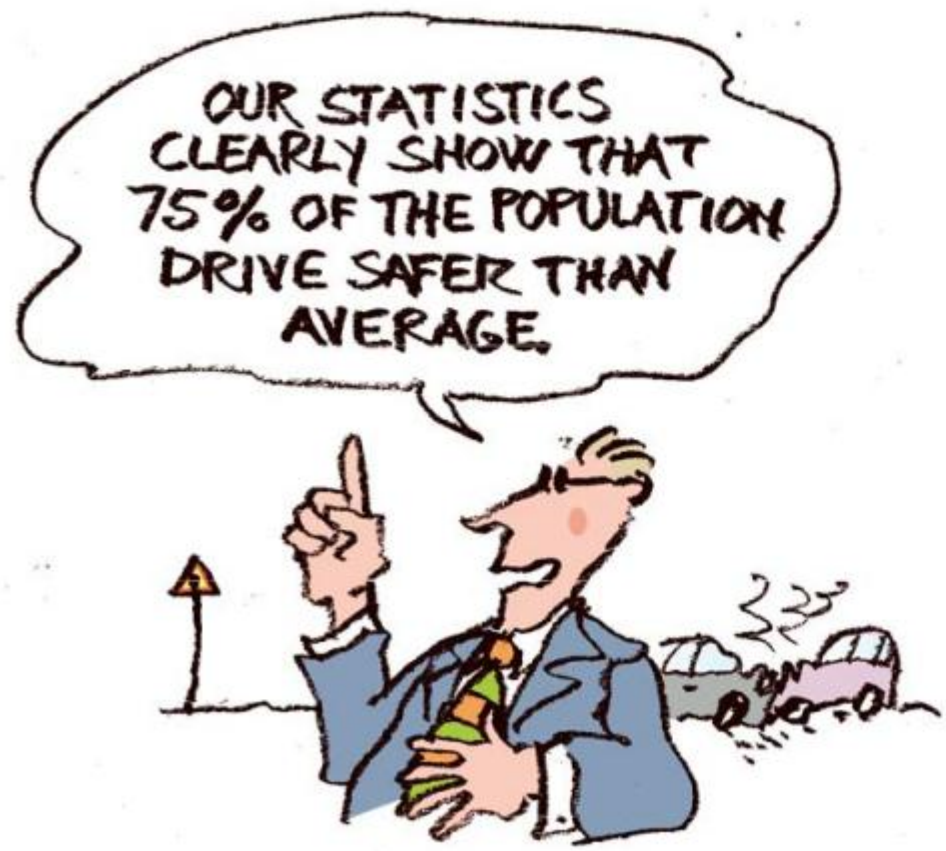
Online Course in Science Journalism

Created by the WFSJ and SciDev.Net



Lesson 9 - Understanding statistics

by Hans Van Maanen

9.1 Introduction

Is it important for science journalists to know something about statistics? Not if they have a statistician friend they can call anytime day or night. All others will want to learn at least the fundamentals of percentages, averages, deviations and significance tests. Maybe not actively — no one expects science journalists to do a chi-square test on their data — but passive knowledge always comes in handy.

This chapter tries to teach you the first principles of numbers and statistics, but it is not 'statistics for dummies' in the usual sense. The subject matter is too extensive for just one chapter, and Wikipedia's statistics entries are remarkably clear and reliable. Rather, **you will learn about the things that can go wrong, and often do go wrong in reporting science — both by scientists and by journalists.**

When you have read this chapter, you will be more wary of all kinds of numerical claims that are made in science, and you will be able to ask more difficult questions. At the very least, you will not be completely defenceless when the numbers pop up. And they pop up often.

9.2 Numbers

What is the size of a dust mite? How many blood cells does a human have? What is the area of a soccer field? How high the moon?

One of the most important skills for a science journalist — for any journalist — is to quickly **put numbers in perspective**. That is the only way to grasp them, to understand the magnitude of things, and to give them meaning. Numbers are abstractions (that is why they are easy to use), but to make them meaningful to your audience you will have to make them concrete again. As the cliché goes: **give them human proportions**.

It is very difficult to understand the size of a garden of 0.735 hectares, but everybody will see the garden in their mind's eye if you say it is 'the size of a soccer field'. A dust mite is about 0.125 millimetres, so on an area 1 square millimetres (smaller than the dot at the end of this sentence) we can fit about 8 times 8 equals 64 dust mites. It is always a good idea in any story to **'re-humanise' these numbers with a good comparison**, but it is also important because it gives you an easy way to check their magnitude. Only in this way will you notice if a number is, say, a thousand times too big or too small.

Checking the numbers

A thousand times is, usually, a lot. We can easily live for a million seconds — that is less than 12 days — but a billion seconds is 32 years and a trillion seconds is 32,000 years — 32,000 years ago, the last Neanderthal died. Remember, the Anglo-Saxon 'trillion' (12 zeros) is a European 'billion', Anglo-Saxon 'billion' (9 zeros) is a European 'a thousand million').

Remember that some large numbers may have been made up on the spot. The number of visitors at a political rally, the number of girls being abducted each year, the television audience at Olympic Games... A good question is always: how do they know? Is it even possible to know? And is it a lot or not, in this context?

9.3 Percentages

Percentages are deceptively simple. One percent is one hundredth, so 8 percent of 80 kilograms is 6.4 kilograms, and if we increase our live stock from 50 to 70 cows, we have expanded by 40 percent.

The first thing to know, is that not everybody knows what ‘40 percent’ means. In Germany, a study has shown that this turns out to be ready knowledge for just a bit more than half the population (54 percent). **What is simple for us, may not be so simple for our readers.**

And percentages are seldom so simple. We divide a numerator and a denominator and express the result as one number, so somewhere, some information gets lost. Especially if the wording is not very exact — which is usually the case. The scientist tells us that antidepressants lead to problems of a sexual nature (like impotence or loss of libido) in 40 percent of the cases, but does that mean that one user will have 40 percent more problems with intercourse, or that 40 percent of users will experience problems?

Perspective

Like with numbers, our first job is to put percentages into perspective. An 8 percent increase can mean a lot (for instance, for your weight) or it can be a negligible fluctuation (with blood pressure or cholesterol). Even a 40 percent decrease in risk of stroke has little meaning if the risk of stroke is very small to begin with. It may sound impressive, but that doesn’t mean it is interesting — or clinically relevant, as they say. The larger the normal fluctuation, the less meaningful a small percentage change is.

Percent point

Sometimes ‘percent’ and ‘percent point’ get confused — either on purpose or by accident. If last year the unemployment rate was 4.8 percent and this year it is 6.0 percent, that’s an increase of (6.0 – 4.8) / 4.8 = 0.25 equals 25 percent. Or is it just 6.0 – 4.8 = 1.2 percent? We can use either, but to avoid confusion, economists use the term ‘percent point’ in the latter case. Here, the government would rather use percent point than percent, the opposition the other way around. There is no rule — except to ask for the underlying data. If only because authorities have been known to change definitions for unemployment and inflation and what not to make them look more gratifying or less disastrous. Sometimes it takes sharp questions to get a look behind the numbers.

9.4 Averages

One boy has 4 coins, the second has 5 coins and the third has 9 coins. On average they have (4 + 5 + 9) / 3 = 6 coins. If the third boy doubles his income, the average goes up to (4 + 5 + 18) / 3 = 9 coins — even though two out of three boys have nothing to be thankful for.

Some people use 'mean' when they mean 'average', some use the one and then the other, and some say there is a subtle difference between the two. Like percentages, averages look simple, and like percentages, this can be deceiving. If 19 labourers get 100 coins and the boss gets 2100, everyone earns on average (19 x 100 + 2100) / 20 = 200 coins, but that is not the way the union sees it.

Median

With the boys, although their average is 6, no boy has 6 coins. If most families have either no or two kids, we don't need many houses for the 'average family'. Almost all people have more than the average number of feet. Two of the boys have less than the average. With averages, we tend to think of them as the middle value, but that is not necessarily true. Some 'distributions' are very skewed — like income in almost all countries — and in those cases, the average is rather meaningless. Here, statisticians prefer the 'median' — that is, the middle value: half have less, half have more. For the boys, the median is 5, both before and after the third one got lucky, so the median is less sensitive to extreme values. In our company, the median is 100 coins.

Averages can be very tricky, and it is easy for nonchalant or unknowing journalists to be fooled by them. Firemen complained that their average income had decreased remarkably last year, and they demanded compensation. Until someone noticed that last year they had hired a lot of new, young firemen — who started with low wages, driving the average down.

In many western countries, the average age of mothers that have their first baby is going up: 'More older mothers', read the headlines. But is the average going up because women get their first child later in life, or because there are less teen pregnancies? Or both? What years do the scientists compare, and why exactly these two years? Averages tend to obscure the natural variation. Sometimes that's good, but sometimes too much information is thrown away — **the variation may be more interesting than the mean or the median**. Of course, scientists devised a way to overcome this limitation: they use the deviation or spread to summarise this variation — yet another number.

9.5 Spread

The boys had 4, 5 and 9 coins and an average of 6 coins. The girls have 5, 6 and 7 coins, also an average of 6 coins. But the 'spread' in the distributions of coins is much larger for the boys than for the girls — there is more variation in their capital.

Statisticians have come up with the '**standard deviation**' to describe this property of distributions mathematically. In almost every science article we have long tables with data like 'Age: average 65 years, standard deviation (SD) 5 years'. The important thing for journalists to know, is that **most distributions are about 4 standard deviations wide**. In this case, 4 x 5 = 20 years. So, most participants were between 55 and 75 years old.

This is usually true even if the **distribution is skewed**. It is not uncommon to see, for instance, 'Alcohol: 15 g/d, SD 15 g/d'. This means, according to our rule of thumb, that alcohol intake in this group varies between –15 grams per day and 45 grams per day. But no one drinks negative g/d, so the distribution is skewed, from 0 g/d to 60 g/d. Most people drink a little, but some drink a lot. With a bit more statistics, and data, it is not hard to deduce what percentage of the participants have more than five drinks a day. This is what we see in smoking as well — for instance, average of 15 cigarettes a day, standard deviation of 25 cigarettes a day. **These things are important to note: it may mean that in this respect, the group is not that 'balanced'.**

About 95 percent of the population will be covered by these 4 standard deviations. If you want 99 percent, you have to take 6 standard deviations, while 2 standard deviations cover about two thirds of the population. So, one compact number, the standard deviation, tells us a lot about reality.

Sometimes '**variance**' is used instead of 'standard deviation': the variance is the square of the standard deviation.

Of course, there are some pitfalls in using the standard deviation and the rule of thumb blindly, too, but handled with care, they are great tools. **If you take the percentage of change and divide it by the standard deviation, you get a very good first impression of the meaning of the result**. The average IQ in a country is 100 points with a standard deviation of 15 points (by agreement). The next year the score has increased to 105. A 5-percent change, to 105 points, equals 5/15 = 0.33 standard deviation (the 'z-score' is 0.33). That is not bad — in the social sciences, a z-score of 0.2 is usually considered a 'weak effect', 0.5 is considered 'medium', and 0.8 is a 'strong' effect. Children grow about 6 centimetre a year, which is about 1 standard deviation — a strong and very visible effect.

9.6 Rates

Scientists love rates and ratios. Relative risks, odds ratios, hazard ratios, standardised mortality rates... What's a poor reporter to do? Again, as with the percentages and averages, keep a cool head, ask for the underlying data, and try to put them in perspective.

Relative risk

Let's start with a real-world example. Taking statins reduces the risk of having a stroke within four years by about 50 percent — the relative risk of stroke versus placebo is 0.5. Of the people who get a placebo for four years, about 3 percent get a stroke. In the group who gets statins, this is reduced to 1.5 percent. In other words, 30 in 1000 people with a placebo drug get a stroke, 15 in 1000 with a statin. So to prevent 15 strokes, 1000 people have to take statins — the treatment number needed to prevent one stroke in four years is 1000/15 = 67. The best way to visualise this effect, is to draw 67 little men, and pick one — that's the one saved by statins. The other 66 only have the side effects of the drug.

Hazard ratio

Hazard ratios are, to all intents and purposes, equivalent to relative risks (if the risks are not too big, let's say 10 percent). Usually, risks are 'corrected' for all kinds of variables, so that investigators do not compare men with women, or women over fifty with women under twenty, or smokers with non-smokers. But especially social-economic status is very difficult to compensate for; something to watch out for when going through the article. An article may be statistically completely sound — if it is biased one way or another, it is very difficult to interpret.

Sometimes it takes a lot of searching to find the right data — sometimes they are nowhere to be found. Many scientific articles, especially on the benefits and dangers of food — coffee, tea, dairy, red meat, but also television or physical activity — come up with important-looking relative risks that, after a lot of calculations, turn out to be **negligible absolute risk** reductions.

9.6 Rates (continued)

Odds ratio

A special trap deserves mentioning here, and that is the odds ratio. 'Odds ratios' are usually interpreted as 'relative risks' but that is dangerous (unless the risks are not too big, let's say 10 percent). For every 100 babies born in one country, 52 are boys. The sex ratio (the 'odds' of having a boy against having a girl) is 52/48 = 1.08. In another country, it's 51 to 49, a ratio of 1.04. The odds ratio is 1.08/1.04 = 1.04. So can we say there are 4 percent more boys born? No: the relative risk is only 2 percent: (52/100)/(51/100) = 1.02. The odds counts 'yes' against 'no', the risk 'yes' against 'total'. Scientists use odds ratios with abandon (they are very easy in their calculations), but **they should not be used in the media**. Or by the media.

Another example. A 'gene for asthma' has been found, according to the news: children with the gene have a 51 percent higher risk than children without the gene. Reading the scientific article, it turns out that 62 percent of the children with the gene have asthma, 52 percent of the children without the gene. The odds ratio is (62/38)/(52/48) = 1.51. The absolute difference between the two groups is only 10 percentage points. The gene is interesting, but not the definitive answer in preventing childhood asthma. **Most people interpret genes as yes/no-affairs, so this is a place where journalists — and scientists — have to be extra careful**.

This is not to say, of course, that small differences in absolute risks are always negligible. In 1995, it was shown that the contraceptive pill of the third generation lead to twice as many cases of thrombosis than the second-generation pill. The difference was tiny — 6/10,000 against 3/10,000 users, and thrombosis is rare, but since so many women use the pill, there is still a considerable and easily preventable effect.

As always, numbers and statistics are important, but real insight is even more important. And as always, try to imagine real people, and what risks they are exposed to.

9.7 Luck

If there is one thing that distinguishes statisticians from ordinary people, it is luck. Statisticians know that luck, randomness, plays a much bigger part in our lives and in all kinds of events than we would think possible. If someone wins the lottery twice in a row, it will make headlines— but the statistician will note that, with all the lotteries going on in the world, this is bound to happen sometime, somewhere. We meet someone we just dreamed about, big surprise and small world — but the statistician will say we all dream of many people every night, so it is not such a big coincidence if it happens to you.

Moreover, **people always tend to see patterns, even where there are none**. That is the way our brains are wired — we see faces in the moon, tigers in the bush, and anything that happens three times in a row is 'remarkable'. Especially if we can think of a causal link: remarkably more cases of leukaemia near a power plant, and the headlines read 'Power plants cause leukaemia'. Even with a fair coin, it is not so rare to get seven heads in a row — but try telling that to the suspicious audience. The soccer team gets a new trainer and right away they win three games in a row — who gets the credit, the trainer or chance?

Science, one could say, is the endeavour to separate real occurrences from chance occurrences, real relationships from chance relationships. The trainer has to do better than chance. In general, it is fair to say that any naive statement about chance and randomness will be false. **Not only do people underestimate the power of chance, they also have no inborn way of dealing with it.**

The same holds for risks. People tend to underestimate risks they are familiar with and have some control over (like driving a car). **They overestimate risks that are unfamiliar and forced upon them (like environmental pollution)**. As with percentages and averages, it is always a good idea to translate risks into a language your readers are familiar with. Do not say '1 in 50', say 'picking the predicted card from a poker deck'; do not say '1 in 10,000', compare it with the number of car casualties, lightning casualties, etc.

9.8 Correlation versus relation

The biggest pitfall of all, of course, is the **mistake of equating correlation with causation**. Every time there's a traffic jam, there's police, but policemen are generally not the cause of the problems. It is easy to think of similar, and similarly silly, examples, but it is surprising to see how many unwarranted conclusions are drawn as soon as the examples are not so silly. People seem to have a special fondness for simple cause-and-effect explanations.

Children with large feet write better than children with small feet, not because shoe size is important, but because older children, who have bigger feet, have had more writing lessons. If you find yourself sleeping with your clothes on and you have a headache, do not think that sleeping fully-dressed causes headaches: both are correlated with a third factor.

Correlation can be completely coincidental. As global temperatures rise, so do the number of pirates. Easy to spot — but what about hurricanes, number of cars, swim-wear sold, population of lizards?

A special case is the **'risk factor' in medicine**. High blood pressure and a high cholesterol are risk factors for heart disease, so people are well-advised to eat healthier. But a high cholesterol is not the cause of heart disease, it's just that in the group of people with high cholesterol more people get heart disease than in the group with normal cholesterol — and since there are many more people with normal cholesterol, most people who die of heart disease have a perfectly normal cholesterol.

Walking with a stick is a risk factor for falls: there are more falls in a group of men with sticks than in a group of men without. Taking away the risk factor, only makes things worse...

Example

Take a look at the article Mobile phone masts linked to mysterious spikes in birth [http://www.guardian.co.uk/science/blog/2010/dec/17/mobile-phone-masts-birth-rate] from the Guardian. It explains how journalists might draw the wrong conclusion from correlated data.

9.9 Significance

These days, 'significance' is the magic word in science. There was a time, not too long ago, when nobody thought of significance, but now everything has to be 'significant' to even get published in the scientific journals. Even so, most scientists do not understand themselves what it really means for a result to be 'significant'. It is a difficult concept.

Significance is science's way of separating real from chance findings. If a scientist has a significant finding, what she is actually saying is: 'If you don't believe me, how else would you explain this, eh?' Whereupon her opponent might say: 'Well, yes, you may be right.' Or: 'That is nice and all, but have you thought of A, B, C and D?'

Testing hypotheses

Significance testing, in the modern sense of the word, is about comparing two hypotheses, and doing an experiment to see which one is right. My hypothesis might be that readers of this course are more intelligent than the population at large (average 100, standard deviation 15). This hypothesis is compared with the 'null-hypothesis' that there is no difference. If my test result clearly cannot be just luck, my result is 'significant'.

Note that, strictly speaking, we must formulate both hypotheses before we do any experiment, agree on what we consider 'just luck', put out a level playing field, and then start the experiment. We can't look at the data, notice a peak somewhere, and say: 'That's a significant finding'. Nor can we look at all our data, and then construct our hypothesis. And we certainly can't keep on gathering data until we finally have a significant result.

So are readers of this course 'significantly brighter' than the rest of the world? Let's say quite a few of them have taken an IQ test, and it turns out they have a mean IQ score of 105, with a standard deviation of 2 points — obviously, some will score a bit higher, some a bit lower than the average, plus there is measurement error, so there is a spread in the distribution of the scores. Again, since 95 percent of the distribution falls within 4 standard deviations— here, from 101 to 109 — we expect only 2.5 percent of our readers scoring below 101 (and 2.5 percent above 109). That is a very small percentage — too small, we think, to be considered a chance result. So we conclude that our readers have a 'significantly higher' IQ than the rest of the population.

We could have used **z-scores** instead: we find a difference from the expected average of 5 points, and the z-score (the difference divided by the standard deviation) is 5/2 = 2.5. A z-score higher than 2 (or lower than –2, of course) is considered 'significant at the 5-percent level'.

This 5-percent level is almost religiously followed in the medical literature, but it is just a kind of tacit agreement. If we wanted to be more confident in our conclusion, we might have chosen a 1-percent level, for a z-score of 2.6. On the other hand, if we are that stringent, we run the risk of rejecting our idea when there is in fact a difference — but not as big as we hypothesised.

9.9 Significance (continued)

Sample size

One complicating factor: the standard deviation of the average of our readers' IQ depends not only on the spread in their scores, but also on the number of people taking the test. If we test more people, our measurement gets more precise. But this also means that any test can be made significant by using a lot of people (patients, mice, electrons) — and vice versa, that a real difference might be missed because of too small a sample size. Sometimes, we need to test large groups to find subtle differences, but sometimes the difference doesn't mean anything anymore — the scientists are measuring noise. Conscientious scientists say beforehand what effect they think is minimally interesting, and then calculate the size of the samples they need. This prevents using too many subjects, and too much money, as well.

Instead of just claiming 'significance' for our findings, it is better to show the **'95-percent confidence interval'**: that is a bit more informative. For our IQ test, we found a point estimate of 105 points with a 95-percent confidence interval of 101 to 109 points. In ordinary language: we found an IQ of 105 in our readers' sample, and we are confident that 19 out of 20 similar samples you take would give an IQ between 101 and 109. Since the confidence interval we found, from 101 to 109, does not include the null-hypothesis of 100, the finding is significant.

9.10 Some final pitfalls

The easiest way to bluff your way into statistics is also the most widely-used: leave out half of the story. In a way, we have come full circle here — again, it is all a matter of context. **Journalists not only need to be careful about what they are being told, but also about what they are not being told.**

Going back to the antidepressants — 40 percent of users will experience sexual problems — the clever question was left out. How many non-users will have problems during intercourse? If they experience problems about 30 percent of the time, things look totally different — that's an increase of only 10 percent.

It is always very important to remember the **control group** — if there is no control group, it isn't science. We read in the paper that a new pill reduces violence (or heart attacks, or risk of cancer) with 24 percent, but how about the reduction in the group that did not get the pill? If a gene doubles the risk of ADHD, how many children with ADHD have the gene, and how many without the gene? Is that 2 in 100 against 1 in 100? That wouldn't justify the claim that 'the ADHD gene' has been found. What if it's 7 against 14 percent?

Here is another pitfall. If impotence is rare, say 1 in 100 (intercourses or individuals), a 10 percent increase (to 1 in 91) may be bearable, but if it is frequent, it may well spoil a relationship. The **'base rate'** is very important, in other problems as well: things that are rare, are hard to find. Even if they are not too rare, a test may pick up more 'false-positives' than 'true-positives', simply because there are too many healthy people. Similarly, mass testing for terrorists in airports is not feasible: you would pick up more innocent people than terrorists. The same problems plague mass screening for breast cancer, prostate cancer etcetera. **Finding a needle in a haystack is impossible, not because you won't find the needle, but because you keep thinking you found it.**

Pilots who did well during test flights were complimented by their instructors, and promptly did less well. So the instructors stopped making compliments. But the pilots were bound to do a bit worse, because they were at their very best — they were good, and had a bit of luck. Next time, they were still good, but had a bit less luck, so did worse. I tend to go to the doctor when my complaints become unbearable — they are bad, plus sometimes they are worse than other times. The doctor just looks, and the next day I feel better — the complaints are still bad, but random fluctuations go the other way. Authorities put cameras on the most dangerous intersections, and think this measure helps drive down accidents. Sons of talented football players are good too, but usually not as good as their daddies.

Some of the best statisticians have made fools of themselves due to this phenomenon, '**regression to the mean**'. It is the main reason why modern patient research always includes a 'placebo': everybody gets better thanks to the regression to the mean, but do people with the real pill get significantly better? Always be on the lookout for regression to the mean, especially when it is claimed that the people worst-off (or the industries in worst shape) benefit most from treatment (or state support).

9.11 Conclusion

We end with another word of warning. Probably no field of science is so dotted and ridden with difficulties, hidden assumptions and traps as statistics. It seems as if it is simply too easy to make a mistake.

So, even though you have taken this course, do not try to become an amateur statistician. **When in doubt, call a statistician**. And if you don't know one, get one to be a friend you can call anytime, day or night. Statisticians are the most helpful bunch of scientists, I have found.

9.12 Self-teaching questions (1-6)

Numbers

1. The nearest star is Proxima Centauri, 4.2 light-years away. A light-year is 10 trillion kilometres. Try to reduce this to a human scale — and do not say 'light from the star takes 4.2 years to reach us'!
2. According to the news agencies, 200,000 people came to listen to Barack Obama in Berlin in 2008. Can we just report that?
3. What is the total area of your lungs, in an understandable unit?

Percentages

4. Students have to pay over one third more on their loans for tuition: the interest rate goes up from 2.7 to 3.7 percent. How would you report this?
5. How much is 1/3 of 27 percent of 405?
6. Decaffeinated coffee causes apolipoprotein B in the blood to go up by 8 percent (http://www.sciencedaily.com/releases/2005/11/051120122949.htm). What numbers do you need to know before deciding on reporting this fact? What else is there to say about the story?

9.12 Self-teaching questions (1-6) answers

Numbers

1. The nearest star is Proxima Centauri, 4.2 light-years away. A light-year is 10 trillion kilometres. Try to reduce this to a human scale — and do not say 'light from the star takes 4.2 years to reach us'!

ANWSER: If the earth is represented by a full stop at the end of this sentence (0.5 mm), the nearest star would be 1570 km away — about the length of the Blue Nile.

2. According to the news agencies, 200,000 people came to listen to Barack Obama in Berlin in 2008. Can we just report that?

ANSWER: In a crowd, we estimate to have 2 to 4 people per square metre, so the crowd would fit in a square field of 300 metres by 300 metres (insert proper comparison here!). However, the avenue in Berlin was only 50 metres wide...

3. What is the total area of your lungs, in an understandable unit?

ANSWER: The total surface area of adult lungs is 70 square metres — let's say half a tennis court.

Percentages

4. Students have to pay over one third more on their loans for tuition: the interest rate goes up from 2.7 to 3.7 percent. How would you report this?

ANSWER: Report the difference in real coins — or better still, what it means for their income.

5. How much is 1/3 of 27 percent of 405?

ANSWER: 1/3 x 27/100 x 405 = (27 x 405) / (3 x 100) = 36.5. Don't give too many decimals in the result: 1/3 is not very exact either.

6. Decaffeinated coffee causes apolipoprotein B in the blood to go up by 8 percent (http://www.sciencedaily.com/releases/2005/11/051120122949.htm). What numbers do you need to know before deciding on reporting this fact? What else is there to say about the story?

ANSWER: Find out the 'normal' value of apolipoprotein B in a population; find out if a higher apolipoprotein B increases the risk of disease; find out if it's not just one reportable finding among many...

9.13 Self-teaching questions (7-12)

Averages

7. In 1982, the famous palaeontologist Stephen Jay Gould was diagnosed with a tumour, and doctors told him the median survival time was eight months. So SJG was fairly optimistic about his chances. Why?

He was right, by the way: he lived for another twenty years.

Spread

8. The average waiting time at two doctor's offices is 30 minutes. One has a standard deviation of 0 minutes, the other of 15 minutes. Which doctor would you prefer?

9. American children play on their computers for an average of 2.6 hours a week, according to one investigation (http://archpedi.ama-assn.org/cgi/content/full/159/7/607). The standard deviation is 5.3 hours. Rephrase this in layman's terms.

Rates

10. Breast cancer screening by mammography lowers risk of dying for women around sixty by 25 percent. About 4 in 1000 women in this age group die of breast cancer. Can you rephrase the percentage? What number gives a better idea of what's going on?

11. The number of women ministers in our government has tripled with the new cabinet. Is that good news?

12. Workers with stress report more common colds (odds ratio 1.20), according to the abstract: http://www.ncbi.nlm.nih.gov/pubmed/11259797 — and many new items all over the world. In the article, we see that 1576 of 2837 stressed workers reported colds, and 1570 of 3059 relaxed workers. What do you report?

9.13 Self-teaching questions (7-12) answers

Averages

7. In 1982, the famous palaeontologist Stephen Jay Gould was diagnosed with a tumour, and doctors told him the median survival time was eight months. So SJG was fairly optimistic about his chances. Why?

ANSWER: Half the people die within eight months, but maybe the other half regain their normal lifespan. Especially if the cancer is diagnosed early, and other factors are advantageous as well.

Spread

8. The average waiting time at two doctor's offices is 30 minutes. One has a standard deviation of 0 minutes, the other of 15 minutes. Which doctor would you prefer?

ANSWER: If the standard deviation is 0, you always wait exactly 30 minutes. If the standard deviation is 15 minutes, sometimes you can walk straight through, sometimes you'll have to wait an hour.

9. American children play on their computers for an average of 2.6 hours a week, according to one investigation (http://archpedi.ama-assn.org/cgi/content/full/159/7/607). The standard deviation is 5.3 hours. Rephrase this in layman's terms.

ANSWER: The average is a bit less than three hours a week, but there must be children playing more than three hours a day.

Rates

10. Breast cancer screening by mammography lowers risk of dying for women around sixty by 25 percent. About 4 in 1000 women in this age group die of breast cancer. Can you rephrase the percentage? What number gives a better idea of what's going on?

ANSWER: Without screening, 4 in 1000 women will die prematurely of breast cancer, with screening, 3 in 1000.

11. The number of women ministers in our government has tripled with the new cabinet. Is that good news?

ANSWER: Change is slow if it goes up from 1 to 3, but substantial if it goes up from 6 to 18 — especially if there are 18 ministers in the cabinet.

12. Workers with stress report more common colds (odds ratio 1.20), according to the abstract: http://www.ncbi.nlm.nih.gov/pubmed/11259797 — and many new items all over the world. In the article, we see that 1576 of 2837 stressed workers reported colds, and 1570 of 3059 relaxed workers. What do you report?

ANSWER: In a group of 100 stressed workers, 56 will report a cold, in 100 non-stressed workers, 51 will.

9.14 Self-teaching questions (13-16)

Correlation versus relation

13. Correlation does not equal causation — the fact that two phenomena are related, does not mean that one is the cause of the other. To practice:

a. Children born with IVF, ('test tube babies') have more complications around birth than other children. Why?

b. Taller people earn more than shorter people. Why would that be?

c. In 95 percent of the deadly bar fights, the victim was the one who started the whole thing. What do you think?

Significance

14. Medicine A relieves pain by an average of 2.5 points and a standard deviation of 1 point (measured on a standard pain scale). Obviously, this is significant: the 95-percent confidence interval goes from 0.5 to 4.5 points. Medicine B relieves pain too, but not significantly so: average 5 points, with a standard deviation of 3 points. The regulatory commission would allow A on the market, but refuse B. Which would you prefer, A or B?

15. Did the goldsmith really use 18-carat gold in the crown he made for the king? Archimedes had to find out. The density of 18-carat gold is 15.5 gram per cubic centimetre, Archimedes found the density of the crown to be 15.1 gram per cubic centimetre. Rather close, but still he advised the king to fire the goldsmith. Why would that be?

In scientific disputes, the error margin is often more important than the point estimate.

16. Taking this course, the WFSJ claims, 'significantly raises IQ, and more than 10,000 students have taken part!' Should you be impressed?

What if only 100 students had followed the course and had also increased their IQ significantly at the 5-percent level?

9.14 Self-teaching questions (13-16) answers

Correlation versus relation

13. Correlation does not equal causation — the fact that two phenomena are related, does not mean that one is the cause of the other. To practice:

a. Children born with IVF, (‘test tube babies’) have more complications around birth than other children. Why?

ANSWER: In general, mothers who have IVF are somewhat older and IVF leads to more multiple births, and there is a correlation between both factors and birth complications. Corrected for these two factors, there is no difference whatsoever. It took scientists a while to figure this out.

b. Taller people earn more than shorter people. Why would that be?

ANSWER: Men are, on average, taller than women, and men earn, on average, more than women. However, if we correct for sex, we still find a difference: taller men earn more than shorter men, and taller women earn more than shorter women. Why would that be?

c. In 95 percent of the deadly bar fights, the victim was the one who started the whole thing. What do you think?

ANSWER: In fatal bar fights, police reports show, the one who started the brawl usually is the one ending up dead. But only because the survivor gets a chance to tell who started it.

Significance

14. Medicine A relieves pain by an average of 2.5 points and a standard deviation of 1 point (measured on a standard pain scale). Obviously, this is significant: the 95-percent confidence interval goes from 0.5 to 4.5 points. Medicine B relieves pain too, but not significantly so: average 5 points, with a standard deviation of 3 points. The regulatory commission would allow A on the market, but refuse B. Which would you prefer, A or B?

ANSWER: Many people would probably prefer B, and run the small risk of more pain against the prospect of much larger pain relief.

15. Did the goldsmith really use 18-carat gold in the crown he made for the king? Archimedes had to find out. The density of 18-carat gold is 15.5 gram per cubic centimetre, Archimedes found the density of the crown to be 15.1 gram per cubic centimetre. Rather close, but still he advised the king to fire the goldsmith. Why would that be?

ANSWER: Because Archimedes found the density of the crown ‘statistically significant’ below 15.5 gram per cubic centimetre. He might have measured the crown quite a few times to be sure, and came up with a standard deviation of 0.15 gram per cubic centimetre, for a 95-percent confidence interval of 14.8 to 15.4 gram per cubic centimetre. It’s the spread that counts, not the single number.

16. Taking this course, the WFSJ claims, ‘significantly raises IQ, and more than 10,000 students have taken part!’ Should you be impressed?

What if only 100 students had followed the course and had also increased their IQ significantly at the 5-percent level?

ANSWER: If 10,000 students have taken part, a difference of 0.2 IQ point would have been significant at the 5-percent level. If 100 students have taken the test, 2 IQ points is significant at this level.

9.15 Assignments

1. Can you find the income distribution for your country on the web? The Netherlands is probably one of the most egalitarian countries in the world, but have a look at their income distribution: http://www.cbs.nl/en-GB/menu/themas/inkomen-bestedingen/cijfers/extra/2009-inkomensverdeling.htm?Languageswitch=on. The average (about 33,400 euro for all families) is much to the right of the modal income class (14,000 to 16,000 euro per year). What's the median income?

Can you find the age distribution for your country?

2. Go to the Statistics site of the World Health Organization (WHO): http://www.who.int/topics/statistics/en/. Discuss with your mentor which statistics for your country you can use as inspiration for an article.

3. Watch the TED talk 'Hans Rosling shows the best stats you've ever seen', http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html. Statistics guru Hans Rosling debunks myths about the so-called "developing world." Afterwards, go to the website he mentions, http://www.gapminder.org. Play around with the statistical data and try to find out as much as you can about your own country and your neighbouring countries. Discuss your findings with your mentor.

4. Visit the six statistics blogs below regularly. You might get inspiration for debunking a statistical myth yourself.

Freakonomics	http://freakonomics.blogs.nytimes.com/
Bad Science	http://www.badscience.net/
Junk Charts	http://junkcharts.typepad.com/junk_charts/
Stats Blog	http://thestatsblog.wordpress.com/
The Numbers Guy	http://blogs.wsj.com/numbersguy/
Statistical Modeling, Causal Inference, and Social Science	http://www.stat.columbia.edu/~gelman/blog/