

Statistics for Journalists

Statistics is the skill of gathering and analysing numbers to find out information. Statistics can help to make sense of a world that can seem too large to comprehend.

Journalists often have to use statistics in reporting stories and this guide will show you how they can support your story and how to present them in a way that isn't misleading.



by Connie St Louis

about the author

Connie is the director of the MA in Science Journalism at City University, London. She is also an award-winning freelance broadcaster, science journalist, writer and scientist.

She presents and produces a range of science and health programmes for BBC Radio 4 and the BBC World Service. She is a recipient of the prestigious Joseph Rowntree Journalist Fellowship.

Her most recent programme on BBC Radio 4 which she produced and presented, investigated the use of racially targeted designer drugs by pharmaceutical companies. She also presented the landmark Radio 4 series 'Life as' which charted the science of life before birth until death.

With thanks to

Andrew Garratt external relations manager Royal Statistical Society and Professor David Spiegelhalter, Cambridge University for being able to include material that they produced.

about the cij

The centre for investigative journalism (cij) came into being in 2003 to address a deepening crisis in investigative reporting.

The cij provides high-level training, resources and research to journalists, researchers, non-governmental organisations, academics, graduate students and others interested in public integrity and the defence of the public interest.

The cij is a non-profit organisation and runs international summer schools, training programmes in basic and advanced investigative techniques and organises public meetings – all designed to raise and sustain the standards of investigative reporting. Our handbooks, archive material, web and audio resources have helped bring additional investigative tools to journalists and the community unable to attend cij workshops and training programmes.

The cij offers particular assistance to those working in difficult environments where freedom of the press is under threat and where reporting can be a dangerous occupation.

contents

introduction	4
questions to ask of each statistic	5-6
statistical terms: the basics	7-8
statistical sureness	9
comparing risks	10
other points	11

Further info

The Royal Statistical Society

www.rss.org.uk

Straight Statistics

www.straightstatistics.org

The UK Statistics Authority

www.statisticsauthority.gov.uk

The Tiger that Isn't: Seeing Through a World of Numbers,
Blastland, M and Dilnot, A. London: Profile Books Ltd, 2008.

How to Lie with Statistics,
Huff, D. London: W.W Norton & Company Ltd., 2003

introduction

We use statistics regularly in our daily lives, for example you might say: "I drink five glasses of water a day". On some days you may drink fewer than five glasses of water and others you may drink more, but on average you estimate that you drink five.

In order not to find statistics too intimidating, it might be easier to think of them as mathematics with a human touch; the way in which data is gathered and the way in which it is later interpreted is very much decided by the individual's judgement.

You will come across statistics that are produced by many different people, in many different forms: there are opinion polls, censuses, death rates, consumer surveys, statistics from the government, political parties, scientific papers and so on. Some will be the starting point or punch line of a

great story. Look at them in the right way and you will often find they are worth a few extra minutes of your time.

A useful approach for journalists is to imagine statistics as sources that are cunningly disguised as numbers. They can give you great stories if you ask them the right questions and terribly inaccurate ones if you don't, particularly if you don't ask any questions at all.

Don't Panic!

statistics can seem intimidating, incomprehensible or to some, uninteresting, they are just numbers with stories behind them. Journalists often panic because they forget how much they already know about numbers and about analysing them - we use them very intuitively in our every day lives – when we are making a judgement about which item is cheaper, whether to cross the road at the crossing or risk taking a shortcut. In these decisions we are using a process often used by statisticians - comparing likely possibilities or a set of numbers and making a decision based on them.

“ Assist readers to become calibrated to effect sizes so that they will be able to discriminate more easily, and to ensure that they do not discount small but important effects.

*Hazel Thornton,
University of Leicester*

questions to ask of each statistic

Where have you come from?
Every statistic has a creator, someone who has put it out there for a reason. One of the most important things to find out is who this is.

Do the creators have a hidden agenda or vested interests that might skew their interpretation?

Does the figure come from a company trying to show that its particular merchandise is the best? Does it come from a political party trying to spin a good image for itself?

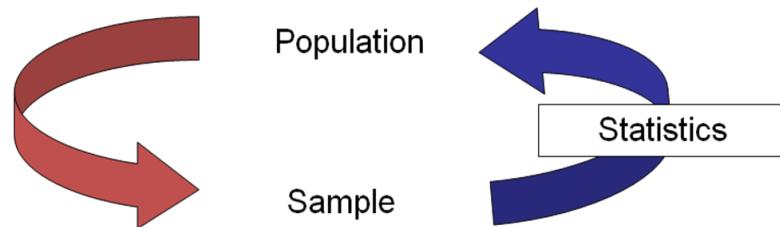
“ Is useful to provide a context that will give some idea of the magnitude of the interventions being suggested/compared.

Hazel Thornton,
University of Leicester

How was the data collected and could it have been collected in a better way?

Find out exactly what is being measured and how it has been measured. Often you cannot get the complete count of what you are looking for – it is too impractical to survey every teenager in Britain to find out what the most popular breakfast cereal is, or every mother to find out the average number of nappies used. So samples of the population are chosen, along with a way of extrapolating the data to speak on behalf of a wider number of people.

This is the most important use of statistics. It allows us to collect a sample of data from a much larger group called a population. Statistics allow us to make inferences about the population.



The key point is to remember that someone has deliberately chosen to collect the data in this way, used a particular method of extrapolation and most importantly, has had to make some assumptions in order to do so. You need to ask whether this particular way is actually the best and what the underlying assumptions might be.

Let's take the example of a press release from a dog food manufacturing company reporting that 100% of owners they surveyed prefer this particular brand. It's important to find out who they surveyed and when they were surveyed. It could have

been 10 of the owner's best friends and they may have only used the product once, when they were given a free sample.

What assumptions have been made? What has been ignored?

Self-selected surveys are an extreme example of these. For example take the magazine the Literary Digest who conducted a poll for the US presidential election of 1936. The magazine had used polls to successfully predict the outcomes of the five previous presidential elections. They predicted that the 1936 Republican candidate, Alf Landon, would comfortably beat the Democrat Franklin D. Roosevelt in 1936. In fact, Roosevelt won a landslide victory, winning 46 of the 48 states.

So what went wrong?

The sample size was certainly large; over 10 million questionnaires were mailed out so no clues or alarm bells ringing there. However, they were mailed to people who had both cars and telephones, unwittingly biasing the poll to the wealthy. A huge numbers of Roosevelt supporters had neither, so their opinions were missing from the data. The moral of this story is that when selecting respondents for a poll, general care is needed.

“ Use frequency statements, not single event probabilities – the latter are prone to being misunderstood. Eg: NOT: “If you take Prozac, you have a 30 to 50% chance of developing a sexual problem, such as impotence or loss of interest.” Instead, state: “Out of every 10 patients who take Prozac, 3 to 5 experience a sexual problem.” ”

*Hazel Thornton,
Department of Health Sciences,
University of Leicester*

What are the statistics reported to mean and what do they really mean?

Don't accept a statistic at face value. In order to find out how most statistics should be interpreted, you need a few definitions and a pinch of maths - you don't need a statistics degree and you don't need to be particularly good at maths.

The most important thing you need to be able to do is spot when something fishy is going on, and to be able to know where to find help if you need to check something out. The original statistician on a scientific paper or the statistical organisation that carried out the research can be a starting point if you have queries. Statistics departments at universities can be very helpful too. If you have any doubts or if something doesn't ring true, there will always be someone you can ask.

statistical terms: the basics

The average

Since most people end up wanting to know how healthy/fit/happy/wealthy they are compared to others, averages of this kind are often used in stories.

There are actually a number of different types of average; three of the most commonly used in journalism are discussed below.

The mean

To work this out you add up all the values in your set of data and divide by the total number of values in your list:

Let's say you are in a seminar of 10 university students with part-time jobs and you want to write a story about how much the people in this particular course are earning.

First you add up each individual salary:

£7000 £6000 £4000 £8000 £12000 £8000 £9000 £7000 £8000 £6000

The total of these salaries is:

£75000

Divide this by the number of salaries (in this case 10), so:

$$\frac{\text{£75000}}{10} = \text{£7500}$$

The average salary is about £7,500 a year.

How the mean could be misleading

Let's imagine that millionaire Bill Gates, the founder of Microsoft, visits your university and joins your seminar. If you were to now calculate the mean, it is somewhere in the region of millions of pounds a year as a result of the impact of Bill Gates' salary. You could now report that the average salary in this university seminar is several millions of pounds. So although the mean can be a useful measure, it is not always the most representative number to use. In this case the **median** would be a more useful value.

The median

To work out the median you take your values (for the example used above that would be the earnings of each person in the room) and put them in a row, from lowest to highest. The median is the number in the middle of all of these.

Working with the scenario in the university lecture theatre, calculating the median would give you a figure near £7,500 a year, a much more useful and accurate figure if you are most interested in the earnings of the typical student in that room.

The mode

This is simply the most common value in your list. Working with the example above, the mode would be £8000 as this is the number that appears the most.

Standard deviation

This is a measure of how much something varies by, in comparison to the mean.

To look at the standard deviation in what you eat each day you can take the mean value of what you eat over a week and compare the calories you took in on a particular day. If you ate the same that day as you do every day you should have a calorie intake which is the same as the mean. If you over indulged one day, you would have a value much greater than the mean, so in this case your deviation from the mean value will be bigger.

You could have two people with the same mean calorie intake but if one alternates between binge eating and not eating at all, they will have a bigger standard deviation; the smaller the deviation the more consistent the calorie intake.

This measure can also be applied to scientific experiments. If there is a small standard deviation the result is more consistent - ie every time the experiment is done, you get results that are close to each other in value. Thus the smaller the standard deviation, the more confidence you can have in the result and the bigger the standard deviation the less confidence you can have in the result.

statistical sureness

In scientific papers the term “**statistical significance**” is often mentioned. This is a way for statisticians to decide how likely it is that a particular finding is down to chance.

You will often see the term $p \leq 0.05$. This means that there is equal to, or less than 5% probability of the event being down to chance alone. This is a commonly used, arbitrarily decided benchmark for “**statistical significance**”. Anything with a higher than 5% (or 1 in 20) probability of being down to chance alone, is not considered statistically significant in the context of many academic papers.

It is worth remembering that just because something is statistically significant it does not necessarily mean it is significant to society.

Risks

We are often bombarded with headlines such as “x causes bowel cancer” and “vaccine causes x”. The

first question most journalists should be asking is; how likely is this to happen to the reader?

Let's imagine, in the first example that x is lipstick (a completely random example - there are no studies to suggest this). Working with this as an example, how likely is lipstick to cause bowel cancer? First we need to know the absolute risk of bowel cancer - that's the risk that an individual has of getting bowel cancer in their lifetime, where no other possible precipitating factors are present. Let's imagine that the risk of getting bowel cancer is 5% for population group we are focusing on.

We now need to figure out what the **relative risk** is - ie the increase in risk of getting bowel cancer if you use lipstick. Let's say the risk is said to increase by 20%. Whilst at first a 20% increase in the chance of getting bowel cancer might seem like a lot, what this works out to, is

an increased risk of 1% - an additional risk of 20% of 5% - that's a lifetime risk of 6%.

If something is more risky or likely than before, it is important to ask how likely it was in the first place. When you are looking at increases and decreases, you need to know what the baseline is. An increase in 50% of someone's wage might seem like a lot, but if the wage is only £2 a day, that's just a £1 increase, which may make little difference to his/her everyday life.

comparing risks

Another murky area of statistics is attempting to compare risks.

Regression to the mean

This is the tendency, after an extreme result, for the rest of the results to be closer to the average. Unless you are in a totally controlled environment there will always be fluctuations from the mean value, but over time your result will always tend back to the mean.

For example: if you count traffic every day over a year you obtain your mean value. In August your traffic may be much higher than the mean value - this could be because lots of people use the road to go on a summer holiday, after this event the traffic on the road will become closer to the mean value again.

The important thing to remember is that although an extreme result may seem like a good story, one extreme result is not a solid basis for drawing

conclusions for the likelihood of future events.

Benefits, risks and harms

Be clear about what you are saying, are you using the word 'risk' when what you mean is 'harms'. The following definitions outline what terms you should use and when:

Harms: The totality of possible adverse consequences of an intervention. These are the opposite of benefits against which they must be compared.

Benefits: The totality of possible positive consequences of an intervention. These are the opposite of harms against which they must be compared.

Risk: The probability of harm taking place. It is not possible to directly compare benefits and risks because one is an outcome while the other is the probability of an outcome.

‘‘ Use absolute risks, not relative risks. ’’

Use mortality rates, not survival rates.

*Hazel Thornton,
Department of Health Sciences,
University of Leicester*

other points

Sample size there is no magic number, use your common sense. Is a sample size of 100 people enough to draw conclusions for a population?

Models such as those used to predict the weather, the course of influenza etc use assumptions. It's worth finding out what the assumptions are and to question them; different assumptions can give different results.

Context is key people often use the most extreme number to make a story more dramatic. For example the Happy Dog trust spent £1m on dog food last year. That may seem like a lot of money, but if they are looking after two million dogs, it works out at only 50 pence per dog. Always put numbers into context.

A disease with a 0.005% fatality rate may sound like a minor concern, but if it was a common disease it would cause millions of deaths each year.

“ I think my biggest tip does not concern the statistics that a journalist is confronted with. It's all the stuff that is hidden from view, all the failed studies and negative results that are not mentioned, but without which the positive results cannot be interpreted. So the tip is 'what am I not seeing, and why am I being told about these results?' ”

*David Spiegelhalter,
Professor of the Public Understanding of Risk, Cambridge*



THE CENTRE FOR INVESTIGATIVE JOURNALISM

Centre for Investigative Journalism
is a registered charity (1118602).

This handbook was made possible by a grant from the Open Society Institute www.soros.org

We would really value your thoughts and feedback on this handbook.
Please write to us at info@tcij.org

If you would like to suggest a new topic for a handbook, or know journalists/authors who could help write one, drop us a line at the above address.

For more in the handbook series, please visit our website www.tcij.org

City University London,
Journalism & Publishing,
Northampton Square,
London
EC1V 0HB

info@tcij.org | +44 (0) 207 040 8220 | www.tcij.org