Dan Sasaki
Institute of Social Science
dsasaki@iss.u-tokyo.ac.jp

Welcome to

Statistics Underground

13 March 2013

A quick sneak preview of contents:

0. Let's start-istics.
Please try warming up your statistical mind by exercise examples in 0.0.

1. How to tell a statistic.

1.1. What is a statistic? Introduction of some of the most basic topics which, however, may not be persuasively taught in standard lectures and textbooks.

1.2. Population and DGP. Sticky and subtle distinction between what samples are and where they come from.

2. How to lie on statistics.

2.1. White lie. Some technical basics which are often misunderstood.

2.2. Statistical abuse. Honesty is hardly ever heard, honesty is such a lonely word...

2.3. Statistical misuse.
Try not to operate firearms, motor vehicles, and statistics without knowing how!

3. How to steal with statistics.

3.1. Creative writing in statistics. Statistics is a language, can be foul.

3.2. Play with numbers for fun. Scrabbling with statistics.

3.3. Be partial, be partisan, reward your friends and punish your enemies!
You have to do what you have to do.

4. How to kill with statistics.

4.1. Statistical Creationism.
Statistics concocted to harm others. Beware not to fall victim...

4.2. Prey on numbers for profit. Awful tales of statistical crimes in history.

4.3. Fourth Reich, George III Bush, Benito XVI, Dishonest Abe, and World War III. Awful truth about statistical crimes, real time.

0. Let's start-istics

0.0. Did you know ?

0.0.1. The population densities and the GDP per capita of the G7 countries have been reported as follows :

countries	US	Canada	UK	France	Germany	Italy	Japan
km^{-2}	31	3	246	110	232	193	339
US\$ 10^3	45	43	46	40	40	36	34

0.0.1.i. Regress the GDP per capita onto the population densities (and the constant, by OLS) to estimate the coefficient.

0.0.1.ii. According to 0.0.1.i., for every additional inhabitant per square kilometre, how would the GDP per capita change ?

0.0.1.iii. Conversely, regress the population densities onto the GDP per capita (ditto) to obtain the estimated coefficient.

0.0.1.iv. Based upon 0.0.1.iii, for every 1000 dollars of additional income per capita, how much population growth per square kilometres should be expected ?

0.0.1.v. Are your expectations from 0.0.1.ii and 0.0.1.iv compatible or not ? If not, where have you gone wrong ?

0.0.2. Regressing Y onto X (including the constant), the estimated coefficient turns out highly statistically significant. Regressing Z onto Y (ditto), the estimate is again statistically significant. Is it implied that X indirectly predicts Z ?

0.0.3. Our allegedly well educated instincts seem to tell us invariably that stupid people outnumber wise people in our troubled world. If so, why is it always the majority opinion, not the minority opinion, that democracy is supposed to heed ? Choose the correct explanation from among the following ⟨a⟩ through ⟨e⟩.

⟨a⟩ Democracy, by definition, is for popularity not for correctness. Were we interested in making correct decisions, we would respect only those opinions of the wisest few and ignore the rest of the public.

⟨b⟩ Should the opinions split sharply betwixt the wise on one side and the stupid on the other, then we should undoubtedly listen to the former in order to achieve a wise decision. In reality, however, wise people disagree amongst themselves, so do stupid people, thus on average those of one opinion differ little in terms of wisdom or stupidity from those of another opinion.

⟨c⟩ Wise people can often influence the public opinion, so that they can in effect exercise more votes than the rest of the crowd.

⟨d⟩ When a sufficiently large majority agree, their opinion is more likely to be correct than the opposing opinion even when the supporters of the latter are substantially wiser than those of the former, and when each individual form her/his opinion independently.

⟨e⟩ All of the above.

0.0.4. It is commonly observed that bad weather on the voting day tends to favour specific parties. Does this mean that it is desirable to do the election on a fine day, or on a stormy day?

0.0.5. Two coins have been tossed.

0.0.5.i. Someone whispers that one is a head. Assuming that this information is reliable, what is the probability that the other is also a head?

0.0.5.ii. You have taken a glimpse at one of the coins, which is a head. What is the probability that the other is also a head?

0.0.5.iii. Someone reports that at least one is a head. Assuming that this report is truthful, what is the probability that both are heads?

0.0.6. You have taken an IQ test, of which the score is known to be distributed approximately symmetrically around the true IQ of the person, and your score is 140. Which one of the following ⟨a⟩ through ⟨e⟩ describes the uncensored truth?

⟨a⟩ Your true IQ is more likely to be below, than above, 140.

⟨b⟩ Your true IQ is approximately equally likely to be either below or above 140.

⟨c⟩ Your true IQ is more likely to be above, than below, 140.

⟨d⟩ Your true IQ is more likely to be 140 than either above or below.

⟨e⟩ Your true IQ is approximately symmetrically distributed around 140.

0.0.7. Two identical sealed envelopes are put on the table. You are told by the instructor that each envelope contains a cheque, one of them exactly twice the amount of the other. However, neither you nor the instructor can foretell which one contains more than the other (both cheques are supposed to carry positive amounts). You are asked to take whichever one of the two envelopes you like.

Determine whether each of the following statements (in *slanted fonts*) is true, or false.

0.0.7.i. After you have picked up one of the two envelopes but *before* you open it, the instructor offers you that you may, if you wish to, exchange the envelopes to take the other one instead. *Your exchanging the envelopes will neither increase nor decrease the expected amount of money you receive.*

0.0.7.ii. Now you have opened one of the envelopes to find out that the cheque inside it is, say, ten euros. Conditional upon this information, *the other envelope contains either five euros with probability one half, or twenty euros with probability one half. The expected amount in the other envelope is therefore* $5 \times 0.5 + 20 \times 0.5 = 12.5$ *euros.*

0.0.7.iii. *In general, if one of the envelopes contains an amount* x *, the conditional expected amount in the other envelope is always* $1.25x$ *.*

0.0.7.iv. Now that you have seen the amount x in one of the envelopes, the instructor tells you that you may exchange the envelopes and have the other one instead. *Insofar as* $x > 0$ *, your exchanging the envelopes will increase the conditional expected amount of money you receive.*

— Please try all the exercise examples on your own, including the above, prior to the intensive sessions scheduled on Wednesday, 13 March. —

0.1. What are we here for ?

Objective, mathematical, technical, and thus cut-and-dry as it may sound, statistics is a man-made construct after all. It is a story written by a red-blooded writer, or a piece of art drawn by a painter, rather than a mechanically taken photograph *à la* fully automatic mode.

Inevitably, anything that is man-made has the following in common. Firstly, it is bound to be flawed in one way or another. Secondly, it is subjective in the sense that it

is made for a specific purpose, whether expressly stated or not. Third and relatedly, it reflects the character and the thoughts of whoever crafts it. Fourth and finally, it also reflects our human nature, the common traits of the entire human race.

Viewed from the flip side, we can utilise statistics as a messaging device, not merely a factual report. Indeed, none of us would bother to collate a statistical report merely for the purpose of reporting *per se*. We take time and effort to report it precisely because we have some purpose, or vested interest, to do so, such as persuading our audience to support whatever opinion we wish to propagate.

It's just like the law. Our legal systems are supposed to be non-partisan, treating everyone of us equally and evenhandedly. The symbolic icon for legal justice features a scale, idealising impartial equality. In practice, however, we all know that lawyers try to bend the system inasmuch as possible. Probably the best-natured amongst us may passively rely upon the legal system in a non-partisan way, whilst the richest and the most profitably successful amongst us are surely those who actively make the best favourable use of the system.

0.2. What are we NOT here for ?

In this intensive lecture we do not intend to deal with such counter-lies that statistics is always fair and impartial, that science is always value-neutral regardless of who reports it, that official statistical publications must have passed strict quality control and thus must be relied upon without any second thoughts, and like.

If you wish to believe in these sorts of false idealism and unrealistic righteousness, we should like to send you to some other (admittedly much more boring) lectures and (mundane) textbooks on one hand, all the while you are the very one who needs our lecture more than anyone else !

Not that we want to turn you into a big-mouthed liar. We are here to give you some hints in sniffing common statistical abuse, so that next time when some real estate agent knocks on your door or you read some sensationalistic news article, you can see those statistical criminals through their square faces.

Science needs us, so does our civil society. Cheers !

1. How to tell a statistic

1.1. What is a statistic ?

(Hint : Note the singular – not yet "statistic**s**" !)

The mathematical definition of a **statistic** is that it is a function of data. Technically, any arbitrary function of any data can qualify as a statistic. For practical purposes, however, we normally discuss only those statistics which are meaningful to us, such as the (sample) moments (mean, variance, skewness, kurtosis, etc), the (sample) quantiles (median, quartiles, quintiles, deciles, percentiles, $\cdots$), various estimators including regression, correlation, determination coefficients, test statistics such as t-, F-, p-values, and also the sample size.

1.1.1. Average and representative figures

Q : Shall we look at the mean, the median, or the mode ?

It is often disputable which statistic best "represents" the population. The (arithmetic) mean minimises the squared distances to all observed data points, whilst the median minimises the absolute distances. These minimands are called **loss function**s. Namely, the loss function for the mean is the squared distances, whereas that for the median is the absolute distances. In words, the former takes **outliers** (those observations which lie far away from the rest) more seriously than the latter.

It is often taught in various textbooks that the choice of loss functions should depend upon the nature of the data in question. This is surely correct on one hand, whilst the choice should also inevitably depend upon our intention on the other hand. That is, we use highly convex loss functions (such as squared distances) when, and only when, we deliberately emphasise those extreme observations.

Sometimes we cite the **mode** as the representative figure. This is even more arbitrarily manipulable than those aforesaid sample moments. Namely, in the case of discrete (or qualitative) data, the way in which the observations are categorised or discretised critically affects where the mode lies. In continuous (or quantitative) data, on the other hand, since strictly speaking no two data points exactly tie-break, nomination of the mode hinges upon the (estimated) density restored from the observed data set. This restoration procedure, however, is arbitrary, as it requires the choice of **kernel**s which is a matter of the handler's discretion rather than something mathematically unique to the data.

Again, categorisation of the qualitative data, or the restored density from the quantitative data, tells the tale about whoever reports the analysis. In turn, when we do the analysis, we can tell our own stories through our discreetly (no pun intended!) chosen ways of discretisation or kernels.

A : It all depends upon what interests us.

Example 1.1.1.i. You are asked to write a news article which highlights how impoverished our national economy is nowadays. Should you cite our GDP per capita, that is our nationals' arithmetic mean income, or our median income, or the mode instead?

Example 1.1.1.ii. Meanwhile, you read an article (written by someone else, not yourself) which claims that it cites the average (the midpoint) between the 1- and the 99-percentiles to represent our "typical middle-class income and wealth levels so as to trim off the extreme effects of outliers." What hidden intention can you read from this claim?

Example 1.1.1.iii. Are people's heights approximately normally distributed? What about our body weights?

1.1.2. Reparametrisation

Q : Why statistical packages like to use log variables? Why, and how exactly, do they "adjust" R-squared?

Aside from quantiles and the sample size (number of observations), most statistics are not **reparametrisation proof**. For example, regression coefficients derived from logarithmic data are estimated **elasticities** which generally differ from raw regression coefficients. The arithmetic mean of log observations is the log of the **geometric**, not arithmetic, mean of raw observations.

As we see from these well-known examples, reparametrisation often has its own purpose. In other words, it is wrong to reparametrise when we do not intend those purposes served by the reparametrisation – even if the calculation is not mathematically wrong, it tells a wrong story and thus sends a wrong message to our audience.

- Do not take logs unless expressly necessary.
- Do not "adjust" R^2 without thinking.
- Do not meaninglessly convert raw numbers into **ordered statistics**.

A digression pertaining to the latter :
The well-known Arrovian impossibility theorem is based upon each participating ballot reporting the preference *rankings* only, not the raw (*à la* von Neumann Morgenstern) preferences. It is not really the axiomatic "universal domain" in this respect.

A : These packages are supposed to cater for specific purposes (which may or may not concur with what you contemplate) !

Example 1.1.2.i. You run a regression using some statistical software which, to your disappointment, spits out an adjusted R^2 that is negative. What does this mean ? (You check the label of the software which says "Satisfaction guaranteed, or your money back"...)

Example 1.1.2.ii. Demographers, labour economists, and other supposedly socialistic economists are vastly fond of analysing the logarithmic income rather than its raw amount. This reparametrisation offers an advantage in taking a closer look at the lower tail of the income distribution : it equates the difference between £100 and £200 with that between £10 million and £20 million. A drawback, however, of this method is : what should you do with those dirt-poor folks with zero income ? Usually, to avoid the $-\infty$ error, we tend to add a penny, or a quid, to the zero income, or to all observations in the data set. What's your take ?

Example 1.1.2.iii. The commonly cited **body mass index**, that is the body weight in kilograms divided by the square of the height in metres, is known to be distributed around the mean approximately 22.5 . What does this really imply ?

1.1.3. Confidence levels and p-values

Q : What is statistically significant ? Any conventional wisdom at all ?

Ages spent. Your dissertation is almost complete at long last, with only one regression left to run. Cross your fingers oh-so tightly : if the t-stat turns out 1.65, you open a bottle of Champagne to celebrate the glory glory hallelujah, well, no, the gloriously successful completion of your much longed-for degree ; or else, if the t-stat proves to be 1.64, you hang yourself – to be or not to be, that is the question !

All this conventional wisdom – actually the lack thereof – reflects nothing but the shortage of science in our minds. We are *scientists*, meaning that our job is to disclose facts not opinions, the latter being left up to our audience. In our aforementioned example, what should be deemed "statistically significant" should be judged by our audience not by us who report our analytical results. Who says it ought to be, say, 5% ?

Well, if you remember (translation: only those of you who stayed awake back then might be able to recall), your 20th-century ~~fox~~ stats teachers and textbooks used to indoctrinate you with all those benchmark threshold t-values: 1.645 for 10%, 1.956 for 5%, 2.33 for 2% and 2.58 for 1%. Once upon the time, there was even this convention that a single star (*) should indicate 10%-significance, double stars (**) 5%, and triple stars (***) 1%.

Luckily, we now live in this enlightened millennium wherein our fellow scientists finally wake up to the ~~sound of music~~ objective fact that these old-fashioned benchmarks mean very little to our most practical purposes. It is nowadays increasingly common that objective facts such as the t-values and corresponding p-values are listed as they are, leaving the judgment up to the audience – the very fundamental attitude in science.

The significance level indicates the likelihood of false accusation. If you wish to minimise it, generously extending the benefit of doubt, you are willing to lower the significance level so as to reject the null hypothesis only when absolutely legitimate. Otherwise, if you take a vested interest in rejecting the null, fishing for every opportunity to advocate the alternative hypothesis, your best bet shall be to pick a very high significance level – up to 100% in theory.

A : The choice is yours in selecting the significance level, or the method of statistical testing altogether. Why not, eh ?

Example 1.1.3.i. By law, a driver's licence needs to be renewed within one month each, either before of after, of the licensee's birthday. Past years, Ally had her licence renewed on 29 March, 1 May, 4 April, and 24 April. What is the 95% confidence interval for her birthday ?

Example 1.1.3.ii. It is mandated by law that every employee undergo an annual health examination, which normally consists of 30-odd checkpoints in the list, each flagging an alert if the result lies outside of 2.5- and 97.5-percentiles. Assuming that you are spotlessly healthy in its true godly sense, what are the odds that you receive an alert or more ?

Example 1.1.3.iii. Storms of anti-inequality protests raging throughout the world, our research interests are also inevitably drawn into the issue of income disparity. Indeed, we increasingly hear the term "relative poverty" which means poverty relative to the average wealth in the society or the national economy in question. In spite of our real-life feelings, however, it turns out that none of us falls more than two standard deviations below our mean national income. Does this really suffice to rest us at ease that there is no one qualifying "relatively poor" in our seemingly troubled society ?

1.2. Population and DGP (Note: Not GDP.)

When we take our samples, where exactly do they come from? The ultimate purpose of sampling is our interest in wherever the sample is drawn from. Generally speaking, it is either drawn from a fixed pool, which is called the **population**, or generated through a certain system, referred to as the **data generating process**.

1.2.1. Census vs sampling

The **census** can be viewed as a special case of sampling, wherein the sample is 100% of the population. Viewed differently, however, the population is already a natural sample drawn via a natural data generating process. For instance, the entire human gene pool is the natural DGP, from which the actual human population has been drawn.

Example 1.2.1.i. By law, public elections must be concluded with every single ballot having been counted. Suppose, for instance, that a certain jurisdiction has an electorate consisting of no more than 2000 voters. 1100 votes have already been opened, out of which 564 are for candidate X and the rest, 536, are for candidate W. What are the odds that W wins over X after counting the remaining, at most 900, ballots? Were you a news reporter, would you pronounce the victory of X immediately, or wait? What if, instead, you lived in a larger city of 2 million voters, 1.1 million having been counted, out of which 564 thousand are for X and 536 thousand are for W?

Example 1.2.1.ii. In some countries, such as Australia and Greece, voting in public elections is mandatory, that is, an eligible citizen can be fined for not voting. This can be viewed that the system regards the election as a census opinion poll.

In other countries such as the US and Japan, voting is a civil right but not quite a civic duty – or only a moral duty at most. Such a system considers the public election as voluntary sampling from the general public opinions. The sampling procedure there is "random" in that it is up to the arbitrary free will of each voter, whilst it does reflect the degree of eagerness each voter feels toward voting.

Would you rather be a proud Aussie, mate, or an ass-kickin' Yank, dude?

1.2.2. "Past records are not indicators for future performances."

Financial institutions' favourite disclaimer turns out illustrative of the relation between the population and the DGP. The realised past performance, and the actually generated

population, is the **factual**. Future performances, on the other hand, may or may not be carbon copies of the past realisation. This is not only because there may be environmental (or "structural" in statisticians' favourite jargon) changes, i.e., alterations to the DGP. Even when the DGP remains intact, the future may not generally be a clone of the past factual, simply because it'll be an independent new draw from the DGP. The latter, i.e., what another draw from the same old DGP could have been, is a **counterfactual**.

Example 1.2.2.i. Assuming that the DGP remains stable over time, do the past footprints really indicate nothing about the future?

Given that the past observations were drawn from the same DGP as the future performances will be, they tell us at least something about the process. However, the past may not always give us **unbiased** prophecy for the future, and/or about the DGP itself.

Stock performances (equity returns) are often said to be distributed with fat tails, as in the Cauchy distribution for instance. What if you calculate past mean and variance of returns of a certain stock? Do these estimates give unbiased predictors for the future?

Example 1.2.2.ii. Should we be interested in the population, or the DGP, when discussing redistributive public finance policies between the retired and the working generations? What if we contemplate what to do for our children?

2. How to lie on statistics

Sometimes unintendedly, sometimes deliberately, we make statistical reports which are misleading. Well, this statement carries double meanings. In general, any statistical report can be potentially misleading. More specifically, you can make your report even more misleading.

2.1. White lie

The most serious cases of all, are those big-time white lies which are mathematically objectively refutable. These include, albeit not exhaustively, the following.

2.1.1. Sample selection bias

Q : How should we sample randomly, and how randomly should we sample ?

Unless expressly specified otherwise, samples are supposed to be drawn randomly, that is, impartially. This, however, is *much* easier said than done.

It is for this reason that questionnaire results are almost always biased and thus scientifically inaccurate : enthusiasts will return the questionnaire without fail, whilst those who are indifferent tend to neglect it. An analogous issue inheres in research by interviews : the more strongly opinionated, the more likely to show up and to speak up.

Panel data often trim off those observations which fail to cover the entire time span of the data set. This surely is necessary in order to maintain the homogeneity of observations within the data set. However, those who enter or exit during the time span, and those who occasionally fail to return their entries, may well share particular traits which may indeed differ statistically significantly from the rest of the sample.

To be phone interviewed, one needs to have a phone, and more practically, needs to be accessible by phone. This precludes those who are too poor to afford a phone, too busy/lazy to answer, too defensive to answer unsolicited calls, and verbal communication impaired. An analogous critique applies to other media of communication such as the Internet.

But then, here is a big question : how random is really random ?

In fact, there is no unified definition of randomness. It is typically quite an equivocal concept, loosely defined according to various alternative criteria such as unbiasedness,

unpredictability, absence of recognisable patterns, serial independence and independence of other variables. More often than not, these criteria are mutually conflicting, so that we may be forced to choose between them.

A : Focus on that specific aspect of randomness which is useful for your purpose at hand.

Example 2.1.1.i. To research socioeconomic gender disparity among our University of Tokyo degree holders, we interview 100 recent graduates. Unfortunately for our purpose, even our relatively new graduates are still disproportionately male dominated, scarcely 20% of whom are women. To maximise statistical accuracy, we should randomly select :

⟨a⟩ 20 women and 80 men, proportional to their respective population sizes.

⟨b⟩ 50 from among female graduates and 50 from among male graduates.

⟨c⟩ 80 women and 20 men, in order to compensate for their unequal population sizes.

⟨d⟩ slightly more women than men (e.g., 52 women and 48 men), reflecting slightly higher mortality rates among males than among females.

⟨e⟩ any 100 graduates irrespective of gender, to maximise randomness.

Example 2.1.1.ii. There is a branch of mathematical statistics pertaining to random number generation. No man-made algorithm, no man-made criterion for randomness, is flawless.

Those of you who happen to be enthusiasts in number theory may well have heard of the following algorithm : starting with an arbitrary n-digit number, squaring it to extract the middle n (or $n+1$) digits, again squaring it to repeat likewise. After a few turns, the sequence generated shall be pseudo-random... really ?

(Let's start with, say, 250. Square it. $250^2 = 62500$. Take the middle three digits, as we have started with a three-digit number. The middle three digits out of 62500 is, well, 250. Let's square it again... what ? Our "random sequence" looks like 250 250 250 250 $\cdots$ forever ?? Well, folks, this ain't going anywhere.

Let's try again. Start, this time, with something like 3792. It's square is... $3792^2 = 14379264$, of which the middle... well... four digits this time, are... yeah, that's 3792 again... Well, well...

Let's be not quite as lazy, start with something slightly longer, that seems safer isn't it ? Oh, sure, most PIN numbers have only four digits or so, which is why so many IDs are hacked by criminals every day. Now we start with, well, half a dozen digits such as 971582, take the square... that is... $971582^2 = 943971582724$. Take its middle half-dozen digits which are... well, 971582, that is. Take the square of it, which makes $971582^2 = 943971582724$, middle six digits are 971582. Square it again... Here's our random sequence : 971582 971582 971582 $\cdots$. Yes, this does

look more random than our earlier ones, does it not ? But, wait a sec... something's missing here : where are 3's ? Are there any ? Actually, 3 ain't the only one absent. Where are 4's, 0's, and 6's ? Where have we gone wrong ?)

To economise time and energy, let us stop nagging you about what algorithm you secretly enlist. However you do it, if you make it random in one way or another, that's a done deal. Now you are asked to produce a random sequence which is, say, 100 digits long. Now the question – no further question asked about how you do it, the only question here is about the end result – is whether you use exactly ten of each cipher. If you don't, then your "random sequence" favours a specific figure which occurs 11% of the time or more at the expense of another which occurs 9% or less. To avoid this sort of imbalance, each figure must appear exactly ten times which, however, comes with a dear price. Namely, even when one of these 100 digits is masked, it can be predicted with certainty based upon the remaining 99. Do we regard such a highly predictable sequence as "random" at all ?

Probably the most sensible response to this sort of question, is to hinge your answer upon what specific purpose you bear in mind when discussing randomness or generating the random sequence. For instance, if you want to maintain unbiased representation of all different figures, then you visit each figure exactly tence. Otherwise, if you wish to steer clear from predictability inasmuch as possible, then you free yourself from such a rigid requirement.

2.1.2. Endogeneity and wrong causality

Q : When, why, and for what purpose, should a regression be run ?

Regression-based empirics and "rat psychology" are the two most pervasive (and invasive) epidemics in our contemporary academia. The most likely reason for such excessive popularity of the regression analysis, is simply because it is easy to run. With every little help offered by various statistical packages, running a regression takes just one button to press. Also, it is more versatile than other methods such as the correlation analysis, in that multiple regressors can be run all in one go, whereas correlation is an inherently *pairwise* concept which would take up to ${}_{\nu}\mathrm{C}_2$ runs to handle ν variables.

There is, however, a price to pay for this seeming convenience. A dear price, it is. Namely a regression, unlike correlation, requires us to pre-specify which variable is **dependent** and which others are **independent** (or **explanatory**). As we know much too painfully, this is indeed the source of all sorts of complicated issues.

Endogeneity, also known as **reverse causality**. When regressing Y on X, if not only X causes Y but also Y causes X, the regression will suffer **endogeneity bias**, that is, the estimated regression coefficients are biased.

Why "regression" ? The aforementioned bias can be illustrated through the following intuition. Regressing (linearly for simplicity) Y on X (and the constant) and regressing X on Y (ditto) generally entail different estimated regression equations, one not being the inverse of the other but flatter than that. This is where the term "regression" originates.

More mathematically precisely, the cosine of the angle between the two estimated regression equations gives the **correlation coefficient** between the two variables.

False causality. It is often misunderstood that regressing Y on X assumes the structure that X causes Y. The fact is, X in this case is called the *explanatory* variable, not the causal variable. Y regressed upon X simply indicates that X is used to predict Y, which need not require X causing Y in the structural sense. When X and Y are concurrent, whether one causing the other or the two being commonly affected by an external cause, one is generally useful in predicting the other.

Indeed, structural causality is inherently a matter outside of numerical data and mathematical technicality. It is the non-numerical background information about the data, not the numbers included in the data set *per se*, that is informative of the structure, including causality. Without this information, the numerical data alone tells us little about the causal structure, all the while enabling us to run regressions and thus to predict a variable based upon another.

A : Regressions are run to quantify sheer mechanical predictability, without undue regard to the structural background.

Example 2.1.2.i. A good number of students have taken a language test and a maths test. When their language scores are regressed onto their maths scores, the estimated regression coefficient turns out highly statistically significant. Does this imply :

- that maths scores are useful in predicting language scores ?
- that maths is useful in improving language scores ?

Example 2.1.2.ii. It is not uncommon that, when we regress the agricultural harvest onto the rainfall, the resultant coefficient proves statistically insignificant, that is, not only insufficiently significant but sometimes the sign is wrong : the more rainfall, the less harvest ! This is paradoxical, as rainfall is absolutely necessary for crops. What is your take ?

(Hint : The regression result here is not just "inconclusive" : The data are ample

and reliable, with large enough sample sizes and little measurement errors, yet nonetheless the regression coefficient proves conclusively insignificant. How is this possible?)

2.1.3. Model specification

Q: What is a good model?

The way in which model specification is taught in statistics textbooks and lectures, tends to focus mostly upon **goodness of fit**. That is, a model which fits the data is typically praised as a good, correct model. As is well known, however, that this notion of goodness of a model has a pitfall – actually more than one.

Firstly, it opens the door for **spurious regressions**. That is, there are those regressions and models which coincidentally fit well but are inherently nonsensical.

Secondly, if we keep adding new **regressors** (explanatory variables), the model is bound to fit better. When the data is small, the degrees of freedom set a limit to the number of regressors we can afford. When the data is sufficiently sizeable, however, there is practically no limit to how many regressors we can introduce without running out of the degrees of freedom. Does this mean that we should keep adding whatever arbitrary regressors after regressors? (A regression which does this is sometimes mocked as a "kitchen sink regression.")

Ofttimes misunderstood as it is, the issue here is not that we should compensate for the addition of extra regressors by suitably penalising the enlargement of the model, as in various **prediction criteria**. These criteria are, after all, indicators for goodness of fit, amongst which are numerous versions of adjusted R^2. Instead, the fundamental issue here is that the goodness of a model and its goodness of fit aren't synonymous at all. The latter is purely mathematical, whilst the former is a contextual matter outside of maths and numbers.

A: A good model for you is a model that aptly serves for your purpose. It may not always be the same as a good model for your neighbour, or that for a mathematician.

Example 2.1.3.i. How many statisticians have you ever met in your life thus far? The number increases as you age. Meanwhile, your blood pressure also tends to rise as you age. If we regress the number of statisticians each of you has met onto your blood pressure, the estimated regression coefficient may well be statistically

significant, the model fitting the data reasonably well. Conclusion: hypertension fosters acquaintanceship with numerous statisticians!? What? Say it again!?

Example 2.1.3.ii. Viewed mathematically, not adding a regressor is equivalent to the constraint that it is identically equal to nought. The unconstrained model, which includes the regressor, performs at least as well as the constrained model. This is always true *independently of whether the regressor in question should really be there* in terms of model specification.

2.2. Statistical abuse

Verbal abuse can constitute a punishable offense. Statistical abuse may not (yet) under our legislations at present. Whatever, genuinely factual data and mathematically correct analyses may still serve to bend the truth. How can this happen?

2.2.1. Data mining

Q: How do we know whether the data has been collected impartially, by looking at the data alone?

If the randomness in sampling the data is questionable, it can be a matter of mathematical accuracy altogether. It simply casts a doubt against the quality of the specific data set at hand.

But what if the data used is reliable and impartial indeed? Does this mean that the analysis is not bending anything at all?

Suppose there is some foregone conclusion to be "empirically supported." Whatever the foregone conclusion may be, even including when you know it's false, if you keep running the same regression using a different data set each time, sooner or later you happen to support the conclusion by chance. Write it up, and publish it! – This is called **data mining**.

Obviously, your unsuspecting audience will find no way to tell how many other data sets you have unsuccessfully tried on, alongside the one and only one you prominently report. And this is hardly any fault of your audience, as the real issue lies not in what they see, but in what's (deliberately) hidden from them.

Ironically, data mining is made possible by the very randomness of sampling procedures. Each data set is sampled randomly, which stochastically implies that there is a chance in

every 100 that the draw falls outside of the 99% confidence set, which is equivalent to a faulty 1-percent statistical significance result.

A : The data, and the statistical analysis thereupon, may not show you the whole picture – that is, no matter how well-educated you are about statistical methods.

Example 2.2.1.i. Contemplate a certain explanatory variable which truly does not affect that dependent variable wherein we are interested. We run the regression, say, at least 100 times, every time using different data sets.
(Sure, why not ? If we used the same data set to run the same regression 100 times, we should simply obtain 100 carbon copies of the same result – Einstein's definition of insanity is to repeat the same thing over and over again and yet to expect different outcomes.)
Then expectedly in one of the 100-odd regressions, the estimated coefficient may happen to be 1%-significant, by the very definition of the significance level. Now, what if we report only that regression with statistical significance, not mentioning the rest ?

Example 2.2.1.ii. If an asset price (be it in raw amounts or logs) truly random-walks, what is the (prior) probability that it exits the predicted 95% band at least once between now and the infinite future ?

2.2.2. "Structural estimation"

Q : Can we statistically estimate the structure ?

As aforementioned, statistics generally does not tell us about the structure of the model. We write the model theoretically, and then use data and statistical methods to examine how plausible the model is.

Possibly on purpose, so-called **structural estimation** can be a misleading mnemonic. Stats, by nature, cannot draft up a story about the structure. It is our theory, that is outside of the data and maths, whereby we decide which variables are **structural primitives**. Structural estimation is simply to estimate the regression coefficients on these variable.

A : Yes and no. Yes in that we run a regression to estimate structural parameters. No, however, in that what parameters are structural can only be decided by theory not by statistics.

Example 2.2.2.i. The world's most legendary weathermen were those in the U.S. military in 1944, predicting the exact two-hour window that the eye of the storm passed the channel, whereby leading the D-day invasion to a landmarking success.

In contrast, our National Meteorological Bureau is a standing joke. Its official weather forecasts are correct approximately only 58% of the time. It is known, meanwhile, that the hit rate by mechanically saying "tomorrow's weather will be the same as today's" day after day, could be as high as 62%.

Example 2.2.2.ii. Repeatedly roll a fair die, and record the outcomes as the X variable. After sufficiently many rolls, regress X^2 linearly onto X (and the constant). What is the expected R^2? If it's considerably less than 1, does it mean that X cannot predict X^2 without nonnegligible errors???

2.2.3. One-side vs two-side tests

Q: How do we choose when to use one-side tests and when two sides?

In theory, the choice directly reflects what null and alternative hypotheses we contemplate. That is, if the null is $\theta \leq 0$ and the alternative is $\theta > 0$, then we use a single-sided test, otherwise if the null is $\theta = 0$ whilst the alternative is $\theta \neq 0$ (this is sometimes referred to as **nested hypothesis** testing, meaning that the null is a special needlepoint case of the alternative) then we use a both-sided test. Well, at least, this was the kind of correct answer whereby we all passed our stats exams and thus successfully completed our degrees.

In practice, however, it is highly uncommon that we hypothesise a parameter with absolutely no premeditation about its sign whatsoever. There is also this logical incoherence problem that when we premeditate and thus use the single-sided test, it is more (twice as) supportive of the premeditated single-sided alternative hypothesis than when we are genuinely agnostic, in which sense the premeditation becomes a self-fulfilling prophecy.

A: Honestly, the choice is arbitrary. Use the single side if you wish to reject the null and thus advocate the alternative; otherwise, if you are reluctant to reject the null and are missioned to claim statistical insignificance, use the both sides.

Example 2.2.3.i. Are women in the University of Tokyo any better (or worse) than their male classmates? To test this at 5% significance, we draw random samples to estimate the regression coefficient on the gender dummy, of which the t-statistic turns out to be 1.75. What conclusion shall we draw?

Example 2.2.3.ii. A rule of thumb for those who are either innumerate or memory impaired, tells you to reject the null iff (if and only if) the t-value lies outside of ± 2. Logically, however, this implies that we adopt a significance level of 4.6% when testing two sides, whilst adopting that of 2.3% in testing one side. Does this sound inconsistent ? Or does this make sound sense to you ?

2.3. Statistical misuse

Albeit possibly less sinful than deliberate abuse, misuse of statistics can lead to serious misinterpretations and wrong recommendations.

2.3.1. Blind belief in numbers

Q : The model is theoretically questionable, but the regression based thereupon runs beautifully, all estimated parameters looking nice and plausible. How shall we make sense of this ? (*Is it possible that numbers are better than theory ?*)

As aforesaid, we are supposed to test our theory by means of statistical data and methods. This means that, if we write a convincing story and it withstands due statistical tests, then the story is scientifically approved. Even if the story sound nice, if it fails statistical tests, then the story needs revision.

If the story is unconvincing from the outset, then there is no use testing it statistically. Even if it happens to pass statistical tests, it is likely to be *spurious*, that is, a numerically coincidental yet otherwise useless model.

A : A theoretically questionable model is questionable no matter what. Resolve the theoretical question first, before crunching numbers (*let alone before publishing, please !*)

Example 2.3.1.i. It is alleged that global warming trails our carbon dioxide emissions by as long as half a century. To confirm this, we try regressing daily high temperatures in central London in 2012 onto those in central Tokyo in 1964. This involves 366 observations. If the estimated regression coefficient turns out highly statistically significant, how will you read it ?

Example 2.3.1.ii. A preliminary doctoral dissertation presentation in Princeton University circa 1993 analysing a panel data with about 17 regressors, of which five were binary dummies taking either 0 or 1, had two slides to project. One was a table showing the raw regression, and the other showed the result taking logarithms of all 17 regressors.

2.3.2. Oh-so fancy hi-tech gadgets

Q : Why every econometrician, every single empirical research paper, seems to use two-stage least squares, instrumental variables, nested logit, and bootstrap ? Where has the good olden OLS gone nowadays ?

In the fewest possible words, the ultimate reason is because we want to use them, which is further because these gadgets enable us to do what we want to do, if not what we objectively should do.

Example 2.3.3.i. Instrumental variables.
The choice of instruments is in our hands, whereby we gain a timely excuse to draw attention to those variables we want to emphasise, even when they are structurally unrelated to the model in question. Indeed, there are few other occasions which allow us to cite a variable that is intrinsically outside of the subject matter.

Example 2.3.3.ii. Nested logit.
Used mostly for computational convenience. In exchange for the specific restriction on the error-term distribution (the **logistic distribution**), the method offers an edge in its invariance property – somewhat akin to so-called **conjugacy** between prior and posterior beliefs. How many of us actually think through this tradeoff ? Isn't it that the method is popular just because it is popularly accepted by many ?

Example 2.3.3.iii. Bootstrap, also known as **re-sampling**.
A flagship in nonparametric statistics nowadays. It has gained its vast popularity mostly based upon its ingenious simplicity and, to put it bluntly, its relative lack of theory.

Different times, different rules. Last century, OLS and linear probability were vastly popular. Now, the same model needs to be handled with all these gadgets, otherwise we wouldn't be able to publish it. Next century, and next millennium, there will be whole new sets of imaginative methods which mandate us to use them, even though there is no scientifically impartial explanation why.

A : Take it easy, dude ! If the use of these gadgets gets your research published, go ahead by all means. None of them is of essential importance, however. What is important is that you think, rather than what you use.

2.3.3. Is significance everything ?

Q : Shall we ignore those parameters which turn out statistically insignificant ?

Our standard statistics textbooks teach us statistical inference first, and then hypothesis testing. Why is this the case, even though we are all freaking out about testing ?

In the case of statistical estimations, inference corresponds to the estimated parameters *per se*, and testing corresponds to their significance. The natural order, therefore, should be that we look at the **estimate**s (the realised values of **estimator**s) first and, only if they interest us, we further look into their significance.

A : Insignificance is not exactly a synonym to negligibility. The first sorting key should be the estimates per se.

Example 2.3.3.i. We intend to inspect the early childhood effects of (A) cod liver oil supplements containing omega-3 acids, and (B) preschool schooling, on the children's IQ ten years later when they are 15. The estimated coefficients turns out to be 0.084 on the dummy variable for (A) with the standard error 0.048 , and 10.01 on the (B) dummy with the standard error 6.37 . Shall we buy our children a cod liver oil supplement first, or rather send them to a preschool academy first ?

Example 2.3.3.ii. Imagine regressing your exam scores onto two explanatory variables : your family income when you were three years old, and the caloric intake from your breakfast on the day of the exam. Suppose it turned out that the t-values associated with the estimated regression coefficients were, let's say, 3.01 for the former, which is 0.25%-significant, and 0.99 for the latter, which is 32%-(in)significant (two-sided). Now, what can you do with these findings ? One is bygone, far too late to change, no matter how highly statistically significant it might be. The other, though, is something you can fully control, even if it's much less significant in the mathematical statistical sense. Which interests you the better : the more statistically significant, or the more *useful* ?

3. How to steal with statistics

Now it's time for learning by doing. Well, what exactly are we going to start? Statistical crimes, that is. Indeed, the best person to lecture in front of detectives, is a former thief.

3.1. Creative writing in statistics

We can concoct all sorts of highly imaginative stories. Their real-life truth isn't really what interests us. We write them not because we want to be honest, but because we want to write them. There are two ways to do so: one is honest, the other is, well, somewhat less honest. The former is to claim openly that the story is fictional. The latter is to bolster the story with statistical data and mathematical analysis.

3.1.1. Prove what you like

Aside from data mining and spurious models, there are various applications (?) in order to prove ostensibly whatever you like.

Example 3.1.1.i. "Our economy is growing fast."

We all doubt it. Surely, if we look at the long-term trajectory of our national economic indices such as GDP per capita, it is hardly shooting upward at all. But look closely. It does.

Well, look at it year by year. Within each year, try plotting the monthly GDP and regress it onto the 12 months, January=1 through December=12, to estimate the coefficient which is monthly growth, and its statistical significance. You'll be pleasantly surprised. Our economy is kickin' the ass, dudes!

Example 3.1.1.ii. "Being short is a health hazard."

Being too fat or too thin may be. But how can the height matter? You'll be surprised, though, it surely does.

Given your height, the more obese you grow, the higher your health risks will be. This means what? Given your weight, the shorter you are, the more obese you are, which is an obvious health threat. Statistics will surely approve this, and it'll be extremely highly significant!

3.1.2. Draw your favourite conclusion

We are mere mortals. Life is but a dream. But life is a dream. Given that you have to live anyway, why not enjoy it?

Well, now, if you are to bother with stats anyway, you might as well draw a conclusion to your liking.

Example 3.1.2.i. "Our nuke plant is absolutely safe, with zero probability of accidents" (even though a major accident has already destroyed it!)

Consider, in abstract general terms, **i.i.d.** (identically and independently distributed) binary trials, each either succeeding with probability $1-\varphi$ or failing with probability φ, ending as soon as a failure occurs for the first time. The parameter we are estimating, of course, is the probability φ of failure. What is the **unbiased estimator** $\widehat{\varphi}$? Is there any at all?

The definition of unbiased estimation is $E[\widehat{\varphi}|\varphi] = \varphi$ for any φ. Therefore, we must first ensure $E[\widehat{\varphi}|\varphi = 1] = 1$. When $\varphi = 1$, however, the system surely ends after only one trial which is bound to fail. Thus, whenever the system fails upon its first trial, we must declare $\widehat{\varphi} = 1$.

We now proceed on to a general φ. The probability that the first trial fails is φ, entailing $\widehat{\varphi} = 1$. Therefore, to abide by the aforementioned definition of unbiased estimation, we need to counterbalance by setting $\widehat{\varphi} = 0$ in all other cases, that is, as soon as the first trial succeeds, regardless of how subsequent trials unfold.

Now we contemplate the serial trials to be our beloved nuke's daily accidents or no accidents – that is, counting only a destructive accident as a failure. The nuke ends its operation as soon as a "failure," that is a destructive accident, hits for the first time. The nuke did survive its first day of operation without devastating failure, so we now have $\widehat{\varphi} = 0$, and this conclusion shall not be altered even if a terminating accident follows later.

Example 3.1.2.ii. "Private schools are crap."

Shall we send our children to expensive private high schools to increase their chance of getting into, say, the University of Tokyo? Or shall we save money with state schools? A recently published empirical study estimates the coefficient for the private schooling dummy with the t-value 1.9 and the associated single-sided p-value 0.028.

Holy crap! There's a three-percent chance that we might be paying a fortune to *reduce* the success probability, all the while scarcely 0.4% of every cohort make their ways to the University of Tokyo anyway!

3.2. Play with numbers for fun

Mathematics is meant to facilitate our analytical thinking. So does it for sure. Can you imagine the nightmare of having to verbalise all our statistical, empirical, econometric exercises without equations and numbers whatsoever ?

Sometimes, however, an untimely mention to numbers may backfire. More often than we notice, that is. There are things better computed than verbalised ; other things better spoken than quantified.

3.2.1. Unduly "descriptive" statistics

A mere mention to descriptive statistics, let alone stats and econometrics, may go a long way, for better or worse.

Example 3.2.1.i. "We have expanded our motorway by 17%."

A three-lane highway underwent a repair, closing one of the lanes, whereby its capacity decreased by 33 percent. When the construction work completed, the lane reopened, widening the highway from two lanes to three, thus increasing the capacity by 50%. (Melbourne, Australia, circa 1998, about the airport access motorway.)

Example 3.2.1.ii. How long will it take on average to drive from London, Ontario to Manhattan, Kansas ? (Hint : Some travellers never make it.)

Example 3.2.1.iii. "100 percent of those who won the prize last year, had bought the lottery." (Public lottery authority, France.)

Example 3.2.1.iv. There is a remarkable formula to create a prime number. And it's so strikingly simple that anyone can easily memorise it. Take any natural number n , and then $n^2 + n + 41$ is a prime number.

Let's try. When $n = 1$, $n^2 + n + 41 = 43$ is indeed a prime number. Tick !
When $n = 2$, $n^2 + n + 41 = 47$ is again a prime number. Tick !!
When $n = 3$, $n^2 + n + 41 = 53$ is also a prime number. Tick !!!

Still suspicious ? Try a few more when you have difficulty falling asleep, and you'll be convinced. Good night.

3.2.2. Regression for aggression, transgression, and digression

Statistics isn't something of which every little helps. Poorly done, or wilfully bent, statistical operations are worse than nothing.

Example 3.2.2.i. The population densities and the GDP per capita of the eight regions in Japan have been reported as follows:

regions	Hokkaido	Tohoku	Kanto	Chubu	Kinki	Chugoku	Shikoku	Kyushu
km^{-2}	67	151	1298	359	692	240	215	355
JPY 10^3	3395	3518	4567	4295	3959	3940	3397	3275

To earn additional 1000 yen per head, how many additional folks shall we "import" per square kilometre of our homeland?

Example 3.2.2.ii. Precedence may or may not imply causation, but causation necessarily requires precedence. This must be logically true.

But then, suppose we are trying to relate the probability of a mother suffering breast cancer during her 60s, and her (biological, not adopted) baby's blood sugar level when he or she was six months old.
Question A: Which causes which? (Hint: Who causes whom?)
Question B: Which precedes which?

3.3. Be partial, be partisan, reward your friends and punish your enemies!

Not always, but sometimes, one can change the way the model is written, whereby quite drastically distinct conclusions can be drawn even based upon the same data set.

Typical examples involve the way in which the null and the alternative hypotheses are rewritten.

3.3.1. Test Tactics

Say it again... What? Did you say "t-statistics" or "test statistics"?

Yeah, that's the question. Doesn't matter, actually, whichever will do.

Data can be God-given. Well, actually no, they're given by humans, but we often obtain data from external sources wherefor we assume little responsibility on our own.

But statistics aren't given to us. We derive them, sometimes concoct them for our own use.

Example 3.3.1.i. We all learn the **permanent income hypothesis** from our macroeconomics textbooks and lectures. We are even requested to memorise it and recite it, in order to pass exams. Allegedly, however, it has repeatedly been "empirically rejected." Bad news, ain't it? Are we all brainwashed to believe in, or to memorise at least, a lie? Hell, no, we don't want to think so. We want the hypothesis to be true, why not?

Looking closely, the way wherein the PIH has been rejected is typically as follows. Regressing the consumption onto the contemporaneous income and the permanent income, the estimated coefficients not only on the latter but also on the former proves statistically significant, whereby the PIH that consumption depends only upon the latter not on the former, is rejected.

Now, let's be a little laid-back about our interpretation of the PIH, taking it that consumption is affected by the permanent income, not only by the contemporaneous income. Sure enough, the estimated coefficient on the permanent income is indeed statistically significant, hence "supporting" the PIH. Bingo!

Example 3.3.1.ii. Some stock brokers try to persuade you that stock returns are distributed normally, whilst others tell you that they're distributed *á la* Cauchy. At least one of them must be telling you a lie here. Whodunit? (Me? Good point, but that's not what we're talkin'bout.)

If you want to "prove" it's normal, write your null as such, and you will be more accepting in its favour – typical odds being 90-10 or 95-5. Otherwise write the null that it's ~~abnormal~~ Cauchy.

3.3.2. Testing statistics

Say it again... What? Do you mean "test statistics" or "statistical testing"?

Some models defy our intention of statistical testing. This time, it's generally not the model's fault, nor are the model writers at any fault. Instead, it really is that the very notion of statistical testing turns out a criminal attempt.

A student from the former communist block used to declare that his research objective was "to test his home-country data using advanced Western theory." There is, after all, a grain of truth in his words. We can't quite test our data, but we surely need to test

statistics – yep, you've heard it right, not that we need test statistics, but that we need testing of statistics.

Example 3.3.2.i. Unit root testing.

In **time series**, we often take our interest in whether random shocks subside according to an **impulse response** coefficient $\rho < 1$ (more precisely, $|\rho| < 1$) to keep the process **stationary**, or shocks are persistent (**unit root**, or "random walk").

Parametrised in ρ, these two hypotheses **nest** in such a way that the unit root $\rho = 1$ is a special case of non-unit roots $|\rho| < 1$. Viewed from the flip side, that is, where the impulse response converges, then the stationary process where the shock tends to nought over time, is a special case of nonstationary processes where the shock tends to persist at any nonzero level.

Example 3.3.2.ii. A certain product claims that the average weight per package is 2 pounds. We are determined to test this claim statistically, buying random samples of the product and weighing them all. If we keep adding more and more observations, what shall our conclusion be?

(A) Is the mean weight statistically significantly different from 2 lbs?

(B) Is the mean weight (not statistically) significantly different from 2 lbs?

4. How to kill with statistics

Nothing in the entire human history has claimed more lives than religious conflicts. Statistics is no exception. When statistics becomes a religion, it kills us.

How can statistics become superstitious, then? In other words, what do abusive statistics and superstitious religion have in common? It's the stubbornly staunch belief in a certain set of foregone conclusions.

Statistics, like any other branch of science, is meant to dissolve such superstitious preoccupations. When abused craftily, however, it can serve adversely, doing more harm than good.

4.1. Statistical Creationism

Statistical analysis is based upon probabilities and likelihoods. Therefore by nature, it is hard to be absolutely certain. Even when something is 0.01% significant, there remains a 1/10000 chance that it is actually false. Huge data sets and astronomical degrees of freedom shall be required in order to eliminate the last microscopic straw of doubt.

4.1.1. Refuse to prove what you don't like

To make matters even worse, the risk we run about drawing wrong conclusions, is not a probability, but a **likelihood**. The statistical significance level is not the probability that we commit type-I errors, but the (supremum of) likelihood of type-I errors when the null is indeed truthful.

Example 4.1.1.i. Hard core believers in tobacco never admit that smoking kills, in spite of all the scientific evidence disclosed publicly. (Sure, they can easily bet their lives can't they!? "I'm immortal, I'll never die. I bet my life on it!")

Curiously, it is not that they consciously trade their lives in for their addiction – which could be economically rational in the sense of subjective utility maximisation. Instead, they are refusing to accept the fact, just like Catholic fundamentalists refuse to admit that the earth isn't flat. This is deadly irrational.

They maintain that no empirical evidence to this date has ever been conclusive. But then, don't they plan a family outing for next weekend? Is it any more "conclusive" that they won't die before next weekend? Sometimes, we must take things for granted even if they're not 100.00000% certain!

Example 4.1.1.ii. As many attempts as twinkle-twinkle little stars up in the sky have been made in empirically testing whether there are any hereditary racial differences in human intellect. Thus far, no conclusive evidence has been discovered to indicate any difference. Would it help if we kept adding millions after millions of observations in our database to run an astronomical regression, assuming we could afford all the logistical costs in running it ?

4.1.2. Twist the conclusion in your favour

We are mere mortals. Not all models we write may run exactly as we want. If the statistical test result is not in your favour, don't easily give up. Try talking around it inasmuch as you can.

Example 4.1.2.i. "Drink up, and drive home safely."

Whilst drunken driving, meaning operating motor vehicles under the intoxicating influence of alcohol, is legally banned in effectively all countries in the world, drinking in moderation before driving is lawful in many countries including highly developed ones. We must question : is it safe ?

We then take a data set to run a single-sided 5% test with the alternative hypothesis that drinking, even in moderation, increases the probability of accidents, and the null that moderate drinking and driving is at least as safe as not drinking and driving. The t-statistic (drum roll, please !) is 1.643 .

Here we are, happier than ever, fondly and proudly toasting : "The landmarking hypothesis that drinking in moderation before driving is actually safer than not drinking, has now been statistically accepted !"

Hooray ! Big cheers, folks ! Would you prefer red, or white ?

Example 4.1.2.ii. "WHO admits passive smoking is harmless after all."

The World Health Organisation cites quite a number of studies on passive smoking, in one and only one of which the empirical result turns out "inconclusive." This, both the study itself and WHO publicly admit, is due to the limitation of data, and they agree that this is not to be (mis)taken as if evidence that passive smoking were safe.

Reactionary old boys around the world, however, all rejoiced and jumped on the bandwaggon misinterpreting the "inconclusive" result. Sure, it is in their God-given rights to misinterpret whatever they wish to. Just don't attribute it to WHO or anyone else who (no pun, dude !) isn't quite as stupid.

4.2. Prey on numbers for profit

Citing unnecessary numbers and data, is one thing. Although it's admittedly a statistical crime, it's not yet a capital offense.

Creating misleading numbers, is another thing. Surely it's more serious an offense. It is surprising, though, to find out how many politicians, authors, governments and other public bodies commit this crime. They are entirely unabashed at the time; neither are they apologetic later.

4.2.1. Framing effect

In economics, the concept of **framing** was introduced by a pioneering behavioural study by Kahneman and Tversky for the first time. The idea is that, in an experimental environment, subjects respond differently to substantively identical questions depending upon how the questions are worded, or "framed."

Example 4.2.1.i. "70% of our nationals support the death penalty."

What a national disgrace, especially now that more than 120 out of 200-odd countries worldwide already have agreed to abolish the death penalty. Are we more stupid, backward, or blood-thirsty than the remainder of the human race?

It turns out that, if there is anything stupid, backward, or blood-thirsty, it is the questionnaire:
"Do you (1) absolutely oppose to, (2) support with some reservation, or (3) absolutely support the death penalty?"

Example 4.2.1.ii. "What do you think should we as a society do with the increasing trend in serious crimes?"

National statistics invariably prove that crimes have been decreasing throughout the postwar period. That is, our forefathers used to commit several time more of almost every kind of crimes than our contemporaries do nowadays.

The fact is, in our present low-crime society, we tend to take public safety for granted and, in turn, to take crimes more seriously than our ancestors did in their high-crime society in the past. Hence, it is likely that our notion of "serious crimes" encompasses a broader range of crimes than that in older generations.

4.2.2. Who says it ?

One of the corollaries to the relativity theory is that time cannot exist by itself. That is, time exists if and only if there is someone to record it, or something, some time-consuming change, to substantiate it.

Somewhat analogously, a fact cannot exist by itself. It is produced by whoever witnesses and reports it.

Example 4.2.2.i. A world-famous physicist in mid 20th century, William Shockley was also (in)famous for his hateful views toward people of African origins (that's all of us, sure, but we're referring to dark-skinned ones here). He cited various statistics to bolster his claim that the "inferiority of the blacks and black Americans must be hereditary."

Well, who was he ? What was his own skin colour ? And who were the ones collating and publishing those statistics he conveniently cited ? Suppose we (contemporary folks like you and me, that is) were deported to pre-modern Africa to spend the rest of our lifetimes as slaves to native African masters there. Only two or three generations later, even if our descendants were freed to enjoy full civil rights for the first time, would they instantly be able to make the same academic achievements as their native African peers ?

Example 4.2.2.ii. It is widely known that Nazis published various statistical results claiming hereditary inferiority of Jews. The U.S. statistics on their wartime enemy, Imperial Japanese soldiers, cited their average height five inches (!!) shorter than that according to the Imperial Japanese Army's own official statistics. Allegedly, however, these statistics were constructed "scientifically," that is, based upon objectively verifiable data and supposedly reliable analytical methods.

Example 4.2.2.iii. There was an economics professor who, apparently by mistake, became the president of one of the world's most prestigious universities. He was openly against the so-called affirmative action, like many others, but unlike them, his view was not based upon fairness or equal opportunities, but male supremacist. More precisely, his claim was essentially that boys' scientific abilities tend to have larger variances than girls', so that higher education such as universities are better off admitting more men than women, as we need to look only at the top cream.

4.3. Fourth Reich, George III Bush, Benito XVI, Dishonest Abe (not Lincoln) and World War III

Statistical crimes are not at all *passé simple*. They are contemporary as well (as ill?) – enclosing and suffocating us. Be alert!

- Doctors recommend us to slim-fast. Insurance companies, on the other hand, report that the weight to minimise mortality rates is actually quite a bit heavier than what the medical profession advocates. Indeed, these insurers have no incentive to lie, because a wrong recommendation would raise their customers' mortality rates which would cost them multi-million payments! Who are the remaining suspects? Yeah, the doctors? They run their businesses when we fall ill. Now, what's our take-home of the day? Supersize me?

- Less than 40% of our youngsters go to universities every year, which defies our real-life intuition.

- Only 1 in 3 of us will be married when we die, ditto.

- Are you the type who gladly unbuckle as soon as your flight lands, or the other type who dutifully wear the seat belt until the plane is parked and you are instructed to alight? Only 1 in 3 who are killed in airplane accidents, are killed in airport premises.

- Political historians even nowadays believe in what they call **democratic peace**, that warfare is unlikely to be launched between two democratic countries. They refer to the United States as a democratic country even before the civil war, in those days when women and non-whites had no votes, and Switzerland before 1971 where women had no civil rights. On the other hand, they regard pre-war Japan as undemocratic, predominantly because women and the poor (those who did not pay income taxes) had no votes... or possibly because non-whites *did* have votes!

- Roughly 5% of Japanese men eventually take their own lives. It is also rumoured that the suicide rate amongst the University of Tokyo graduates is thrice the national average. – Didn't wanna know this, did you?

- In Japan, men live seven years less than women. In the States, blacks live seven years less than whites. In Australia, Aboriginals live twenty years less than whites. Implications?

- Eager as fickle-minded folks are in rallying against the nuke after the mega-quake in 2011, how many death tolls have ever been conclusively attributed to radiations

from the nuke plant ? A big fat round 0, that is. On the other hand, disenchanted as financial news articles are about the reduced car production by Japanese manufacturers, how many death tolls are recorded every year due to car accidents ? In 000s please ?

- "Guns don't kill. People do." (National Rifle Association, the Confederate States.)

- According to a recent Japanese source, more than 20% of men and less than 10% of women never marry. Wait a sec, the numbers don't add up... unless millions commit criminal bigamy !
  (Just in case you suspect more Japanese women marry non-Japanese men than Japanese men marrying non-Japanese women, the fact is the exact opposite, the latter being nearly thrice the former.)

- It is officially claimed, even in history textbooks, that the Axis lost the World War II because they had less resources than the Allies, often citing various statistics about natural, human, and financial resources. Sorry, folks, this is a myth, though. Statistics are irrelevant in this case, only serving to obfuscate the essential picture. The true reason why the Axis were bound to lose the war from the outset, was because they were stupid.

  For example (1), the Imperial Japanese Navy simulated several major battles in the midst of the Pacific separately from one another. In one of the simulated battles, three out of, say, 15 warships would be sunk and lost. In another, planned a month later, the same 15 warships would be enlisted.

  For example (2), the Imperial Japanese commander would urge their soldiers to die for the Emperor, whilst the American commander would encourage their soldiers to survive and come home as war heroes. Put together, the only outcome wherein both of these wishes can come true, is that the Americans unilaterally slaughter all the Japanese. Everyone happy ? Any complaint ?

- Democratic or not, our society is a class society, a house divided along university degrees. An American source tells us that fully 93% of married couples are either both university degree holders or neither – leaving scarcely 7% being "mixed marriages." How can this be justified ? Let's boil this down to two questions ?
  (1) *Were* you smarter than your classmates who did not go to universities ?
  (2) *Are* you smarter than those who did not go to universities ?

  To save time – I know you're plenty busy, dudes – let's distill these into one question :
  (3) Are you smarter than you were ?

Certainly, in your not-yet-so-old glory, you were once capable of doing those problems as listed below. With ease. That was no more than a few years ago.

How have you been since?

See if you are truly more respectable now than when you were in high school...

— (*What follows is solely for your possible intellectual curiosity;*
need not necessarily be done as homework prior to our 13 March session.) —

4.3.0. Prove that

4.3.0.i. there are infinitely many positive **prime numbers**.

4.3.0.ii. $\sqrt{2}$ is an **irrational number**.

4.3.0.iii. there are **uncountably** many real numbers between 0 and 1.

4.3.1. Into how many regions can a ball be subdivided by n mutually intersecting planes? Express the maximum number of regions as a function of n.
(Hint: One region when $n = 0$, two regions when $n = 1$, four regions when $n = 2$, eight regions when $n = 3$, fifteen regions when $n = 4$, and so on.)

4.3.2. Identify $10^{10^{100}}$ modulo 97, that is the **remainder** when $10^{10^{100}}$ (known as a **googolplex**) is divided by 97.

4.3.3. A square glass panel is subdivided into four small square cells, each of which is to be tinted in red, orange, yellow, green, blue, indigo, or violet. Assuming that adjacent cells cannot have the same colour (diagonal ones can), how many colour configurations are admissible altogether? We do not distinguish those configurations which can be made identical by mirroring and/or rotation.

4.3.4. Each of four identical balls is touching the other three. We now consider two concentric balls, the smaller of which is positioned in between the aforementioned four balls touching all four of them, whilst the larger is the smallest large ball enveloping the aforesaid four. Compute the ratio between the volumes of these two concentric balls.

4.3.5. Consider a sequence consisting of n bits. Each bit can be either 0 or 1, provided that the *product* of any two consecutive bits must be zero. How many different sequences are admissible? Express the number of such sequences as a function of the length n.

4.3.6. A unit cube rotates around one of its longest diagonals. Volume the region passed by the *surface* of the rotating cube.

4.3.7. By drawing $n+1$ vertical lines $x = \dfrac{2^k-1}{2^n-1}$ where $k = 0, 1, \cdots, n$ and $n+1$ horizontal lines $y = \dfrac{2^\ell-1}{2^n-1}$ where $\ell = 0, 1, \cdots, n$, there are $({}_{n+1}\mathrm{C}_2)^2 = \left(\dfrac{n(n+1)}{2}\right)^2$ rectangles created altogether.
Formulate the sum of the areas of all of these rectangles as a function of n.

4.3.8. Factorise $x^{31} - x$ into as many as possible **polynomials** with integer coefficients only.

4.3.9. Write, as a function of k, the surface area of the region created by the filled circle $x^2 + y^2 \le 1$ rotating around the axis $x = k$. Without loss of generality $k \ge 0$ may be assumed.

4.3.10. Identify the remainders when the following **hexadecimal** integer is divided by 5, 7, and 9 respectively:

$$122333333444\cdots\mathrm{FFFFF}$$

that is, 1! = 1 hexadigit of 1 followed by 2! = 2 hexadigits of 2's followed by 3! = 6 hexadigits of 3's and so forth, ending in F! hexadigits of F's, where A, B, C, D, E, and F represent ten, eleven, twelve, thirteen, fourteen, and fifteen.

4.3.11. Volume $\{(x, y, z) |\, x^2+y^2 \le 1,\ y^2+z^2 \le 1,\ z^2+x^2 \le 1\}$, that is the intersection of three perpendicularly positioned unit cylinders.

4.3.12. Six vertices of a unit regular hexagon are serially numbered from 1 through 6. When three dice are rolled to call three vertices, what is the expected area of the triangle generated by these vertices? The area is zero when the same vertex is repeated. (*University of Tokyo, admission, 1981.*)

4.3.13. Let P, Q, and R be the first trisectors of the respective edges AB, BC, and CA of a triangle ABC (that is, $\mathrm{AP} = \frac{1}{3}\mathrm{AB}$, $\mathrm{BQ} = \frac{1}{3}\mathrm{BC}$, $\mathrm{CR} = \frac{1}{3}\mathrm{CA}$). Then, the area of the small triangle bordered by straight lines AQ, BR, and CP is a constant fraction of the area of the original triangle ABC. Find the fraction.
(*University of Tokyo, admission, 1962.*)

4.3.14. Express

4.3.14.i. each element of $\begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}^n$ as a function of m_{11}, m_{12}, m_{21}, m_{22}, and n.

4.3.14.ii. $f[n]$, such that $f[n+2] = \ell_1 f[n+1] + \ell_2 f[n]$ where $n = 0, 1, 2, \cdots$, as a function of ℓ_1, ℓ_2, $f[0]$, $f[1]$, and n.

4.3.14.iii. $g[n]$, such that $g[n+1] = \dfrac{k_{11}g[n] + k_{12}}{k_{21}g[n] + k_{22}}$ where $n = 0, 1, 2, \cdots$,
as a function of k_{11}, k_{12}, k_{21}, k_{22}, $g[0]$, and n.

4.3.15. A rectangular floor is completely covered with unoverlapping rectangular tiles, each of which has at least one integer dimension (i.e., either its length or its width is divisible by the unit of measurement, or both are). Prove that the whole floor must also have at least one integer dimension.