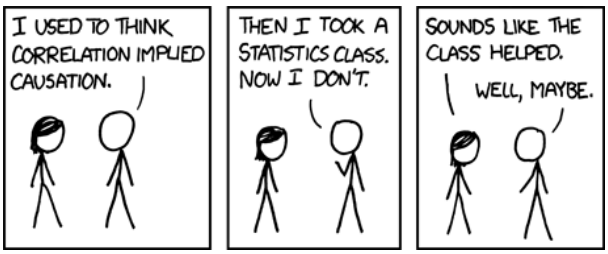
  > When I perform a regression with 100 variables, none are significant, but when I perform pairwise correlation with the response, I find 10 significant ones. I would like to report those significant correlations.

  Increasing the number of hypothesis tests (**multiple comparisons**) that we apply increases the number of false positive results we obtain.
- Related to this problem is the mistake of **data-snooping**, or formulating a hypothesis test **after** observing the data. This increases the Type I error and greatly reduces the reproducibility of the result.

1.2.3 Misinterpretation of Results (intentional or otherwise)

In this category, I have included several issues pertaining to the presentation and reporting of results.

- *Statistics show that more people die in hospital than at home. Also, there is a strong association between dying and being in bed.* These are nonsensical. In general, we should not **mistake correlation for causation**, or even make the suggestion.



- Sometimes, we perform several analysis and only show the ones that strengthen our case. This is very bad. The **selective presentation of data** is deceitful; unfortunately the pressures of our time and the increasing number of analyses done are making this more common. We must have integrity in our work and watch out for this possibility when we study others'.

- Today, more and more statistics are being invented and presented. Here is one such statistic, the symmetric Mean Absolute Percentage Error (sMAPE) in time series analysis:

$$\frac{1}{h}\sum_{t=1}^{h}\frac{200|y_t-\hat{y}_t|}{y_t+\hat{y}_t}$$

  Although they can seem complicated, we should make an effort to understand these metrics before we make a decision based on them.

- Lastly, we must be aware that a **visual representation** of the data can be influential. By manipulating the scale, or omitting outliers, a graph can appear to strongly support a desired theory. Here is an example (from [21]) of the opposite: a stunningly accurate record of data that led to a better understanding of how cholera was spread, and kickstarted the field of epidemiology.

1.3 Polio Vaccine Trial

In the 1950s, polio was a great concern for America. It hit young children the hardest and left many of them crippled, including some who could only survive on a respirator. Its inexplicable epidemic behaviour led the government to spare no effort to eradicate it. In this section, we shall review how the government conducted a vast experiment to determine the efficacy of a vaccine. Further details can be found in [13].

Here is some important information about polio and the trial:

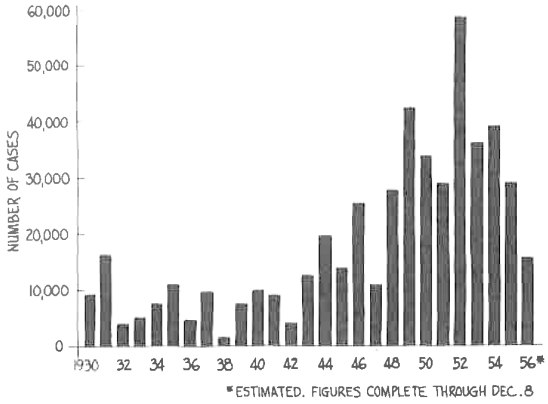
- It is caused by a virus. There are 3 main types involved.
- Countries or communities where the hygiene level was highest were hardest hit by the epidemics.
- Once an individual has been infected by the virus, he or she is immune to another attack.
- In 1954, the government wished to test the use of a killed virus preparation to inoculate people. This was known as the Salk vaccine.
- The desire was to show that the vaccine was at least 50% effective; the number of subjects involved was more than a million.

1.3.1 Approach I

A first idea was to distribute the vaccine as widely as possible, through the schools, and then check if the rate of reported polio was appreciably less in the next season (year).

Here's one problem with that approach:

We need to consider the inherent variation in polio incidence from year to year. We have no **control** group in this approach!



Another problem with this loose approach is that it is difficult to diagnose the mild cases of polio. A doctor might be influenced by his general feeling about how widespread polio is in his or her community at that time.

1.3.2 Approach II

A second idea was to **offer** vaccination to all children in the second grade of participating schools, and to follow the polio experience not only in these children, but also in the first and third grade children. The vaccinated children would constitute a treatment group, and the first- and third- graders would make up the control group.

- Suppose you were a doctor, and a vaccinated second-grader came to see you with mild fever. Would you be more inclined to diagnose him as having polio or not?
- Is it fair to assume that the characteristics of the volunteer group is different from those who do not volunteer?
- Is the issue of difficult diagnoses removed?

1.3.3 Approach III

- Children were randomly assigned to the polio vaccine or a placebo.
- Action was taken to eliminate any possible observer biases from administration to diagnosis of polio by applying a double-blind protocol.

These are the results of the experiment:

Table 1 Summary of study cases by diagnostic class and vaccination status (rates per 100,000)

				Poliomyelitis Cases									
		All Reported Cases		Total		Paralytic		Non-paralytic		Fatal polio		Not Polio	
Study Group	*Study Population*	*No.*	*Rate*	*No.*	*Rate*	*No.*	*Rate*	*No.*	*Rate*	*No.*	*Rate*	*No.*	*Rate*
All areas: Total	1,829,916	1,013	55	863	47	685	37	178	10	15	1	150	8
Placebo control areas: Total	749,236	428	57	358	48	270	36	88	12	4	1	70	9
Vaccinated	200,745	82	41	57	28	33	16	24	12	—	—	25	12
Placebo	201,229	162	81	142	71	115	57	27	13	4	2	20	10
Not inoculated*	338,778	182	54	157	46	121	36	36	11	—	—	25	7
Incomplete vaccinations	8,484	2	24	2	24	1	12	1	12	—	—	—	—
Observed control areas: Total	1,080,680	585	54	505	47	415	38	90	8	11	1	80	7
Vaccinated	221,998	76	34	56	25	38	17	18	8	—	—	20	9
Controls†	725,173	439	61	391	54	330	46	61	8	11	2	48	6
Grade 2 not inoculated	123,605	66	53	54	44	43	35	11	9	—	—	12	10
Incomplete vaccinations	9,904	4	40	4	40	4	40	—	—	—	—	—	—

*Includes 8,577 children who received one or two injections of placebo.
†First- and third-grade total population.

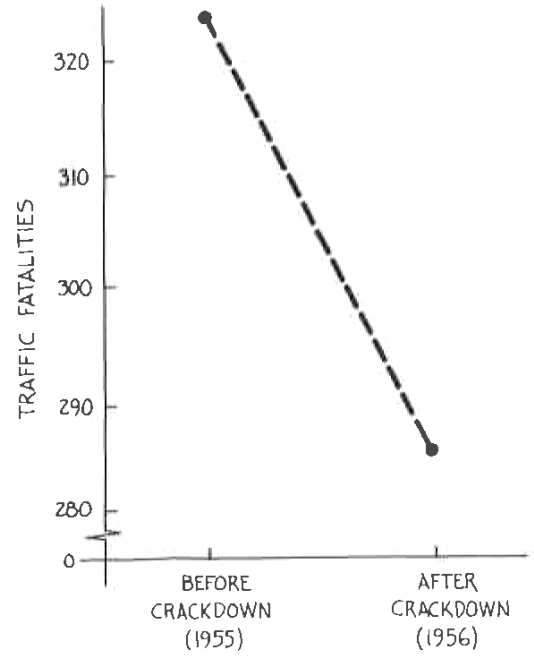
Source: Adapted from T. Francis, Jr. (1955), Tables 2 and 3.

1.4 Measuring the Effects of Social Innovations

We have just seen that it in order to make a valid comparison, we need to have a legitimate control group. What happens if we don't? It is easy to be misled when there appears to be no proper control. In such scenarios, we must be on the lookout for the lack of a control group, and to find a proxy if we can.

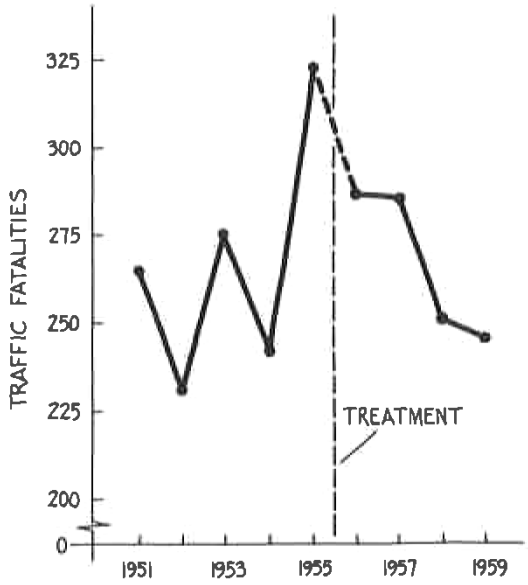
1.4.1 Crackdown on Speeding

In the mid-1950s (see [4]) , the state of Connecticut instituted a state-wide crackdown on speeding. To demonstrate the effectiveness of the program, a simple before-and-after chart was presented, accompanied by the proud governor claiming that the program was definitely worthwhile.



Without any supporting context, it appears to be a convincing argument.

There is a saving of 40 lives upon implementation of the crackdown!



With the larger context, the governor's claim appears almost certainly incorrect.

The drop in fatalities in 1956 might have been part of a steady drop that was ongoing.

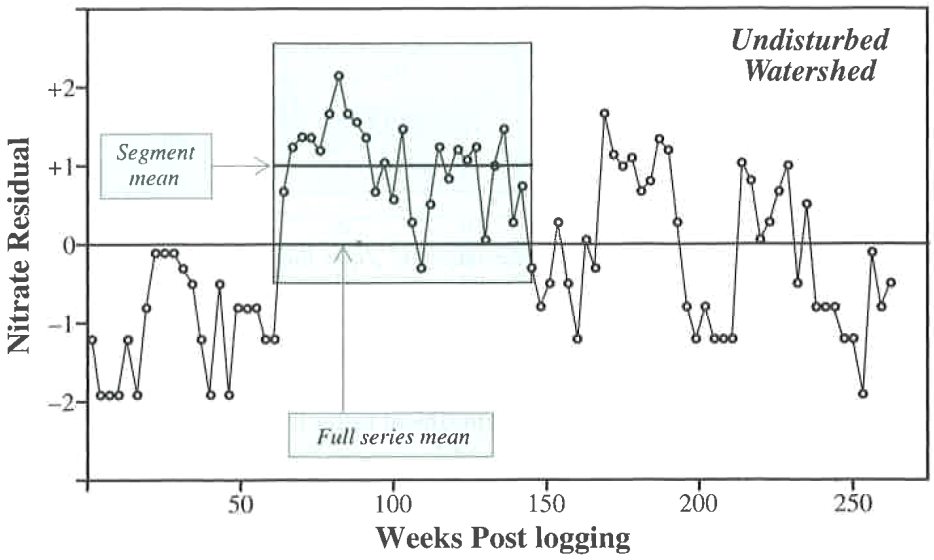
There are several troubling issues with the official statement:

- There is no proper comparison or control group to assess the effect of the crackdown.
- The selective presentation of the data amplifies the positive impact of the crackdown.

Suppose we try to fix this, by testing the mean number of traffic fatalities before and after the crackdown. To compare means, a commonly used procedure is the two sam-

ple t-test. If we were to apply this, however, we would be completely wrong, because one of the assumptions of this test is the independence of observations. We have a time series of observations, and these are (almost by definition) not independent.

Consider the time series in the figure below, taken from [14].



If we were to extract one segment, compute its mean, and compare that to another segment after an intervention of sorts, we would be likely to get a false positive precisely due to the **autocorrelation** among observations:

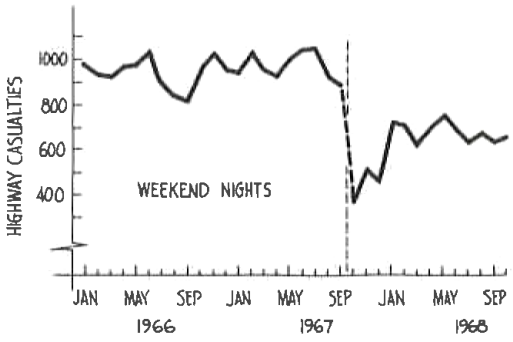
- If an elevation in one of the segments is elevated (or depressed), neighbouring values will also be elevated (or depressed).
- Elements near to each other will be close to each other; the variance within each segment will be under-estimated, leading to a larger test statistic than what should be.

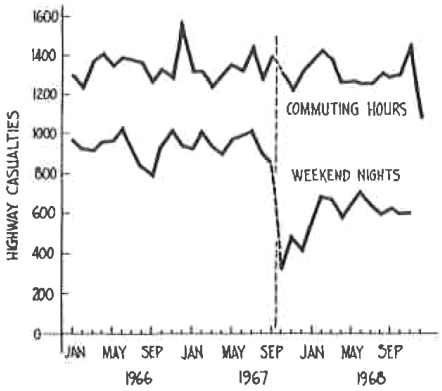
Before we proceed, here is another reason why the governor is being too optimistic in his analysis. The number of fatalities in 1955 was an all-time high. The governor felt he had little choice but to introduce a crackdown. However, there is a phenomena referred to by statisticians as **regression to the mean**. You can read more about it in this nontechnical article [3]. It basically contends that after an extreme point, subsequent ones will be, on average, nearer the general trend. This is why we must avoid overfitting. Following this principle, even if there had been no intervention, traffic fatalities in 1956 would probably have reduced.

1.4.2 Breathalyser Crackdown in Britain

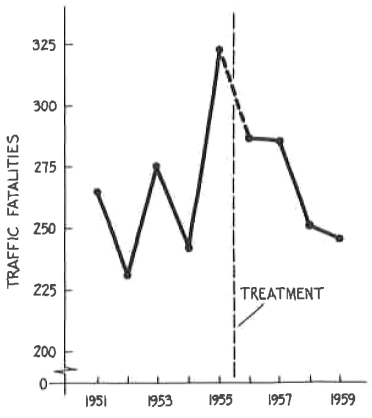
About a decade later, the British government introduced a toughened stance on drink driving (see [15] for full details). The testing was made more stringent and the punishment more severe. If we take a look at the effect on serious casualties on Friday and Saturday nights, it is almost undeniable that this was a successful measure.

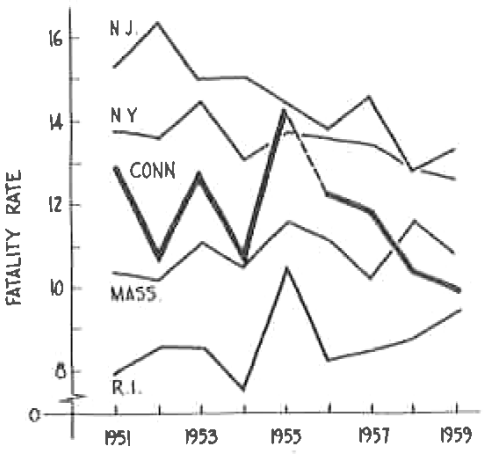
This can be made even more apparent by introducing a comparison group. The ideal control group is not available, but consider the number of casualties during hours that the British pubs are not open:





Let us see if this approach works for the Connecticut crackdown. Instead of comparing fatalities in a different hour, let us compare them to fatalities in nearby states, where the crackdown was not applied.





1.4.3 Summary and Fixes

We have been discussing situations where there was no control group until after all the data collection had been carried out. In these cases, we should try to be creative and find similar populations. These can serve as the basis for comparison.

It is imperative that we decide on the evaluation criteria before the innovation is implemented.

What other ways can be used to assess if the change is significant?

1.5 The Choice of Control Group

In the previous cases, we encountered situations where the control group was not present. Now let us study a case where it is not clear what the control group should be, and where the choice of control group leads to different outcomes! For more details on the following analyses, the reader is referred to [19].

1.5.1 Employment Discrimination

In the 1970s, the US government sued the Hazelwood school district for discriminating against African American (AA) teachers. The decisions went back and forth, but statistics played a significant role in the arguments. Note that Hazelwood is a district in St. Louis County, which includes the city of St. Louis.

(a) First, in the district court, the percentage of AA teachers in Hazelwood district was compared to the percentage of AA students. Since they were roughly the same, the decision ruled in favour of the Hazelwood school district.

(b) Next, the Court of Appeals ruled that the comparison with students was irrelevant and reversed the decision.

(c) Then the case was tried in the Supreme Court.

(d) Here, the Court noted that

- the percentage of AA teachers in the school district was 3.7%.
- the percentage of AA teachers in the encompassing St. Louis County (excluding the city of St. Louis) was 5.7%.
- the percentage of AA teachers in the encompassing St. Louis County (including the city of St. Louis) was 15.5%.

The comparison of 3.7% to 5.7% was not significant, but the comparison of 3.7% to 15.5% was.

It all boiled down to a question of the relevant labour market. To get around this gray area, the argument in courts shifted towards a different question:

> Given the actual pool of AA applicants for the job, how did they do in comparison to the non-AA applicants?

In 1977, the Washington Hospital Center was sued for employment discrimination based on the following data:

	Selected	**Rejected**	**Pass Rate**
AA	4	5	44%
non-AA	26	0	100

This difference in proportions was statistically significant, even accounting for the small sample sizes.

1.5.2 Simpson's Paradox

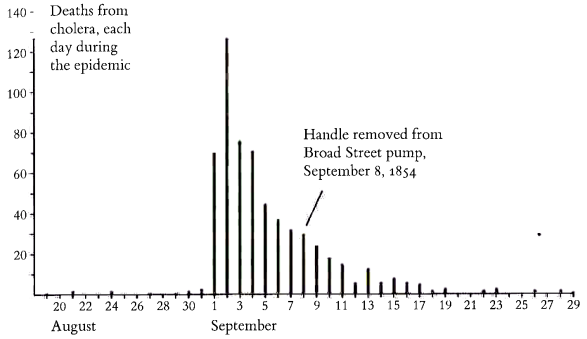
In many corporations, we deal with data at the department or sub-department level. This innocuous choice of the basis for comparison can lead to spurious correlations known as Simpson's Paradox.

In the 1970s, UC Berkeley was accused of discrimination against females when considering admissions to graduate school: the overall proportion of females was much lower than the proportion of males admitted. Paradoxically, this observation at the aggregate level could have (and did) arise due to differing numbers of applicants to departments. Here is a simplified example of how it could happen:

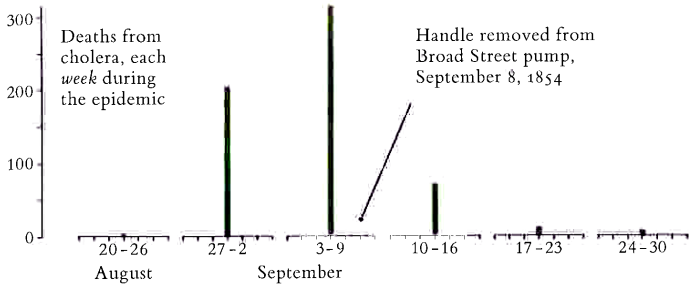
	Mathematics			English			Combined		
	Admit	Deny	%	Admit	Deny	%	Admit	Deny	%
Males	90	10	90	1	9	10	91	19	83
Females	9	1	90	10	90	10	19	91	17

1.5.3 Related Problems With Aggregation

In 1854, John Snow contributed to the discovery that cholera is spread via water, not air. Was he also responsible for the end of that epidemic, and the saving of hundreds of lives? The truth is this:



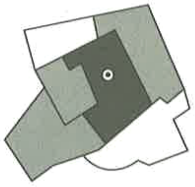
With a choice of aggregation into weeks, we can make him out to be a bigger hero than he was:



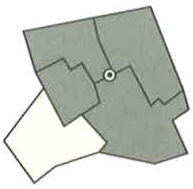
Through his careful and painstaking analysis and investigation, John Snow identified the Broad Street pump as the one causing the transmission of cholera in that epidemic.

If he had only dealt with aggregated data, his conclusion would have differed based on the aggregation!

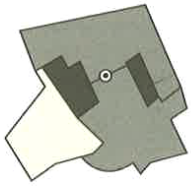
For more details on the comprehensive detective work he did, take a look at chapter 2 in [20].



In this aggregation of individual deaths into six areas, the greatest number is concentrated at the Broad Street pump.



Using different geographic subdivisions the cholera numbers are nearly the same in four of the five areas.



In this aggregation of the deaths, the two areas with the most deaths do not even include the infected pump!

1.6 Regression Modeling Strategies

Regression models are usually of the form

$$\text{response} = \text{weight}_1 \times \text{predictor}_1 + \text{weight}_2 \times \text{predictor}_2 + \cdots + \text{weight}_p \times \text{predictor}_p + \text{error}$$

When we fit such a model to the data, we certainly want a good fit. However, some standard and accepted practices actually lead to overfitting to the data, leading to irreproducible results.

For a nontechnical introduction to these ideas, please refer to [1]. For technical explanations, please take a look at [6] and [7].

1.6.1 Dubious Practices

One such practice is known as *automated stepwise variable selection*. In this procedure, we perform a hypothesis test one stage at a time and either reject or include a new variable, finally stopping when no change occurs. These multiple comparisons inflate the possibility of Type I errors and thus result in the inclusion of useless variables!

Another common practice is the *pre-screening of variables* - to determine what transformation to use for each variable, to drop/retain variables, etc. Although we have not performed a hypothesis test, we will still end up over-fitting to the data.

Finally, a common practice of dichotomizing continuous variables and including them as factors in a linear model can lead to overfitting when two of these variables are correlated.

1.6.2 Some Suggested Solutions

Stay away from automated stepwise variable selection. When you fit a model, you might have some pre-planned hypothesis you wish to test; stick to those when reporting p-values. By all means, do go ahead and explore the data to find which is the "best-fitting" model and pre-screen the variables. Data are too expensive to restrict ourselves to just those pre-determined questions. As Tukey said,

> Even here, restricting one's self to the planned analysis – failing to accompany it with exploration – loses sight of the most interesting results too frequently to be comfortable.

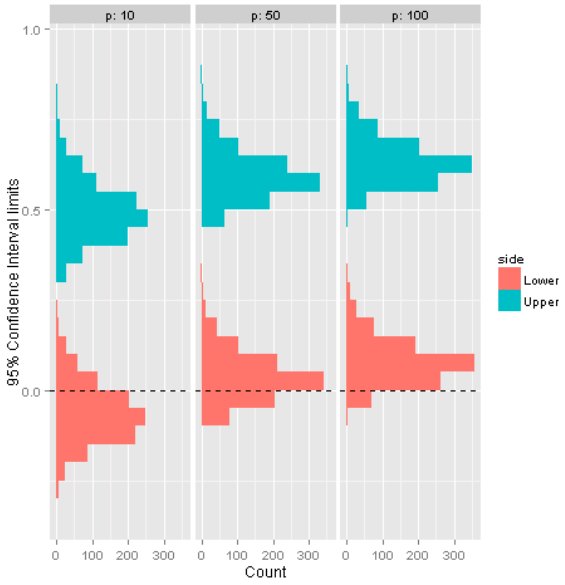
However, do not over-generalise the findings from exploring the data. Confirm them, perhaps with a follow-up experiment, or using a validation set of data.

If you have correlated variables, try to combine them (using PCA perhaps) or leave out all except one.

Finally, ensure that you have enough power to perform your tests. Empirically, studies have found that we need a minimum sample size of 15 per regression variable. If you wish to explore the functional form for a particular variable, then you would need more.

1.6.3 Selection Problem

Suppose that we were to run an experiment to compare battery brands. We buy 10 batteries each of 4 brands, then run them on the same type of device to see how long they last. We apply an ANOVA test and find there is a significant difference between the mean battery life-times. We pick the brand which yielded the longest life, and estimate the mean lifetime using a 95% confidence interval constructed using the sample data for that brand. Is there anything wrong with this?



The problem with this approach is that we have not taken into account the selection procedure!

The maximum of a group will tend to be large, so our final estimate of that group's mean should be shrunk slightly.

On the left, we have a simulation experiment with 10, 50 and 100 groups, each with 50 observations, where the means of all groups are identical. The mounds represent the upper and lower confidence intervals for the mean of the group with the maximum sample mean, over 1000 simulations.

1.7 Visualisation Principles

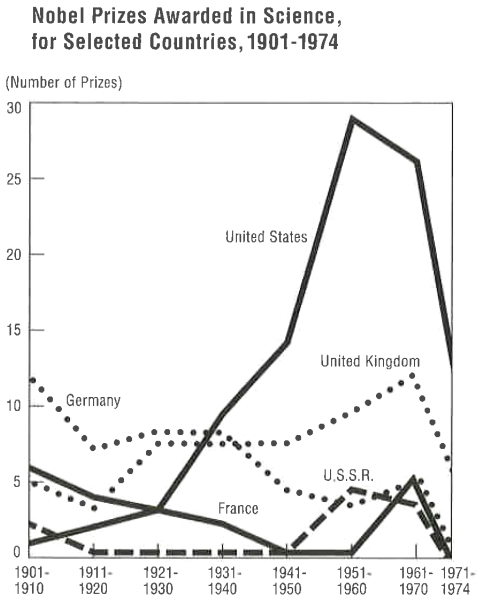
Graphics are invaluable when we present an analysis. As John Tukey said in [22],

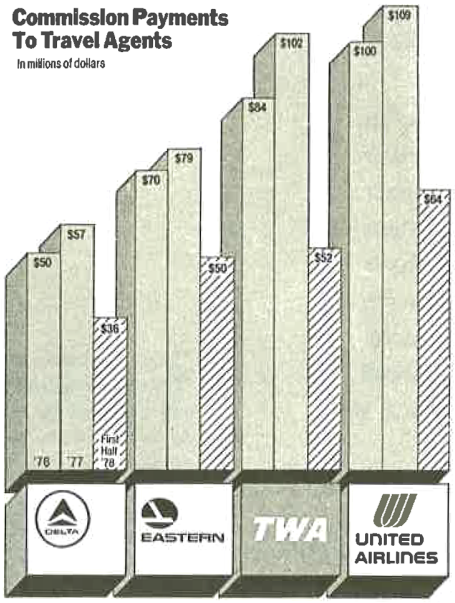
> **The greatest value of a picture** is when it *forces* us to notice **what we never expected to see**.

However, there are many graphics that are untruthful to say the least. When we present data in a graphic, we could try to abide by these principles:

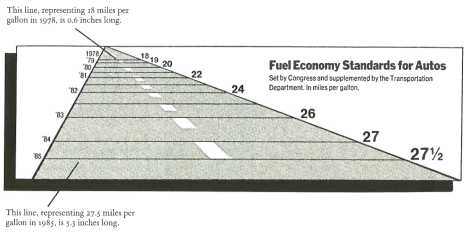
- The representation of numbers should be directly proportional to the quantities represented.
- Show data variation, not design variation.
- In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.
- Graphics must not quote data out of context.

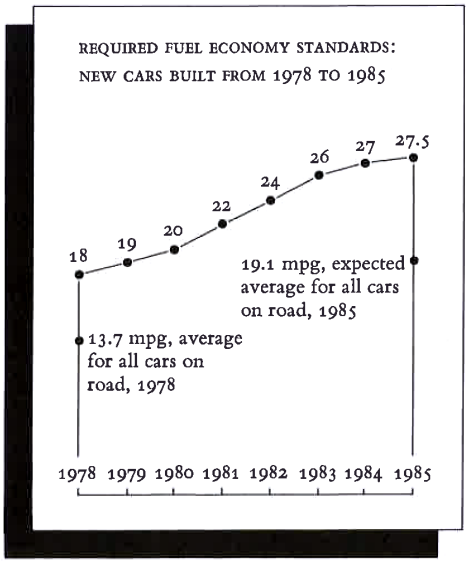
Here are some examples where these principles have been violated.





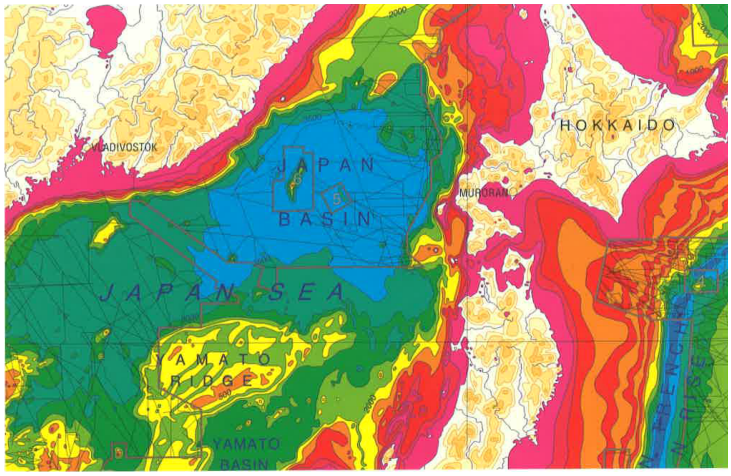
It is easy and striking enough to present the data truthfully:



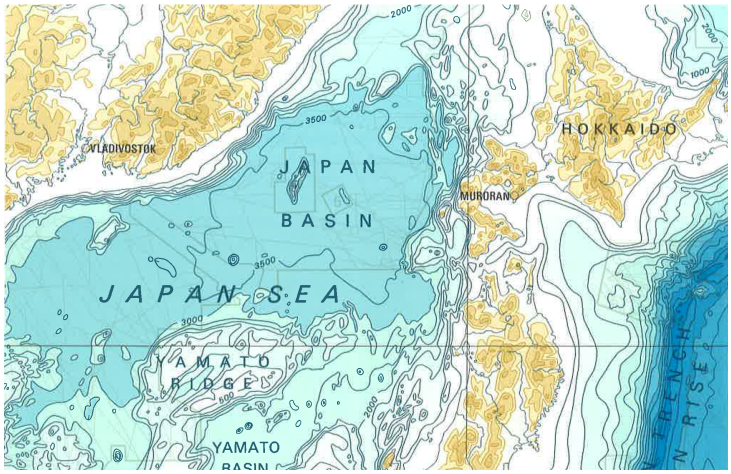


When dealing with time series data of money, it is worth considering adjusting the prices for inflation before presenting it.

A garish choice of colours can also lead to a very poor understanding of your data.



In how many ways does this improve on the previous image?



1.8 Enamoured of Large Models and Small p-Values

As more and more complex models are being developed, it becomes more and more difficult to judge whether their contribution is worth the additional computing resources, time and intelligence to run.

Bear the following empirical rules in mind. These were outlined after two very extensive forecasting competitions involving a number of experts in their domain [12].

- Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simple ones.
- The relative ranking of the performance of various methods varies according to the accuracy measure being used.
- The accuracy when various methods are being combined outperforms, on average, the individual methods being combined and does well in comparison to other methods.

- The accuracy of various methods depends upon the length of the forecasting horizon involved.

Related to this point is the one that we are simply performing too many hypothesis tests these days. In almost all of them, the p-value is simply being compared to 0.01 or 0.05. Why? This pressure to publish, or just to find something significant, has led to highly irreproducible results. It has come to a point where a large coalition of very famous statisticians are petitioning for the default α-level to be 0.005 (see [2]).

1.9 Summary

If we had to make a list, here are the things I would ask/watch out for in a data analysis:

- Has any data been left out? Were they outliers?
- Why was the data aggregated in this way?
- Were too many comparisons made?
- Could there have been other reasons for this data to appear so, other than the conclusions stated here?

Chapter 2

Risk Analysis for Complex Systems

2.1 Characteristics of Complex Systems

Classical science tends to dissect a system into smaller and smaller isolated parts in an effort to reduce the problem to essential elements. Each of these elements are then studied separately. Quite often, we search for a physical law that explains the process we are studying. An example of this is an explanation of planetary motion.

On the other hand, we now find we are dealing with systems so complex that we cannot afford to study their components in isolation. Here are some examples of complex systems:

- The swarming of locusts.
- Traffic jams.
- The 2010 BP oil spill.

There is no single definition that is agreed upon for **complex systems** (see the interpretations reviewed in the essay [10]), but it is possible to derive certain properties that these systems exhibit. For a deeper explanation of these traits, take a look at the book [11].

- The system contains a collection of many **interacting objects** of agents.
- The objects' behaviour is affected by **feedback**.
- The system exhibits **emergent behaviour** which are generally surprising, and could be extreme.
- The emergent behaviour arise in the **absence of a central control**.
- The system **oscillates** between orderly and disorderly states.

Complexity science is a new field that is emerging to deal with models with the above traits. It focuses on what new phenomena can emerge from a collection of **relatively simple components, interacting together**.

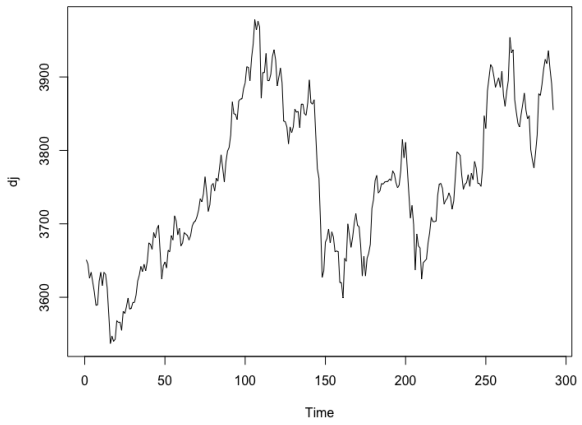
2.2 The Financial System

We would all love to predict what will happen in tomorrow's financial market. Methods and strategies that claim to do so sell for exorbitant prices. In time series methods of forecasting, we would typically we would try several models, assess them via their prediction errors, and decide to use the one that returns the smallest prediction errors.

One of the most popular models for financial time series is a random walk model:

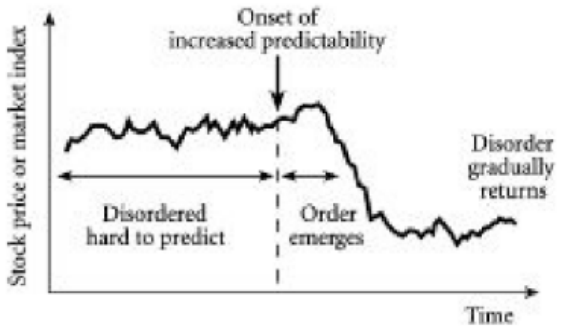
$$y_t = y_{t-1} + e_t$$

where e_t is typically assumed to be Gaussian errors. This sort of model can predict a time series of this sort quite well:



There is no monotonic trend in the series, and the differenced series looks a lot like white noise. It can in fact be shown that the optimal prediction is the most recent observation.

This, however, is the reductionist view of the financial world. We consider one time series at a time, and try to determine its behaviour as best we can. However, this is rather simplistic a view of the financial world. This model will never be able to produce a crash like this:



This is an example of the financial system, a complex system, moving in and out of pockets of order. When the order emerges, what is happening is that huge groups of traders are selling at the same time. Then the groups fizzle out and we are back in the state of disorder, which is the norm. The emergent phenomenon is the pocket of order.

The crucial element lacking in traditional models is that of feedback. Note that all traders in the market are privy to historical prices, news updates, and expert forecasts. In addition, these agents adjust their behaviour **according to how their strategy is performing**.

Suppose then, that we simulate the financial market in the following way:

(a) Traders have to pick a buy or sell signal at each unit of time.

(b) Each trader knows the true behaviour of the market for the past, say 10 units of time.

(c) Each trader has a simple strategy to decide, based on the 10 recent values, what to do. For instance, he could just look at when the same behaviour occurred and see what happened then.

This is almost enough to produce crashes like we see in reality, but not quite. It turns out that a slight modification will do it:

> Traders should only enter the market if their strategy has been **successful in the recent past**.

If we believe we can model the financial market with some degree of confidence, the next natural question is whether we can predict these pockets of order and disorder. An important paper that tries to answer this question is [17]. In that paper, the author purports that we can create **prediction corridors**, that suggest in which direction the system is moving. The width of these corridors change with time; when the system is moving into a pocket of order, the corridors narrow.

2.3 Mathematical Models for Complex Systems

2.3.1 Cellular Automata (CA)

A cellular automaton is a model of a world with very simple physics. The space is divided into cells. Each cell typically has 2 states - on or off. Time is also divided into discrete steps. Rules specify how to compute the next state of each cell, based on the current state.

Probably the most popular CA to be studied is known as the Game of Life. It was developed by John Conway in 1970. The cells in GoL are arranged in a 2D grid, and each cell is either alive or dead. The next state (in time) depends on the current state of itself and its eight neighbours. This behaviour is loosely analogous to cell growth: cells that are isolated or overcrowded die, but at moderate densities they flourish.

Number of neighbours	Current state	Next state
2 – 3	live	live
0 – 1, 4 – 8	live	dead
3	dead	live
0 – 2, 4 – 8	dead	dead

There are a number of stable patterns that emerge from relatively simple starting states. A visualisation of these patterns can be found `https://bitstorm.org/gameoflife/` and `https://www.youtube.com/watch?v=bTPN3spiq1I`.

2.3.2 Agent Based Models

Agent-based models depict the interacting components in the system as agents that

- are intelligent, based on a simple set of rules.
- have local, imperfect information.
- have differing behaviours.

One of the first agent-based models was put forward by Thomas Schelling in [16] to explain how racial segregation arose in cities. Suppose we have an array of cells, where each cell represents a house. A house could be occupied by a blue agent, a red agent, or it could be unoccupied. At any time, an agent may be unhappy or happy. If there are at least two neighbours like themselves they are happy. Otherwise they are unhappy. If an agent is unhappy, he chooses one of the unoccupied cells at random and moves there.

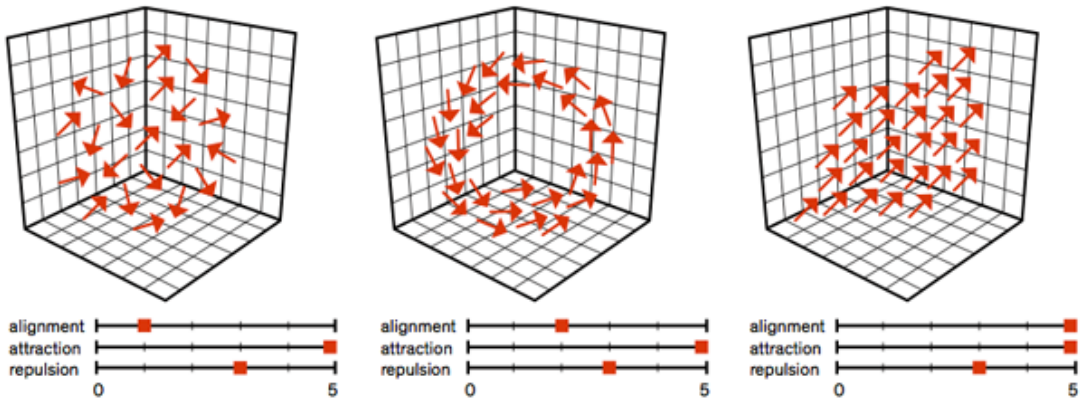
Surprisingly, if we start with a simulated city that is entirely unsegregated, clusters of similar agents will appear very quickly. As time passes, the clusters grow until there are a small number of large clusters and most agents live in homogeneous neighbourhoods.

One of the most common arenas for agent-based modeling is traffic jams. Suppose we consider a one-lane road that forms a circle. We start drivers at random positions and speeds, but they follow these rules:

- If the following distance to the next car is too short, the driver brakes. Otherwise, he accelerates.
- If the current speed would cause a collison, the driver stops.

This is enough to create the emergent behaviour of traffic jams!

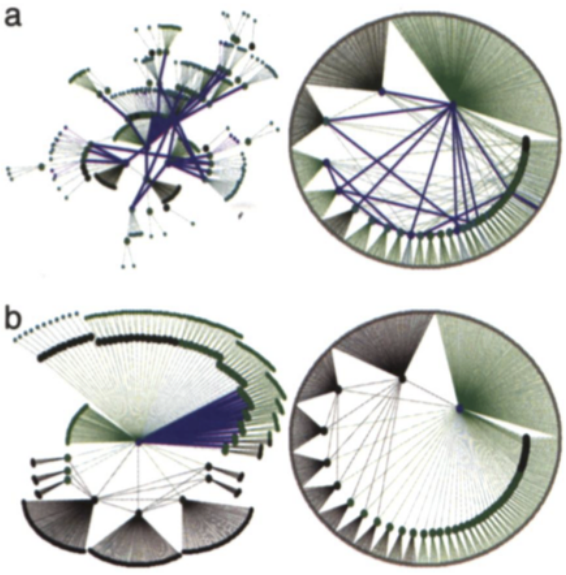
Another domain where agents have been successfully used to model a real-life process is the study of swarms. The following image was taken from a story in WIRED magazine: `https://www.wired.com/2013/03/powers-of-swarms/`.



2.4 Risk Modeling for Complex Systems

As you can see, most of these techniques involve simulation, and rely on the model being a good representation of the true system. At present, most of the work is being done on generating or identifying the rules that result in emergent behaviour of the system.

As part of the simulation, we track adverse outcomes and the frequency with which they they occur. We must be careful to compare the model we are using to the observed data. There have been cases when the output of the resultant model does not adhere to observed data.



For instance, when modeling the structure of the internet, a common assumption is that the degree structure of vertices follows a power law or Zipf Law, which results in what is known as a scale free network. This leads to a robust yet fragile network of nodes, where random attacks can be withstood, but not targeted ones.

At left, the graphs in the top row correspond to scale-free networks, while the graphs below correspond to what real-world internet nodes look like.

However, it has been shown that this is not true and that a different set of modeling assumptions should be used (see [5])

2.5 Introduction to Probabilistic Risk Assessment

A complex system, such as the company that writes a popular operating software, or a large hospital, or a nuclear power plant, consists of a multitude of people, processes and technologies working together.

In order to ensure a satisfactory operating standard, it is not sufficient to ensure that each component, on its own, operates at an acceptable level. We have to be aware of how components work together, and the impact of their simultaneous failure.

In this session, we shall discuss a tool known as Probabilistic Risk Assessment (PRA). It provides a formal framework for modeling the combinations of multiple failures that lead to a specific undesirable outcome. It is used to systematically identify and review all of the factors that can contribute to an event.

The main goal of a PRA is to quantify risk by answering the following questions:

(a) What can go wrong?

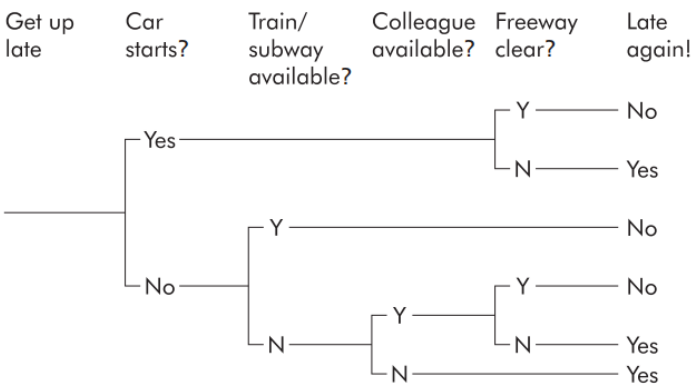
(b) How likely is it?

(c) What are the associated consequences?

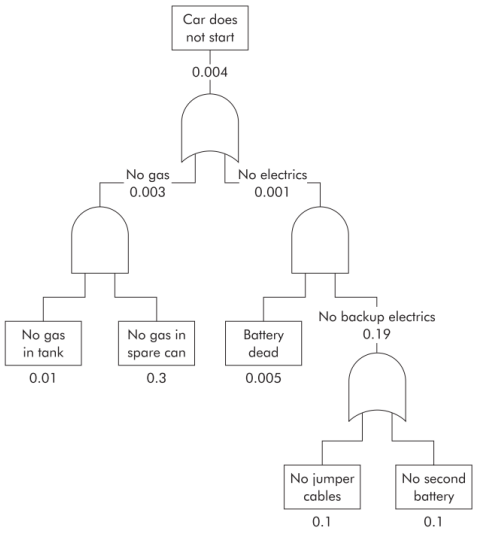
We can use the answers from a PRA to decide on resource allocation, or to review specific processes that contribute to **undesired scenarios** with high probability of **system failure**.

2.6 Late for Work!

Let us begin with a simple example that we can all relate to [23]. The outcome of being late for work is undesirable. Our first task is to identify **all possible scenarios** that lead to this outcome. At this stage, the PRA process uses a graphic known as an Event Tree (ET) to visualise the relationship between the scenarios and the adverse outcome. Here is one possible ET for being late. The event at the extreme left is an Initiating Event (IE). It kicks off the sequence of events that could lead to the undesirable outcome. Between the IE and the outcome are chronologically arranged Pivotal Events (PE) that ultimately decide if we are late or not. Can we add on to this diagram, based on our local knowledge (of Singapore, of ourselves, etc.)?



The next step is to extend this tree, to model each of those PEs. This extended tree is known as a Fault Tree (FT). Modeling a PE amounts to arranging events that culminate in the PE occurring in a top-down tree. A FT utilises AND and OR logic gates to reflect that certain events must occur together in order in order for the PE to happen, while other sets of events only need one member to occur.



At the base of each FT are the basic events, for which we have to assign probabilities. This is one of the most difficult parts of PRA. The difficulty stems from the fact that these are typically rare events, that we might not have observed at all! How then do we put down a reasonable number? What if we are wrong? Would we even be able to tell if we are far from the truth?

Supposing that we can overcome the anxiety and put down some plausible numbers. How do we then compute probabilities further up the FT? There are software that will do it for you, but the ideas are not to complicated. Consider one of the AND gates:

$$\begin{aligned} P(\text{no gas}) &= P(\text{no gas in tank AND no gas in spare can}) \\ &= P(\text{no gas in tank}) \times P(\text{no gas in spare can}) \\ &= (0.01)(0.3) = 0.003 \end{aligned}$$

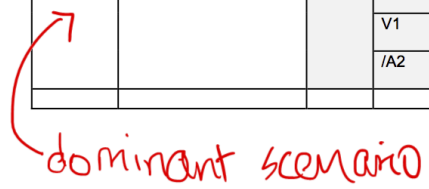
Assuming the basic events are independent, the computations for the OR gates are only slightly more complicated:

$$\begin{aligned} P(\text{no backup electrics}) &= P(\text{no jumper cable OR no second battery}) \\ &= 1 - P(\text{jumper cable AND second battery}) \\ &= 1 - (1-0.1)(1-0.1) \\ &= 0.19 \end{aligned}$$

Following the FTA, we can generate minimal cut sets – minimal sets of events that result in the adverse outcome, and their overall probability of occurrence. At this point, we are in a position to use sensitivity analysis to decide which event we should target to eliminate or reduce, and so on.

Here is an example of minimal cut sets, from the NASA manual on PRA ([18]).

				unsuccessful isolation		
6	Hydrazine Leak, Detection Succeeds but Isolation Fails, Avionics Failure	5	IE	Leak	1.0E-2	1.0E-4
			L	Leak occurs upstream of isolation valves	1.0E-1	
			/A2	Avionics fail after unsuccessful isolation	1.0E-1	
		6	IE	Leak	1.0E-2	9.0E-7
			/L	Leak occurs downstream of isolation valves	9.0E-1	
			V2	Isolation valve V2 fails to close	1.0E-3	
			/A2	Avionics fail after unsuccessful isolation	1.0E-1	
		7	IE	Leak	1.0E-2	9.0E-7
			/L	Leak occurs downstream of isolation valves	9.0E-1	
			V1	Isolation valve V1 fails to close	1.0E-3	
			/A2	Avionics fail after unsuccessful isolation	1.0E-1	
					Total	1.02E-4



The typical follow-up, a risk assessment of the dominating scenario, would look something along these lines:

	Structure of Dominant Scenario			
	IE: Leak occurs	**Leak occurs upstream of isolation valves**	**Leak damages critical avionics**	**Frequency**
OPTIONS	IE	L	/A2	
Do nothing	0.01	0.1	0.1	1.0E-4
Option 1: Reduce the likelihood of leak between the propellant tank and isolation valves (e.g., change in piping design)	0.01	0.05 (see note below)	0.1	5.0E-5
Option 2: Reduce susceptibility of avionics to leak (e.g., rerouting of wires and fortify wire harnesses)	0.01	0.1	0.01 (see note below)	1.0E-5
Option 1 and 2	0.01	0.05	0.01	5.0E-6
Note: The numerical values shown in this table are hypothetical.				

2.7 Limitations of These Approaches

Assessing risk through a model-based approach, or through a PRA approach relies on having a sufficiently detailed representation of the real-world sytem. In my opinion, we should start with a simple model, and then slowly add complicated modifications.

One possibility that could result from these simulation-based probabilities is that we ignore behavioural traits, or more specifically, how people react to policy changes. Suppose we find that the probability that the pharmacist and doctor both make a mistake when dispending medication is unacceptably high. Maybe we try to solve this with a third person checking the medication. This may not reduce the mistake probability, because the mind set that "the other person will check this again, so I do not need to spend so much time on it" may materialise.

The other, more obvious limitation of these approaches is that we do not know the probabilities of certain events. How do we ascertain the probability that the doctor prescribes the wrong medication? One way is to acknowledge the uncertainty in our probability model; this is known as the epistemic uncertainty (as opposed to aleatory) uncertainty. Once we acknowledge this, we can assess the probability of an adverse event in a better context.

2.8 Summary

At the end of the day, any model is only an approximation to the real process. If we do not utilise an adequate model, we will not be alerted to the important safety issues we were looking for.

There are serious questions that require us to think carefully about our own system. There is no intelligent black box that will autonomously identify the extreme emergent behaviour that we will see in the real system.

In short, here are my thoughts on risk analysis and management:

(a) Come up with a model for the system, be it an emergent one, or one using PRA. Identify the main risk areas.

(b) When developing this model and fixing probabilities, for instance, consult the people from all involved departments.

(c) Use online monitoring to alert you as early as possible to out-of-control behaviour. These tools can also feedback into your model to refine it.

Bibliography

[1] Michael A Babyak. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3):411–421, 2004.

[2] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature Human Behaviour*, page 1, 2017.

[3] J Martin Bland and Douglas G Altman. Statistics notes: some examples of regression towards the mean. *Bmj*, 309(6957):780, 1994.

[4] Donald T Campbell and H Laurence Ross. The connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review*, pages 33–53, 1968.

[5] John C Doyle, David L Alderson, Lun Li, Steven Low, Matthew Roughan, Stanislav Shalunov, Reiko Tanaka, and Walter Willinger. The robust yet fragile nature of the internet. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14497–14502, 2005.

[6] Julian J Faraway. On the cost of data analysis. *Journal of Computational and Graphical Statistics*, 1(3):213–229, 1992.

[7] Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[8] Clay Helberg. Pitfalls of data analysis. eric/ae digest. 1996.

[9] Abhaya Indrayan. Statistical fallacies in orthopedic research. *Indian journal of orthopaedics*, 41(1):37, 2007.

[10] Christopher W Johnson. What are emergent properties and how do they affect the engineering of complex systems?, 2006.

[11] Neil Johnson. *Simply complexity: A clear guide to complexity theory*. Oneworld Publications, 2009.

[12] Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476, 2000.

[13] Paul Meier et al. Safety testing of poliomyelitis vaccine. *Science*, 125(3257):1067–1071, 1957.

[14] Fred Ramsey and Daniel Schafer. *The statistical sleuth: a course in methods of data analysis*. Cengage Learning, 2012.

[15] H Laurence Ross, Donald T Campbell, and Gene V Glass. Determining the social effects of a legal reform: The british" breathalyser" crackdown of 1967. *American Behavioral Scientist*, 13(4):493–509, 1970.

[16] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.

[17] David Smith and Neil F Johnson. Predictability, risk and online management in a complex system of adaptive agents. *arXiv preprint physics/0605065*, 2006.

[18] Michael Stamatelatos, Homayoon Dezfuli, George Apostolakis, Chester Everline, Sergio Guarro, Donovan Mathias, Ali Mosleh, Todd Paulos, David Riha, Curtis Smith, et al. Probabilistic risk assessment procedures guide for nasa managers and practitioners. 2011.

[19] Judith M Tanur, RS Pieters, Frederick Mosteller, and GR Rising. Statistics: a guide to the unknown. 1978.

[20] Edward Tufte. Visual explanations. 1997. *Chesire Connecticut: Graphics Press Google Scholar*, 1997.

[21] Edward Tufte and P Graves-Morris. *The visual display of quantitative information.; 1983*. 2014.

[22] John W Tukey. Exploratory data analysis. 1977.

[23] J Wreathall and C Nemeth. Assessing risk: the role of probabilistic risk assessment (pra) in patient safety improvement. *Quality and Safety in Health Care*, 13(3):206–212, 2004.