

# **Data and statistics for journalism**

Thomas Lumley

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”

– Samuel Wilks, paraphrasing H.G. Wells

"The statistician's task is to go into the light and spread darkness"

– Scott Emerson

Everybody says that statistics should be taught. But how ?

Statistics are not simply figures. It is said that nothing lies like figures except facts. You want statistics to tell you the truth. You can find truth there if you know how to get at it, and romance, human interest, humor and fascinating revelations as well. The journalist must know how to find all these things truth, of course, first. His figures must bear examination. It is much better to underestimate than to overstate his case, so that his critics and not himself may be put to confusion when they challenge him to verify his comparisons.

He must not read his statistics blindly; he must be able to test them by knowledge and by common sense. He must always be on the alert to discover how far they can actually be trusted and what they really mean. The analysis of statistics to get at the essential truth of them has become a well-developed science whose principles are systematically taught. And what a fascinating science it is!

-- Joseph Pulitzer, 1904

# Me

- Main author of StatsChat
- Theoretical and applied medical statistician
- *Herald* and *Dom Post* and *Spinoff* subscriber

# You

- Journalism interests
- Particular data interests?

# Outline

## Context: compared to what?

- Denominators
- Trends

## Uncertainty

- Sampling error
- Random variation

## Visualisation

- Compared to what?

**Compared to what?**

# The fundamental question: compared to what?

NEW ZEALAND

## Ho ho, oh no: ACC festive injury claims show steady increase

24 Dec, 2019 05:00 AM

3 minutes to read



25 Jun, 2020 01:24 PM

Quick Read

As for which day costs ACC the most?

That would be New Years Day, which is slowly and steadily costing more in claims over the past five years, with \$3380 on Christmas Day and \$4741 on New Years Day in 2014/15.

ADVERTISEMENT

SP

Sponsored by Loan Market - NZ Herald

**up 11% over four years:  
inflation: 7.9%  
population growth: 8.8%  
down 7% per capita in constant dollars**

**The whole holiday period has lower accident costs than a typical working day**

# Denominators

**idnap  
bar**

Figures show that during the past two years a total of 9703 teenagers were found guilty of drink driving in New Zealand. Fifteen teens have been convicted at least six times in that period.

Auckland teenagers had the highest number of convictions for drink-driving, with 1553 in the past two years.

..... Christchurch teenagers weren't far behind, with 1383, while 793 Hamilton teenagers were convicted and 728 Wellingtonians.

[stuff.co.nz](http://www.stuff.co.nz), 29 October 2012

Rate of convictions per capita is *lowest* in Auckland  
Auckland + Wellington have about *half* the national average rate  
Christchurch rate is high.

# More students cheat in exams, and most are in Auckland

4:26 PM Thursday Jun 30, 2016

NZ Qualifications Authority

SHARE: [f](#) [t](#) [G+](#) [in](#)



More students are cheating in NCEA exams. Figures show. Photo / iStock

159/290 means Auckland is high

25/290 means Northland is *much* higher.

Also, rate is 2 per 1000, so we're probably missing a few

## 2015 cheats by region:

- Auckland - 159
- Bay of Plenty - 6
- Canterbury - 23
- Central Plateau -
- East Coast - 3
- Hawke's Bay - 12
- Manawatu - 7
- Nelson/Marlborough - 5
- Northland - 25
- Otago - 5
- Southland - 2
- Taranaki - 4
- Waikato - 20
- Wairarapa - 2
- Wanganui - 1
- Wellington - 27
- West Coast - 4
- Cook Islands -

Wednesday, 21 October 2020

## Rate of assaults in Wellington 10 times higher than nation's average

An analysis of the country's crime statistics by Dot Loves Data shows the rate of assaults in Wellington is 10 times higher than the national average.

It considered all reported crimes over a five-year period, and found in the capital there were 2056 counts of assault and 176 counts of sexual assault reported.

Those crimes were centred around the Cuba-Courtenay precinct, comprising most of the city's nightlife and bars.

"In Wellington, assaults and sexual assaults stick out," said Dot Loves Data's

# Denominator? Also too, comparison?

# Special investigation: Auckland house prices

5:00 AM Saturday Jul 11, 2015

Auckland Region

China

NZ Labour Party

National



9.1k

54

108

10

**Exclusive: Leaked figures support claims that Chinese investors are a big influence on Auckland's overheated property market.**



Labour claims the data shows for the first time the scale of an issue that is pricing first-home buyers out of the Auckland market. Photo / Doug Sherring

The first picture has emerged of Chinese buying patterns in Auckland's pressure-cooker housing market — and it suggests a powerful, big-spending influence.

**What do you think the Government should do about overseas-based investors buying houses in Auckland?**

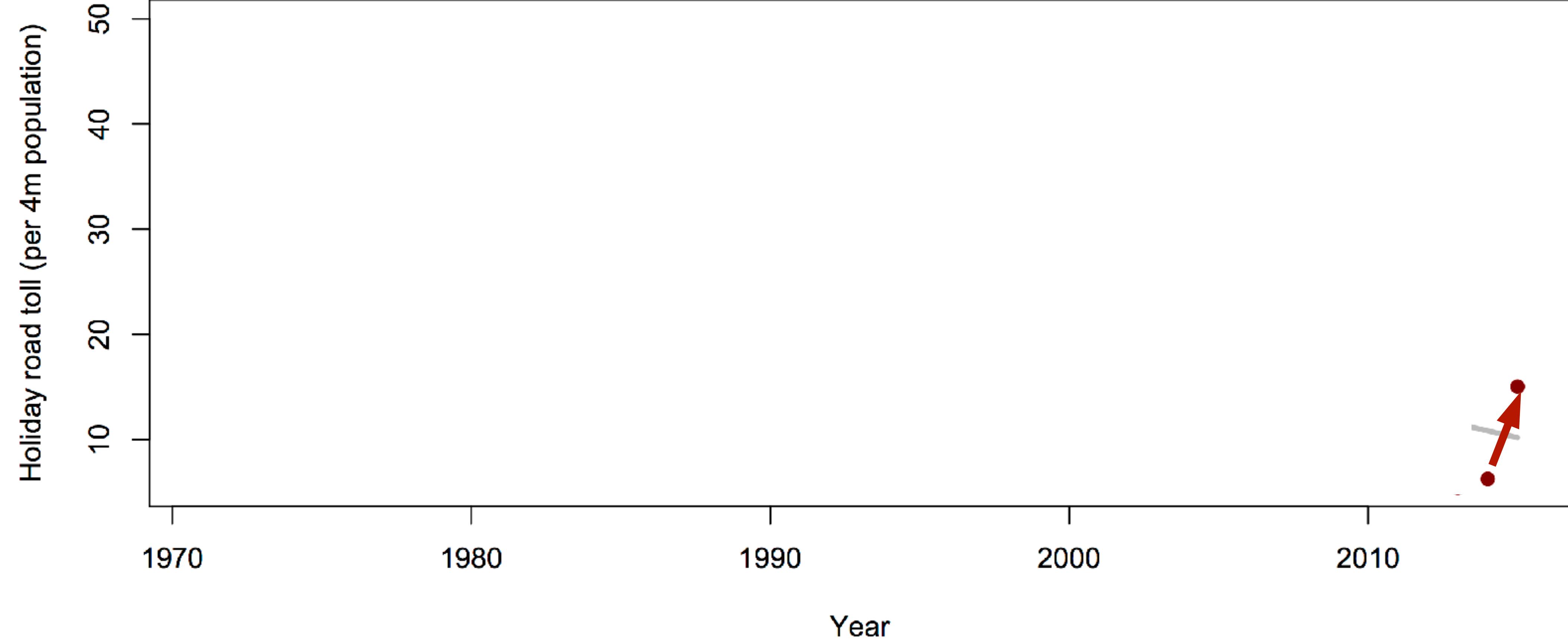
Herald, 11 July 2015

# Your turn

## From the pre-reading

1. Read **Infidelity: it's a right-wing, meat-eaters' thing** [NZ Herald, January 2012](#). This starts “*If your partner supports National, has a PC, drinks Coke, eats meat, has a tattoo, smokes and is a Christian, be warned - they could be a cheater.*” Why is this misleading? How different was the proportion supporting National from the population proportion? How meaningful do you think the differences mentioned between cities are? How do you think this story came to be written?
2. Read **Auckland student brings gun to university to ‘use as a film prop’** [Stuff, August 2020](#). Roughly what proportion of tertiary students are subject to disciplinary action for cheating? Do you think that’s a large fraction of those who cheat? Based on the numbers here, what can you say about the rates of sexual harassment and sexual assault at NZ universities?

# Trends



## *Building a safer New Zealand*

*Crime is down*

**2008**

**1011\***

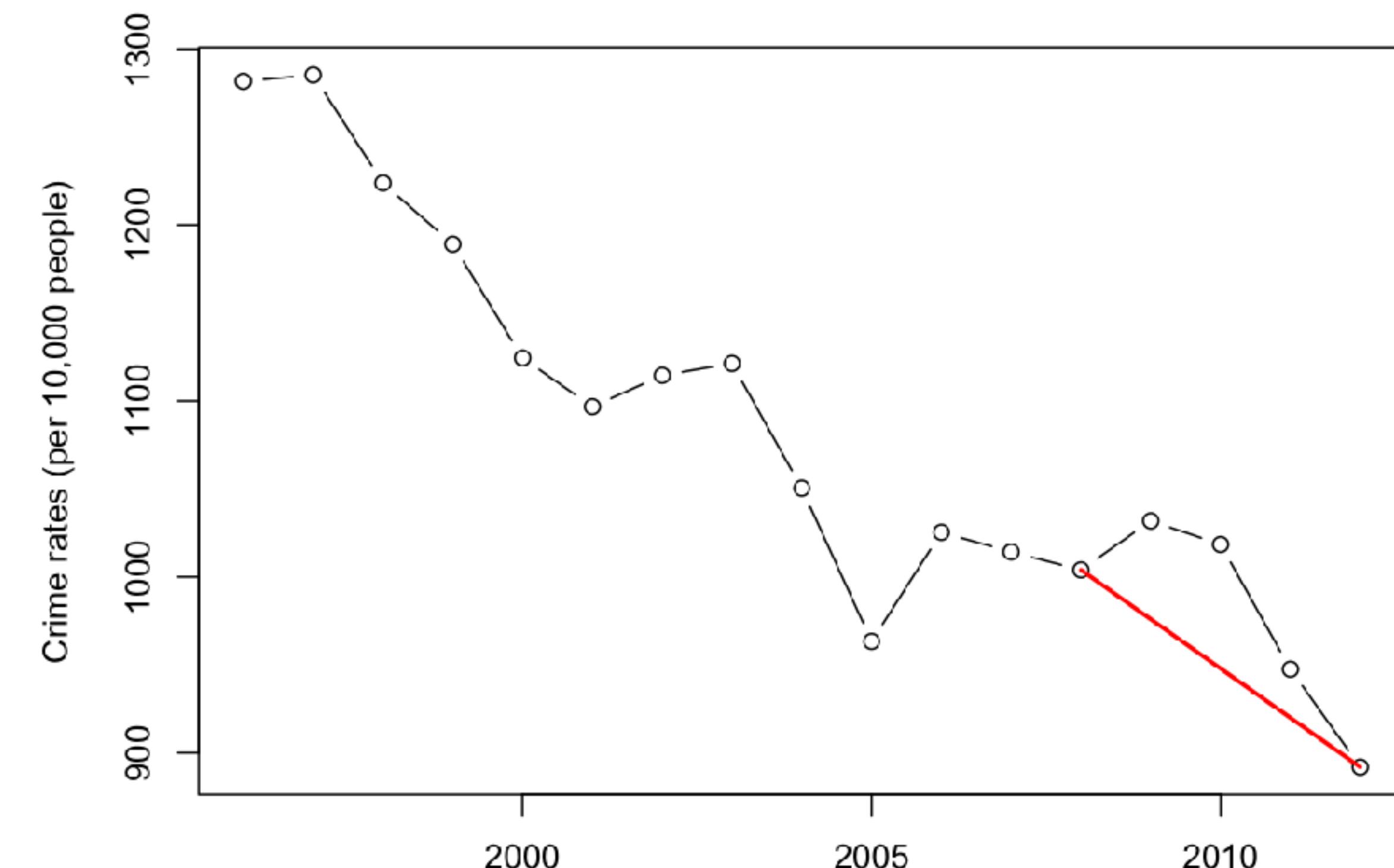
\*Recorded crimes  
per 10,000 people

**2012**

**848\***

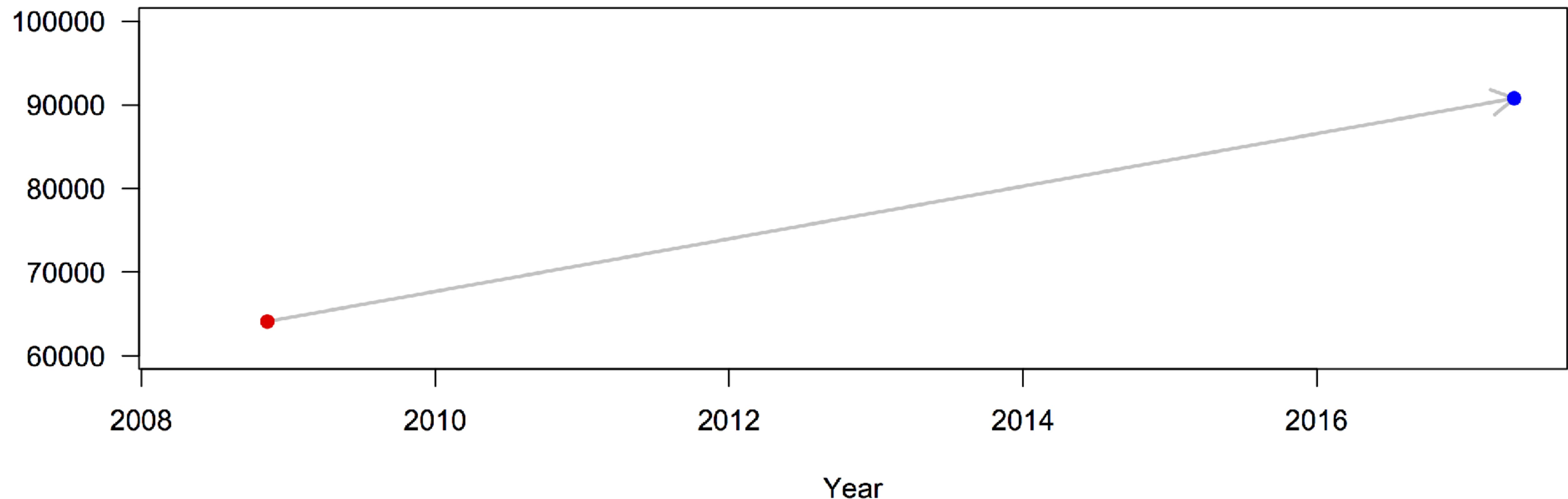
*The lowest crime rate in 30 years*

Authored by John Key MP Financial Wing, Parliament, Wellington Street, Wellington



**“Under National, the number of young people not earning or learning has increased by 41%”.**  
**(Labour ad)**

### **Young people not earning or learning under National**



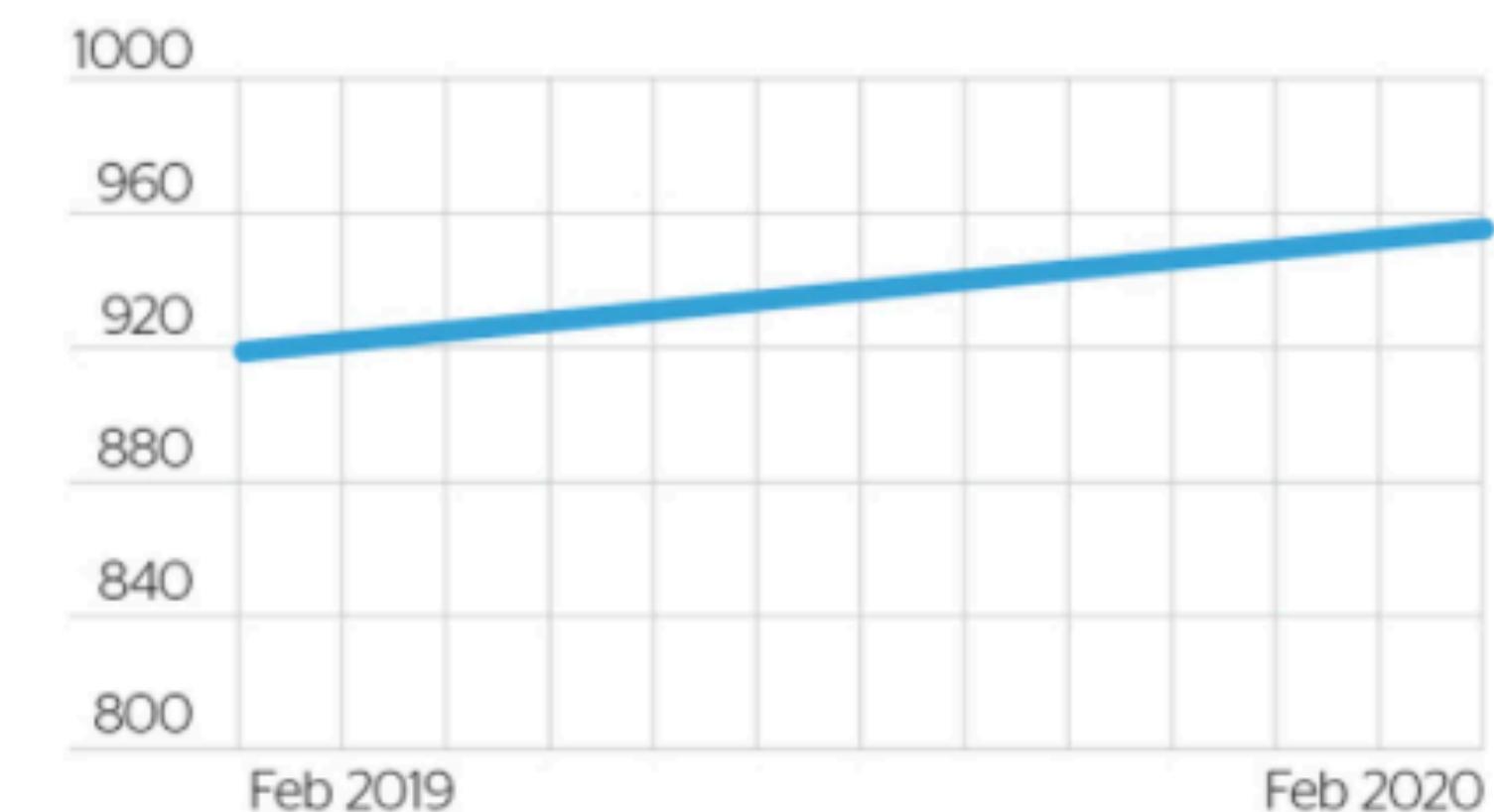
# Your turn

Dr Shane Reti wrote in the Northern Advocate and then in the Herald, arguing that Covid lockdowns had harmed mental health, with these graphs of hospital admissions for intentional self-harm as evidence?

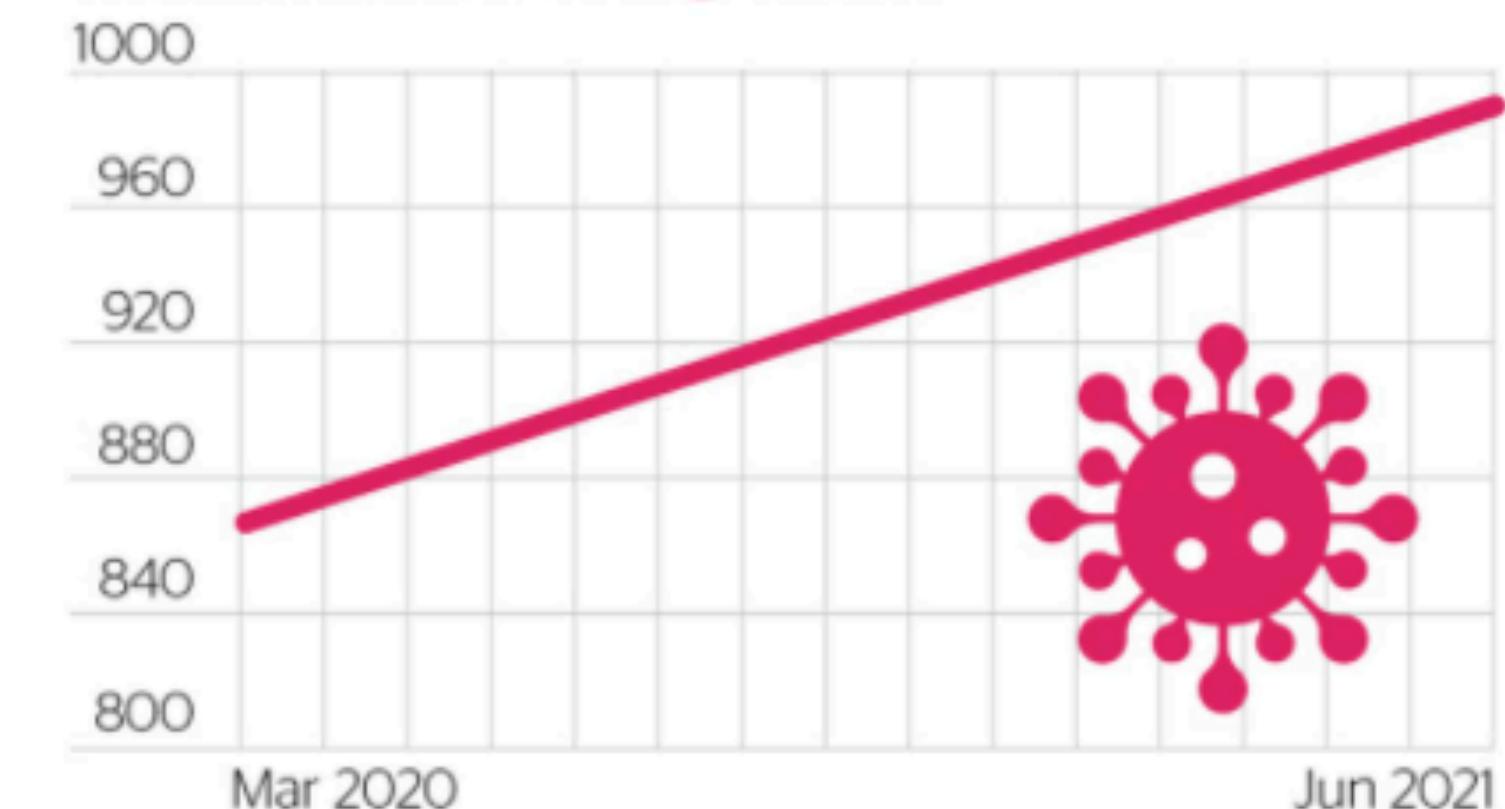
- What's wrong with these charts?
- Assuming the data are correct, what do they say about the impact of Covid restrictions?
- What data do you need to comment?
- What major non-statistical questions arise?

## Intentional self harm

Admissions Pre Covid



Admissions During Covid



Herald Network graphic

# Break

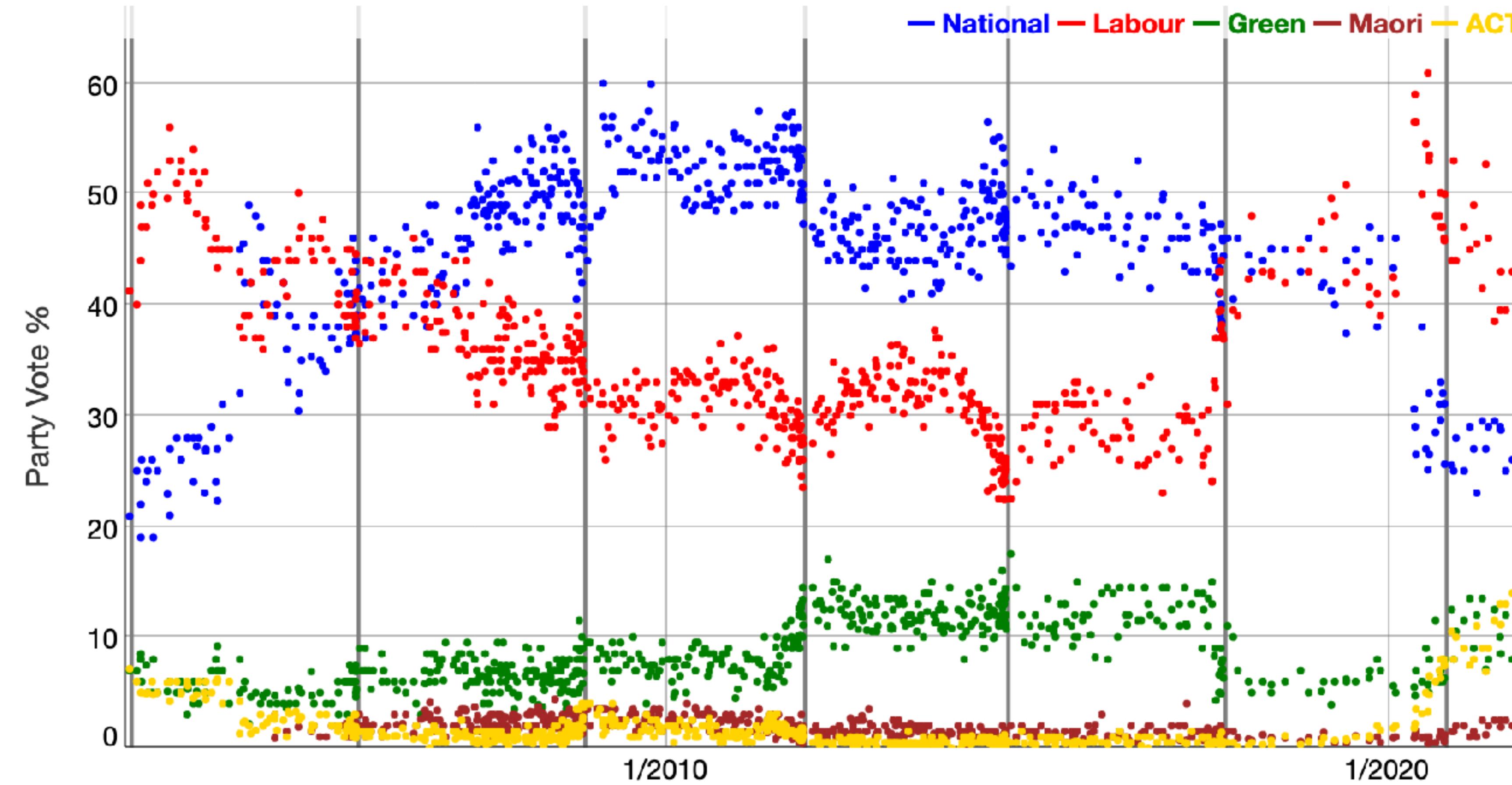
See you in a few



# Uncertainty

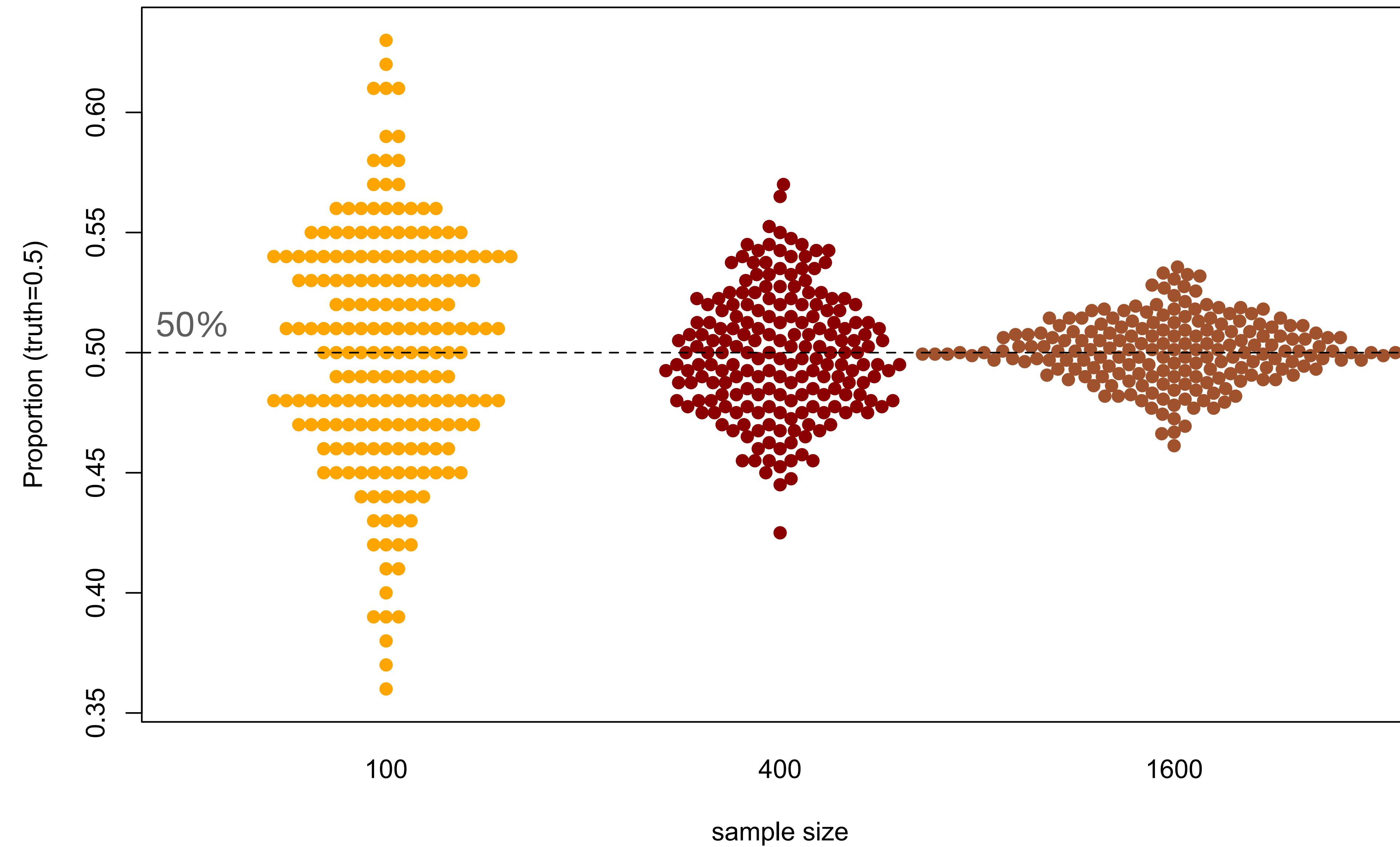
'Statistics can be reduced to the question "Why aren't all these numbers the same?"'  
Richard Arnold, VUW

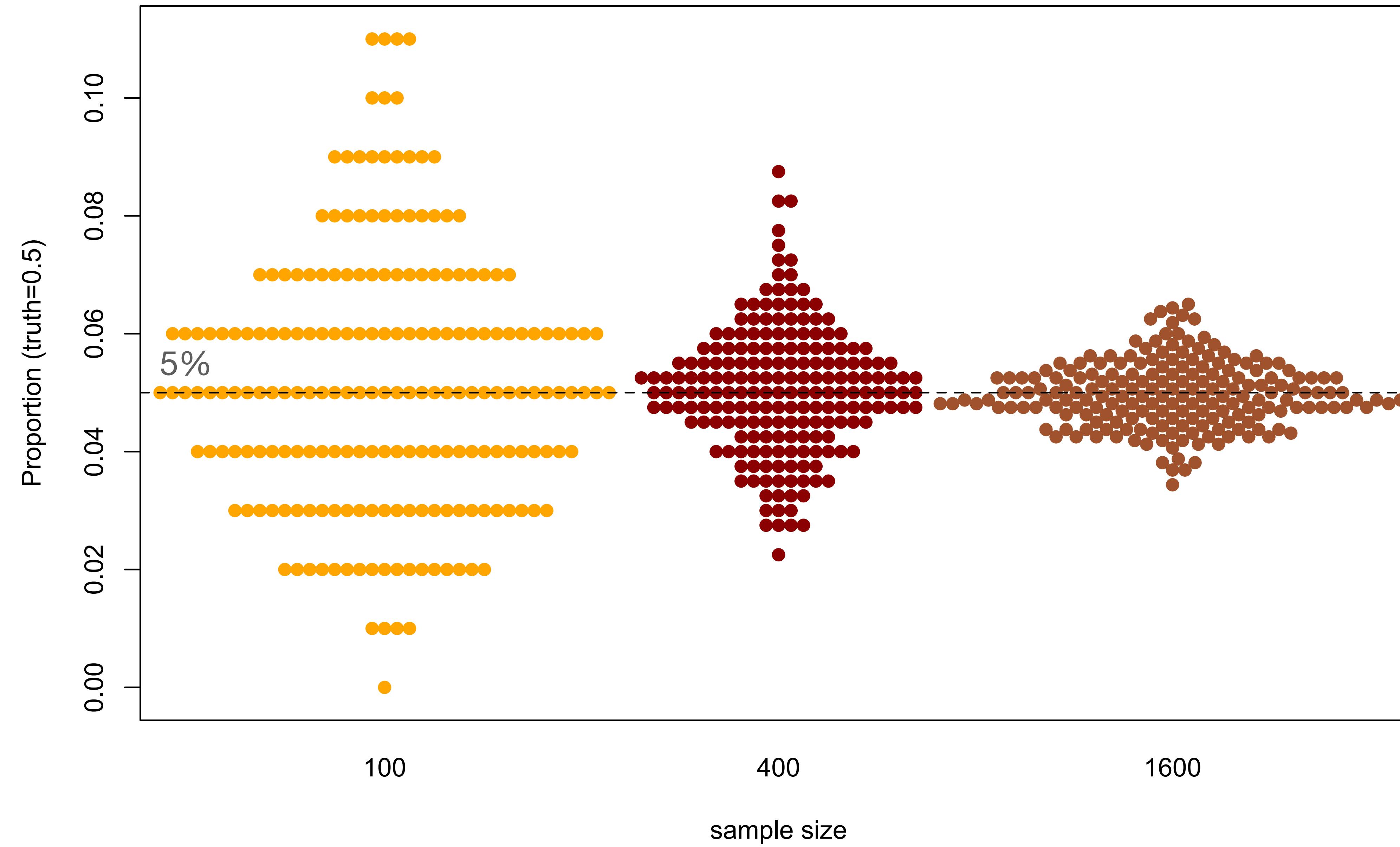
# New Zealand Political Opinion Polls

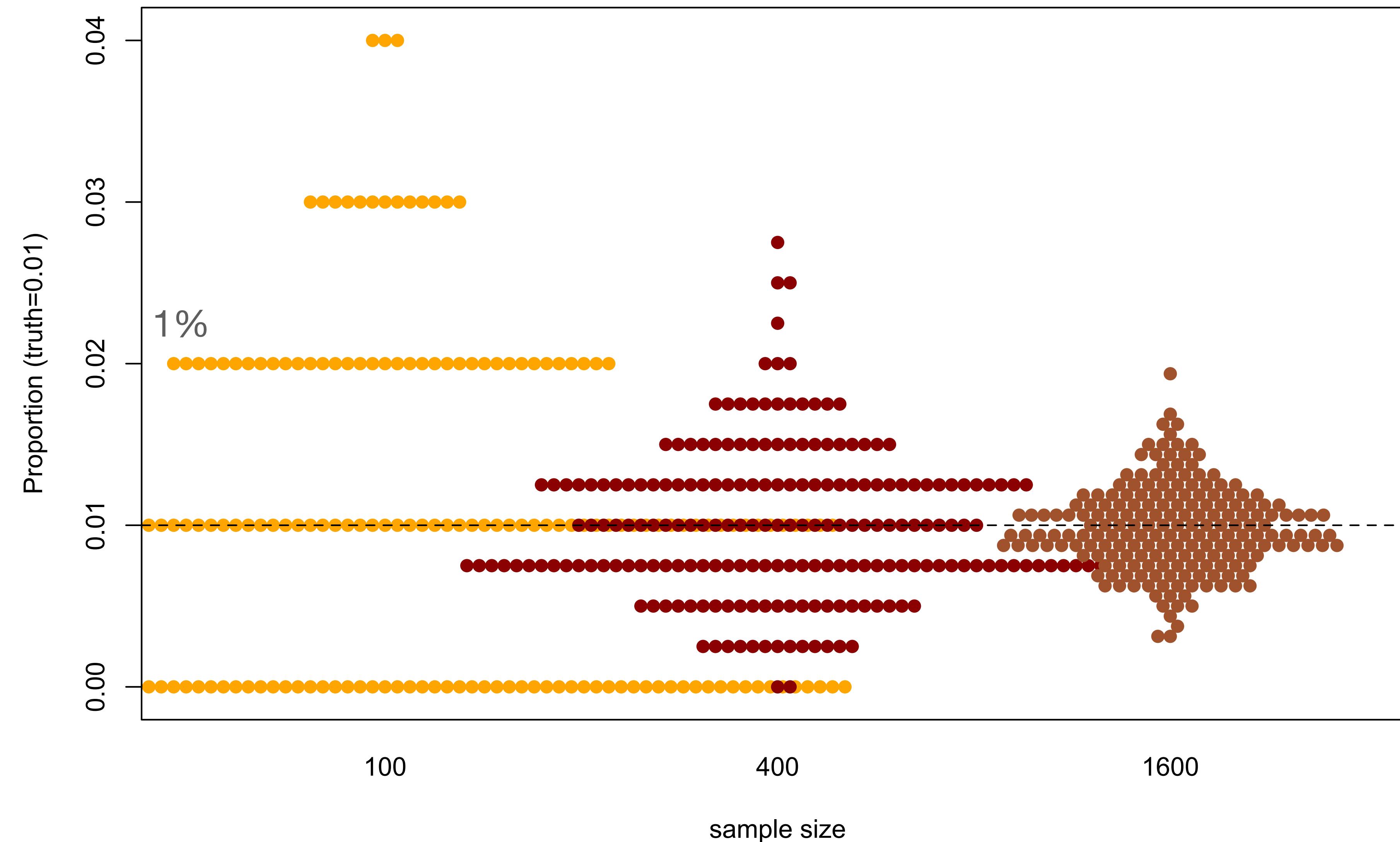


<https://www.andrewchen.nz/polls>

# Sampling uncertainty







# Mathematically ideal surveys

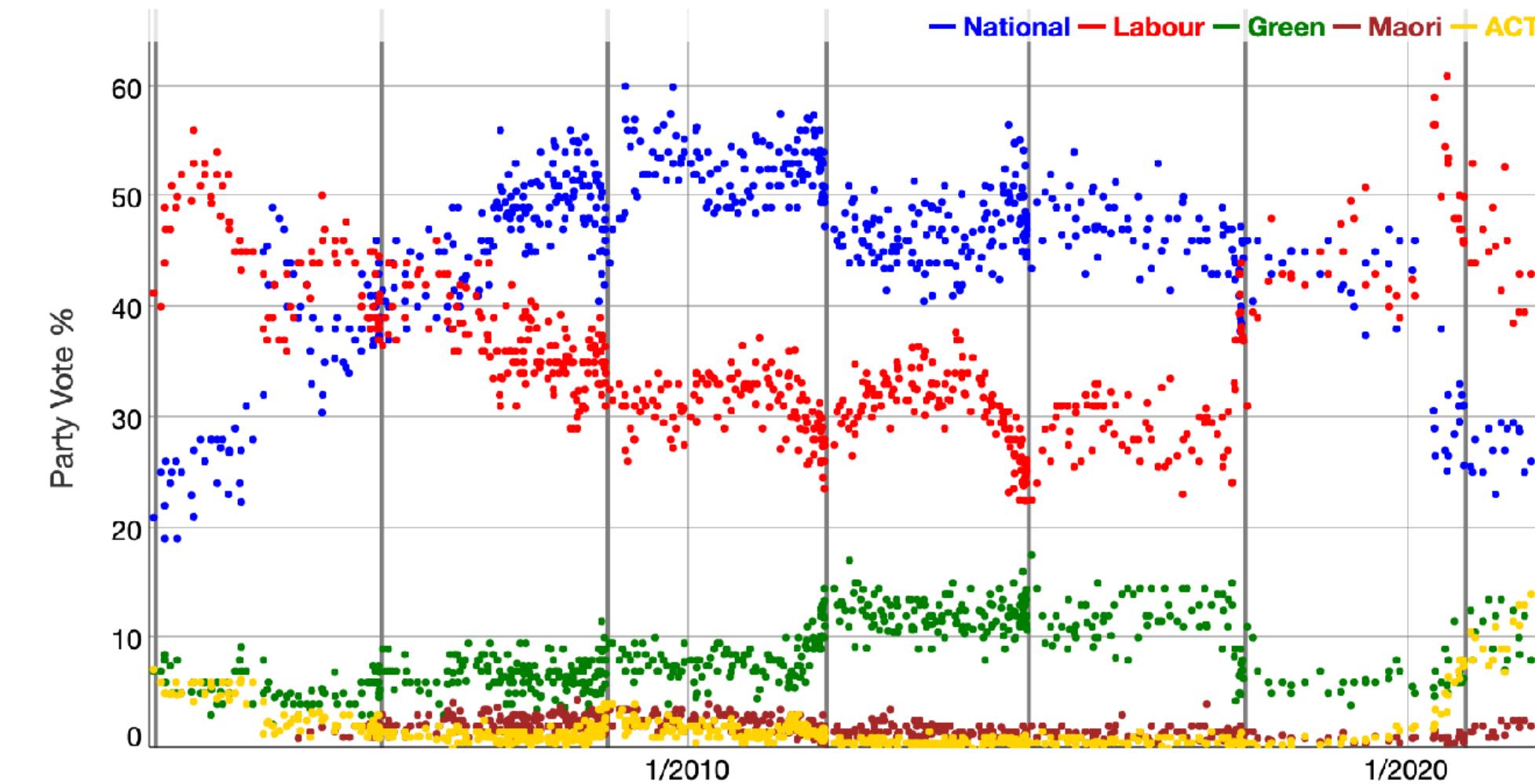
- Absolute error is smaller for small groups (eg minor parties)
- Relative error is larger for small groups
- You need four times the sample size to get half the error
- 1000 people gives 3% 'margin of error' for proportions near 1/2
- Margin of error for poll-to-poll change is 4.5%
- Magic formula for small groups: (or [here](#))
  - take **count**, square root it
  - add and subtract one to get two numbers,
  - square them, then divide by total sample size to get proportions.

# Real opinion surveys

- Not a perfect random sample: most people refuse/aren't contacted
- Data are reweighted to look like the whole population
- Error is still **larger** than for ideal surveys: maybe 1.5 times higher

Most 5% changes from poll to poll  
don't mean anything

New Zealand Political Opinion Polls



# Bogus polls

- Self-selected online clicky things with no reweighting
- Completely uninformative. Don't even try to interpret them quantitatively

We asked motorists why they thought the road toll had climbed, what police should target and what was the worst driving they saw last year



**Douglas Barker, 43, army officer, Waiouru**  
1. "The road toll is probably a statistical function of more people being on the road."  
2. "I don't know if it's the police's problem, the problem is one of individual responsibility."  
"The other day a guy overtook us while I was going 90kmh. It was on a two lane to get one."

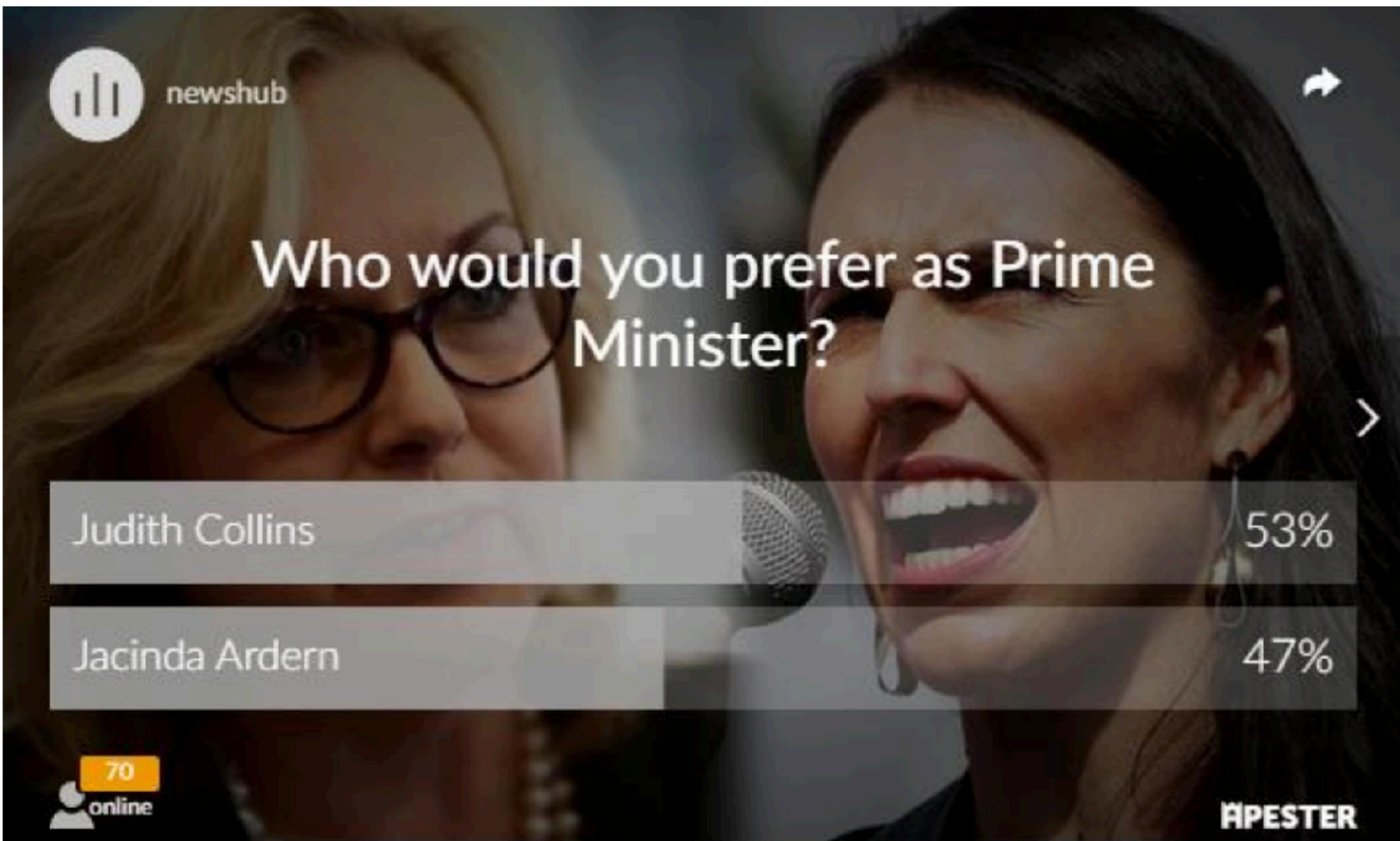
**Mike Courtney, 51, computer engineer, Otaki**  
1. "I would imagine a lot of those deaths are youths, I guess it comes down to experience and driving skills."  
2. "The Government should introduce a mandatory defensive driving course for anyone getting a licence. It's about education and

**Lindsay Ross, 48, analyst, Kaiwharawhara**  
1. "I think the road toll's just one of those random things, it just ebbs and flows."  
2. "Slow drivers. People doing 70kmh in a 100kmh zone – people are going to want to pass them."  
3. "Just last week a guy overtaking a campervan. He ran out of road

**Paul McKee, 48, tradesman, Stokes Valley**  
1. "There's going to be a natural variation but it's ironic that when the police target speed the road toll seems to go up."  
2. "Basically people who don't know the road rules."  
3. "I met someone coming the wrong way down a one way street in

**Lloyd van der Krog, 52, IT, Karori**  
1. "I think it's a little bit random but hopefully it'll come back down."  
2. "Binge drinkers and recidivist offenders."  
3. "Serious red light runners. Buses on Lambton Quay are the worst red light runners."

Dominion Post, 2015-1-2



**1 News/Colmar Brunton**  
54% Ardern  
20% Collins

**3/Reid Research:**  
62% Ardern  
15% Collins

Newshub, 15 July 2020

## Who won tonight's leaders' debate?

3250–3300 votes

David Cunliffe. 49%

John Key. 48%

It was a draw. 3%

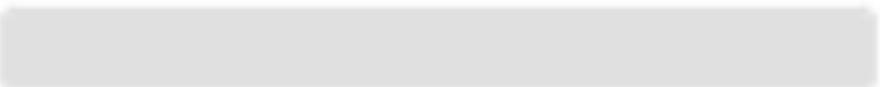
NZ Herald

## Who won the debate?

John Key  36 %

David Cunliffe  63 %

Neither/It was a tie  1 %

Unsure  0 %

Newstalk ZB

## Which leader impressed you most in tonight's debate?

1 David Cunliffe 39%

2 John Key 61%

The question drew 45,898 responses.

TVNZ

28 August 2014

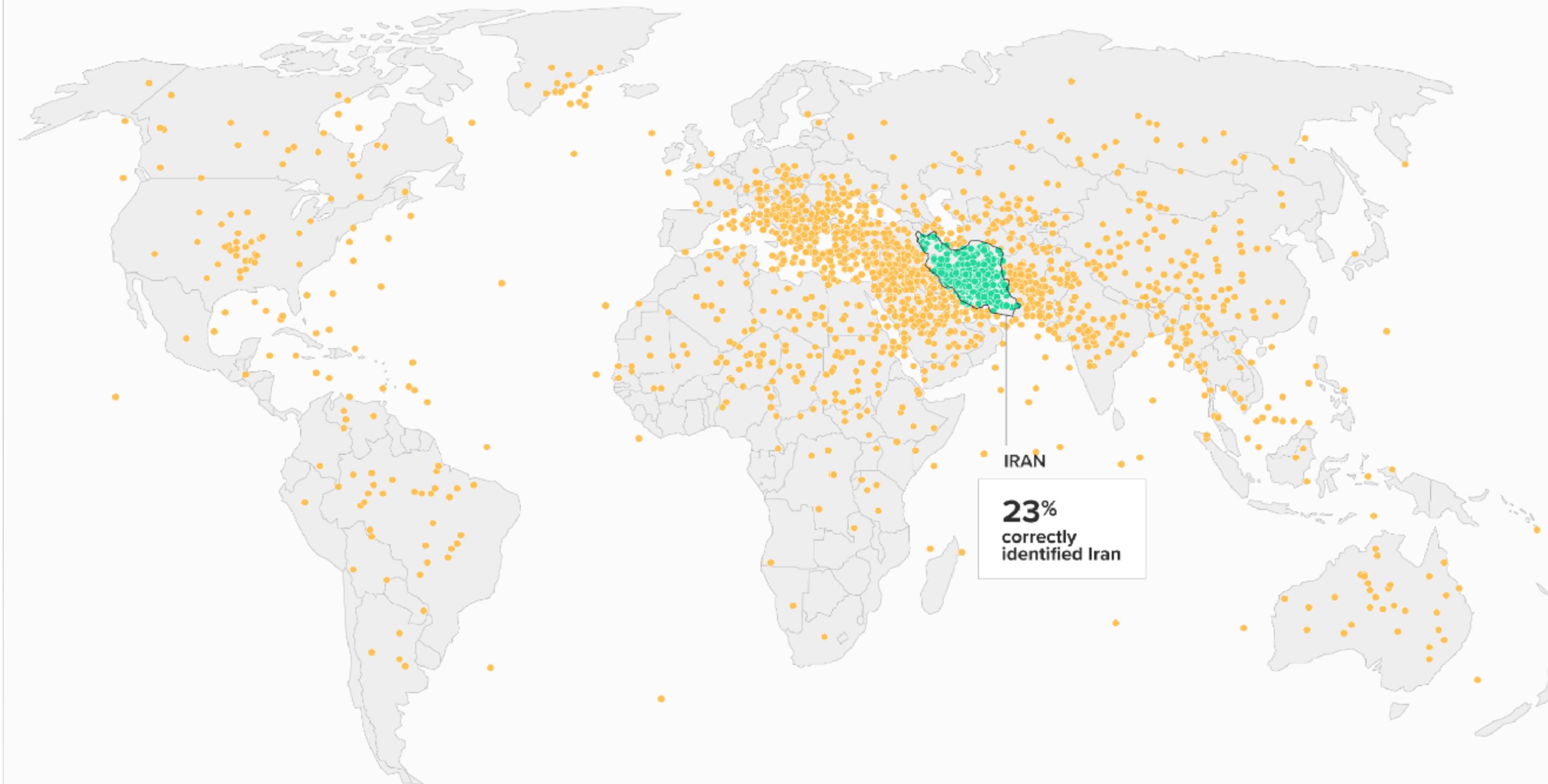
*While the Middle East saw definite clustering, some respondents believed — among dozens of wild responses — that Iran was located in:*

- *The U.S.*
- *Canada*
- *Spain*
- *Russia*
- *Brazil*
- *Australia*
- *The middle of the Atlantic Ocean*

-- Rashaan Ayesh, Axios

(January 2020)

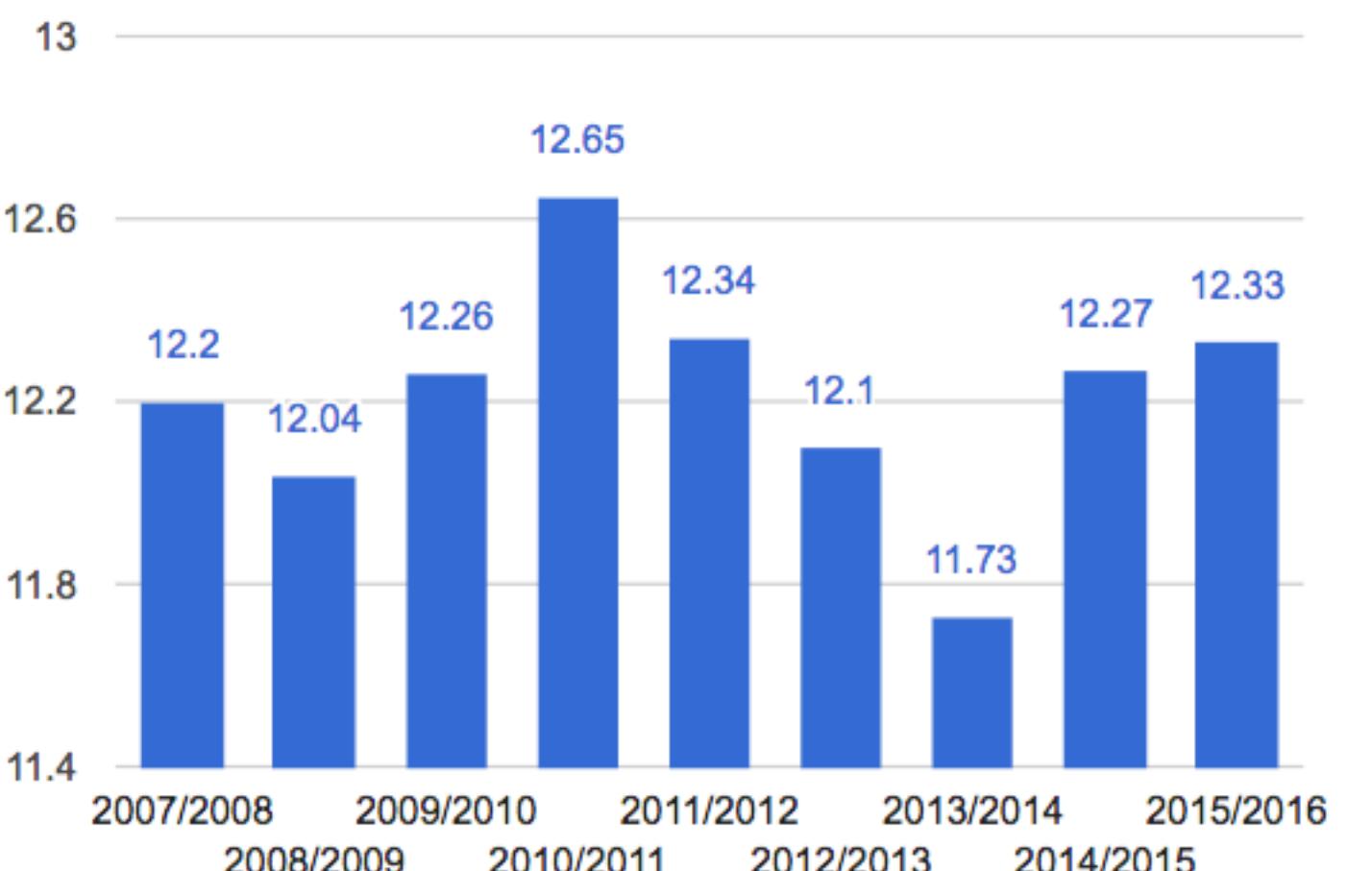
Respondents were asked to identify Iran on the map



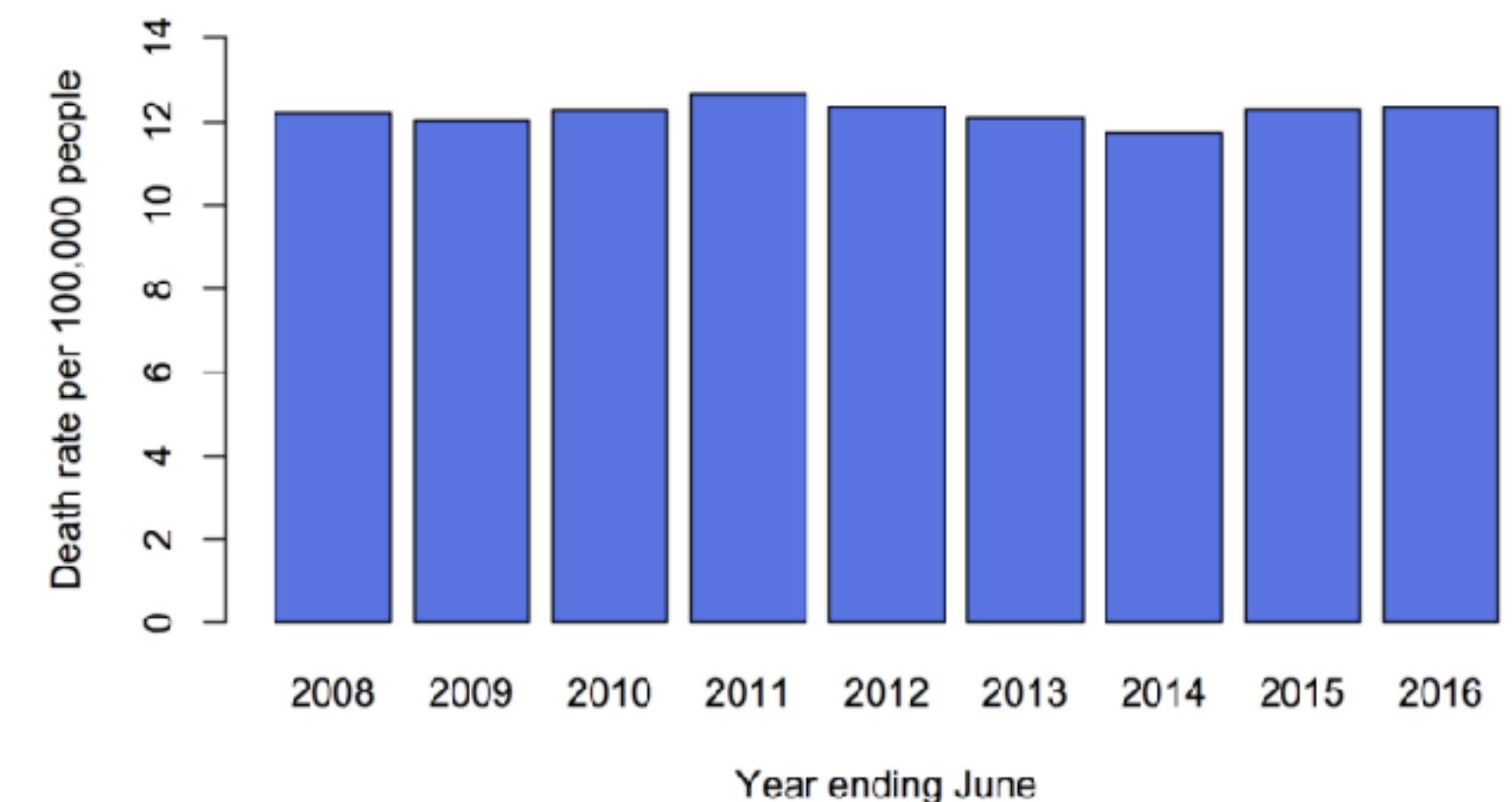
Each ● represents a respondent's guess. Each ● represents correct placement.

# Random variation

**New Zealand provisional suicide death rate per  
100,000 people**



Herald, 2016-10-18



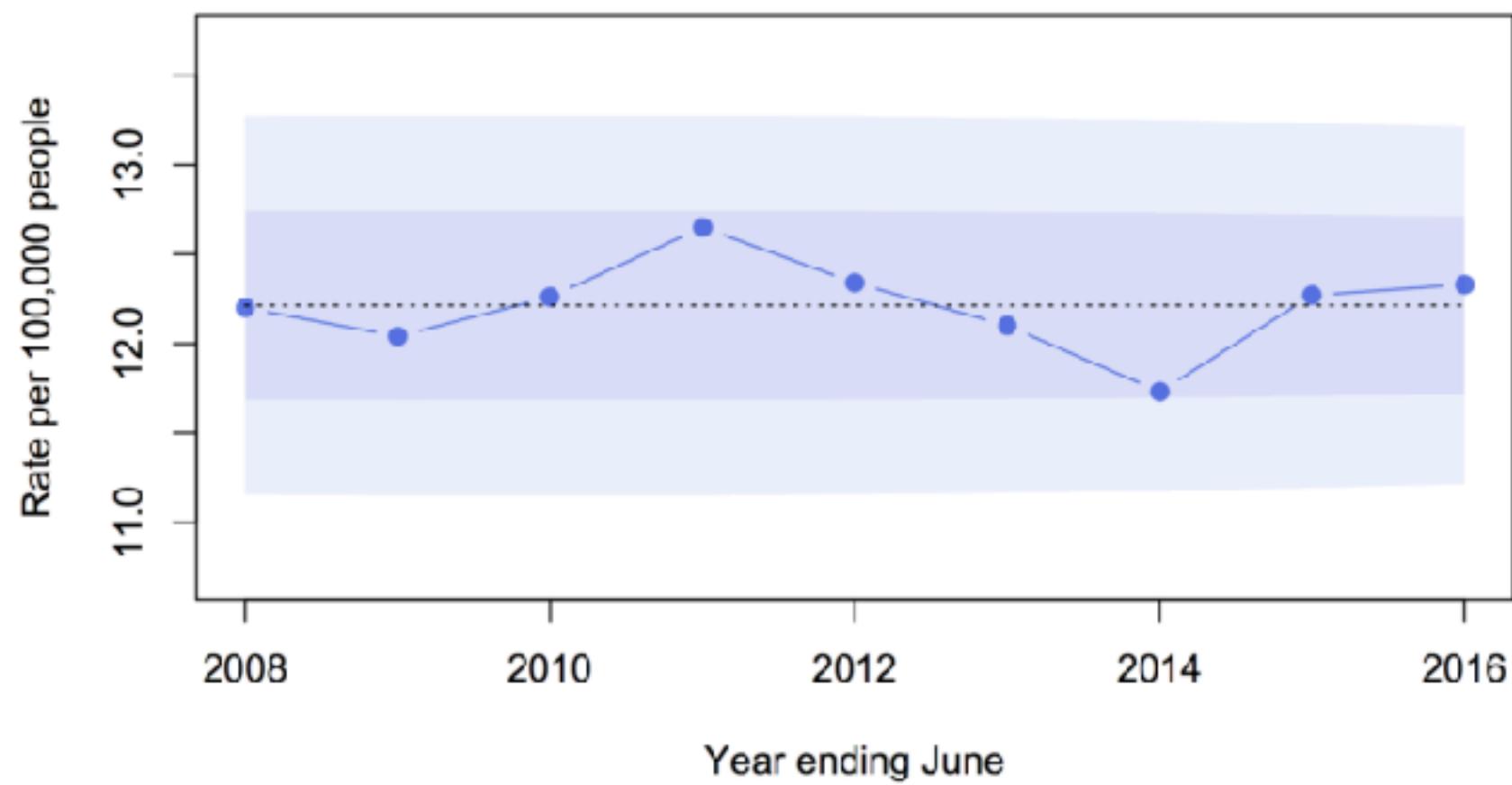
Simplest mathematical model for count data: "Poisson"

Standard deviation = square root of mean

+/- 1 standard deviation: about 2/3 of points

+/- 2 standard deviations: about 95% of points

Real variability is usually bigger than this



# Rising road toll alarms police

MICHAEL DALY

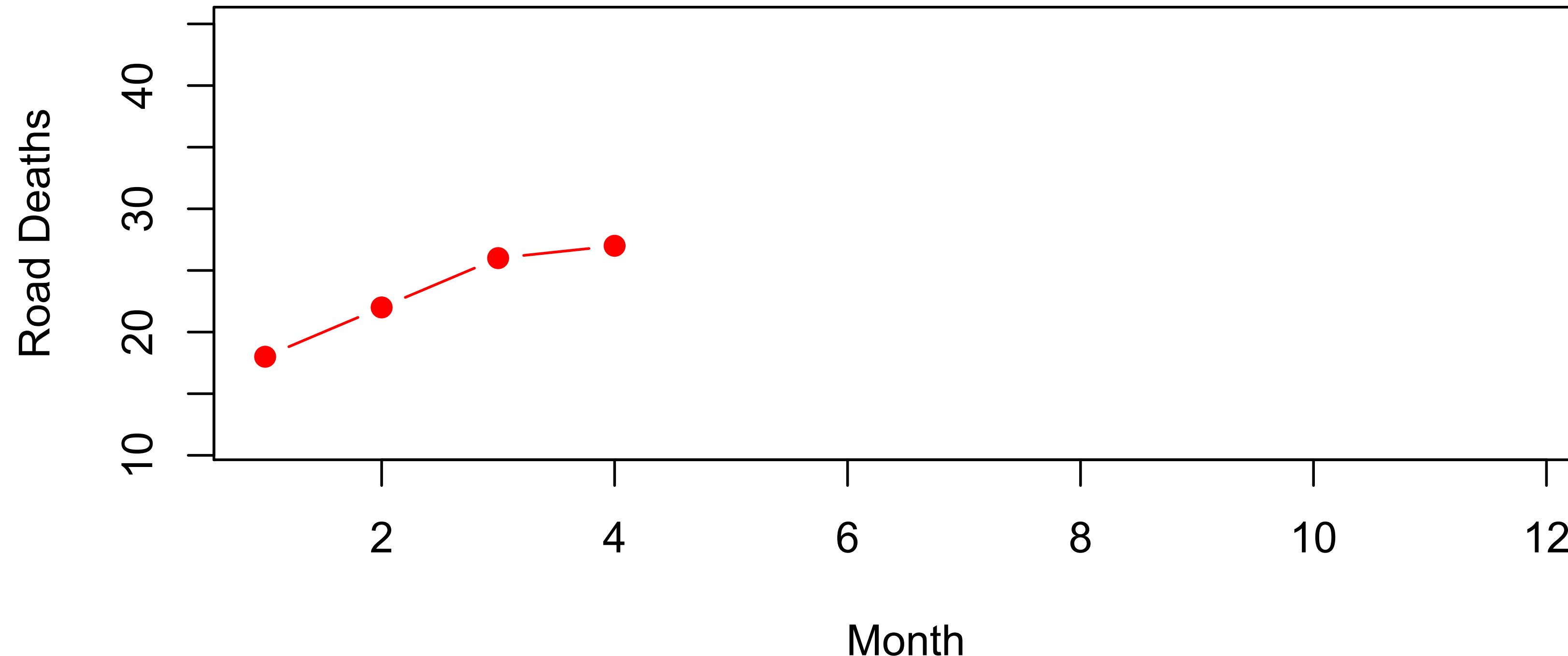


Last updated 11:10 07/05/2014

Like 92

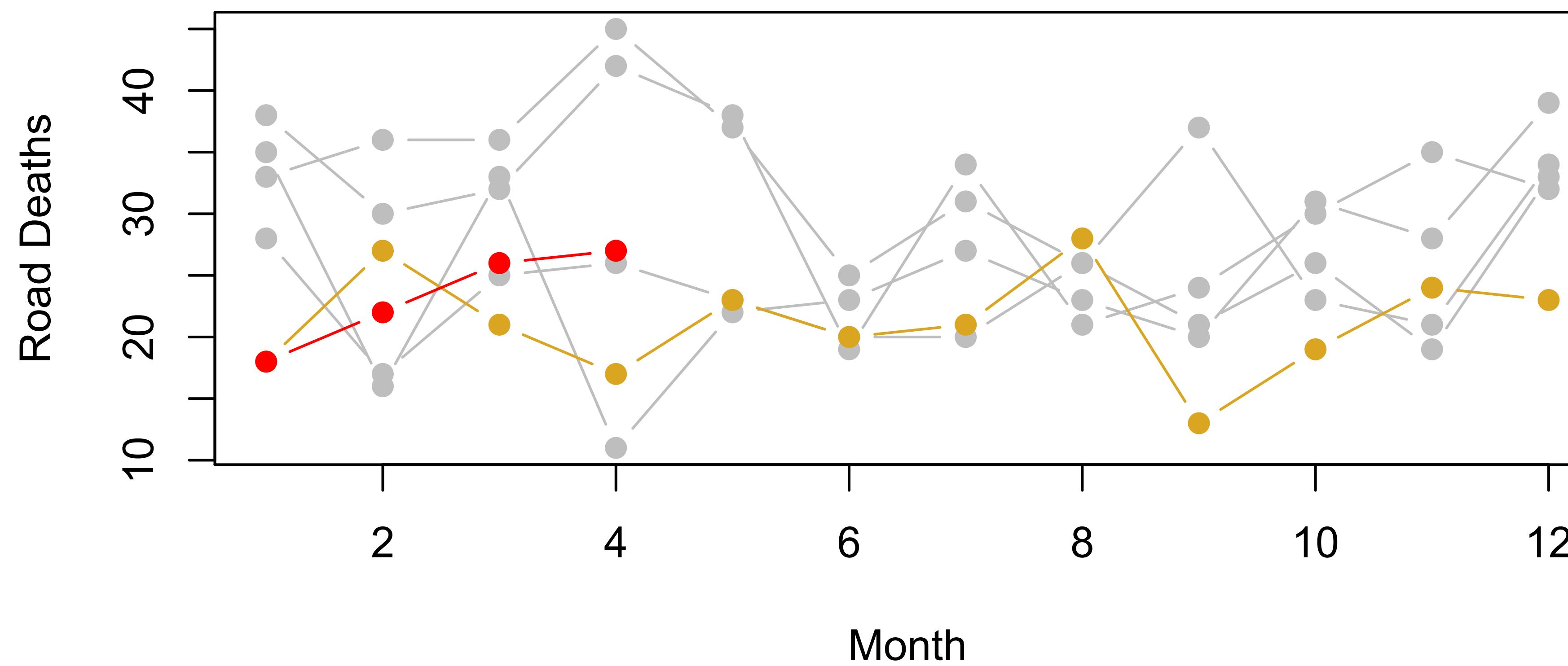
Tweet 9

g+1 Share



"Poisson" variability is about  $\pm 5$  for 1 standard deviation,  $\pm 10$  for 2

# Historical variability



# Uncertainty: summary

- Quantifying uncertainty tells you if the data can support a story
  - There could still be a story if they can't, it's just not based on the data
- For well-conducted surveys, we know the uncertainty fairly precisely
  - (bogus polls aren't data)
- For data that count things, we can get a useful lower bound
- Historical variability is also a good guide

People tend to underestimate random variability in almost every field.

**Break**  
See you in a few

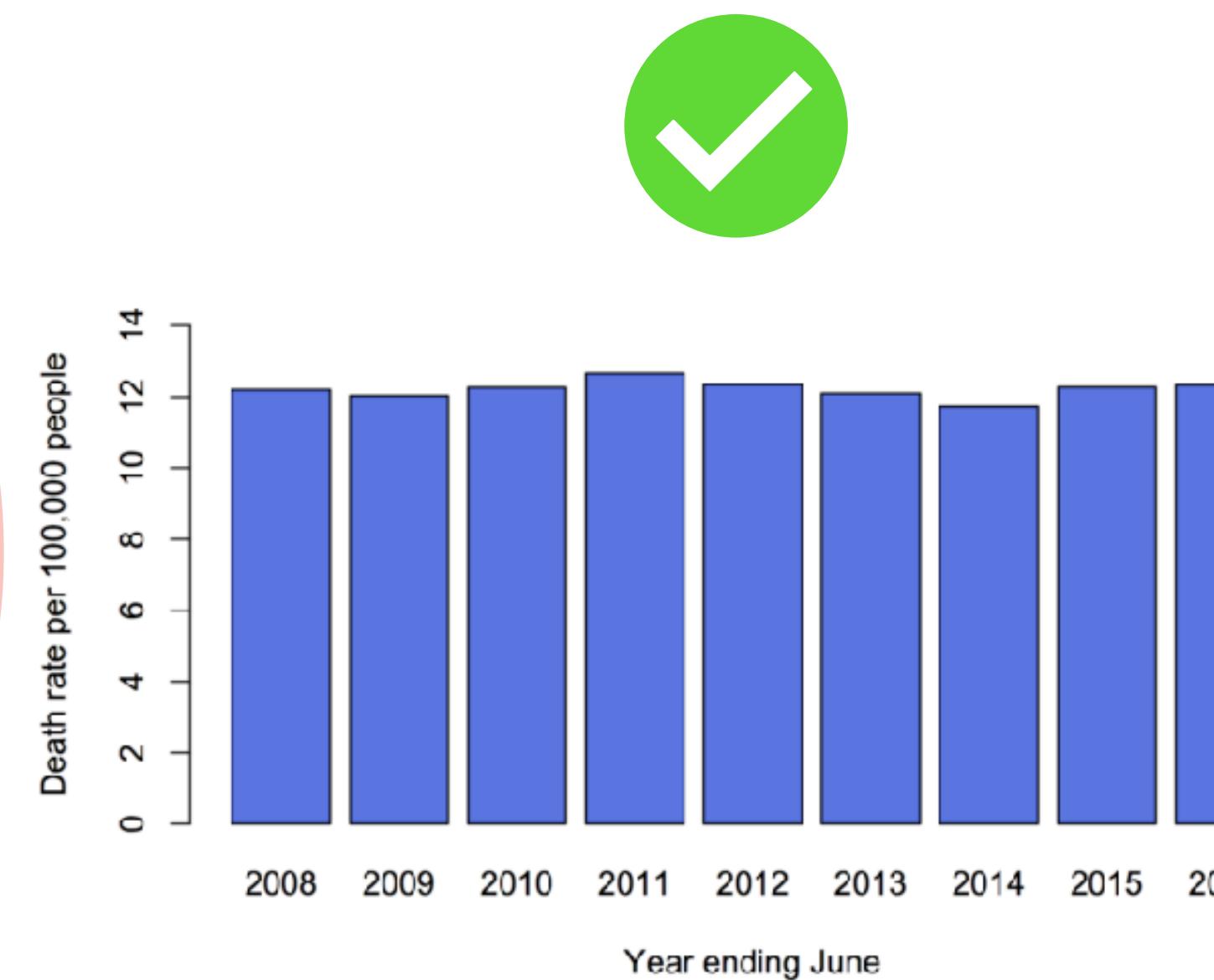
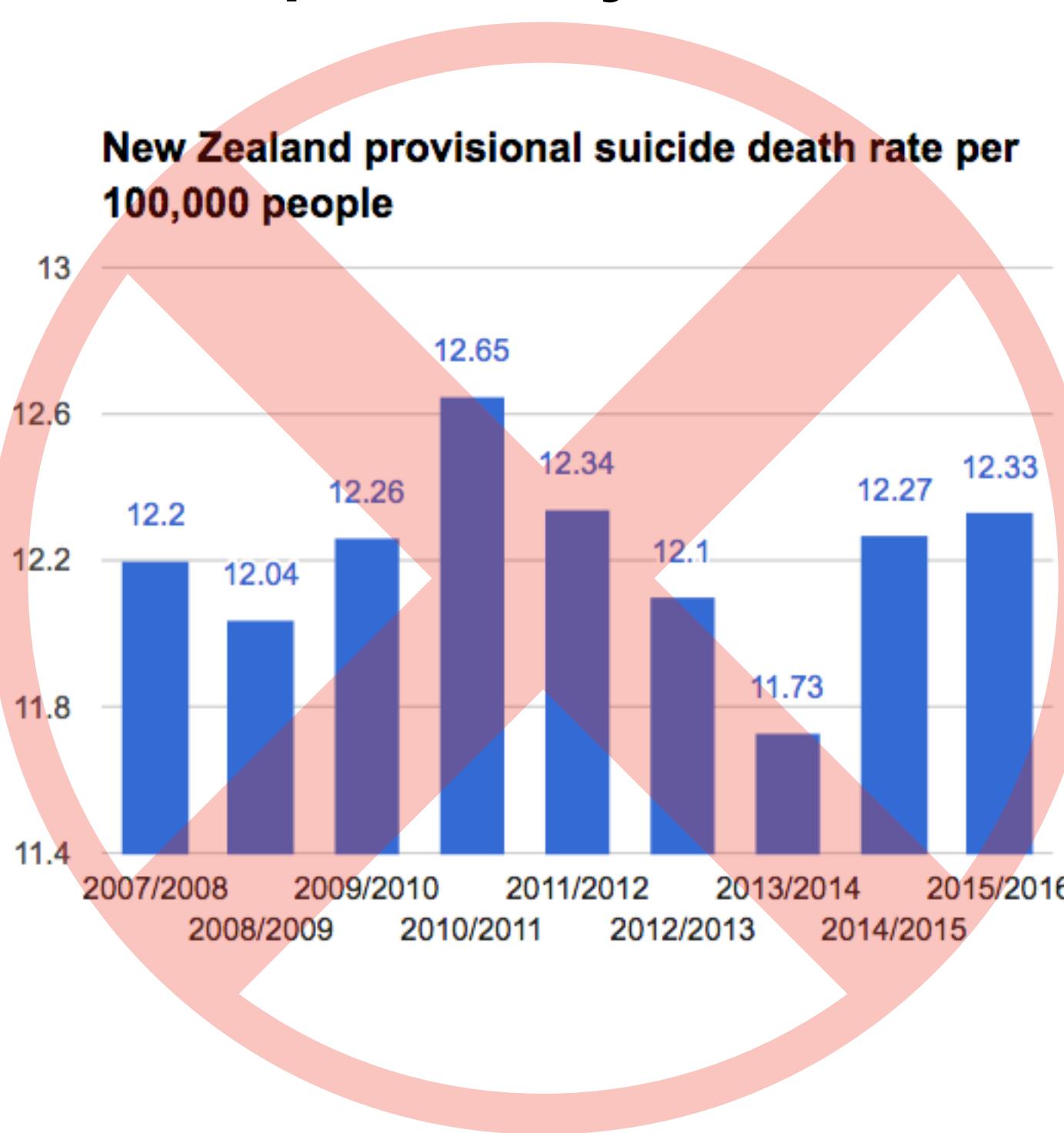


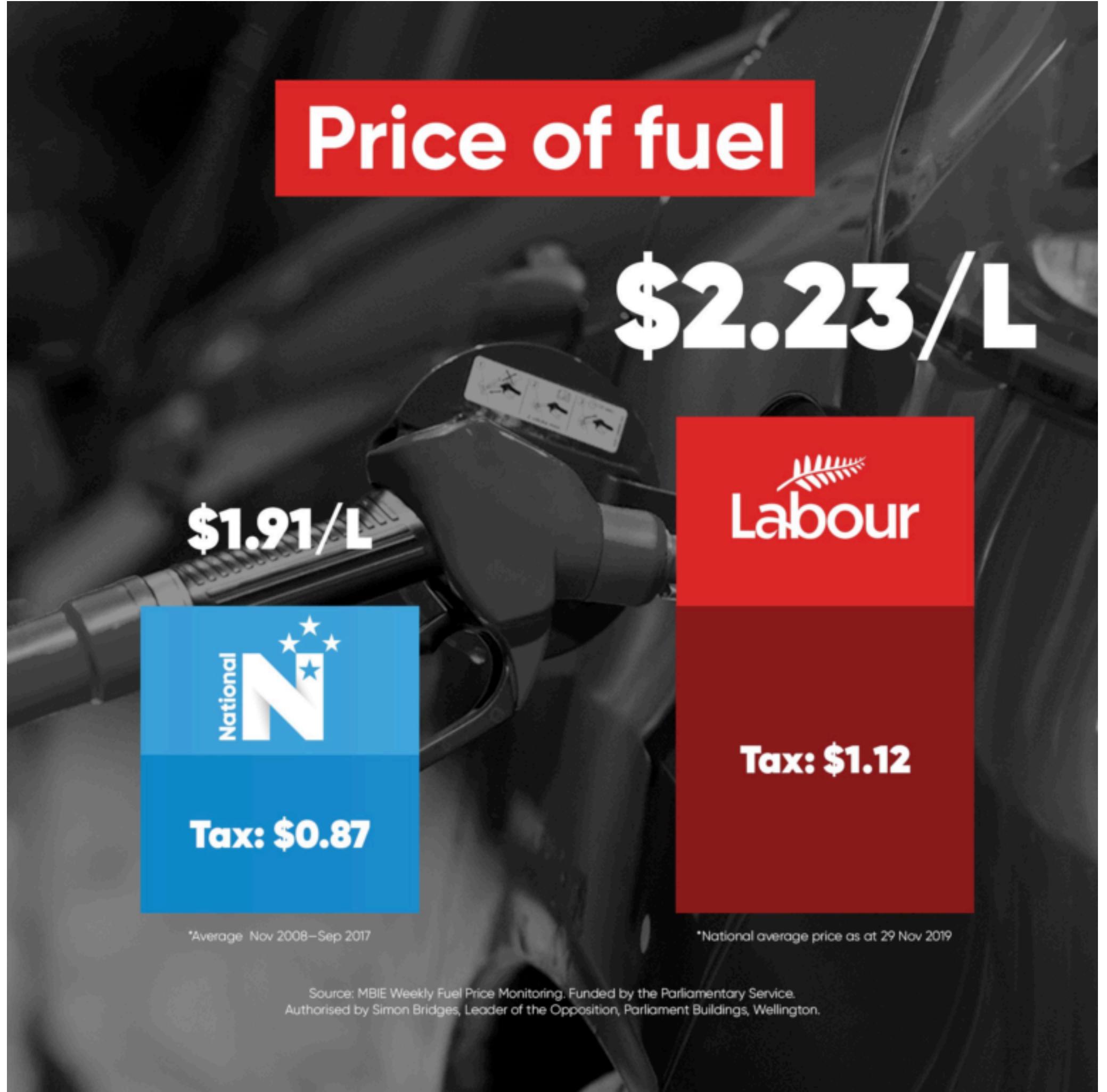
# Data visualisation

"A language is something you can use to lie"  
-- paraphrase of Umberto Eco

# Principles

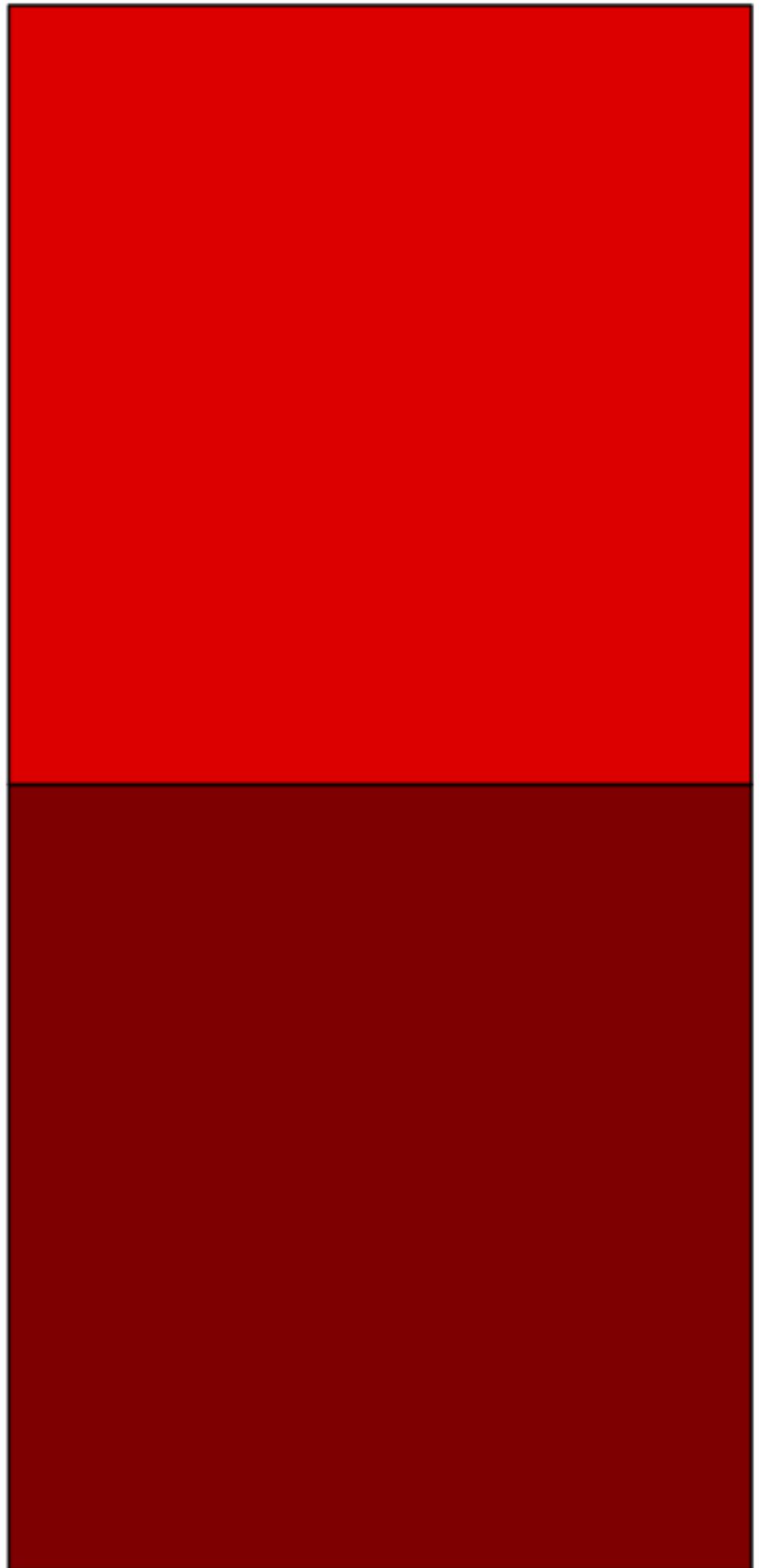
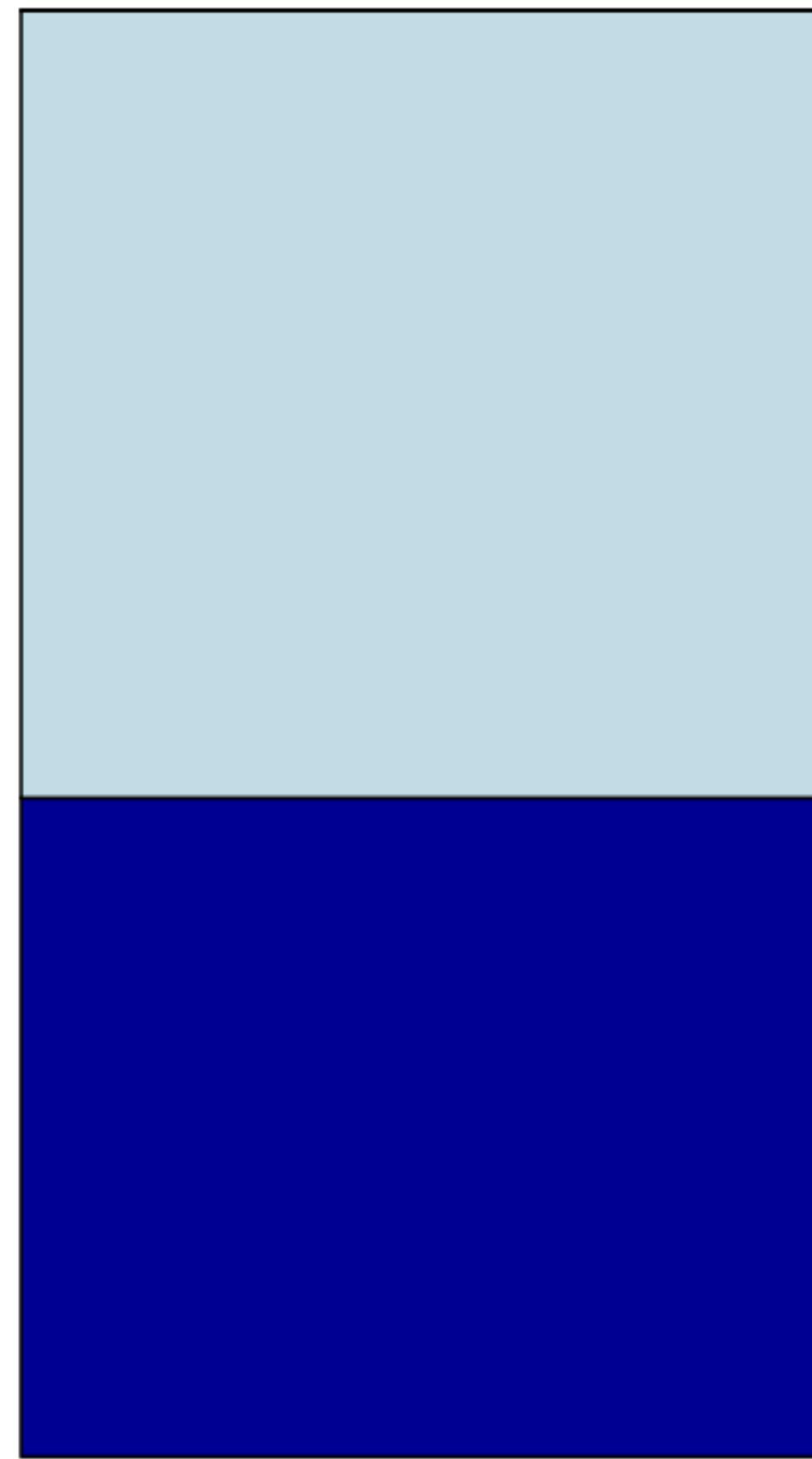
- Data visualisation is for communicating numbers, not decoration
- Communicating numbers is usually about comparisons
- The primary choice is how numbers are coded in the graph





Nominal \$/L

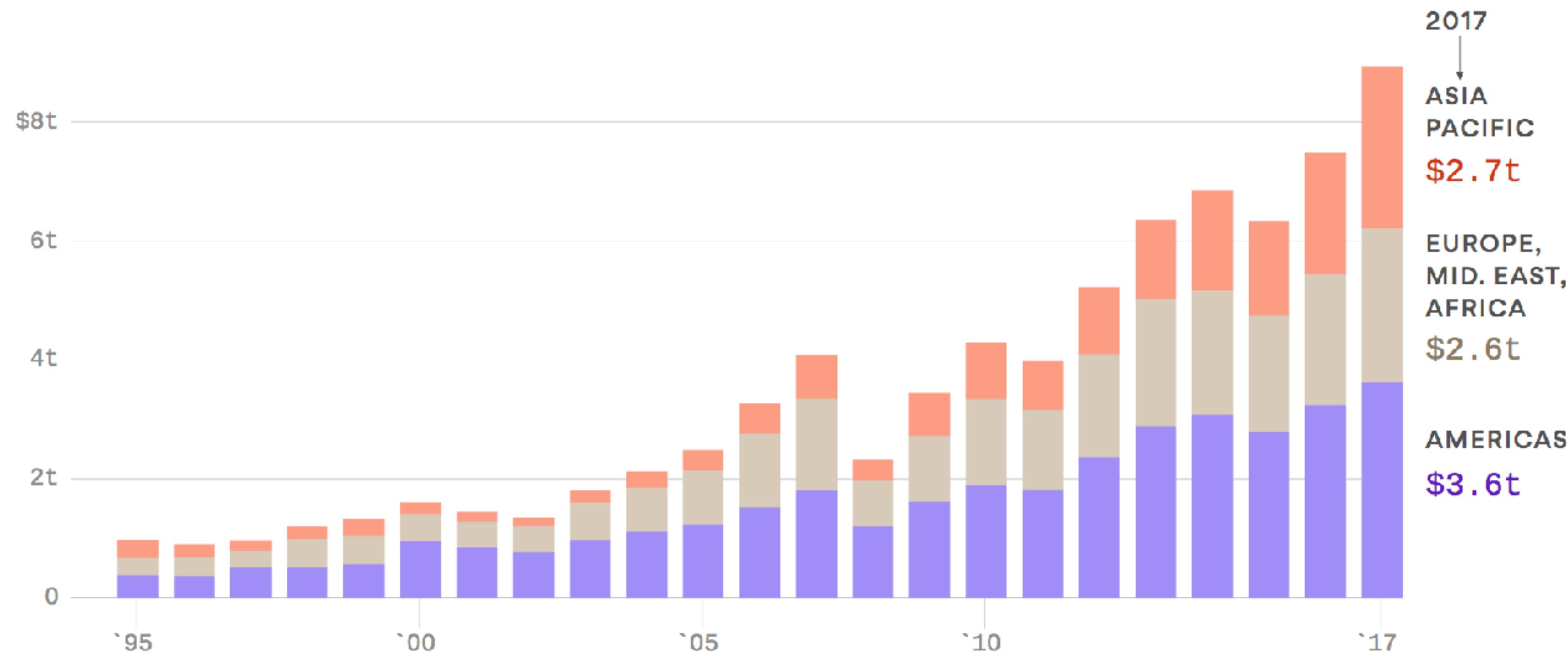
2.0  
1.5  
1.0  
0.5  
0.0



*The wealth of the world's billionaires rose by \$1.4 trillion in 2017, the largest annual increase ever.*

***The details:*** *Nearly all of that increase was driven by the Asia-Pacific region, and specifically China, where billionaire wealth rose 39%.*

### Annual combined wealth of world's billionaires



# Which age group receives the most working-age benefits?

09:46, Aug 11 2017



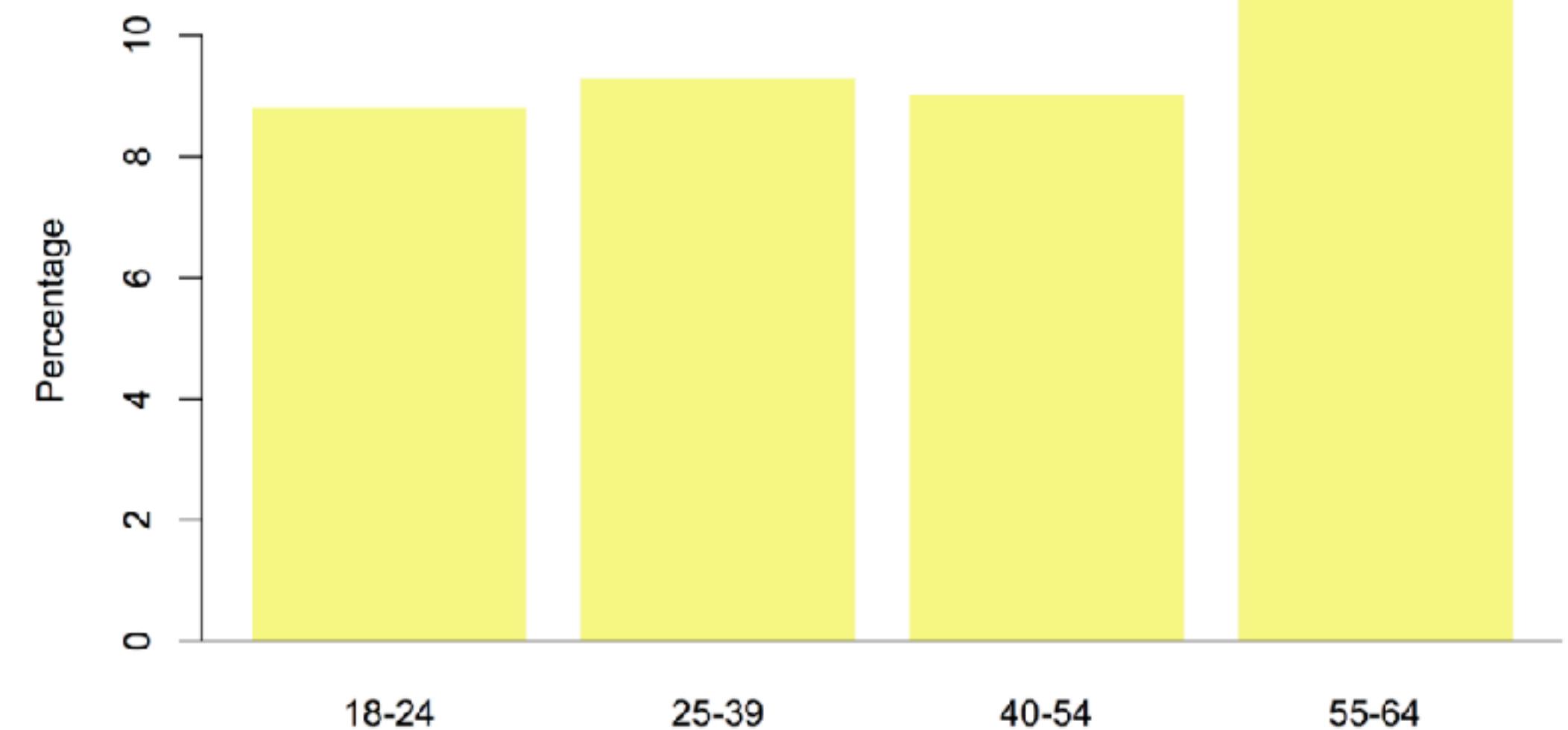
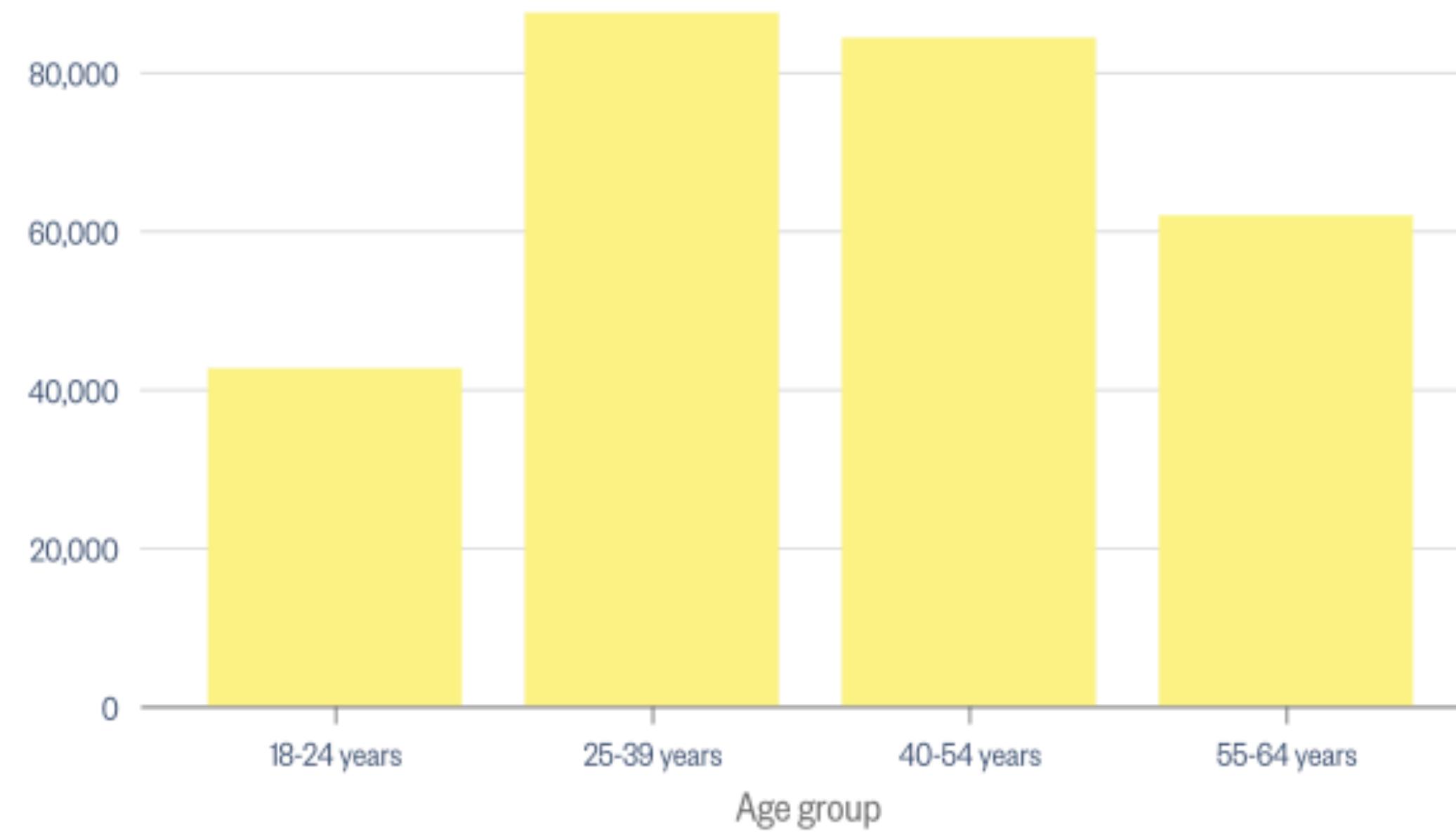
Stuff

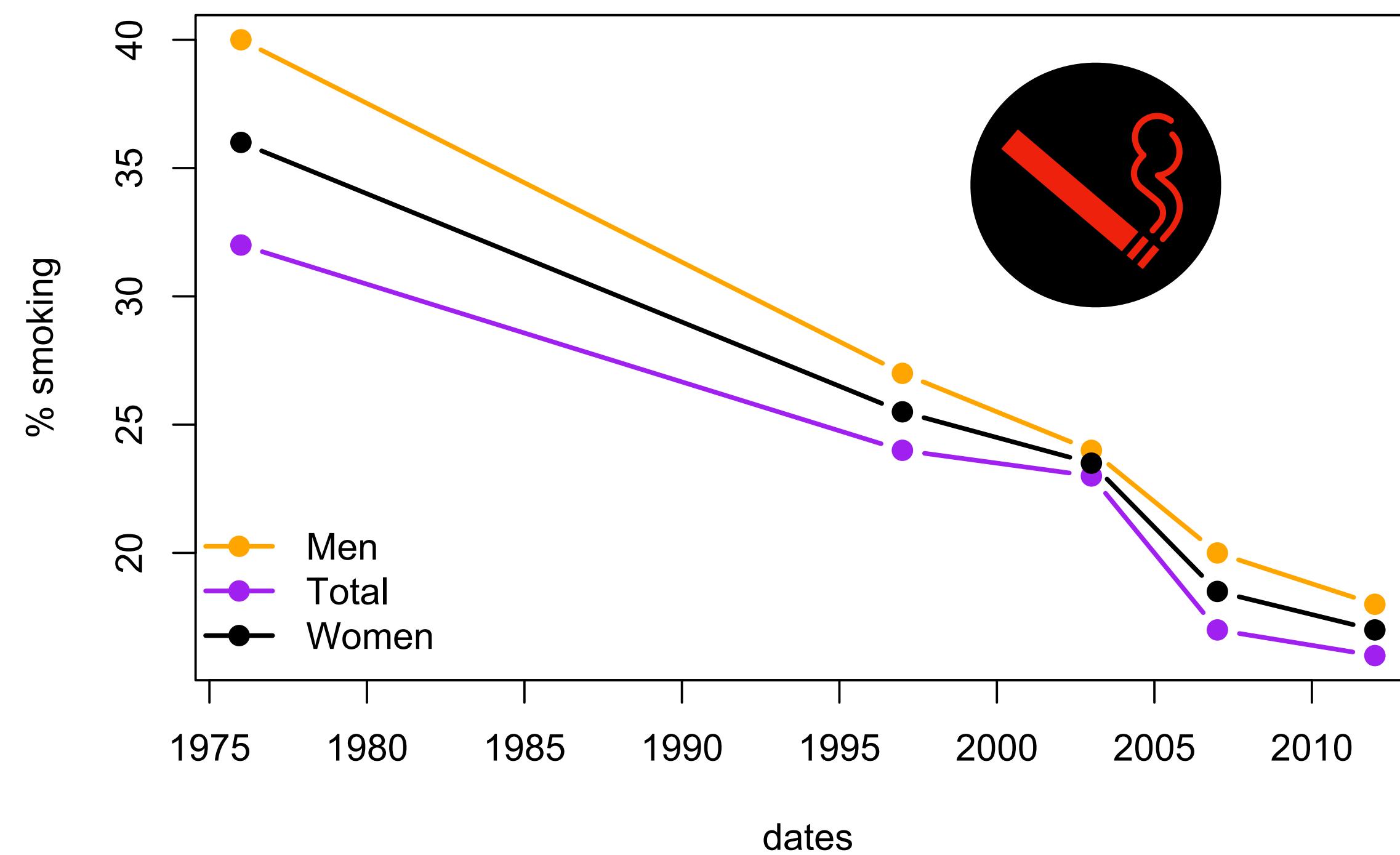
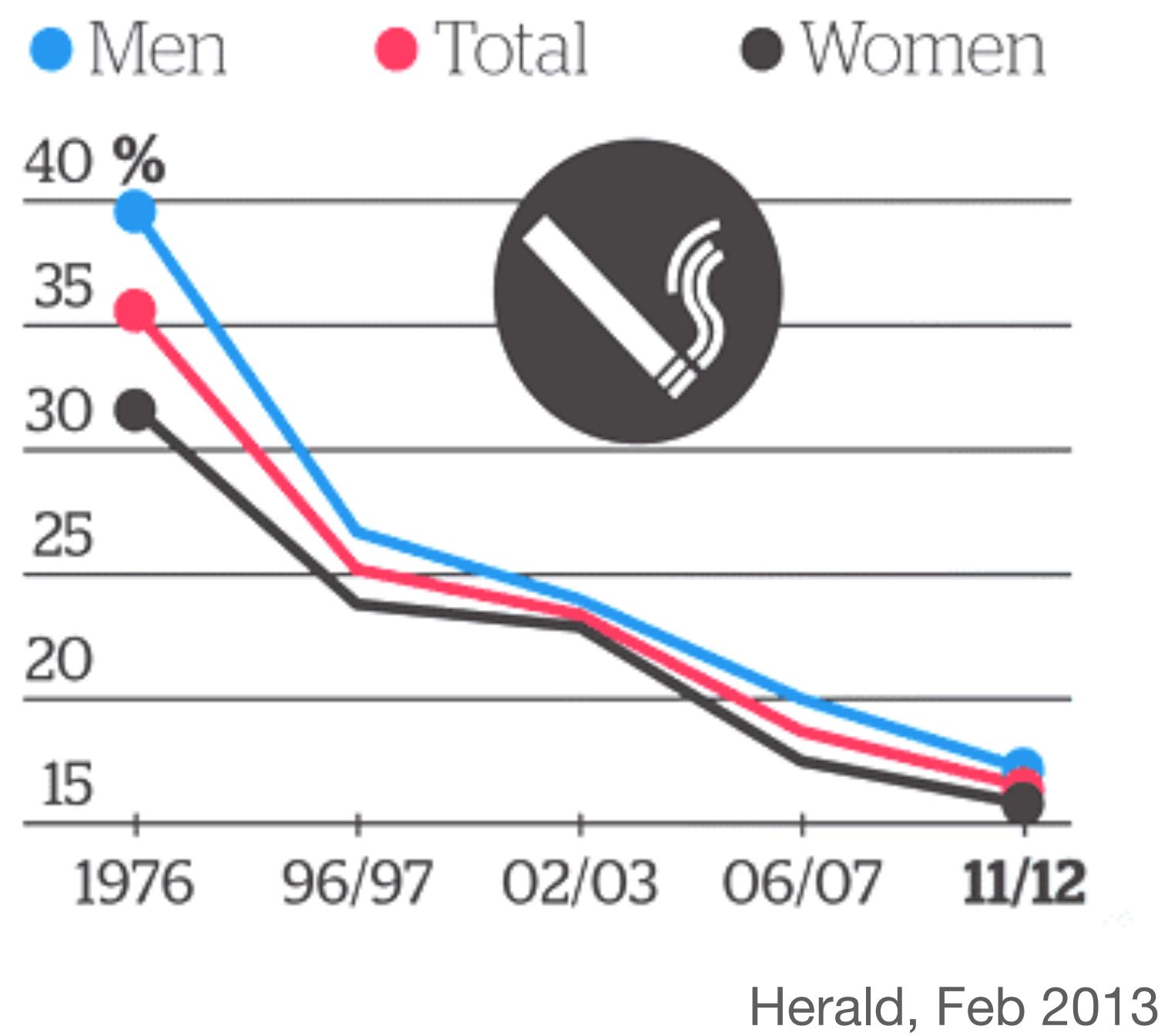
People receiving benefits by age group in New Zealand

2017 Q2, number of people

Provider: Ministry of Social Development

figure.nz





# **Doing it yourself**

# Software tools

- ❑ Prototype in Excel then draw it in eg Adobe *Illustrator*
- DataWrapper*
- Flourish*
- R+ggplot2* (+ touch up in *Illustrator*)

# Flourish

- Upload spreadsheet data (no useful data editing in app)
- Pick graph type, customise it (mostly good defaults and pretty)
- Supports individual graphics, stories
- *The Spinoff* uses it. Alberto Cairo likes it. (I hate it a lot less than most non-code systems)
- Private or public publishing: [example](#)

# Examples

Interruptions of US Supreme Court judges by other judges during court presentations. Source: [Empirical SCOTUS, 2016](#)

NCEA results by school in NZ. Released by Ministry of Education, processed a bit by [Luis Apiolaza](#)

# Your turn

selfharm.csv has monthly hospital admissions for self-harm over time.

Draw a better graph than the Herald drew for Dr Reti.

**Questions.**  
You ask.



**Bye**  
See you around

