Sampling

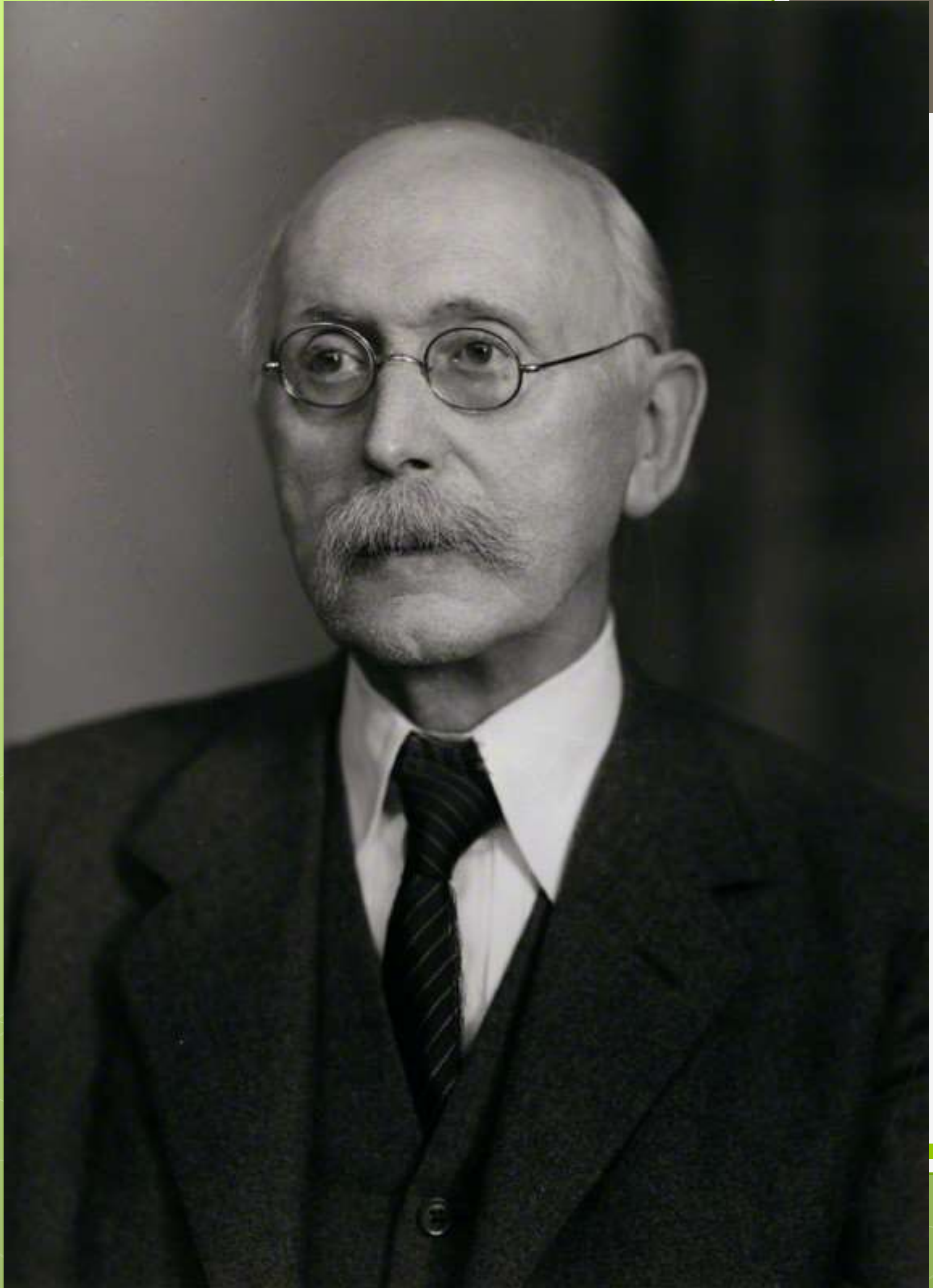
People have used random sampling for a long time

- Sampling by lots is mentioned in the Bible.
- People recognised that it is a way to select fairly if every individual has an equal chance of being chosen.

Statisticians took longer to recognise that random sampling is likely to get a sample that represents the population.

Adolphe Quetelet introduced the idea of "the average man".

Statisticians in the 1800s tried to find the individuals who were most average to use as a sample.



Arthur Lyon Bowley

Pioneered the use of random sampling in statistics from 1906.

It took a while for people to accept that a random sample could be representative of the population.

A 1936 turning point: In the USA, The Literary Digest asked 10 million people who they would vote for president. Based on 2 million responses they predicted Landon would win 57% of the votes. Roosevelt actually won 61% of the votes.

People realised that they needed to change the way sampling was done.

Teaching sampling has changed a lot too.

- Ten years ago we taught students that simple random sampling was difficult unless you had a small population size.
- Now we use it in the classroom to sample from populations of 30000 or larger.

AS91264 Use statistical methods to make an inference

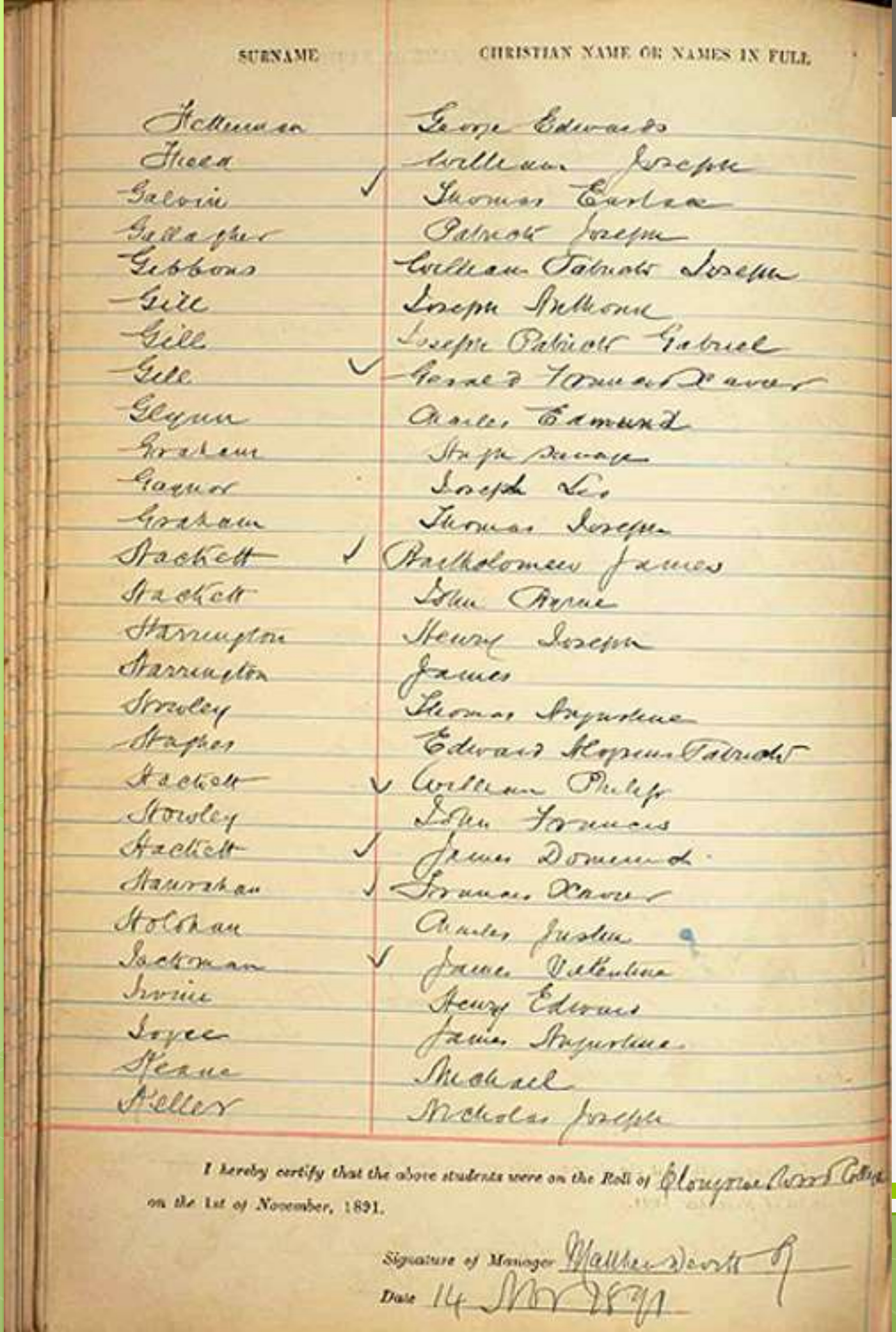
- This standard was obsolete before it was implemented.
- Statisticians would never take a sample from a population already recorded as a dataset.
- We would do the analysis on the whole dataset.

Sampling in "the classroom world"

- Sometimes in the classroom we do things that make no sense, just for practice.
- We practice sampling techniques on readily available populations.
- We still want the students to learn how to sample and to understand the use of sampling techniques.

Starting with sampling by hand

All the Lake Taupo Trout lesson plans and full data sets are available at http://schools.reap.org.nz/advisor/



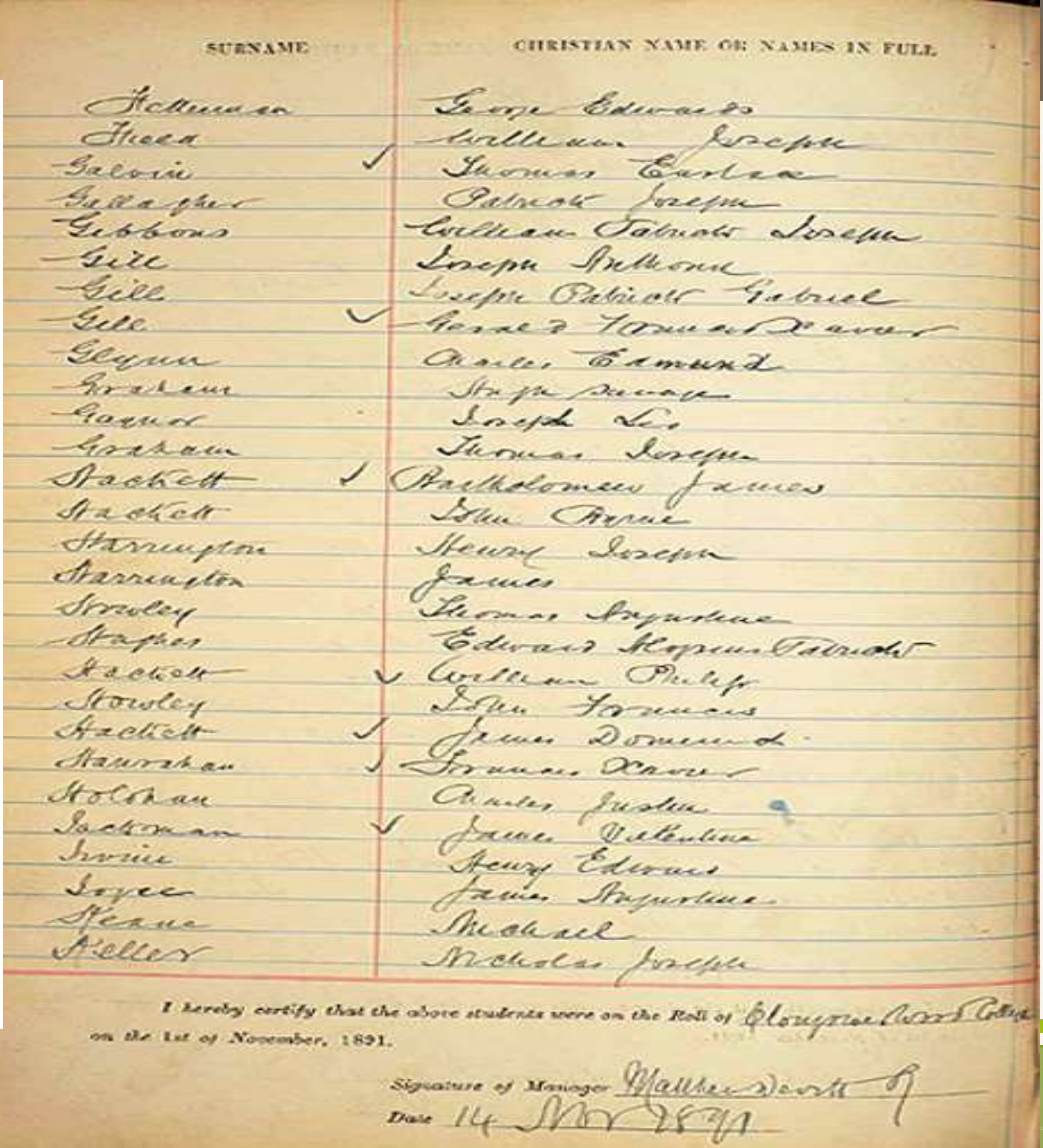
SURNAME		CHRISTIAN NAME OR NAMES IN FULL
[illegible]		George Edwards
Flood		William Joseph
Galvin	✓	Thomas Charles
Gallagher		Patrick Joseph
Gibbons		William Patrick Joseph
Gill		Joseph Anthony
Gill		Joseph Patrick Gabriel
Gill	✓	Gerard Thomas Xavier
Glynn		Charles Edmund
Graham		Stephen [illegible]
Gaynor		Joseph Leo
Graham		Thomas Joseph
Hackett	✓	Bartholomew James
Hackett		John [illegible]
Harrington		Henry Joseph
Harrington		James
Hawley		Thomas Augustine
Hayes		Edward [illegible] Patrick
Hackett	✓	William Philip
Howley		John Francis
Hackett	✓	James Dominick
Hanrahan	✓	Francis Xavier
Holohan		Charles Justin
Jackman	✓	James Valentine
Irvine		Henry Edward
Joyce		James Augustine
Keane		Michael
Keller		Nicholas Joseph

I hereby certify that the above students were on the Roll of Clongowes Wood College on the 1st of November, 1891.

Signature of Manager Matthew Devitt

Date 14 Nov 1891

Sampling from the class roll



	SURNAME		CHRISTIAN NAME OR NAMES IN FULL
1	Fetherston		George Edward
2	Flood		William Joseph
3	Galvin	✓	Thomas Charles
4	Gallagher		Patrick Joseph
5	Gibbons		William Patrick Joseph
6	Gill		Joseph Anthony
7	Gill		Joseph Patrick Gabriel
8	Gill	✓	Gerald Thomas Xavier
9	Glynn		Charles Edmund
10	Graham		[illegible]
11	Gaynor		Joseph Leo
12	Graham		Thomas Joseph
13	Hackett	✓	Bartholomew James
14	Hackett		John [illegible]
15	Harrington		Henry Joseph
16	Harrington		James
17	Hawley		Thomas Augustine
18	Harper		Edward [illegible] Patrick
19	Hackett	✓	William Philip
20	Howley		John Francis
21	Hackett	✓	James Dominic
22	Hanrahan	✓	Francis Xavier
23	Holohan		Charles Justin
24	Jackman	✓	James Valentine
25	Irvine		Henry Edward
26	Joyce		James Augustine
27	Keane		Michael
28	Keller		Nicholas Joseph

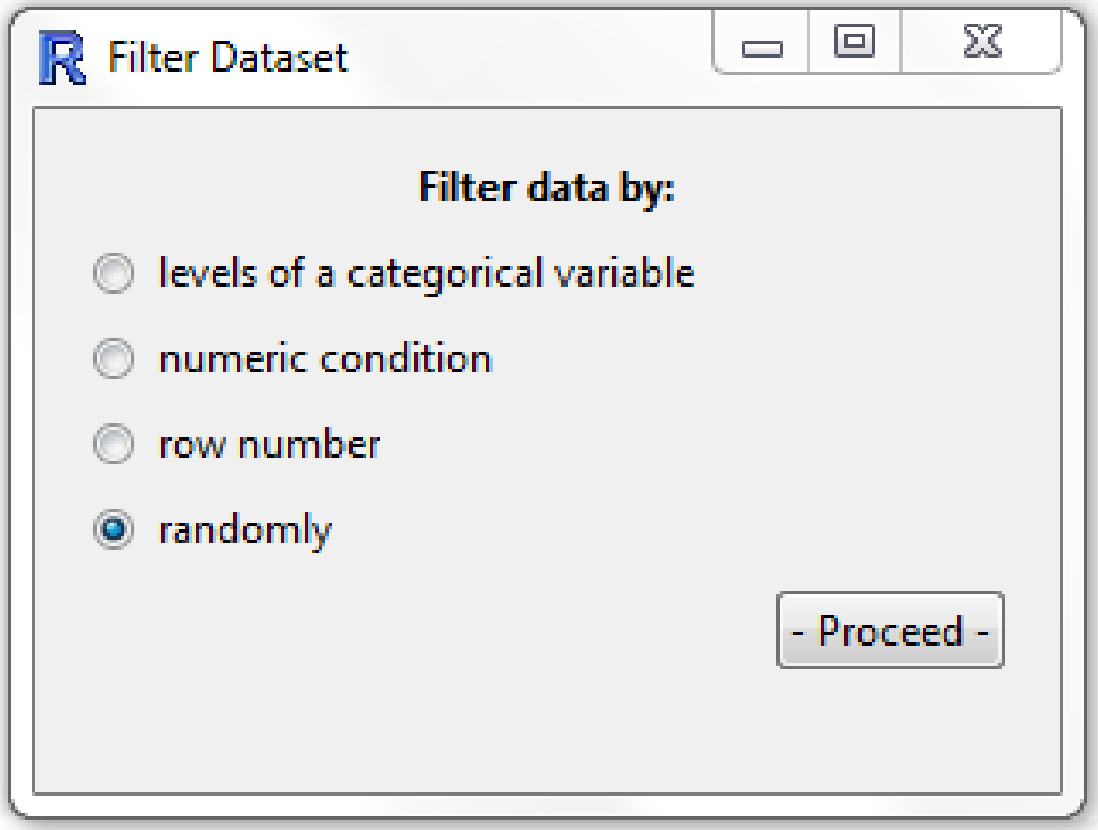
I hereby certify that the above students were on the Roll of Clongowes Wood College on the 1st of November, 1891.

Signature of Manager Matthew Devitt SJ

Date 14 Nov 1891

Sampling in excel

Sampling in Census@School and iNZight (and Fathom?)



Filter Dataset

Filter data by:

○ levels of a categorical variable

○ numeric condition

○ row number

● randomly

Is it time to bring back the big statistics project where students have to take a real sample?

Some ideas related to sampling

Lots of texts and homework books have errors at the moment, especially for things like defining sampling and non-sampling error which are defined differently in different countries.

The NZ definitions from statistics NZ:

Sampling error arises due to the variability that occurs by chance because a random sample, rather than an entire population, is surveyed.

Non-sampling error is all error that is not sampling error.

Non-sampling error is error due to:

- Choice of sampling frame
- Survey design
- Sampling method
- Sampling process
- Behaviour of population being sampled (eg response rate)

Understanding centre and spread (more difficult than you might think)

centre	spread
one best number to describe the group	how different members of the group are from each other
position	variation
signal	noise
central tendency	dispersion
prediction	precision
estimate	uncertainty

Describing centre (position)

- Shift and overlap are both measures comparing centres
- Shift: which one is bigger?
- Overlap: by how much?
- Students need to observe, then back up their observations by referring to the median in context.
- Demonstrating understanding is not about knowing the formula, it is about describing what the median tells us about the population eg “tend to”, “on average”.

Describing centre

In the graph of my sample I notice that the graphs for men's height is shifted upwards from women's heights, showing that men tend to be taller than woman in my sample. There is a lot of overlap of the middle 50% of heights. The median height for male students in my sample was 170cm, while the median height for female students was 165cm. This confirms that in my sample the men tend to be about 5cm taller than the women.

Describing spread (variation)

- We want a measure for the whole sample or population, describing how different the individuals are from each other.
- The interquartile range is a measure of spread for the whole sample or population. It should be given as the calculated IQR without reference to “middle 50%” or values of the quartiles, so it is not confused with describing position.
- Describing shape can be done with spread.

Describing spread

In the graph of my sample the men’s heights are spread out more than the women’s heights, with a lot more tall men than there are tall women. Both sexes have most heights close to the middle with fewer tall or short people. In my sample the IQR for male height is 12cm, while the IQR for female height is 8cm (4 cm narrower), confirming my observation that in my sample men’s height is more variable than women’s height.

One final thought for year 13

- The mean is a more efficient measure than the median. A sample mean is a better estimator of the population mean than the sample median is of the population median.
- But…mean and standard deviation are affected by skew and outliers.
- Choose your variable with understanding.

I'M FILLING OUT A READER SURVEY FOR *CHEWING* MAGAZINE.
SEE, THEY ASKED HOW MUCH MONEY I SPEND ON GUM EACH WEEK, SO I WROTE, "$500." FOR MY AGE, I PUT "43," AND WHEN THEY ASKED WHAT MY FAVORITE FLAVOR IS, I WROTE "GARLIC / CURRY."
CHEWING
THIS MAGAZINE SHOULD HAVE SOME AMUSING ADS SOON.
I LOVE MESSING WITH DATA.

Source: Calvin and Hobbes comic strip, by Bill Watterson, 23 August 1995