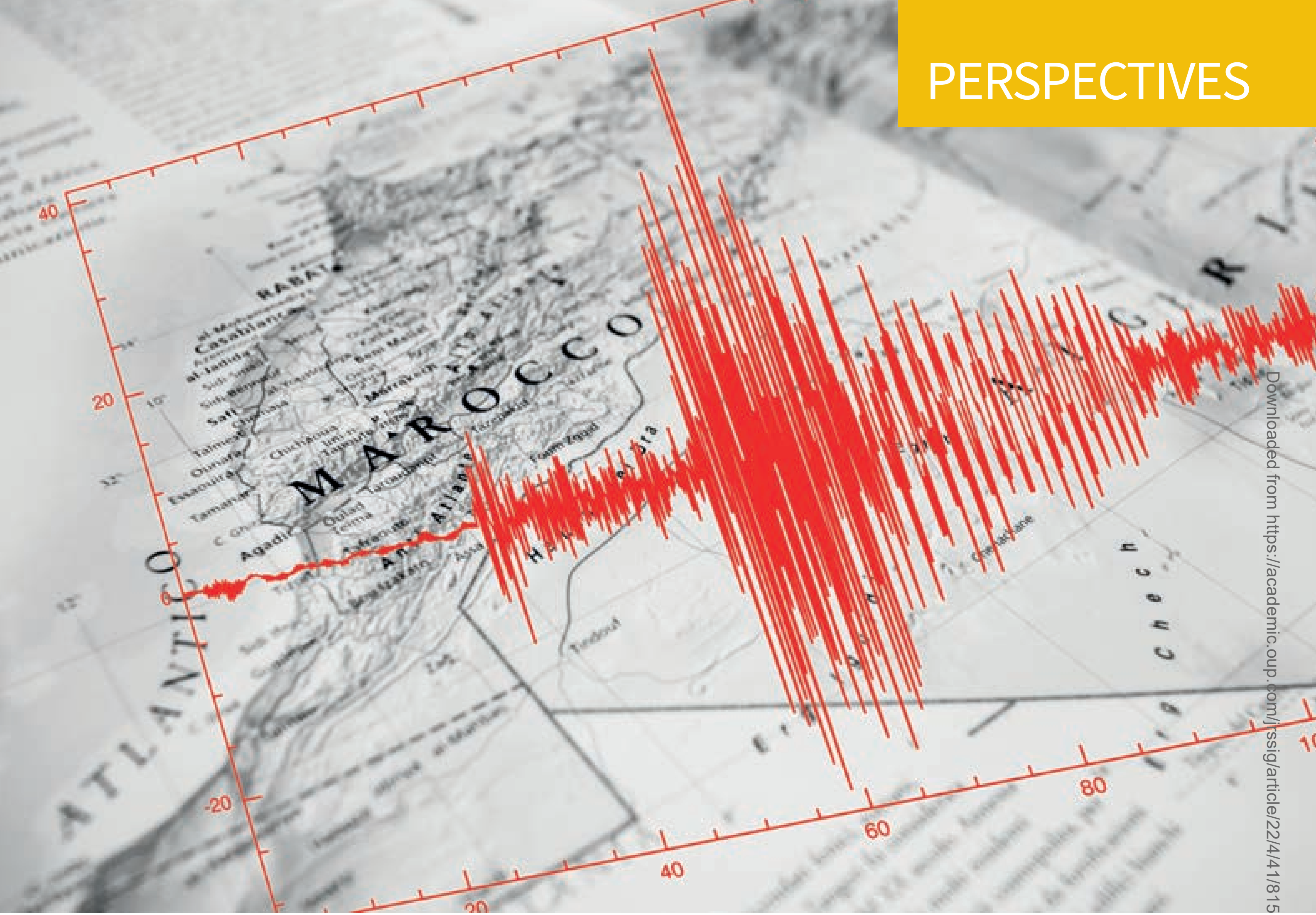

Tonello Photography/shutterstock.com

Bad stats

A regular series exploring slip-ups, snafus and salutary lessons from the world of statistics

A statistical fallacy of seismic importance

A middle-aged woman attends a routine breast cancer screening. She has no known risk factors but when the result comes back she's shocked to learn it's positive. So, what are the chances she really does have breast cancer: 75%, 85% – or higher still?

Confronted with this question, many people may think of the amazing abilities of medical technology and guess somewhere around 85%. And they'll feel vindicated when told that mammography does indeed detect cancer in around 85% of those who have it.

Many *Significance* readers, in contrast, may just roll their eyes, having recognised this familiar "gotcha!" question about probability. It's the textbook example of the so-called base-rate fallacy, a notorious trap awaiting anyone trying to make sense of diagnostic tests. Warnings about its ability to lead people into making faulty judgements date back at least to the mid-1950s, and it came to prominence following experimental studies of its impact by psychologists Amos Tversky and Daniel Kahneman in the 1970s. Their findings suggested people fall into the trap because of a tendency to judge evidence such as diagnostic test results in isolation, rather than setting them in proper context. And to do that requires knowledge of the *base rate*, the general prevalence of whatever is being tested for.

In the case of the cancer test, calculating the odds that a positive test really does imply cancer depends on knowing three numbers: the chances that the test gives false alarms and misses genuine cases, and the base rate, which in this case is the prevalence of breast cancer among those screened.

As the original question fails to give all three, many readers will conclude that nothing concrete can be said about the chances of the woman having cancer. Well, almost nothing: ▶

Robert Matthews has spent 40 years reporting on science for international media outlets. He is currently visiting professor in statistics at Aston University, Birmingham, UK and a member of the *Significance* editorial board.

Perhaps the most profound manifestation of the base-rate fallacy is its role in the age-old quest to predict devastating earthquakes. So far this century these natural disasters have claimed the lives of over 700,000

Diagnosing: the problem

The cancer screening question is more than a "gotcha!". It's a real-life example of reasoning with *conditional* probabilities, where the chances of one event taking place depend on others. Questions about conditional probabilities usually involve *Bayes' theorem*, which captures the relationship between them. This states that the odds of event A, *given* an event B has taken place, are

$$\text{Odds}(A \mid B) = LR \times \text{Odds}(A),$$

where LR is the "likelihood ratio" $\Pr(B \mid A) / \Pr(B|\neg A)$, with $\neg A$ denoting that A does not take place, and $\text{Odds}(A) = \Pr(A)/[\Pr(\neg A)]$, $\Pr(A)$ being the probability of A.

Bayes' theorem reveals the base-rate fallacy lurking in the cancer screening example. We want to know $\text{Odds}(C \mid T+)$, the odds of the woman having cancer *given* a positive test result. Note that this is not the same as $\text{Odds}(T+ \mid C)$, which are the odds of getting a positive result, *given* the presence of cancer (i.e., the "true positive" rate). The theorem shows that to get what we want we need *three* quantities: this true positive rate $\Pr(T+ \mid C)$, along with the false positive rate $\Pr(T+ \mid \neg C)$ and also $\Pr(C)$, the chances that any similar woman chosen at random has cancer before the test was carried out, (i.e. the *base rate*). We need this to put the outcome of the test into context using the general prevalence of breast cancer among the population to which the woman belongs. Failure to do so constitutes the base-rate fallacy.

Plugging all three numbers into Bayes' theorem will give the answer – which is easy enough *if* one knows the formula. A more insightful approach has been advocated by psychologist Gerd Gigerenzer of the Max Planck Institute for Human Development, Berlin. It's called the *natural frequency method*.[1] Imagine the woman is one of 1,000 women screened. Then if $\Pr(C)$ – the base rate of cancer among such women – is 0.5%, there will be just 5 with cancer while the other 995 will be cancer-free. Given a true positive rate of 85% and a false positive rate of 10%, that means screening will detect around $0.85 \times 5 \approx 4$ of the genuine cases, but also around 100 false positives. So of all the 104 positive results, only $4/104 \approx 4\%$ will be genuine.

This might suggest that screening is pointless, but Bayes' theorem helps put the result in context. The likelihood ratio captures how much stronger the evidence of cancer has been made by the screening. Plugging in the figures gives a likelihood ratio of 8.5. So, far from being pointless, the screening has boosted the odds of cancer being present almost ninefold. The theorem also shows that this likelihood ratio and those of follow-up tests multiply together, so the strength of evidence quickly mounts.

As Xiao-Li Meng has recently shown,[2] Bayes' theorem also leads to several rules of thumb for when we are trying to diagnose rare diseases – or predict rare events like major earthquakes (see main text). For example, if the base rate is lower than both the false positive and false negative rates of the diagnostic or predictive method, a positive test result or prediction is more likely wrong than right. Whether this condition holds can often be decided using basic knowledge about the test. It certainly holds for general breast cancer screening – and also for putative "precursors" of major earthquakes. Bayes' theorem also shows that, surprisingly, even if a diagnostic or predictive method can catch every genuine event, its false positive rate must still be below the base rate to ensure most alerts aren't false alarms. That is not plausible for precursors of major earthquakes.

being familiar with this gotcha question, they'll know that whatever the correct answer may be, it will be much lower than the seemingly reasonable guesstimate of 85%.

That's the effect of the base-rate fallacy, which in this case means failing to take account of the prevalence of breast cancer among screened women. Simply put, as this is low for such women – around 0.5% – even a test that correctly spots 85% of those with cancer will still struggle to find many. On the other hand, as the vast majority are disease-free, even if the test also correctly rules *out* cases with 90% reliability, the 10% of false alarms will still be a large number. Crunching the numbers (see box), it turns out that the chances the woman really does have breast cancer given her positive result are just 4%. It's a stunning result, and one which still provokes debate about the pros and cons of screening programmes (tinyurl.com/NCI-falsepos).

Readers who fell into the base-rate fallacy trap can console themselves by knowing that senior physicians can make the same mistake even when given all the figures.[1] But those who spotted the trap and concluded that nothing concrete could be said shouldn't feel too smug – as it's not true. The same theory that shows that three numbers are needed for the right answer also shows that just knowing that the chances of the test fouling up (i.e., the false alarm rate and the false "all clear" rate) are likely higher than the base rate implies that a positive result will most likely be wrong. And that's often pretty easy to check. For the cancer screening test, the false positive and false negative rates are pretty low, but they're still much higher than the breast cancer base rate of less than 1%. So that means it's odds-on that a positive test result will be wrong.

It's a handy rule of thumb – and it surprised a lot of statisticians (and the author) when Harvard University statistician Xiao-Li Meng tested it on the audience at the famous American Statistical Association "Woodstock of Inference" meeting in Bethesda in 2017.

Bayes' theorem shows that other traps await those trying to make sense of such evidence. For example, it warns us of the danger of blithely accepting figures for the "accuracy" of, say, a new type of lie detector. A 95% "success rate" may sound impressive, but what does it mean? Is it the true positive rate? If so, what is the corresponding false positive rate, which is needed to work out the likelihood ratio which captures the detector's ability to add evidential weight to what we already know. And what is the base rate? How does one even go about estimating the "prevalence" of the dishonesty of an individual? If it's very low, Meng's rule of thumb may well apply – and a miscarriage of justice looms.

The dream of predicting major earthquakes on timescales appropriate for evacuations was always going to be just that: a dream

Perhaps the most profound manifestation of the base-rate fallacy is its role in the age-old quest to predict devastating earthquakes. So far this century these natural disasters have claimed the lives of over 700,000, with many millions more left injured or homeless.

Attempts to find reliable tell-tale "precursors" of earthquakes date back millennia, the hope being that these would buy time to lead people to safety. Over the centuries a host of precursors have been claimed, ranging from outbreaks of small tremors and releases of natural radioactive gas to the strange behaviour of animals. On 3 February 1975 a combination of changes in groundwater levels, suspected foreshocks and snakes emerging from hibernation prompted the evacuation of the city of Haicheng, 500 km east of Beijing. The next day a devastating 7.3 magnitude earthquake struck. Despite damaging or destroying 90% of the city's buildings, all but around 2,000 of the million-plus population survived.

Yet what was initially hailed as the first-ever successful earthquake prediction proved to be the outcome of gut feeling, local actions and sheer luck.[3] Neither the timing nor strength of the earthquake was accurately or precisely predicted. There had also been multiple false alarms in the run-up to the earthquake itself.

A terrible reality check for believers in the dream of earthquake prediction came the following year when an even stronger quake struck Tangshan, around 400 km from Haicheng. This time there were no foreshocks or "anomalous events" to convince officials to take action, and at least 240,000 perished.

Despite this, the 1970s saw a surge of excitement about the possibility of just-in-time earthquake prediction. A key driver was the development of a theory suggesting that reliable precursors may actually exist. Known as the dilatancy-diffusion hypothesis, it promised to replace anecdotal tales of bizarre "anomalies" with precursors based on laboratory studies of rocks under extreme stress. Over the years, attempts were made to see if such precursors turned up in the field. The results were mixed. Unsurprisingly, extrapolating findings made in the lab to the behaviour of colossal slabs of rock in the Earth's crust proved problematic, and the hypothesis is now regarded as another false dawn.

An influential review published in 1997 by University of Tokyo seismologist Robert Geller is widely regarded as the obituary for earthquake prediction: "The idea that there must be empirically identifiable precursors before large earthquakes is intuitively

Gorodenkoff/shutterstock

appealing but studies over the last 120 years have failed to support it".[4] Yet for anyone familiar with the base-rate fallacy, the wonder is how the dream stayed alive so long. For suppose that, against all the evidence, a precursor *did* exist that infallibly predicts a major earthquake arriving in the next few days – that is, its true positive rate is 100%. Bayes' theorem warns us that even this is not good enough unless the precursor's false positive rate is lower than the relevant base rate. Historically we know that the base rate for a major earthquake striking a city in any given week is far less than 1%. So the false positive rate must be even lower still to ensure most alerts are genuine. And nothing we know about the behaviour of rock under stress suggests this is plausible.[4]

In short, the dream of predicting major earthquakes on timescales appropriate for evacuations was always going to be just that: a dream. Given that the base-rate fallacy has been known about for at least 70 years, it is bizarre that its relevance to earthquake prediction seems to have been overlooked until the mid-1990s.[5]

Yet while the base-rate fallacy may not have stopped vast sums being spent on a mirage, the abject failure it predicted ultimately did – to life-saving effect. Today millions of people in seismically active areas benefit from alerts based on the ultimate earthquake precursor: the quake itself. When rock ruptures, it sends out seismic waves travelling at different speeds. The fastest are to-and-fro primary ("P") waves, which travel at around 6 km per second, followed by up-and-down secondary ("S") waves. These travel around half as fast but are far more destructive. So detecting the P-waves can give early warning of the impending arrival of the S-waves.

First used on the famous Shinkansen "bullet train" network in the 1960s, early warning systems exploiting this phenomenon are now operational in parts of the Americas, Asia, Australasia and Europe. They are not perfect. At best, they provide only a minute or two's warning to take shelter, and anyone close to the epicentre won't even get that. Not all earthquakes produce strong S-waves, which can lead to false alarms. Even so, these systems have already saved countless lives.

The vision of precise earthquake prediction has also given way to broad-brush mitigation policies, where the ever-present threat in specific regions is countered through robust building and retrofitting codes, public education and drills.

All these measures make sense when seen through the prism of probability theory, and it is a tragedy this was not recognised many decades ago. Statistics is notorious for its many inferential traps, but few have the power to shock like the base-rate fallacy, which continues to stalk the assessment of evidence in many areas beyond those considered here. An ability to detect its presence should surely be part of everyone's mental toolkit. ■

References

1. Gigerenzer, G. (2002) *Reckoning with Risk*, pp. 39–87. London: Allen Lane.

2. Meng, X.-L. (2022). Double your variance, dirtify your Bayes, devour your pufferfish, and draw your kidstrogram. *New England Journal of Statistics in Data Science*, **1**(1), 4–23.

3. Wang, K., Chen, Q.-F., Sun, S. and Wang, A. (2006) Predicting the 1975 Haicheng Earthquake. *Bulletin of the Seismological Society of America*, **96**(3), 757–795.

4. Geller, R. J. (1997). Earthquake prediction: A critical review. *Geophysical Journal International*, **131**(3), 425–450.

5. Matthews, R. A. J. (1997). Decision-theoretic limits on earthquake prediction. *Geophysical Journal International*, **131**(3), 526–529.

From the archive

If you enjoyed this issue's feature on Chinese astrology (page 14), check out these other articles related to China from the *Significance* archives.

18 March 2009 Too Many Males in China: The Causes and the Consequences

We expect that, roughly, as many boys will be born into the world as girls. However, in some places, social pressures combined with modern medicine seriously distort the ratio of the sexes. In China, there are a million excess male births each year. **Thérèse Hesketh** looks at what this will mean for the generation that lacks women. tinyurl.com/ywf76am9

29 May 2019 Ask a Statistician: Does Manchester United Really Have 100 Million Followers in China?

Manchester United has claimed to have "100 million followers" in China, based on a survey by a market research company. **Rob Mastrodomenico** examines the veracity of this figure and argues that the key question is, what do we mean by "follower"? tinyurl.com/4ruhrnxm

■ Did you know *Significance* articles become free to read one year after publication, and remain so for ten years? Explore the full archive at: academic.oup.com/jrssig