Day 05

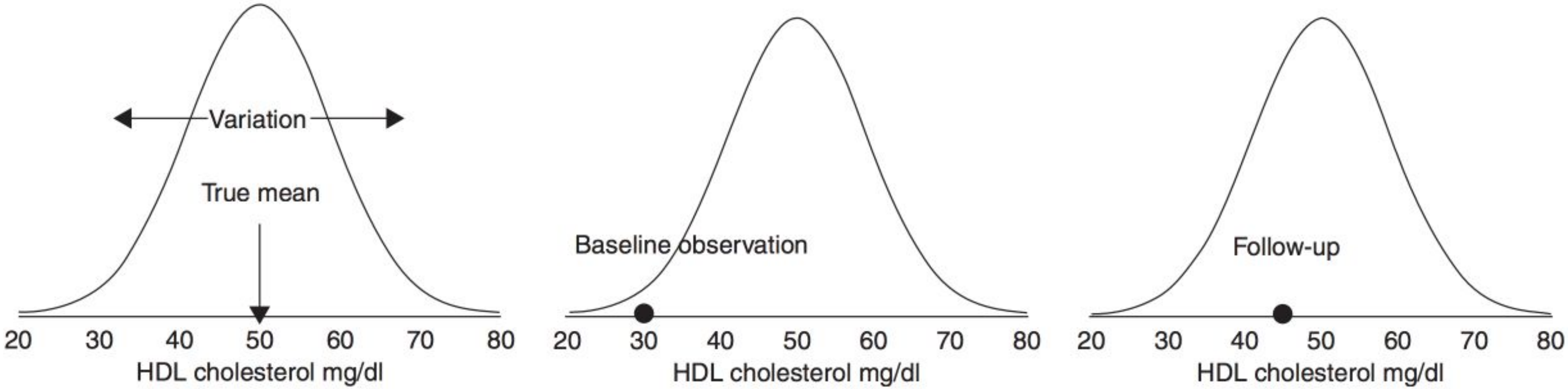
RTM, Circularity, Sampling biases

- Regression to the mean
- Circular analysis / Double dipping
- Sampling biases

Regression to the mean

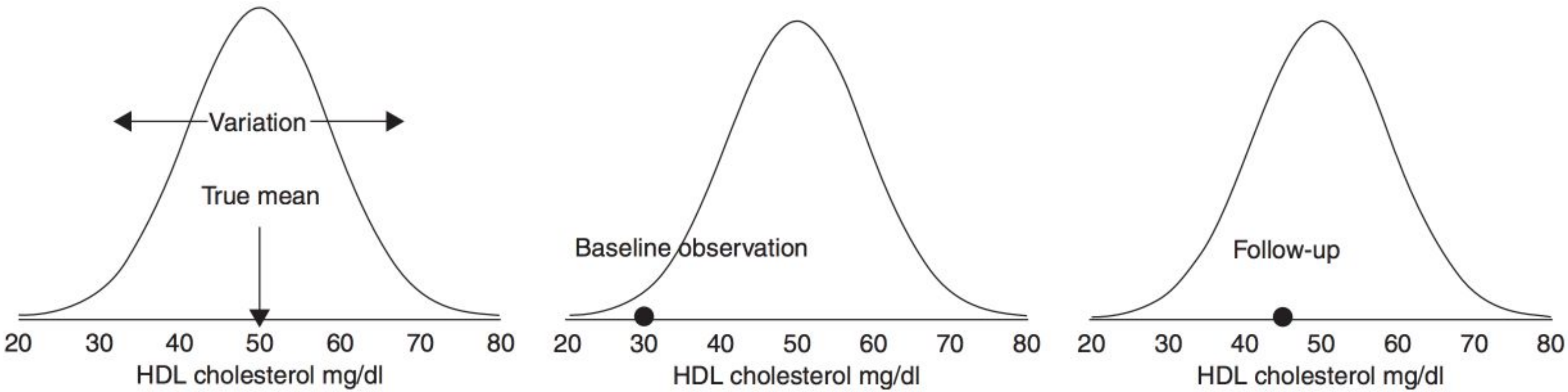
Regression to the mean (RTM) is a statistical phenomenon that is characterized by the fact that unusually large or small measurements tend to be followed by measurements that are closer to the mean.

- Occurs when repeated measurements are made on the same subject or unit of observation.



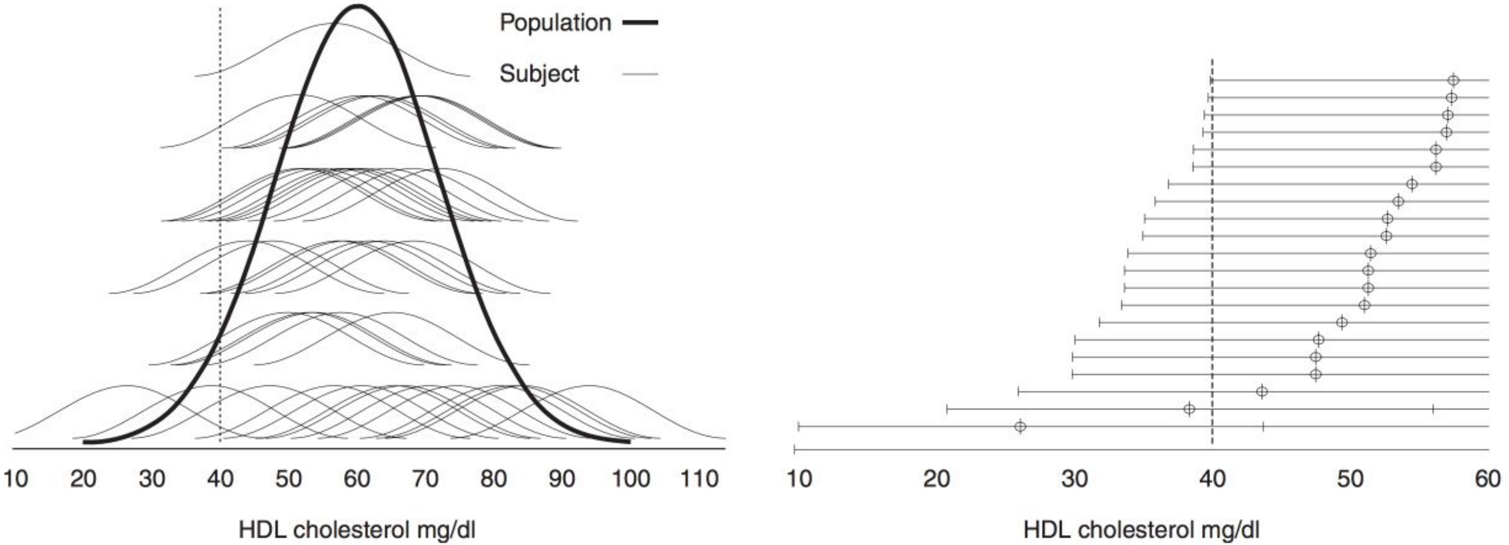
Regression to the mean

- RTM can make natural variation in repeated data look like real change.
  - Values are observed with random error (non-systematic variation like random measurement error or random fluctuations in a subject).
  - Since there is almost no data without random error, RTM is common!



Regression to the mean

Effect of RTM is compounded by categorizing subjects into groups based on baseline measurements.

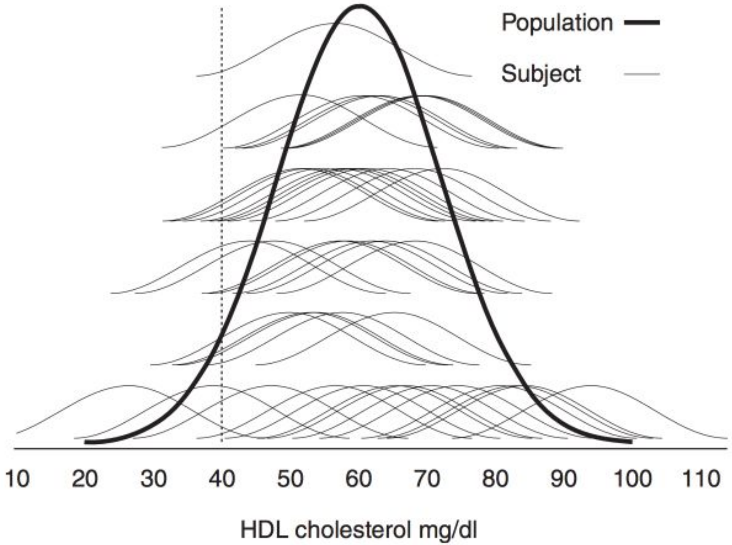


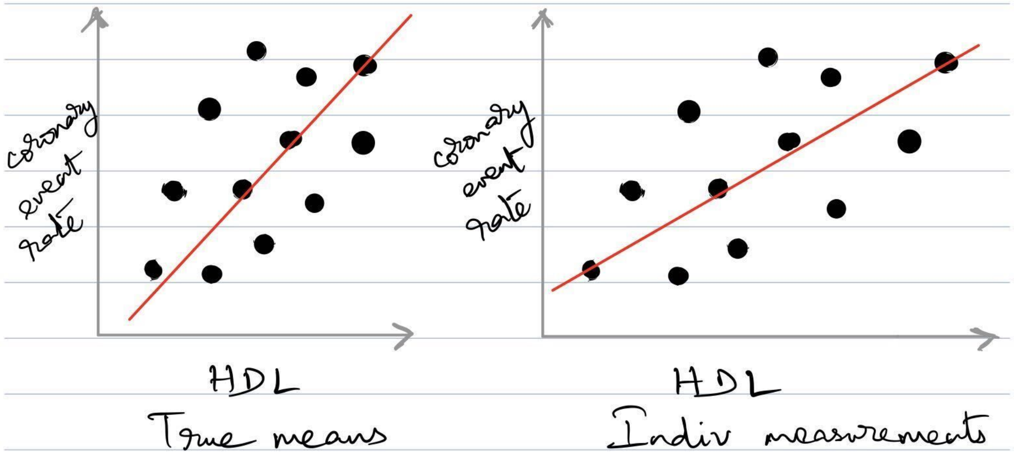
https://academic.oup.com/ije/article/34/1/215/638499

Regression to the mean

Variability in individual measurements > variability in the true means

→ Attenuation of association (regression dilution bias).



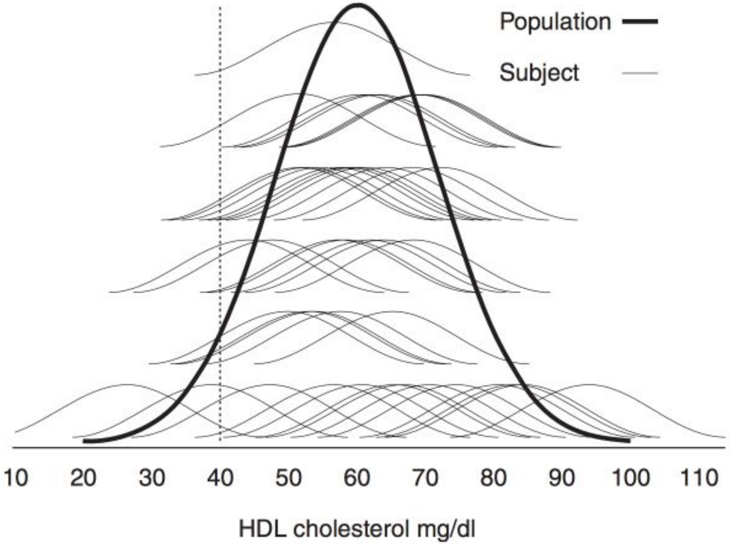


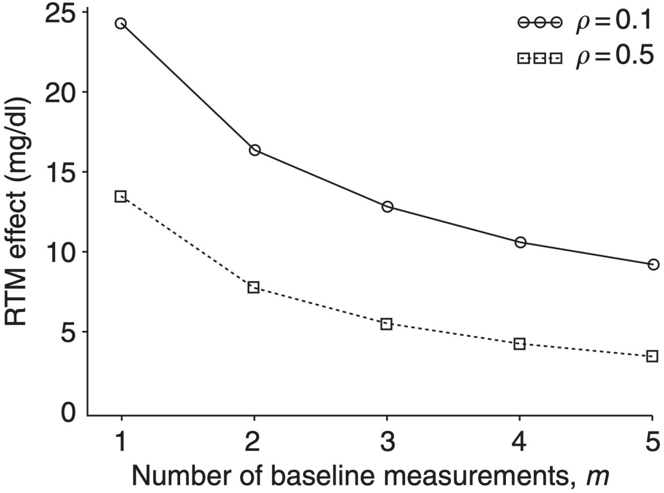
https://academic.oup.com/ije/article/34/1/215/638499

Regression to the mean

Categorizing subjects into groups based on baseline measurements

Variability in individual measurements > variability in the true means



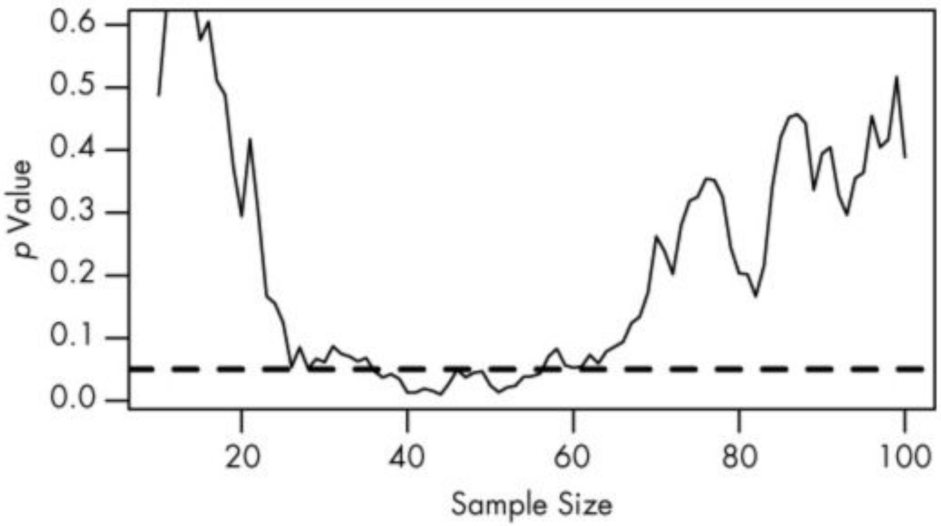


https://academic.oup.com/ije/article/34/1/215/638499

Regression to the mean

Longitudinally tracking the effect of drug.

- Terminate the study early if there if it is clear that the drug has an effect.
- In fact, it is *unethical* to withhold the drug from the control group.



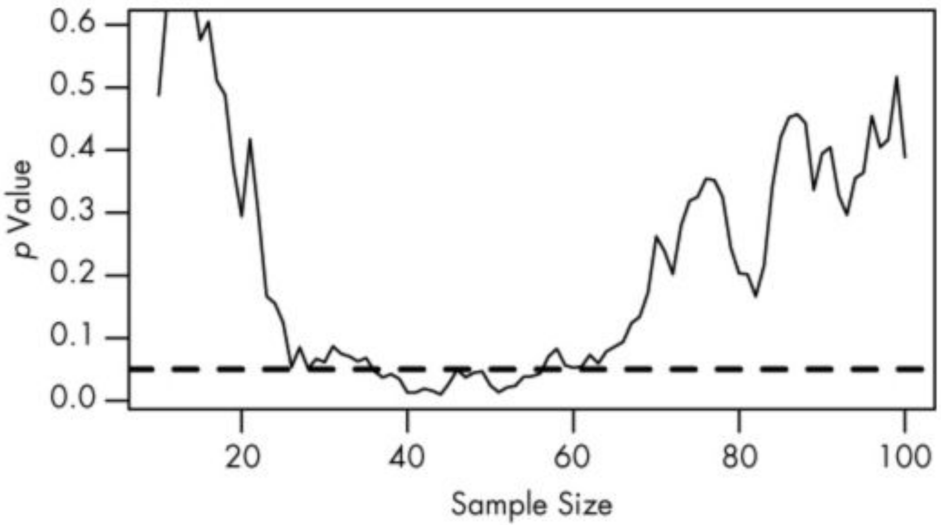
Issues:

- Null hypothesis should take varying group size into account.
- Truth inflation (lucky patients, not brilliant drugs): stopped trials exaggerate their effect by 29% more than trials that run their full course.

Regression to the mean

Longitudinally tracking the effect of drug.

- Many published studies do not publish their:
  - Original intended sample size or
  - The stopping rule used to justify terminating the study
- Preregistration!
  - Statistical protocols
  - Few pre-selected evaluation points

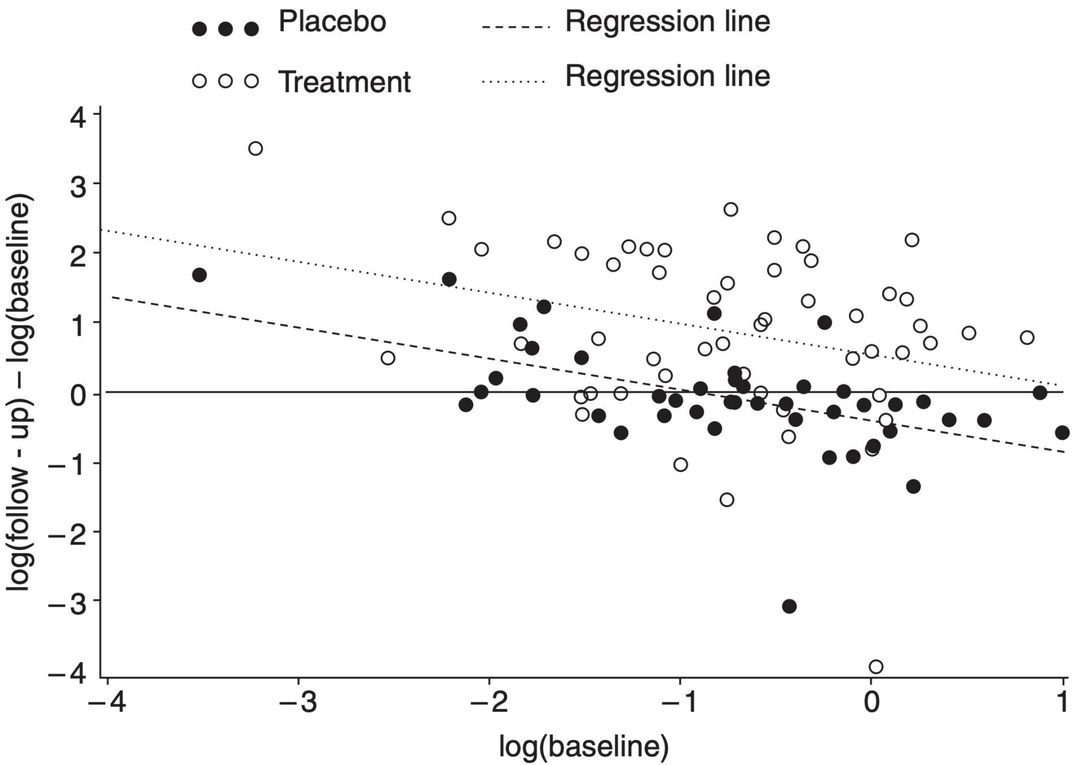


Regression to the mean

Assigning a narrative or causal reasoning from observed data is often very hard:

- Tall parents had (on average) children who were shorter than them, and short parents had (on average) children who were taller than them.
- “Norway had a great first jump; she will be tense, hoping to protect her lead and will probably do worse”; “Sweden had a bad first jump and now he knows he has nothing to lose and will be relaxed, which should help him do better.”
- Depressed children treated with an energy drink improve significantly over a three-month period.

Regression to the mean



https://academic.oup.com/ije/article/34/1/215/638499

Circular analysis / Double-dipping

- Statistical analysis is often exploratory: no hypothesis is advance.
    - Collect data → Poke around to see if there's something interesting → New hypotheses → Perform new experiments / Collect new data → Test the hypotheses.
- But, scientists take shortcuts.
    - E.g. Collect data → Poke around to see if there's something interesting → New hypotheses → Take the subset of the original data that appears to show signal →Test the hypotheses.
- Happens all the time in neuroimaging (apparently 40% of the literature), genetics, epidemiology.

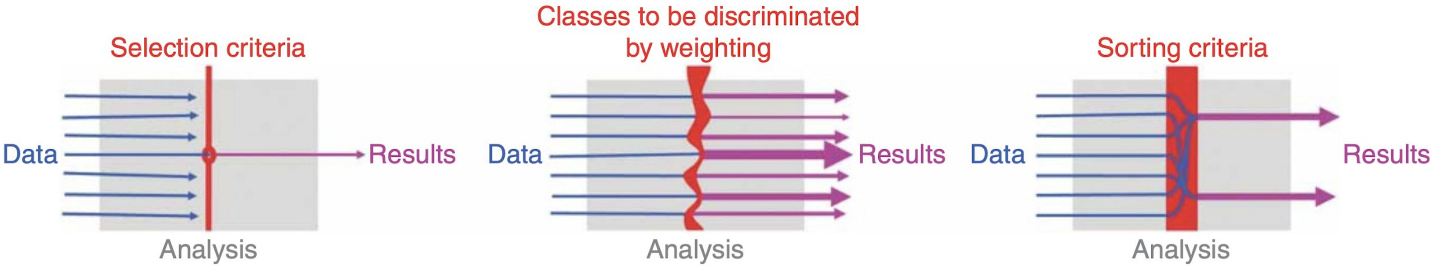
https://elifesciences.org/articles/48175

Circular analysis / Double-dipping

Typical mistake:

- **Dividing** (e.g. sub-grouping, binning)
- **Reducing** (e.g. defining a region of interest, removing ‘outliers’)
- Variants: **Weighting** or **Sorting**

... the complete dataset using a selection criterion that is retrospective and inherently relevant to the statistical outcome.



https://elifesciences.org/articles/48175; https://www.nature.com/articles/nn.2303

Circular analysis / Double-dipping – Example

A study of a neuronal population firing rate in response to a given stimulus.

- A researcher compares the population as a whole & finds no significant differences are found between pre and post stimulation groups.
- However, they observe that some of the neurons respond to the stimulation by *increasing* their firing rate, whereas others *decrease* in response to the stimulation.
- So, they split the population to sub-groups, by binning the data based on baseline activity levels, and compare pre & post stimulus with each sub-group.

Substantial interaction effect:

- Neurons that initially produced low responses show response increases.
- Neurons that initially showed relatively increased activity exhibit reduced activity following the stimulation.

Plus, statistical artefacts!

https://elifesciences.org/articles/48175

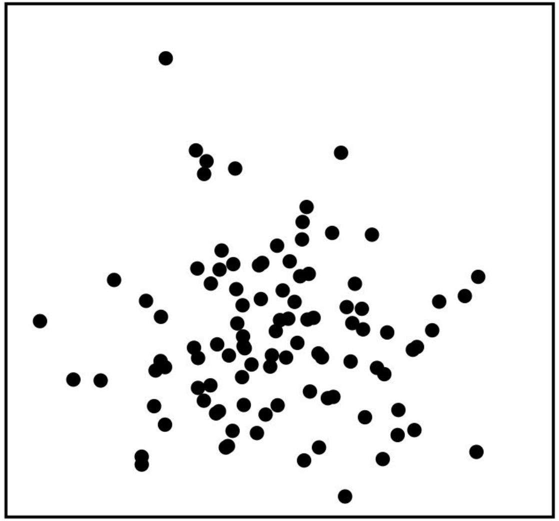
Circular analysis / Double-dipping – Example

Another common form: dependencies are created between the dependent and independent variables.

- A researcher might report a correlation between the cell response post-stimulation and the difference in cell response across the pre- and post-stimulation.
- But both variables are highly dependent on the post-stimulation measure.
- Therefore, neurons that by chance fire more strongly in the post-stimulation measure are likely to show greater changes relative to the independent pre-stimulation measure, thus inflating the correlation.

https://elifesciences.org/articles/48175

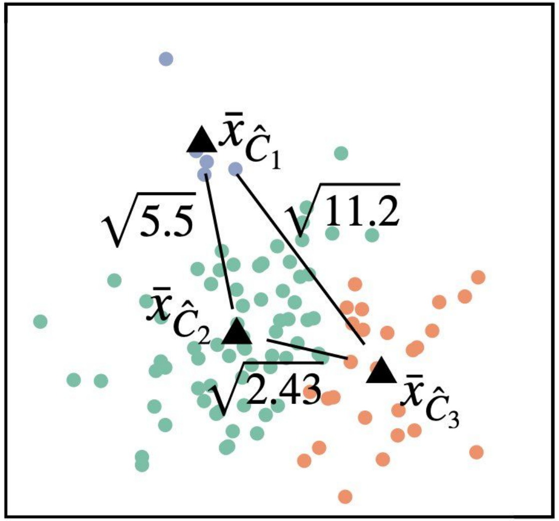
Circular analysis / Double-dipping – Example



Step 1: Sample 100 observations

https://twitter.com/lucylgao/status/1299763664644313089

Circular analysis / Double-dipping – Example



Step 1: Sample 100 observations

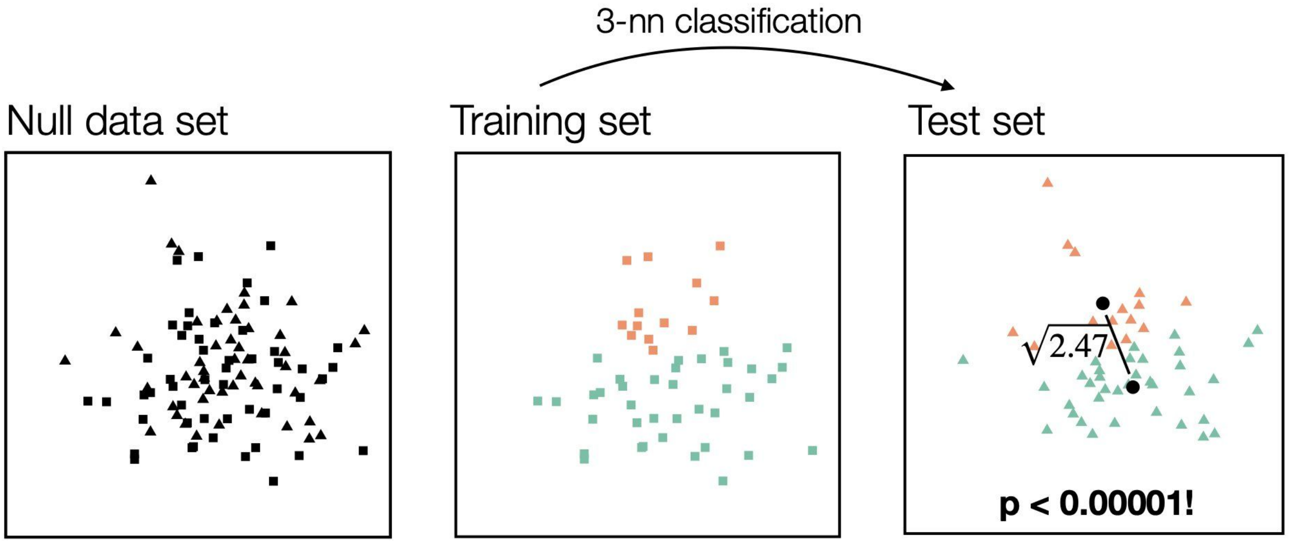
Step 2: Cluster the observations

Step 3: Compute p-values for a difference in means

All three p-values < 0.000001!! 😱

https://twitter.com/lucylgao/status/1299763664644313089

Circular analysis / Double-dipping – Example



Step 1: Divide the observations in half.

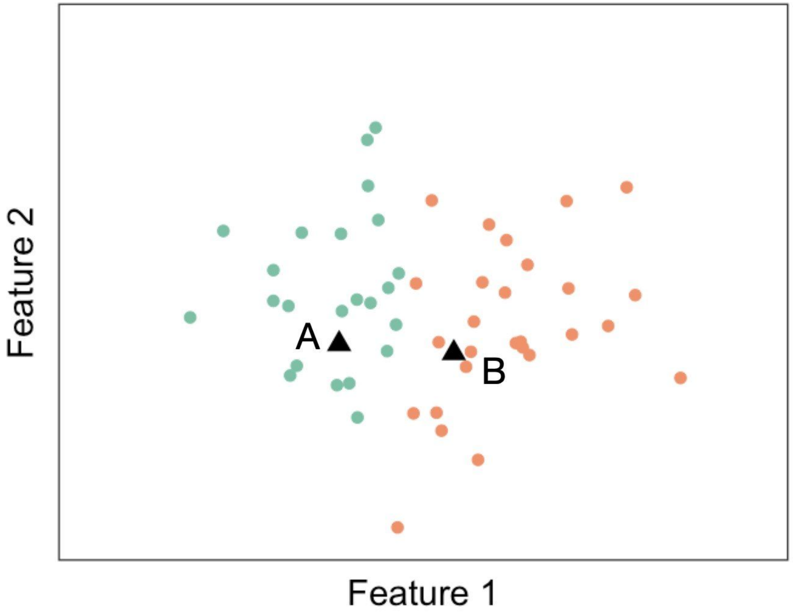
Step 2: Cluster the training set.

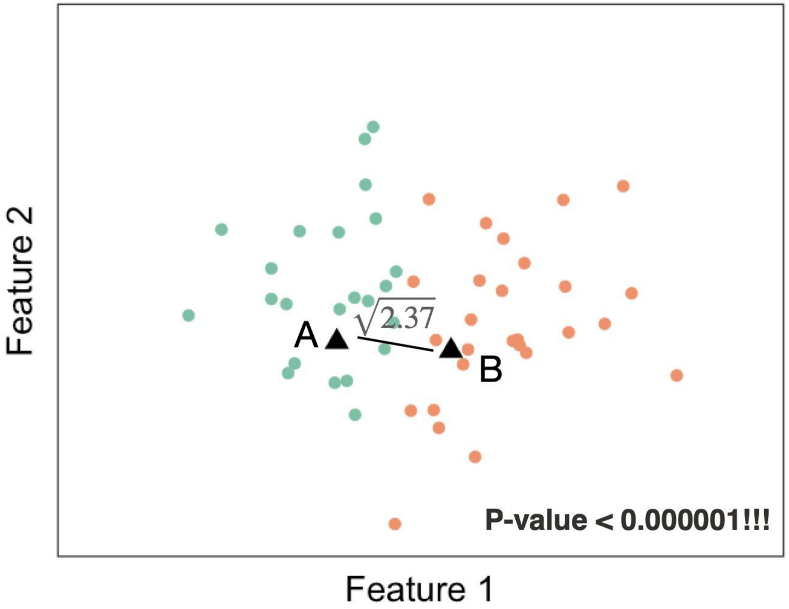
Step 3: Label the test set using a classifier trained on the training set.

Step 4: Test for a difference in means on the test set.

https://twitter.com/lucylgao/status/1299763664644313089

Circular analysis / Double-dipping – Example





https://twitter.com/lucylgao/status/1299763664644313089

Circular analysis / Double-dipping

Selective analysis is perfectly justifiable when the results are statistically independent of the selection criterion under the null hypothesis.

However, circular analysis **recruits the noise** (inherent to any empirical data) **to inflate the statistical outcome**, resulting in distorted and hence invalid statistical inference.

- Null hypothesis based on random chance is wrong at the final stage.
- Only signals with the strongest random noise make it into further analysis.

https://elifesciences.org/articles/48175

Detecting circular analysis / double-dipping

Whenever the statistical test measures are biased by the selection criteria in favour of the hypothesis being tested.

- (Obvious) If the analysis is based on data that were selected for showing the effect of interest, or an inherently related effect.
- Many times require more nuanced understanding of co-dependencies across selection and analysis steps.
  - Impossibly high effect sizes because they are theoretically impossible and/or based on relatively unreliable measures (e.g. if two measures have poor internal consistency, the potential to identify a meaningful correlation is limited).

If there is any selection, look for a justification for the independence between the selection criteria and the effect of interest.

https://elifesciences.org/articles/48175

Mitigating circular analysis / double-dipping

- Defining the analysis criteria in advance and independently of the data will protect researchers from circular analysis: **Pre-registration**!

- Use a different dataset (or different part of your dataset) for specifying the parameters for the analysis (e.g. selecting your sub-groups) and for testing your predictions (e.g. examining differences across the sub-groups).
  - Independent split-data analysis (using a different group to identify the criteria for reducing the data) or independent analysis using all data (using different trials but from all participants). Can be achieved w/o losing statistical power.

- If suitable, run a simulation to demonstrate that the result of interest is not tied to the noise distribution and the selection criteria.

- Acknowledge circular results!

Sampling biases: Will Rogers phenomenon

Moving elements from one set to another set raises the average values of both sets!

This happens when the elements being moved is:

- below average for its current set.
- above the current average of the set it is entering.

Example: Cohort of patients with lung cancer first treated in 1977 had higher six-month survival rates than a cohort treated between 1953 and 1964 at the same institutions.

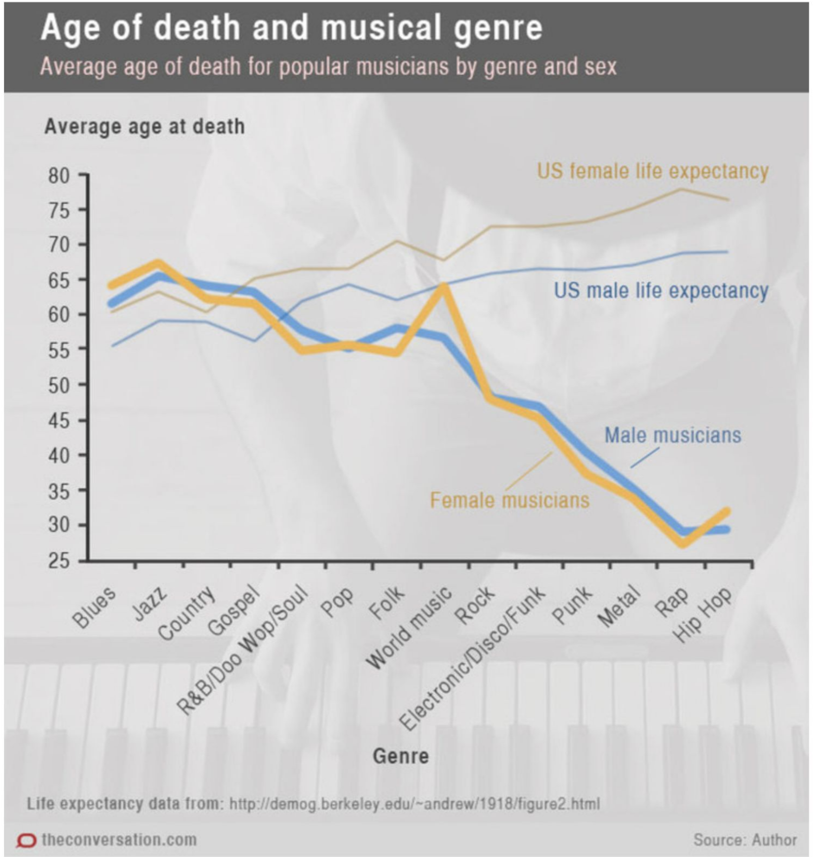
https://www.nejm.org/doi/full/10.1056/NEJM198506203122504

Sampling bias: Ascertainment bias

Data for a study or analysis is collected (or surveyed, screened, or recorded) such that some members of the intended population are less likely to be included than others.

- More intense surveillance or screening for the outcome among exposed individuals than among unexposed individuals, or differential recording of the outcome.
- In screening, where take-up can be influenced by factors such as cultural differences.
- In case-control studies, in the initial identification of cases and controls, which can be skewed by relevant exposures, leading to biased relationships.
- In a clinical trial, if allocation concealment and blinding are lacking, then outcome ascertainment can be overtly influenced by knowledge of the allocation.

Sampling bias: Right censoring



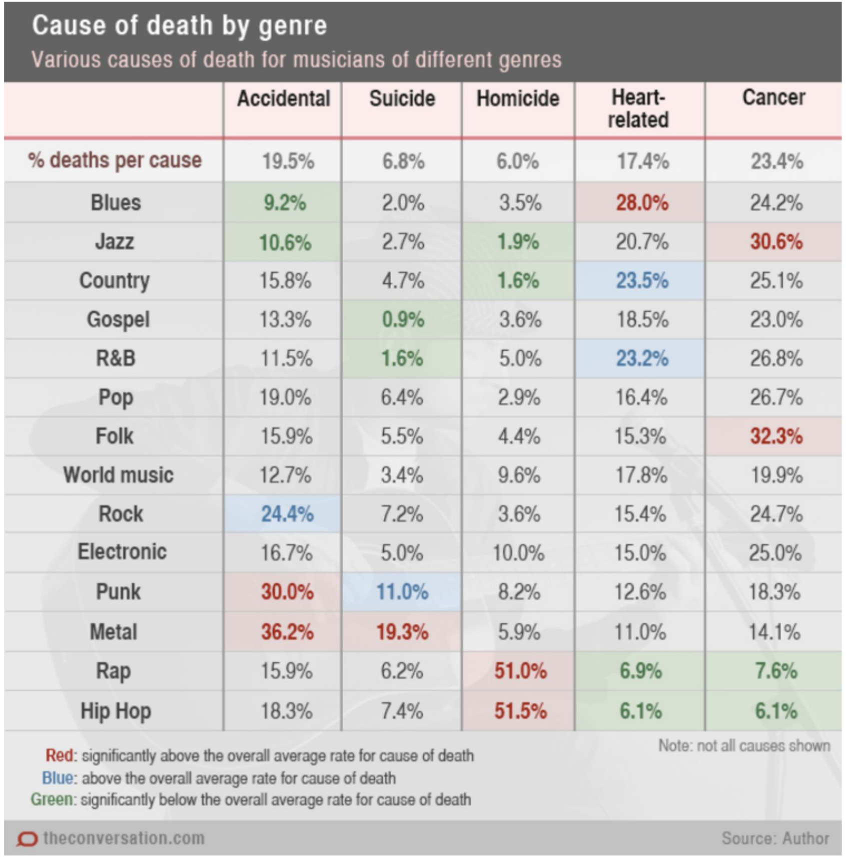
Age of death and musical genre

Average age of death for popular musicians by genre and sex

Life expectancy data from: http://demog.berkeley.edu/~andrew/1918/figure2.html

theconversation.com

Source: Author



Cause of death by genre

Various causes of death for musicians of different genres

	Accidental	Suicide	Homicide	Heart-related	Cancer
% deaths per cause	19.5%	6.8%	6.0%	17.4%	23.4%
Blues	9.2%	2.0%	3.5%	28.0%	24.2%
Jazz	10.6%	2.7%	1.9%	20.7%	30.6%
Country	15.8%	4.7%	1.6%	23.5%	25.1%
Gospel	13.3%	0.9%	3.6%	18.5%	23.0%
R&B	11.5%	1.6%	5.0%	23.2%	26.8%
Pop	19.0%	6.4%	2.9%	16.4%	26.7%
Folk	15.9%	5.5%	4.4%	15.3%	32.3%
World music	12.7%	3.4%	9.6%	17.8%	19.9%
Rock	24.4%	7.2%	3.6%	15.4%	24.7%
Electronic	16.7%	5.0%	10.0%	15.0%	25.0%
Punk	30.0%	11.0%	8.2%	12.6%	18.3%
Metal	36.2%	19.3%	5.9%	11.0%	14.1%
Rap	15.9%	6.2%	51.0%	6.9%	7.6%
Hip Hop	18.3%	7.4%	51.5%	6.1%	6.1%

Red: significantly above the overall average rate for cause of death
Blue: above the overall average rate for cause of death
Green: significantly below the overall average rate for cause of death

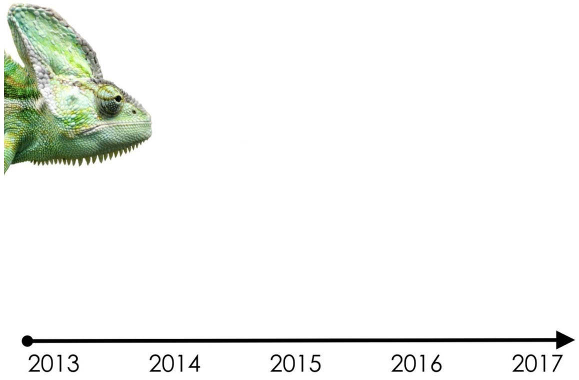
Note: not all causes shown

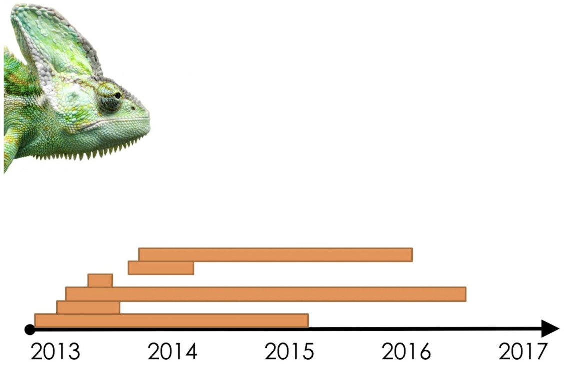
theconversation.com

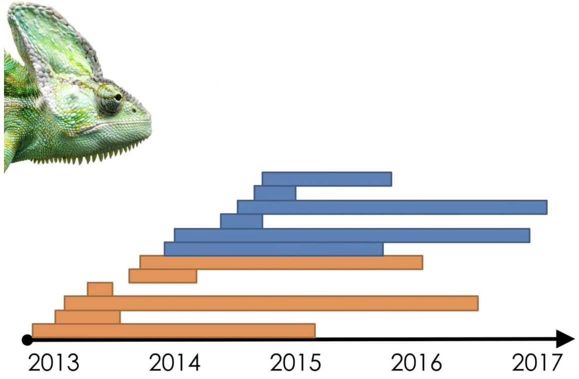
Source: Author

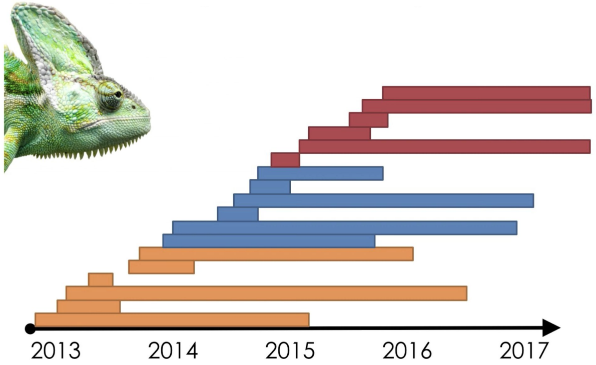
https://www.callingbullshit.org/case_studies/case_study_musician_mortality.html

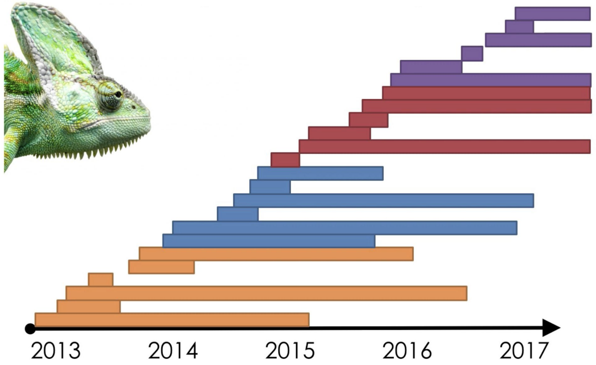
Sampling bias: Right censoring





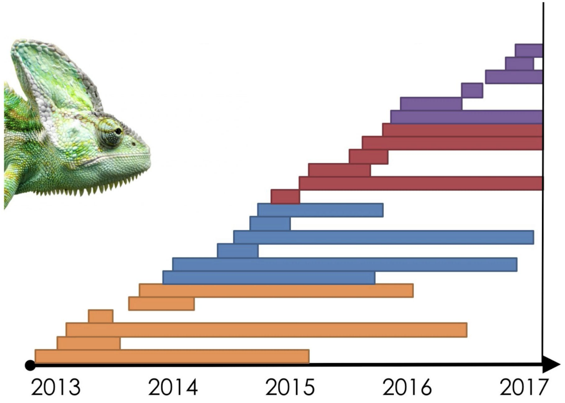


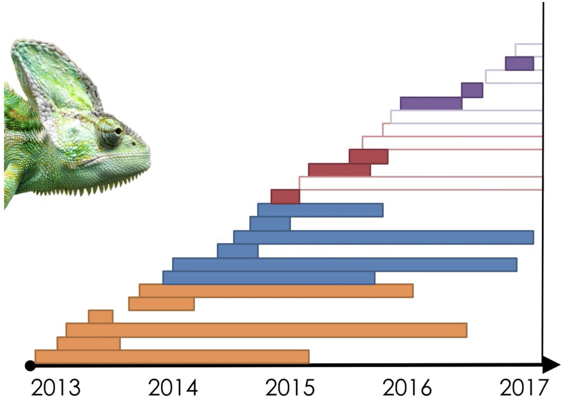


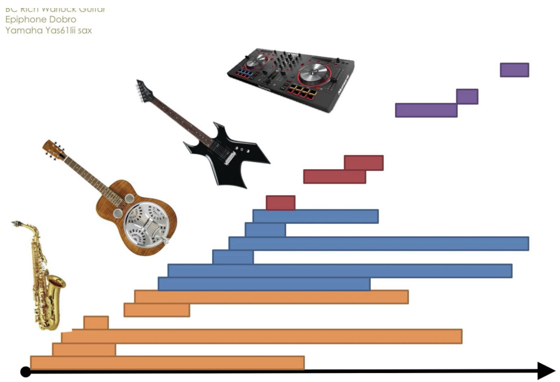


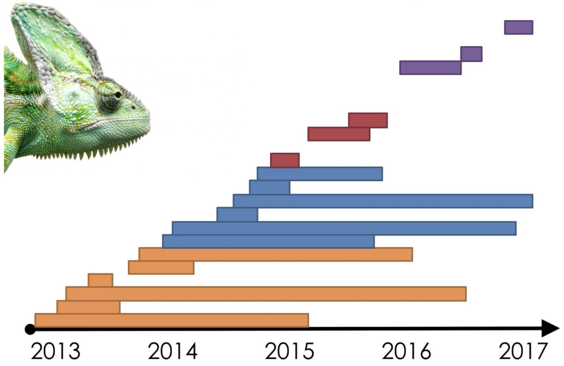
https://www.callingbullshit.org/case_studies/case_study_musician_mortality.html

Sampling bias: Right censoring



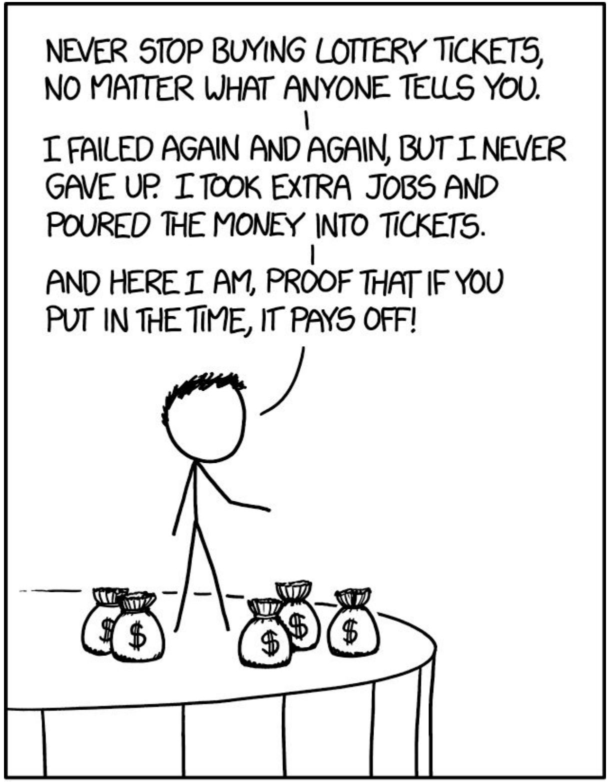




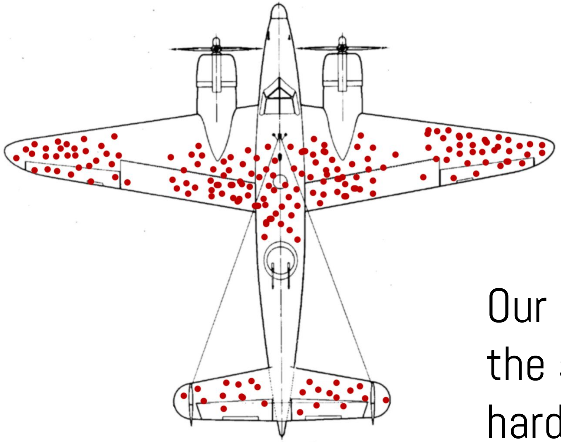


https://www.callingbullshit.org/case_studies/case_study_musician_mortality.html

Sampling bias: Survivorship bias


NEVER STOP BUYING LOTTERY TICKETS, NO MATTER WHAT ANYONE TELLS YOU.
I FAILED AGAIN AND AGAIN, BUT I NEVER GAVE UP. I TOOK EXTRA JOBS AND POURED THE MONEY INTO TICKETS.
AND HERE I AM, PROOF THAT IF YOU PUT IN THE TIME, IT PAYS OFF!
$
$
$
$
$

EVERY INSPIRATIONAL SPEECH BY SOMEONE SUCCESSFUL SHOULD HAVE TO START WITH A DISCLAIMER ABOUT SURVIVORSHIP BIAS.



To minimize bomber losses to enemy fire, add armor to the areas with most damage. [Abraham Wald said no!]

Our online survey showed that all of the students had the required hardware and internet bandwidth to participate in the online courses.

Dramatic rise in field hospital admissions of severe head injury victims → redesign! [Unknown statistician said no!]

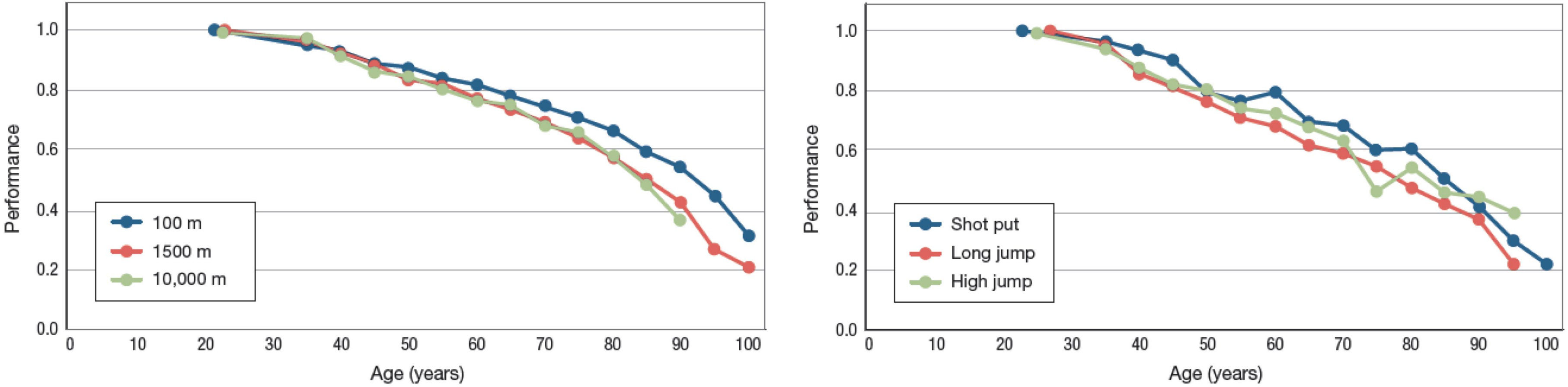
https://xkcd.com/1827/; https://en.wikipedia.org/wiki/Survivorship_bias

Sampling bias: Outliers and sample size

Bergstrom and Dugatkin (2016) *Evolution*: Not only does mortality increase and fertility decrease with age, but individuals undergo decline in physical performance with age as well.

World record performances in six track and field events, for different age groups.

Performance is scaled relative to the world record for any age. In track events, performance is quantified as average speed; in field events, performance is quantified as distance or height.



https://www.callingbullshit.org/case_studies/case_study_track_records.html

Sampling bias: Outliers and sample size

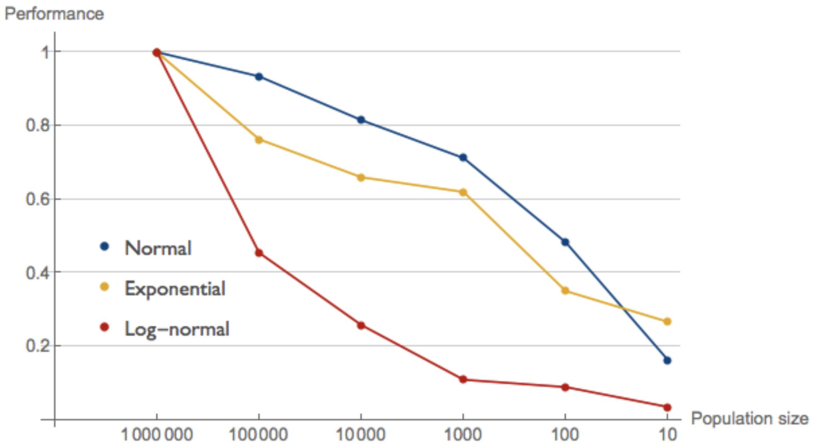
The issue is that: sample sizes are very different for different ages and world records are outliers.

Simulation:

- Draw scores from the same distribution for populations of size 10^6, 10^5, etc., down to 10^1.
- For each population, plot the highest score (scaled relative to the highest score overall).

The scores fall off least rapidly in the normal distribution (blue), more quickly in the fat-tailed exponential distribution (yellow) and even faster in the fatter-tailed log-normal distribution (red).

For all three distributions, the top performer in the population of size 10 scores less than 1/3 as high as the top performer overall, despite being drawn from the same distribution.

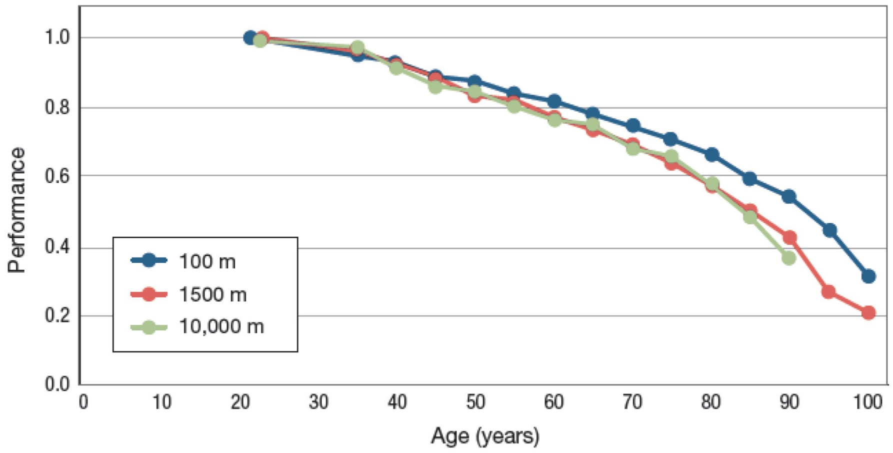


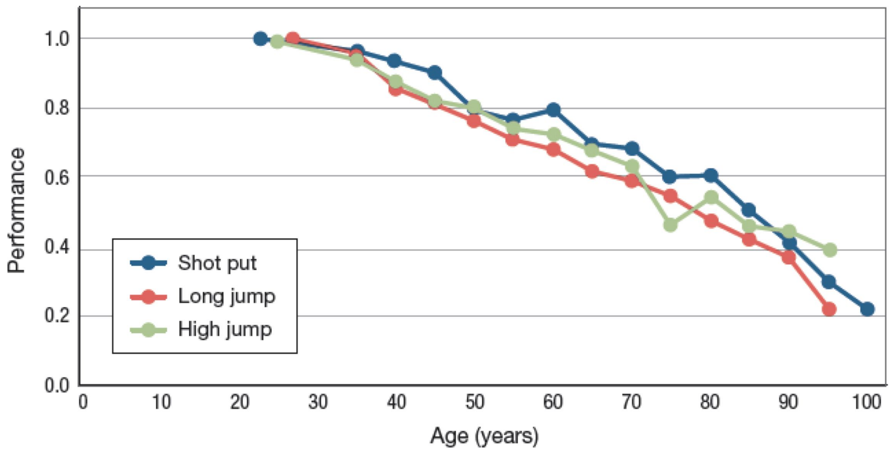
https://www.callingbullshit.org/case_studies/case_study_track_records.html

Sampling bias: Cohort effects

Cohort effects occur when changing environmental conditions over time result in different age groups having experienced different environmental effects on the trait values under observation.

Example: we may observe higher rates of lung cancer in 90-year-olds than in 50-year-olds not only because cancer rates increase with age, but also because of the cohort effect that 50-year-olds are less likely to have smoked in their 20's.



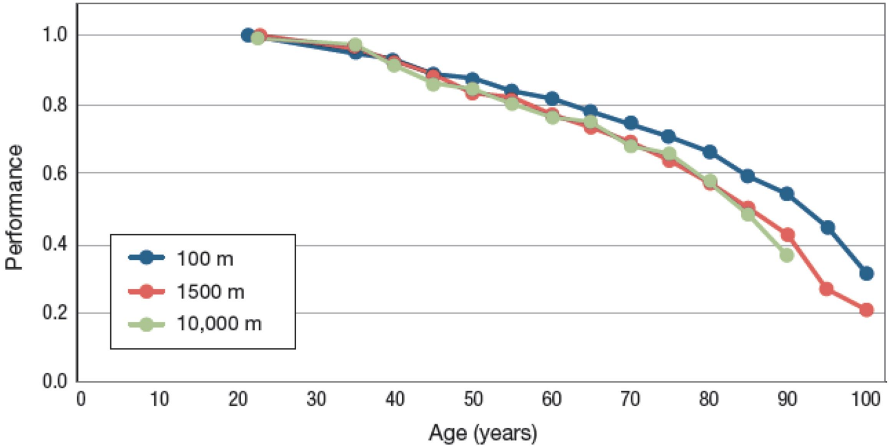


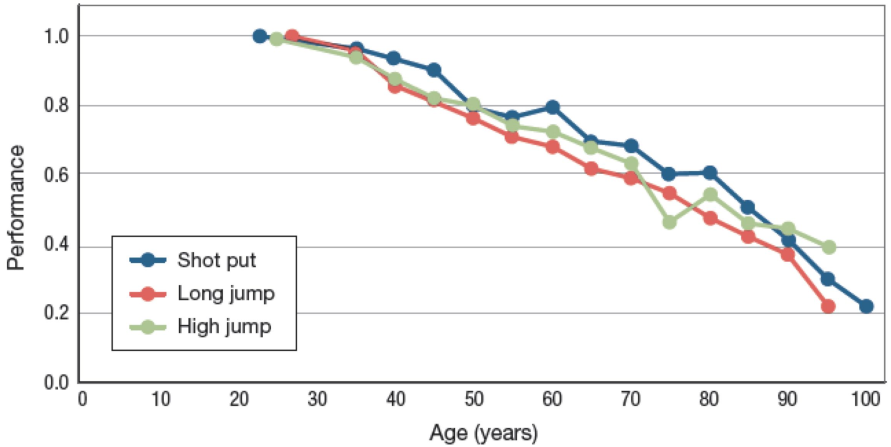
https://www.callingbullshit.org/case_studies/case_study_track_records.html

Sampling bias: Cohort effects

The 80-year-old record holders are not only older than members of other cohorts.

They were born earlier and thus differ in a number of ways including training regimens, nutrition, size of the pool able to participate, possible use of performance enhancing drugs, etc.





https://www.callingbullshit.org/case_studies/case_study_track_records.html