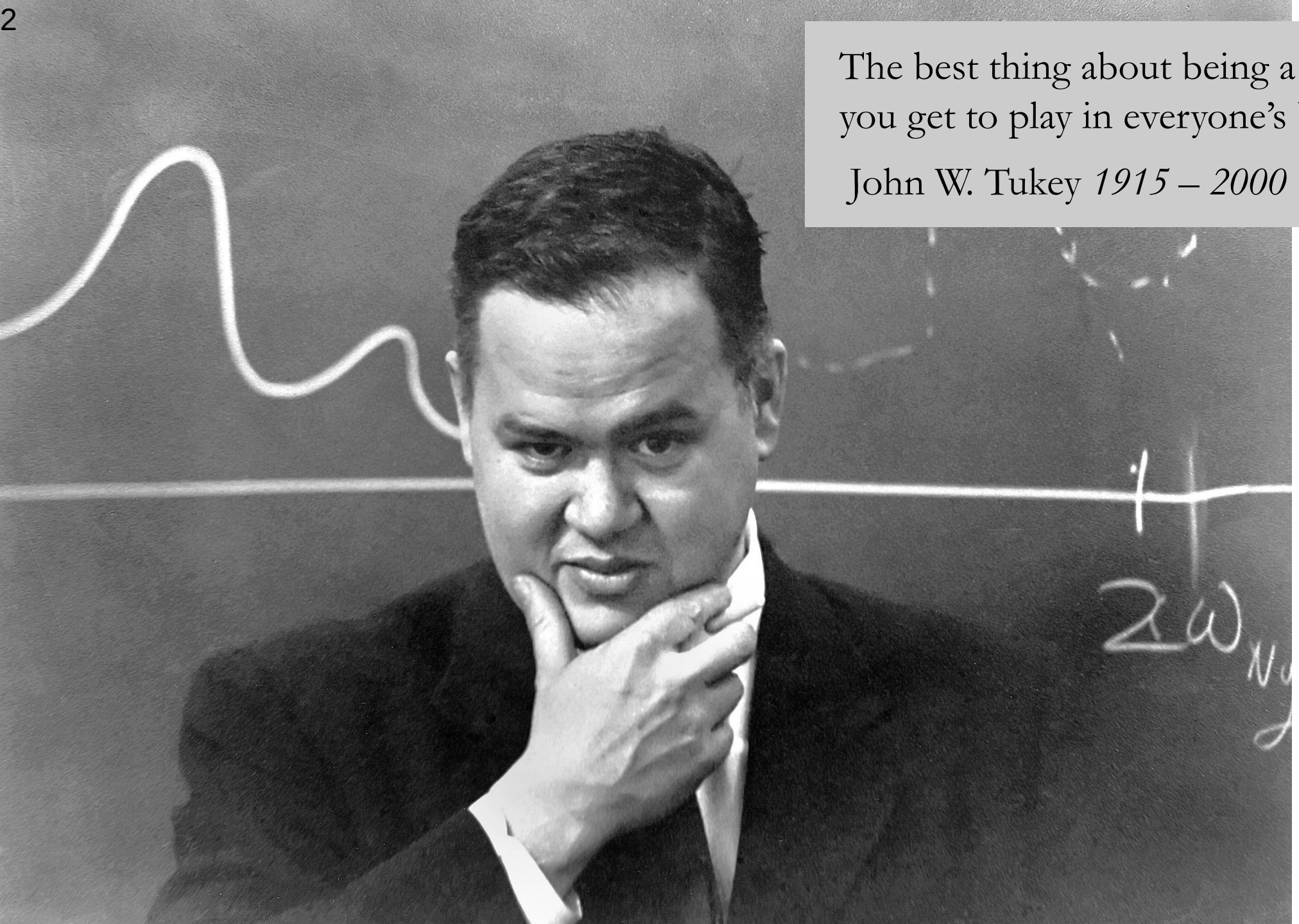
Statistics 101: what to care about and why

Kevin McConway
Emeritus Professor of Applied Statistics, The Open University
for ABSW Summer School, 27 June 2023

The Open University



The best thing about being a statistician is that you get to play in everyone's backyard.

John W. Tukey *1915 – 2000*

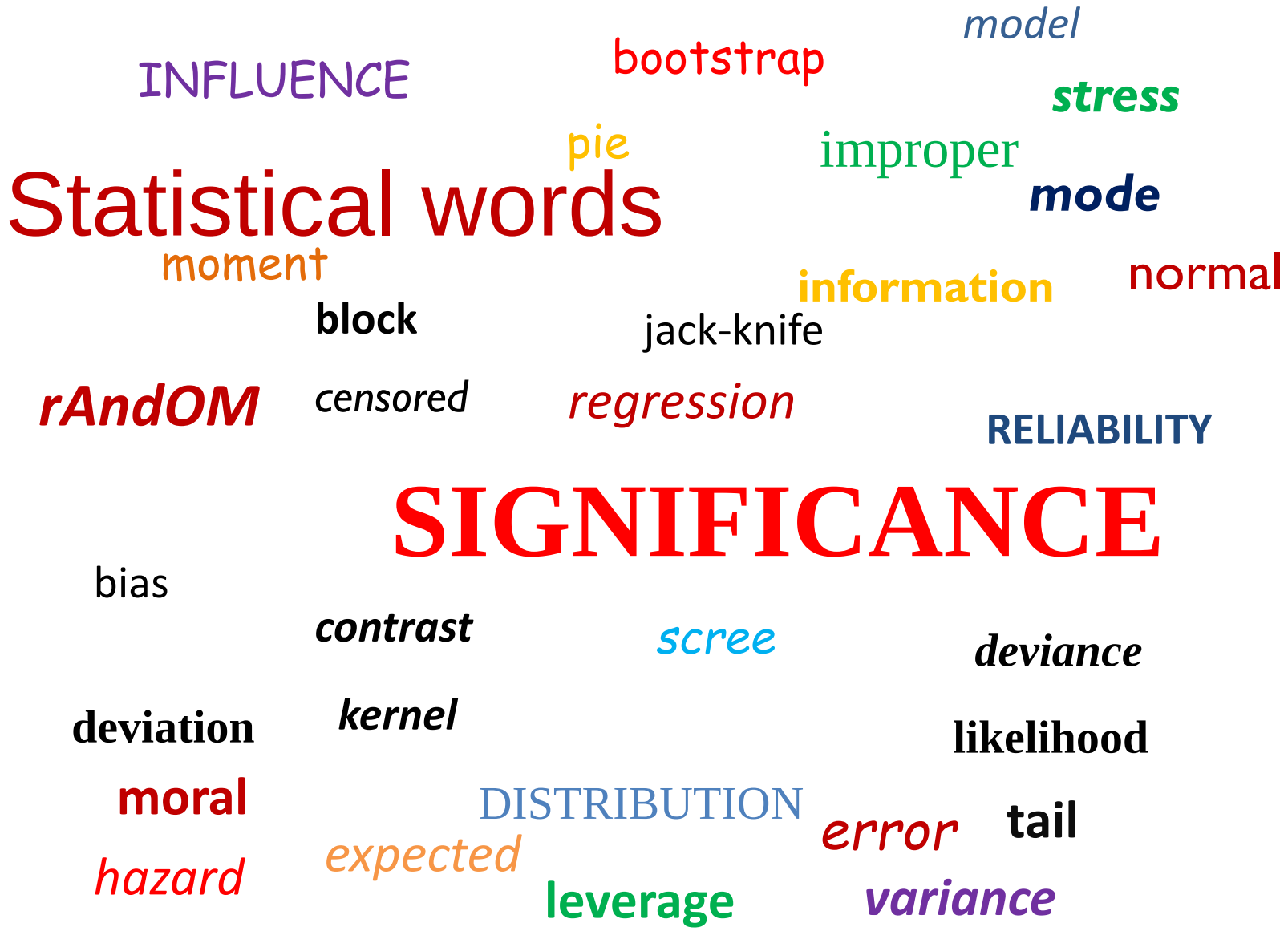
Timetable for today

- Statistical significance, *p* values, related concepts
- Types of study
- Confounders and causality
- Absolute risk, relative risk
- Reading scientific papers
- ‘Red flags’
- Books worth a look
- Q+A

What I *won't* do today

- More advanced topics
  - Regression and correlation
  - Bayesian statistics
  - Causal inference and mendelian randomisation
  - Diagnostic and prognostic modelling and tests
  - Survival analysis
  - More on meta-analysis
  - Time series
  - Surveys
- Three hours of me going on about some of these things at https://www.absw.org.uk/resources/kevin-mcconway-s-statistics-reading-resources-list-for-journalists

STATISTICAL SIGNIFICANCE AND *P* VALUES


model
INFLUENCE
bootstrap
stress
pie
improper
Statistical words
mode
moment
information
normal
block
jack-knife
rAndOM
censored
regression
RELIABILITY
SIGNIFICANCE
bias
contrast
scree
deviance
deviation
kernel
likelihood
moral
DISTRIBUTION
error
tail
expected
hazard
leverage
variance

Statistical significance

- *It doesn't mean what you (possibly) thought it meant.*
  - Nothing directly to do with real-world **importance**
- A (flawed?) way of judging whether some finding can be explained away by chance variability.
- …usually based on the *P value*.

P values

- *It **almost certainly** doesn't mean what you thought it meant.*
- Number between 0 and 1 (or 0% and 100%).
- The smaller *P* is, the less likely it is that a finding is just chance (*more or less!)*
- Common convention: if *P* is less than 0.05 (that is, 5%) a result is statistically significant.
- *But it's just a convention* (and doesn't work well with big data)
- Not (necessarily) used in other contexts: e.g. the *P* value for Higgs boson discovery was 3×10^{-7}.

- Controversy about it – papers saying use a much smaller cut-off for 'discovery', don't use *P* values at all, justify your choice of threshold in the paper

P values

- E.g. “people on the new drug had, on average, lower blood pressure than people on the old drug (*P* = 0.03)”.

- *DOESN’T mean:*
  - There’s a 0.03 (or 3%) probability that the result was due to chance.
- ***DEFINITELY*** *DOESN’T mean*
  - There’s a 0.97 (or 97%) probability that the result was NOT due to chance.
- *DOES mean:*
  - Assume that the new drug has the same average effect on blood pressure as the old drug. (In the jargon, this is the *null hypothesis*.)
    Then there’s a 3% probability that a difference like this (at least as extreme as this) will be observed.

- *So what IS the probability that the result was due to chance?*
  - **You can’t say from this information (and often you can’t say at all)**

Statistical significance still matters – *but be careful*

- *If the treatment (or whatever) has absolutely no effect, there's a 5% chance that each statistical test will report a significant effect.*
- Doing multiple statistical tests is like having lots of shots on goal – increased chance one will score (be significant).
- Hence risk of "data dredging", existence of (numerous) statistical methods to correct for multiple testing, etc.
  (This is a form of *P-hacking*.)
- **Wrong** assumption by many researchers that $P < 0.05$ does mean there's a real effect, $P \geq 0.05$ means there definitely isn't one.

- If *P* value is too big – it's plausible there's no effect, just chance variation.
- But still don't know the probability that it *is* just down to chance.

Health warnings for *P* values

- There are 2 ways that a significance test result can give a misleading answer:
  - Significant result but no true difference (false positive)
  - Non-significant result but there *is* a true difference (false negative)
- Looking at *P* values doesn't on its own deal with probabilities when there's a true difference (null hypothesis false).
- *Pr*(significant result, given that there's a true difference) is the (statistical) *power* of the test.

- *Sample size* should *be chosen to give adequate power, and the details of how it was chosen* should *be reported.*
- If the result is *not significant* (null hypothesis is *not* rejected) and no account is taken of power, you *can't say much* about what is going on.

P values

What to do instead?

- Concentrate on estimation (and confidence intervals) instead.
- Power calculations – how big an effect *could* the study actually detect?
- Bayesian analyses. If you must test, at least find the probability you are *really* interested in.

Confidence intervals

- These give a range of plausible true values of some quantity of interest (risk difference, population mean or median, etc. etc. etc.)
- There is (usually) a technical connection with significance testing, but confidence intervals tell you something much easier to interpret about the *size* of effects and *uncertainty* about their size.
- They have a number attached (usually 95%).
- If one calculates 95% confidence intervals on many occasions, then the interval should contain the true value on 95% of occasions.

Air pollution could increase stillbirth risk

WARNING: Exposure to air pollution could 'increase risk' of STILLBIRTH

EXPECTANT mothers are being warned exposure to air pollution could increase the risk of the 'neglected tragedy' of stillbirth.

Pregnant women 'should consider moving to the countryside' because air pollution may raise the risk of stillbirth, doctors warn

- Worldwide, for every 1000 total births, 18.4 babies were stillborn in 2015
- Researchers identified a very strong link between stillbirth and pollution
- Scientists called for tighter curbs on vehicle fumes and industrial waste
- Also said pregnant women should consider moving to greener areas

Science

Air pollution may raise risk of stillbirth and pregnant women should consider leaving cities, say scientists

Air pollution and stillbirth?

- Systematic review and meta-analysis of 13 studies worldwide (though most results depended on just 3 studies).
- Paper abstract says "*Although not reaching statistical significance*, all the summary effect estimates for the risk of stillbirth were systematically elevated[…]".
- The press release did mention a few caveats, but *not* the lack of significance.
- The press release gave the relative risk for just one pollutant, $PM_{2.5}$, but no absolute risks, and it's unclear how big that relative risk is anyway.
- (For later: Everything is based on *observational* studies so we can't conclude that the pollution *causes* the stillbirths anyway.)

Messages about statistical significance, *P* values and confidence intervals

- Statistical significance is a (flawed?) way of judging whether some finding can be explained away by chance variability, usually based on the *P* value.
- Other things being equal, which they hardly ever are, the smaller the *P* value, the less likely it is that a finding is due to chance alone.
  - Common (but flawed) convention: if $P<0.05$, the result is statistically significant.
  - But you still don't know how likely it is that the result is due to chance.
- Calculating many *P* values increases the chance of finding false positive results.
- Often better to use *confidence intervals* – ranges of (statistically) plausible values for a quantity of interest.

TYPES OF STUDY

Important because this determines strength of findings and confidence in results

Study types

Roughly in decreasing order of strength of evidence:

- Randomized controlled trial (RCT) (or better still, a good systematic review of RCTs)
- Prospective cohort study
- Retrospective cohort study
- Case-control study
- Cross-sectional study on individuals
- Ecological study (doesn’t mean what you might think it means)
- Animal studies (don’t fit on the same strength of evidence scale, really)

Study types

Roughly in decreasing order of strength of evidence:

- Randomized controlled trial (RCT) (or better still, a good systematic review of RCTs)
- Prospective cohort study
- Retrospective cohort study
- Case-control study
- Cross-sectional study on individuals
- Ecological study (doesn't mean what you might think it means)
- Animal studies (don't fit on the same strength of evidence scale, really)

"observational"

Early baldness higher heart disease risk factor than obesity, says study (November 2017)

Pollution wipes out the benefits of exercise, study suggests (December 2017)

Drinking hot tea linked to lowered glaucoma risk (January 2018)

Taking paracetamol while pregnant 'could harm your daughter's fertility' (January 2018)

Air pollution harm to unborn babies may be global health catastrophe, warn doctors (December 2017)

One portion of spinach a day can fend off dementia (December 2017)

Randomized Controlled Trials

RCTs

- Why controls?
  - Why not historic controls?
- Why randomize?
  - Avoid bias
  - Blinding

- But not all RCTs are good
- Equivalents of RCTs outside medicine and health *still* controversial

Why are systematic reviews better than a single RCT?

- More data
- (Usually) more settings
- Can look for heterogeneity
- RCTs aren't always very good on "which specific kinds of patient is this treatment good for?"

- The statistical approach to putting the data together is *meta-analysis*.

Statistical analysis in RCTs

- Possibility of “cheating” – carry out lots of tests, some will be significant even if there’s no real effect.
- RCTs registered in advance, stating *primary* and *secondary* outcomes
- You might want to check the registration…
- Things *not* registered: *Post hoc analysis.* Be careful how you write about these! (And the press release might not tell you which they are.)

Observational studies in epidemiology (etc.)

The general idea in epidemiology

- When you can't do an RCT (e.g. epidemiology of things that are possibly bad for people – or just the expense and complication).
- There's exposure to some (potential) risk
- There's an outcome (disease, early death, etc.)
- Want to know whether exposure changes the chance of the outcome.
- But all you can do is *observe* what people do or did. No experimental manipulation (as in an RCT). So these are **observational studies**.

Cohort studies, case-control studies, jargon…

- Lots of different technical descriptions for studies, depending on:
  - whether people were chosen on the basis of the exposure to risk or the outcome,
  - whether people were followed up over time from exposure to outcome, or everything was done *after* the outcome from records or recollection.

Meta-analysis is common for observational studies too

- A systematic review should follow a standard procedure.
  - Ideally, with a pre-registered protocol…
- But not all meta-analyses are good science; problems with:
  - *Study quality*
  - *Levels of evidence*
  - *Publication bias*

Early baldness higher heart disease risk factor than obesity, says study	Case-control study (November 2017)
Pollution wipes out the benefits of exercise, study suggests	Randomized controlled trial (December 2017)
Drinking hot tea linked to lowered glaucoma risk	Cross-sectional study (January 2018)
Taking paracetamol while pregnant 'could harm your daughter's fertility'	Animal experiment (January 2018)
Air pollution harm to unborn babies may be global health catastrophe, warn doctors	Retrospective cohort study (December 2017)
One portion of spinach a day can fend off dementia	Prospective cohort study (December 2017)

Messages about study types

- The strength of scientific evidence from a study depends on how it was done. There are many types of study. Understand those used in the fields you write about.
- Randomised controlled trials (RCTs) provide high quality evidence *if done well*.
- Sometimes an RCT is impossible. Alternatives may include *observational* studies.
  - In observational studies it can be difficult to tell what is causing what.
- A *systematic review* combines the results from several studies. The statistical method involved is *meta-analysis.*

Confounders and causality

Air pollution linked to much greater risk of dementia

Dementia warning: Toxic air pollution could INCREASE risk of developing dementia

Air pollution linked to much greater risk of dementia, London-based study suggests

Air pollution could be responsible for 60,000 cases of dementia in the UK, with people exposed to dirty air 40% more likely to develop the disease

WORLD / SCIENCE & HEALTH

Air pollution linked to higher risk of dementia

Air pollution, traffic noise and dementia

- Observational study, in London. (doi:10.1136/bmjopen-2018-022404)
- 140,000 adults aged 50-79 (in 2005), followed up till 2013.
- Recorded air pollution and traffic noise measures at their home address in 2004 (exposure)
- Outcome was diagnosis of dementia.
- Conclusions:
  - Higher chance of dementia for people in areas of higher air pollution. (Chance of dementia was about 40% higher in most polluted areas than in least polluted.)
  - Effect of traffic noise very small and not statistically significant.

Air pollution, traffic noise and dementia

- If it was an RCT, would *allocate* people randomly to live in areas of high or low pollution.
- But this was an *observational* study – maybe the people in polluted areas were older, or smoked more, or [something?]…
- Possible 'somethings' are called *(potential) confounders*.
- One can make statistical adjustments to allow for confounders, but you have to *know* about them and *have data* on them.
- These researchers adjusted for age, sex, ethnicity, smoking, body mass index, level of socio-economic deprivation of the area, and whether people had certain other diseases.
- Press release: 'These associations were consistent and unexplained by known influential factors, such as smoking and diabetes.'
- **Is this enough to avoid the confounding?** You can't tell, though reading the paper might give you a better idea.

Causality

- Crudely: you can infer causal effects from an RCT, but you can't from an observational study (cohort, case-control, cross-sectional, whatever).
- "Correlation is not causation"
  - But statisticians haven't always been always so good on saying what *is* causation.
- So how do we know smoking causes lung cancer?
  - "Causal narrative" from differing sources
  - Bradford Hill criteria (e.g. effect size, plausibility, repetition and reproducibility, etc.)
- Increasing use of causal inference methods. E.g. mendelian randomisation.

Causality

- Wording in the dementia study press release: "*This is an observational study, and as such, can't establish cause, and the findings may be applicable only to London.*"
- BMJ press releases usually say that about cause, when appropriate. But not everyone is so careful.

- A well-written scientific paper will mention it, e.g. in the Limitations discussion. But that's not always done.

- On (bio)medical papers, press offices might use SMC/AMS labelling recommendations: http://press.psprings.co.uk/AMSlabels.pdf

Confounders

- A confounder is a factor that might explain why *A* is *related* to *B* but *A* does *not* cause *B*.
- Can often "adjust" statistically for confounders that you know about.
- Press releases often say they adjusted for confounders but don't say which. This might cause you to smell a rat.
- Use your imagination! If you can think of a possible confounder that isn't mentioned, ask the researchers whether it's important and what they did about it.
- *Ask experts!*
- **Confounders mean that there's a *possible* alternative explanation (not necessarily that that's the *true* explanation).**

Messages about confounders, causality and correlation

- Confounders – quantities other than those being studies – can be the true cause of observed effects, and generally can muddy the waters.
  - This is particularly an issue in observational studies, but can also arise in poorly controlled experiments.
- *Be careful* not to make causal claims from observational studies (probably even if the researchers make them).
- **"Correlation is not causation."**
- If there is *any* suspicion that confounders might be involved (and there almost always is in observational studies), *mention it* in your story.

ABSOLUTE AND RELATIVE COMPARISONS

What do you want from a study of a potential risk factor?

- Ideally, want to know:
  - How likely is the outcome in exposed people? (*Absolute* risk)
  - How likely is the outcome in unexposed people? (Baseline for comparison)
  - How do these compare? (And how, statistically, to compare them?)

Interpreting the results: absolute or relative comparisons?

- I said, “Chance of dementia was about 40% higher in the most polluted areas than in the least polluted.”
- Sounds scary, but actually what you want to know is how great the risk is in *absolute* terms, not just *relative* to the risk in the ‘safest’ group.
- The paper doesn’t say!
- 1.7% of participants had a dementia diagnosis during follow-up (2.4 people, on average, per 1000 participants per year).
- But that is across all the pollution levels.
- My dubious back-of-envelope calculation estimates the % in the least polluted areas as about 1.5%.

The absolute risk isn't always in the research paper

- There can be good reasons for that.
- They might give the *relative risk* a.k.a. *risk ratio,* or the *relative risk reduction* (e.g. vaccine efficacy/effectiveness).
- Sometimes, again (usually) for good reasons, the paper reports *odds ratios* (**OR**) or *hazard ratios* (**HR**) instead. These are also *relative.*

Why absolute risk matters

- The impact of a new treatment, or a harmful exposure, depends on the absolute risks involved, *not just on the RR*.
- If the baseline risk (in unexposed people) is very small, doubling or halving it *may* not be very important.
- Twice not very much is still not very much.
- But this can be a big issue in relating the statistical results to individuals in a story.

What if the paper reports only relative measures?

- **Ask the scientists** to give you useful numbers. Or ask the **Science Media Centre**.
- Or use the online **RealRisk** tool at https://realrisk.wintoncentre.uk/

- Absolute risks might be in the paper somewhere, just well hidden.
- But are you the best person to be digging in the paper for them? And they might not be there anyway.

- *Press the researchers for an answer! Your readers want the numbers, or at least some clear indication of importance.*

Is it worth agonising over the uncertainty?

- It depends! Not always.
- But:
  - Paper on “Association between breastfeeding duration and educational achievement in England,” June 2023, doi: 10.1136/archdischild-2022-325148
  - Press release: “Only around a fifth (19.2%) of children who were breastfed for at least 12 months failed their English GCSE compared with 41.7% of those who were never breastfed.”
  - That’s true, but those are *unadjusted* figures, hence potentially misleading.
  - **What did RealRisk show?**

Results

Risk for never breast fed

Out of 100 *children aged 16 in England, enrolled in Millennium Cohort Study*, we would expect around 42 to *fail english gcse*

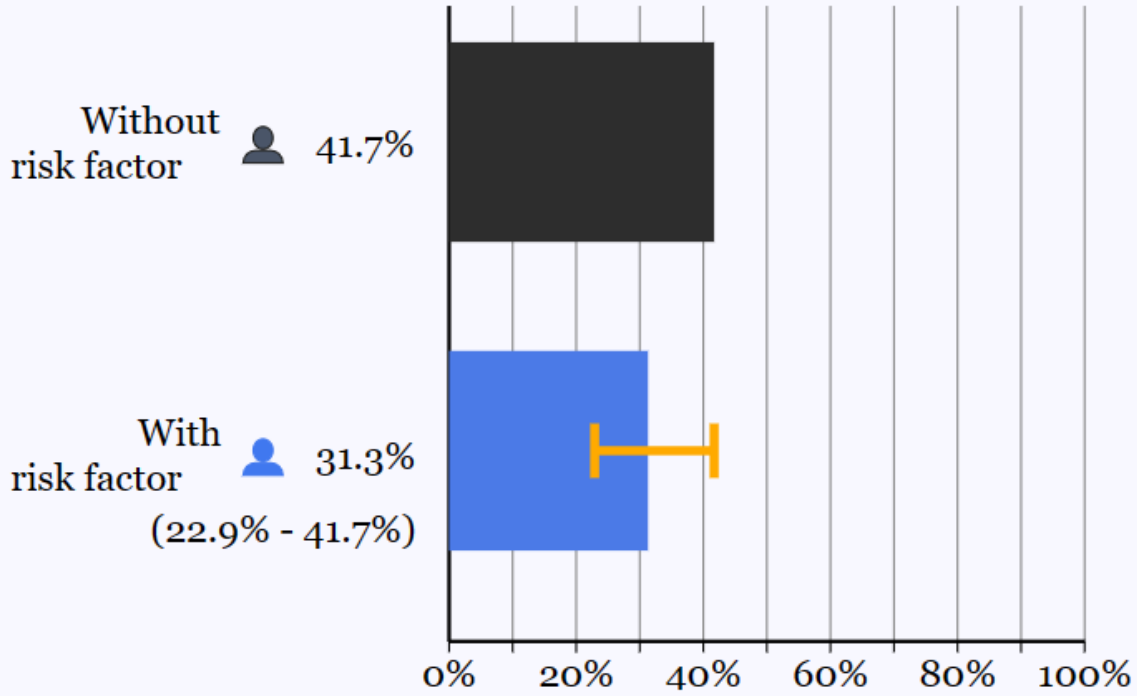
Edit Text

Risk for breast fed for at least 12 months

Out of 100 *children aged 16 in England, enrolled in Millennium Cohort Study*, we would expect around 31 to *fail english gcse*

Edit Text

Barchart | Icon Array



Results

Risk for never breast fed

Out of 100 *children aged 16 in England, enrolled in Millennium Cohort Study*, we would expect around 42 to *fail english gcse*

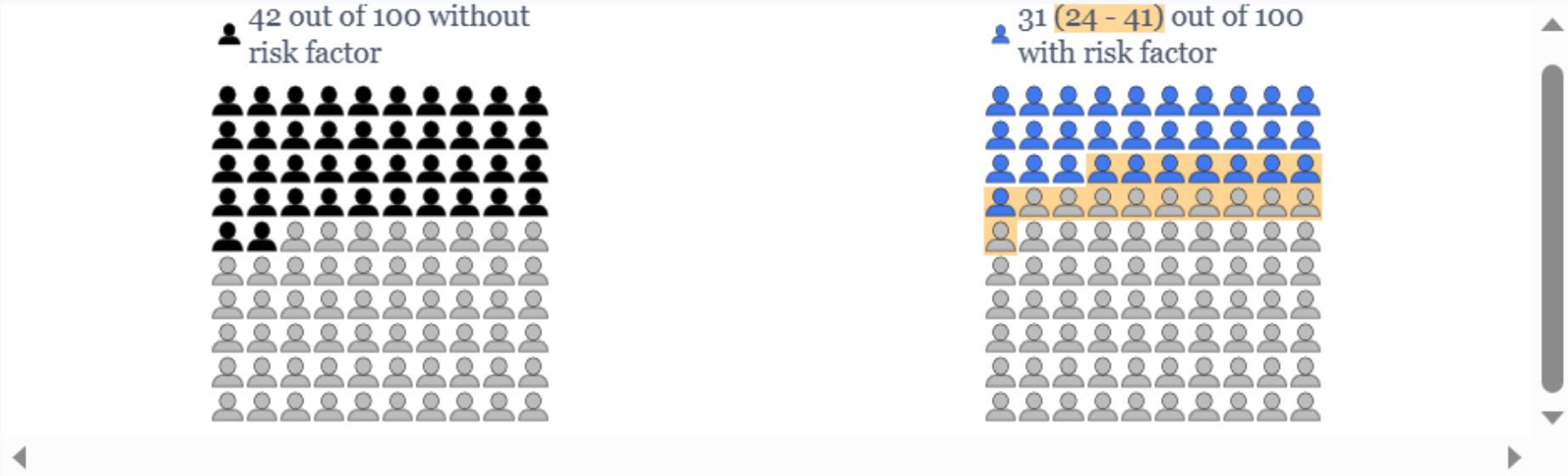
Edit Text

Risk for breast fed for at least 12 months

Out of 100 *children aged 16 in England, enrolled in Millennium Cohort Study*, we would expect around 31 to *fail english gcse*

Edit Text

Barchart | Icon Array



!RealRisk make sense of your stats

Results

Risk for people living in areas with the lowest fifth of nitrogen dioxide concentration

Out of 100 *adults aged 50-79 living in London*, we would expect around 2 to *had a dementia diagnosis over 8 years*

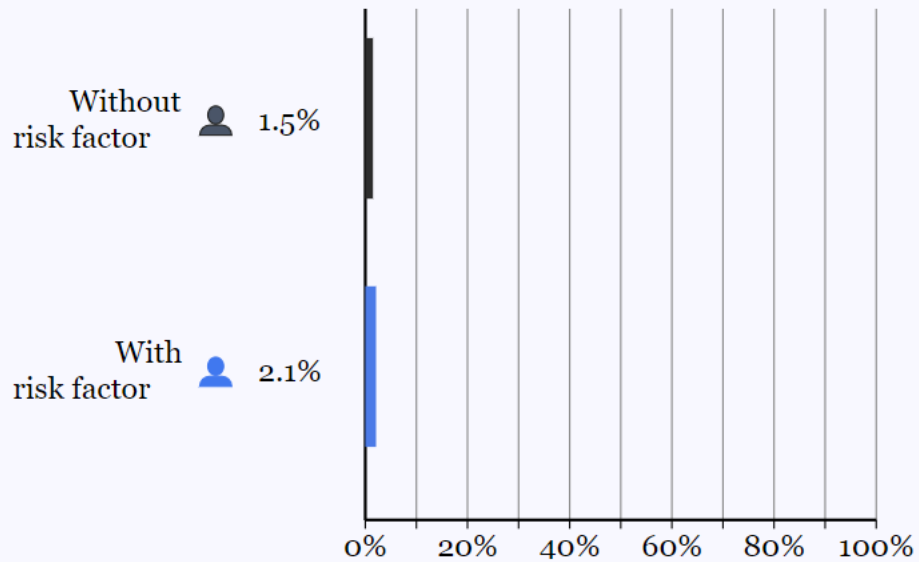
Edit Text

Risk for people living in areas with the highest fifth of nitrogen dioxide concentration

Out of 100 *adults aged 50-79 living in London*, we would expect around 2 to *had a dementia diagnosis over 8 years*

Edit Text

Barchart | Icon Array



RealRisk make sense of your stats

Results

Risk for people living in areas with the lowest fifth of nitrogen dioxide concentration

Out of 100 *adults aged 50-79 living in London*, we would expect around 2 to *had a dementia diagnosis over 8 years*

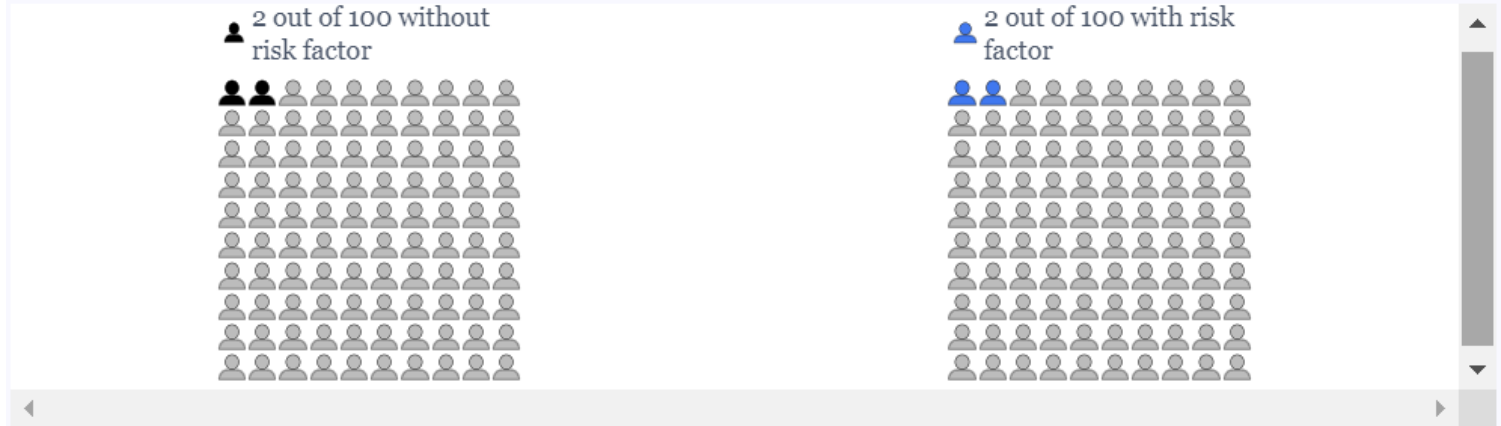
Edit Text

Risk for people living in areas with the highest fifth of nitrogen dioxide concentration

Out of 100 *adults aged 50-79 living in London*, we would expect around 2 to *had a dementia diagnosis over 8 years*

Edit Text

Barchart | Icon Array



Quiz!

A dietary supplement reduced the risk of having at least one cold in a year by 20%, according to an RCT.

In a given year, people who don't take the supplement have a 50% chance of getting a cold.

If a large group of people take the supplement, what's their chance of having at least one cold next year?

a) 30%
b) 40%
c) 44%
d) I'm a journalist, not a numbers person. You're the numbers person – you tell me!

Study links heavily processed foods to risk of earlier death

Modern diets could be killing us, suggests major study on ultra-processed foods

TAKING THE BACON Junk food diet 'DOES increase risk of cancer, heart disease and dying young'

Ready meals and snacks linked to deadly diseases

Eating 'ultra-processed' foods like pizza and cake knocks DECADES off your life

Messages about absolute and relative comparisons

- You can give, or get, quite a different impression depending on which you use.
- So you can slant what you write, deliberately (*but please don't!*) or accidentally…
- …or you can be taken in by a misleading comparison in a press release.
- Giving only relative comparisons, e.g. of risk, can be very misleading. Double a very small risk is still very small.
- Use absolute comparisons if at all possible.
  - You may have to use RealRisk, or ask the researchers, or a statistician, or the Science Media Centre.
  - If the study is about a disease or condition (medical, environmental…), at least write *something* about how common it is.

Reading a scientific paper: What to read and why.

Reading the paper

- Takes time, has jargon (statistical, and subject area)
- But, typical structure (don't need to read whole thing)
  - Abstract – summary (group studied, study design, findings, stats results)
  - Title – useful but sometimes like news headline
  - Discussion – the researchers' interpretation. Usefully has 'strengths' and 'limitations'
  - Introduction, or (simpler!) 'What is already known on this topic'/'What this study adds'
  - Usually less useful (for journalists):
    - Methods – for confounders (but usually listed in Results too)
    - Results – usually enough in Abstract, but sometimes a release concentrates on others

- Usual order: Title, Abstract, Introduction, Methods, Results, Discussion
- … but sometimes Methods comes last.

RED FLAGS FOR DUBIOUS STATISTICS

AMBER

~~RED~~ FLAGS FOR DUBIOUS STATISTICS

More amber flags

- Problems on standard journalistic questions – who's telling me this, why, what's in it for them, why now, what aren't they telling me? And if it's too good (or bad) to be true, it probably isn't true.
- Not peer reviewed. Review isn't perfect but is usually better than nothing.
- Generally, claiming importance for non-significant differences. (Watch out for 'league tables' or small changes in regularly-published statistics.)
- Going beyond the data/evidence: e.g. extrapolation to unobserved ranges; claiming that results apply to populations very different from those in the research; unsupported extrapolations to humans from animal studies.
- Selective reporting – think about what the researchers actually did and check it's all mentioned.
- Inadequate presentation/discussion of uncertainty.

Presenting numbers – some books

- *News and Numbers: A Writer's Guide to Statistics*. Victor Cohn and Lewis Cope with Deborah Cohn Runkle. 3rd edn, 2012. Wiley-Blackwell
  - The only book I know of on this topic written by journalists. Good (IMHO) and short.
- *How Charts Work: Understand and explain data with confidence*. Alan Smith. 2022. Pearson Education
  - Recent, very good indeed on charts and graphics.
- *The Art of Statistics: Learning from Data*. David Spiegelhalter. 2019. Penguin
  - If you think *I* put a lot in … no, it's really good, though not aimed specifically at journalists.
- *How to Read Numbers: A Guide to Statistics in the News (and Knowing When to Trust Them).* Tom Chivers and David Chivers. 2021. Orion Publishing Group
  - Aimed at general public, but (I think) better for journalists than most, because Tom C is one (and his cousin Dave is an academic economist who uses statistics).
- *Statistics Behind the Headlines*. John Bailer and Rosemary Pennington. 2022. CRC Press
  - Also by a statistician (Bailer) and a journalist (Pennington) – aimed generally but, I suspect, good for journalists.
- *Practical R for Mass Communication and Journalism*. Sharon Machlis. 2019. CRC Press
  - A lot is fairly standard stuff on how to use the R software, but probably more useful to data journalists than (other) science writers, and at least tries to use relevant examples, even if they are almost all American.
- For more on info sources, see https://www.absw.org.uk/resources/kevin-mcconway-s-statistics-reading-resources-list-for-journalists

More amber flags

- Problems on standard journalistic questions – who's telling me this, why, what's in it for them, why now, what aren't they telling me? And if it's too good (or bad) to be true, it probably isn't true.
- Not peer reviewed. Review isn't perfect but is usually better than nothing.
- Generally, claiming importance for non-significant differences. (Watch out for 'league tables' or small changes in regularly-published statistics.)
- Going beyond the data/evidence: e.g. extrapolation to unobserved ranges; claiming that results apply to populations very different from those in the research; unsupported extrapolations to humans from animal studies.
- Selective reporting – think about what the researchers actually did and check it's all mentioned.
- Inadequate presentation/discussion of uncertainty.

Books on presenting numbers

- *News and Numbers: A Writer's Guide to Statistics*. Victor Cohn and Lewis Cope with Deborah Cohn Runkle. 3rd edn, 2012. Wiley-Blackwell
  - The only book I know of on this topic written by journalists. Good (IMHO) and short.
- *How Charts Work: Understand and explain data with confidence*. Alan Smith. 2022. Pearson Education
  - Recent, very good indeed on charts and graphics.
- *The Art of Statistics: Learning from Data*. David Spiegelhalter. 2019. Penguin
  - If you think *I* put a lot in … no, it's really good, though not aimed specifically at journalists.
- *How to Read Numbers: A Guide to Statistics in the News (and Knowing When to Trust Them).* Tom Chivers and David Chivers. 2021. Orion Publishing Group
  - Aimed at general public, but (I think) better for journalists than most, because Tom C is one (and his cousin Dave is an academic economist who uses statistics).
- *Statistics Behind the Headlines*. John Bailer and Rosemary Pennington. 2022. CRC Press
  - Also by a statistician (Bailer) and a journalist (Pennington) – aimed generally but, I suspect, good for journalists.
- *Practical R for Mass Communication and Journalism*. Sharon Machlis. 2019. CRC Press
  - A lot is fairly standard stuff on how to use the R software, but probably more useful to data journalists than (other) science writers, and at least tries to use relevant examples, even if they are almost all American.
- For more on info sources, see https://www.absw.org.uk/resources/kevin-mcconway-s-statistics-reading-resources-list-for-journalists

Thanks!

- kevin.mcconway@open.ac.uk
  Twitter: @kjm2