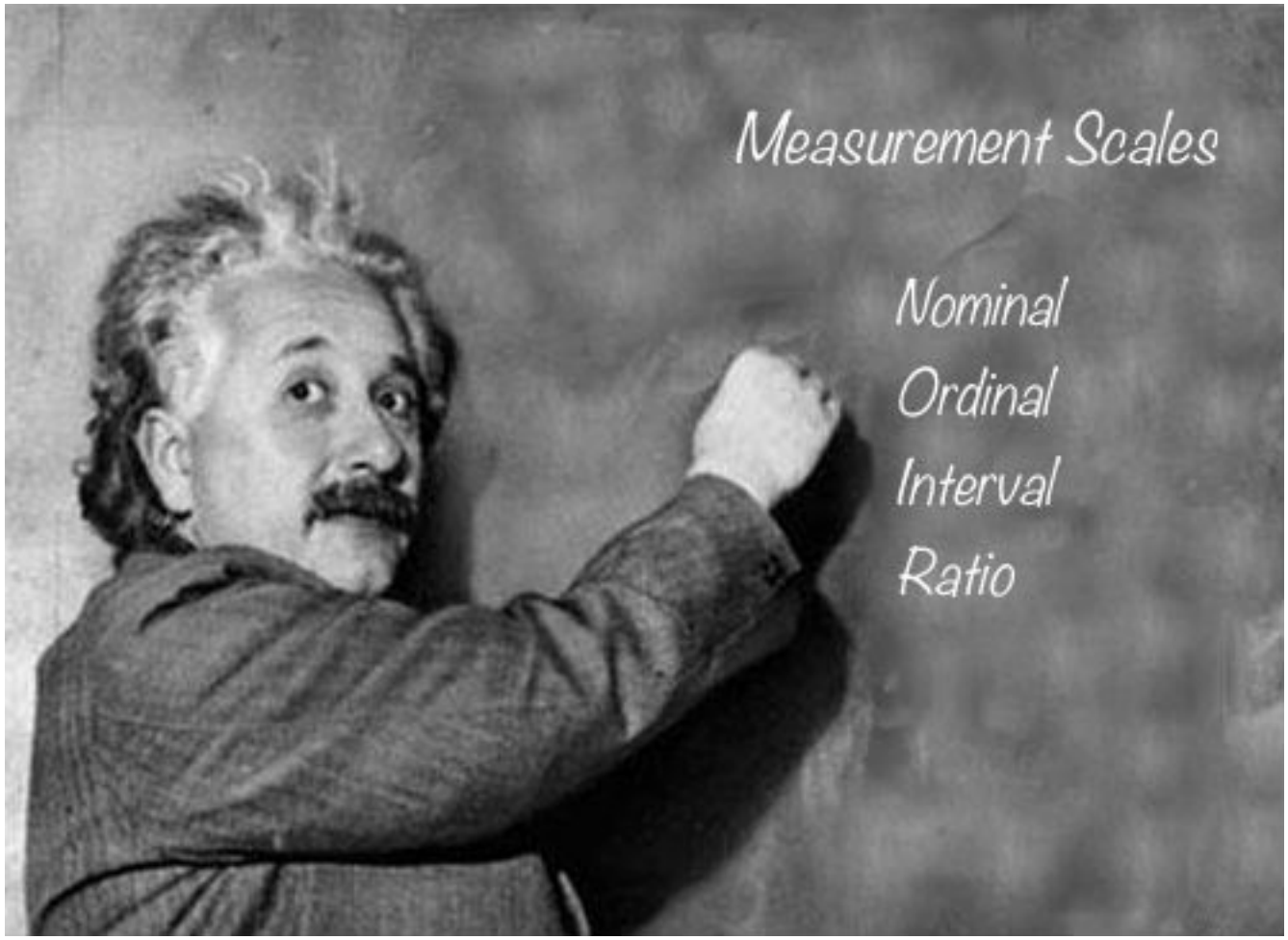
Why there's no such thing as an ordinal test

Thomas Lumley
Department of Statistics
University of Auckland
@tslumley
http://notstatchat.tumblr.com

In general, would you say your health is …

96	1. Excellent
93	2. Very good
76	3. Good
35	4. Fair
19	5. Poor
0	Dead


Measurement Scales
Nominal
Ordinal
Interval
Ratio

If you have decided at the psychometric stage that your scale is ordinal, you are likely to employ some sort of nonparametric test at the inference stage, not only because of the distribution-free nature of such tests, but because they tend to be more appropriate for hypotheses that are meaningful for ordinal variables.

Treating Ordinal Scales as Interval Scales: An Attempt To Resolve the Controversy.
Knapp, 1990

The t-test is to the mean as the Wilcoxon rank-sum test is to….

- median?
- median pairwise mean?
- something more complicated?

Non-transitive dice

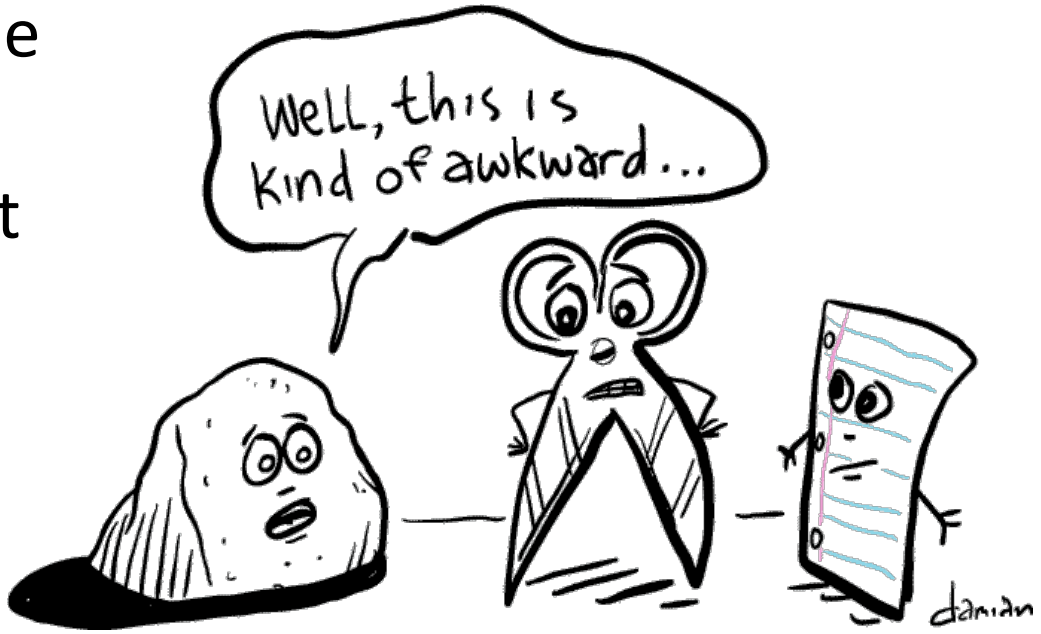
Bradley Efron (circa 1973)
A: 4, 4, 4, 4, 0, 0
B: 3, 3, 3, 3, 3, 3
C: 6, 6, 2, 2, 2, 2
D: 5, 5, 5, 1, 1, 1

http://www.grand-illusions.com/

Each die beats the next one at least 2/3
You can let the sucker choose first, and still win.

Why do we care?

- Dice generate probability distributions
- Comparing dice by pairwise chance of winning is non-transitive
- Comparing probability distributions by pairwise chance of winning is the Mann-Whitney U aka Wilcoxon rank sum test



Why do we care?

- There is no ordering on **probability distributions** that is consistent with the Wilcoxon test, even asymptotically
- No one-dimensional summary statistic agrees with the Wilcoxon test, even asymptotically

- Rank tests are like that.

How general is the problem?

Theorem: any (sane) transitive test is a test for a univariate real-valued summary statistic

Proof outline:

A transitive test defines ordered equivalence classes of distns where power=level.

The classes can be labelled with real numbers unless the order topology is 'too big'

[Debreu, 1960s, for preference relations
Lumley & Gillen (submitted), for tests]

In general, would you say your health is …

1. Excellent
2. Very good
3. Good
4. Fair
5. Poor

You have data for a set of treatments on a large sample of people from a population

The data is purely ordinal: within-person rankings of treatments, with no numerical values.

You need to choose which single treatment is best for new people from the population

Sanity conditions

- You have to make a choice
- It can't just depend on one person's data
- For each treatment there is some set of data that would led to it being chosen
- Making the result for a non-chosen treatment *worse* will not lead to it being chosen
- Adding a new treatment options will not make a different *existing* treatment get chosen

Ordinal Testing

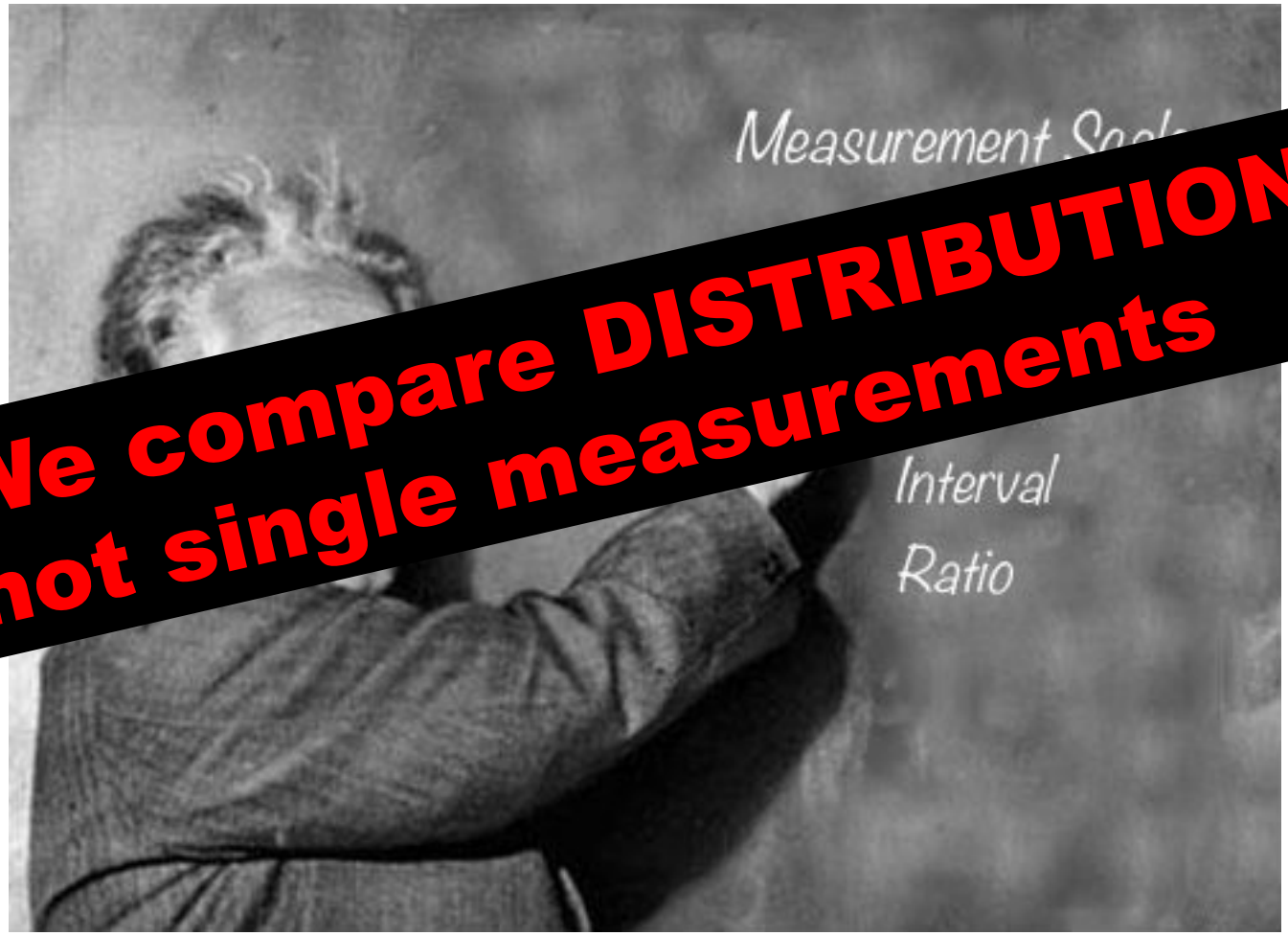
A Difficulty in the Concept of ~~Social Welfare~~

Kenneth J. Arrow

The Journal of Political Economy, Vol. 58, No. 4. (Aug., 1950), pp. 328-346.

Stable URL:
http://links.jstor.org/sici?sici=0022-3808%28195008%2958%3A4%3C328%3AADITCO%3E2.0.CO%3B2-R

The Journal of Political Economy is currently published by The University of Chicago Press.



We compare DISTRIBUTIONS, not single measurements

Potentially, each treatment is better for some people and worse for others.

You can't possibly evaluate the tradeoffs without knowing **how much** better or worse

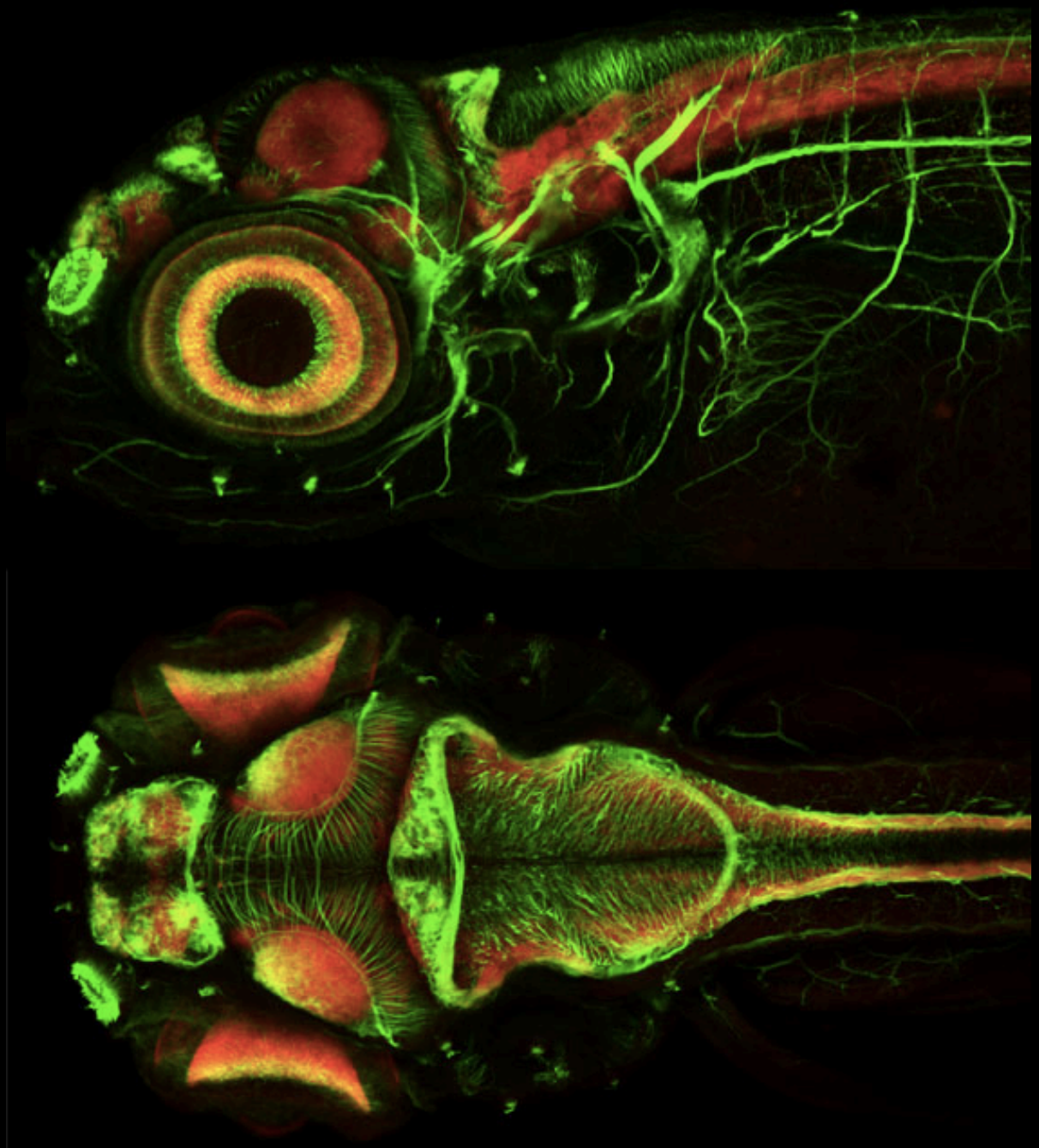
Any method that purports to, **must** be wrong.

When does it work?

Definition: $F(t) \leq G(t)$ for all t

- Under stochastic ordering all location tests will agree on the direction of a difference.
- Wilcoxon test is transitive on stochastically ordered sets of distributions

Basically only one-dimensional families are stochastically ordered



http://www.isciencemag.co.uk/blog/the-secret-life-of-zebrafish/

Beyond transitivity

- Non-transitivity just the extreme case
- Easy for **different** statistics to order distributions differently
- Disease prevention (eg inhaled steroids/asthma)
  - increases **median** medical cost
  - decreases **mean** medical cost
- Not just an efficiency issue: different hypotheses

The poor performance of the t-test, particularly for distributions with heavy tai
can be seen in comparison with n ic
tests, such as the Wilc
tests.

ASSUMING A LOCATION SHIFT ALTERNATIVE

f finite variance, the
ive efficiency relative to *t* is
≥ for Wilcoxon and ≥ 1 for normal scores.

Diaconis & Lehmann, JASA, 2008

Teaching

We have a bad habit of silently assuming location shift alternatives

“If you don’t even know whether an intervention makes X go up or down, how can you know it has the same effect on every individual?”

-Scott Emerson

Introductory teaching

- Descriptive summaries lead to confidence intervals, which lead to tests for those same summaries
  - no Wilcoxon test, but test for median is ok
- Present t-test initially as test for mean, not test for Normal
  - mention good small-sample performance on Normal data later, if you like

Introductory teaching

- Show students that different statistics order groups differently
  - median income, mean income, % in poverty
  - mean tweets/friends/followers vs median
- Choice of summary is not value-free, and is not determined by the data
  - what do you care about?
  - what is likely to be affected?
  - fallback: what is easy to estimate precisely?

Math teaching

- Non-transitive dice and voting paradoxes are fun and easy at high school level

- Non-transitive tests useful in math stat to clarify limits of efficiency results
  - cf Hodges superefficient estimator

Questions?