INM430
Principles of Data Science

Week 04
Investigate relations (& structures)

Aidan Slingsby, giCentre

CITY
UNIVERSITY OF LONDON
EST 1894

On the menu today

- Null-hypothesis testing and “The New Statistics”
- Relationships between variables (columns)
  - Correlation
  - Regression
- Finding **causal relations**

From last week …

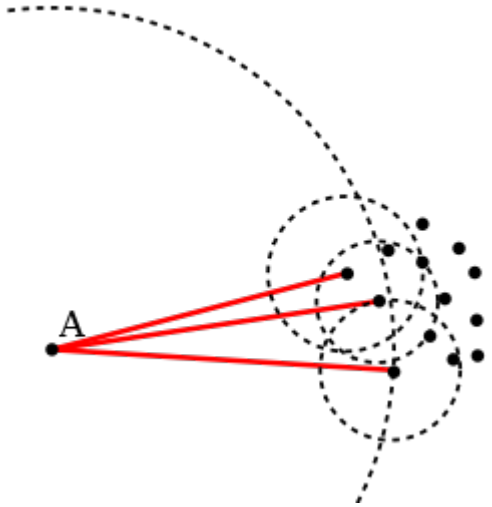
A recap -- outliers

Outliers – friend or foe?

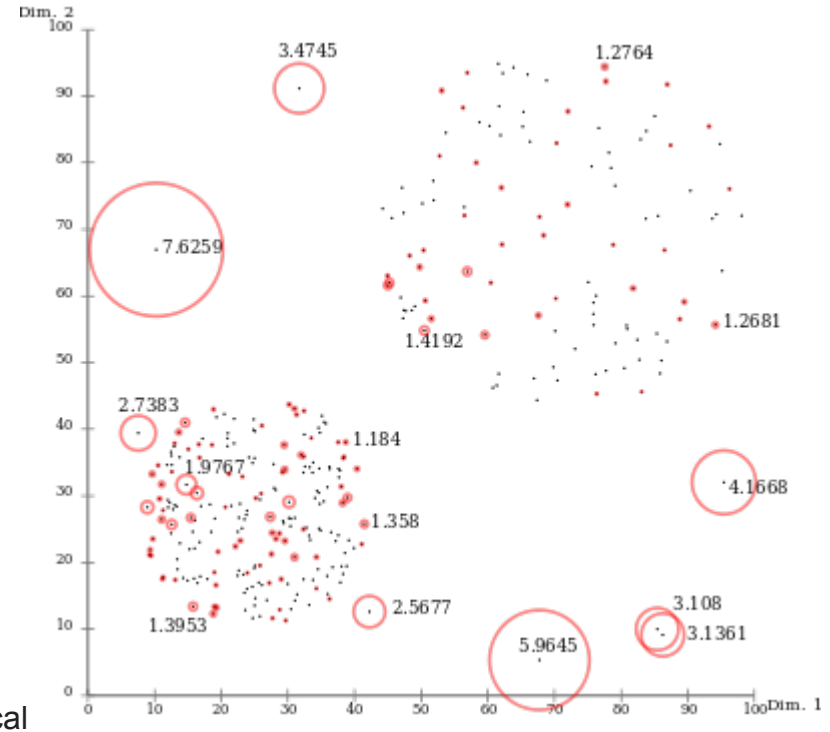
- might be **problematic values**
  - faulty readings, measurement errors, missing data, …
- might be **what you are after**
  - fraud detection, network intrusion detection, ….
- might be **something unexpected**
  - valuable analytical finding
  - might be filtered by automated methods

Outlier Detection – Density Based Approach

- Local outlier factor (Breunig et al. 2000)
- Find outliers by measuring the **local** deviation of a given data point **with respect to its neighbours**



A has lower density compared to neighbours



LOF scores

Breunig, Markus M., et al. "LOF: identifying density-based local outliers."*ACM sigmod record*. Vol. 29. No. 2. ACM, 2000.

Living with outliers -- Robust statistics

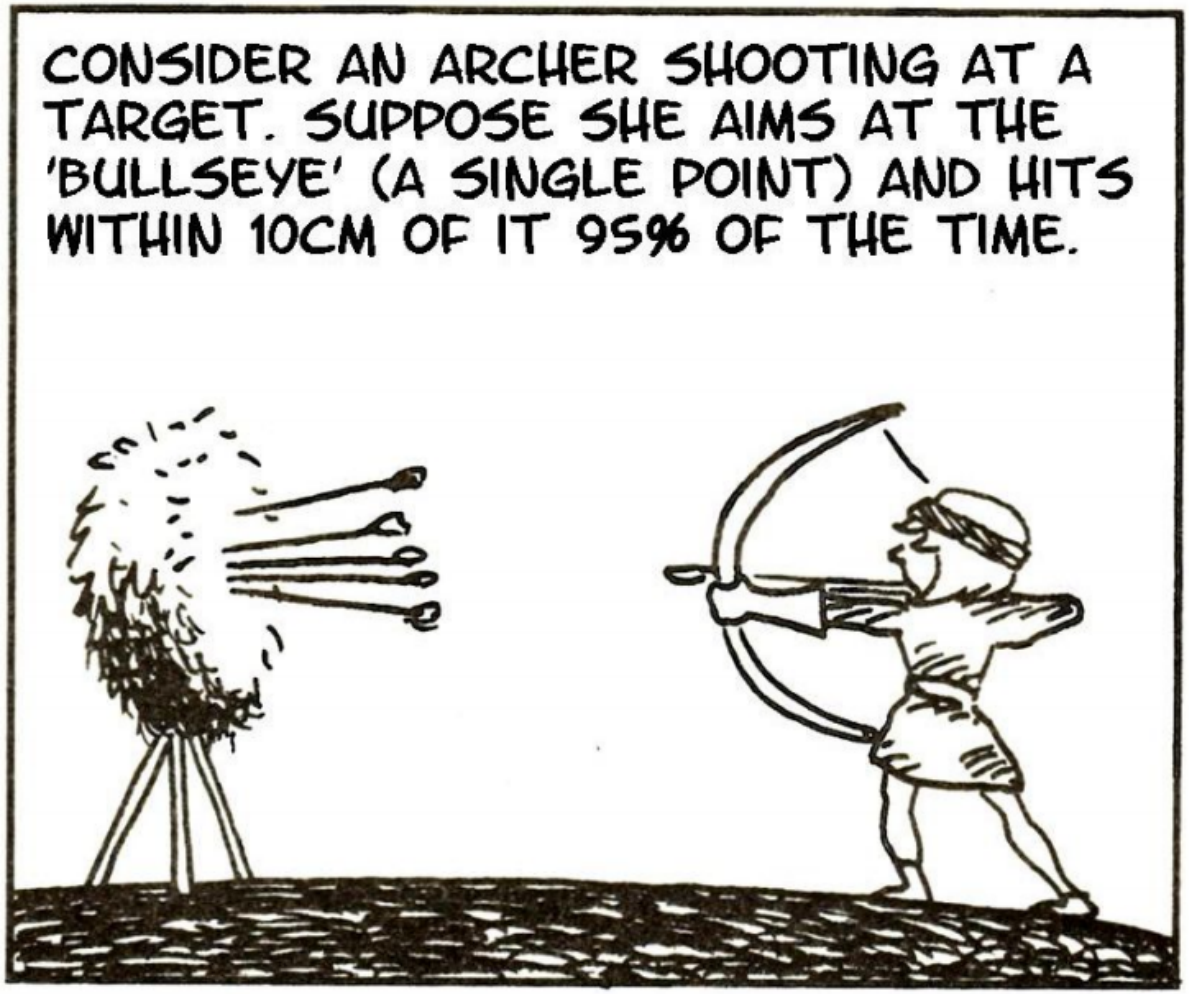
- Statistics / methods that are resistant against outliers
- No need to remove outliers, in theory
- Focus on finding better statistical estimates
- Can use robust statistics in parametric methods to "robustify" them, e.g., fit a regression line using robust μ and σ

FREQUENTIST INFERENCE
NULL HYPOTHESIS TESTING
THE NEW STATISTICS

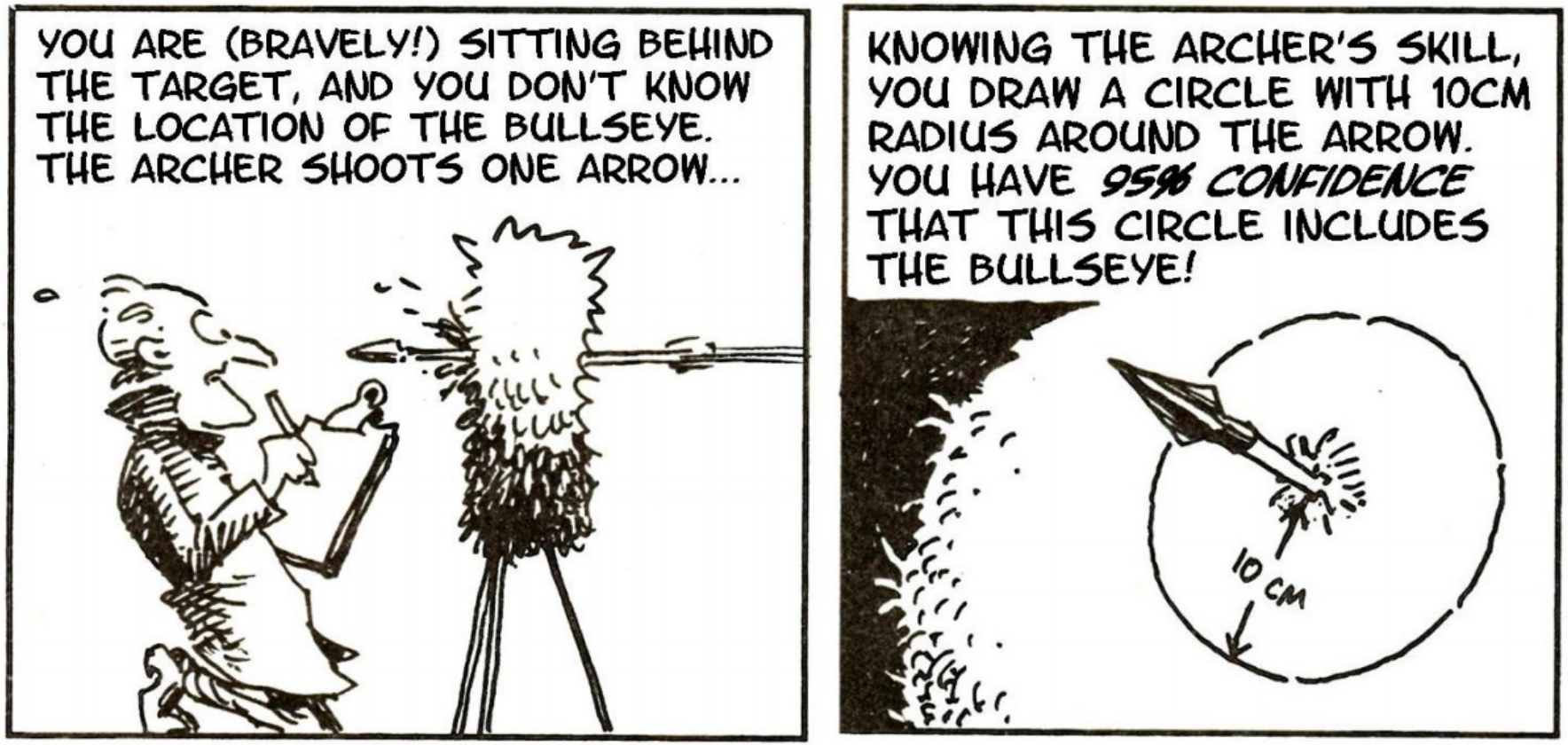
Inferential statistics

- (Descriptive statistics simply describe data you have sampled)
- Inferential statistics allow you to make inferences to the population from which you are sampling

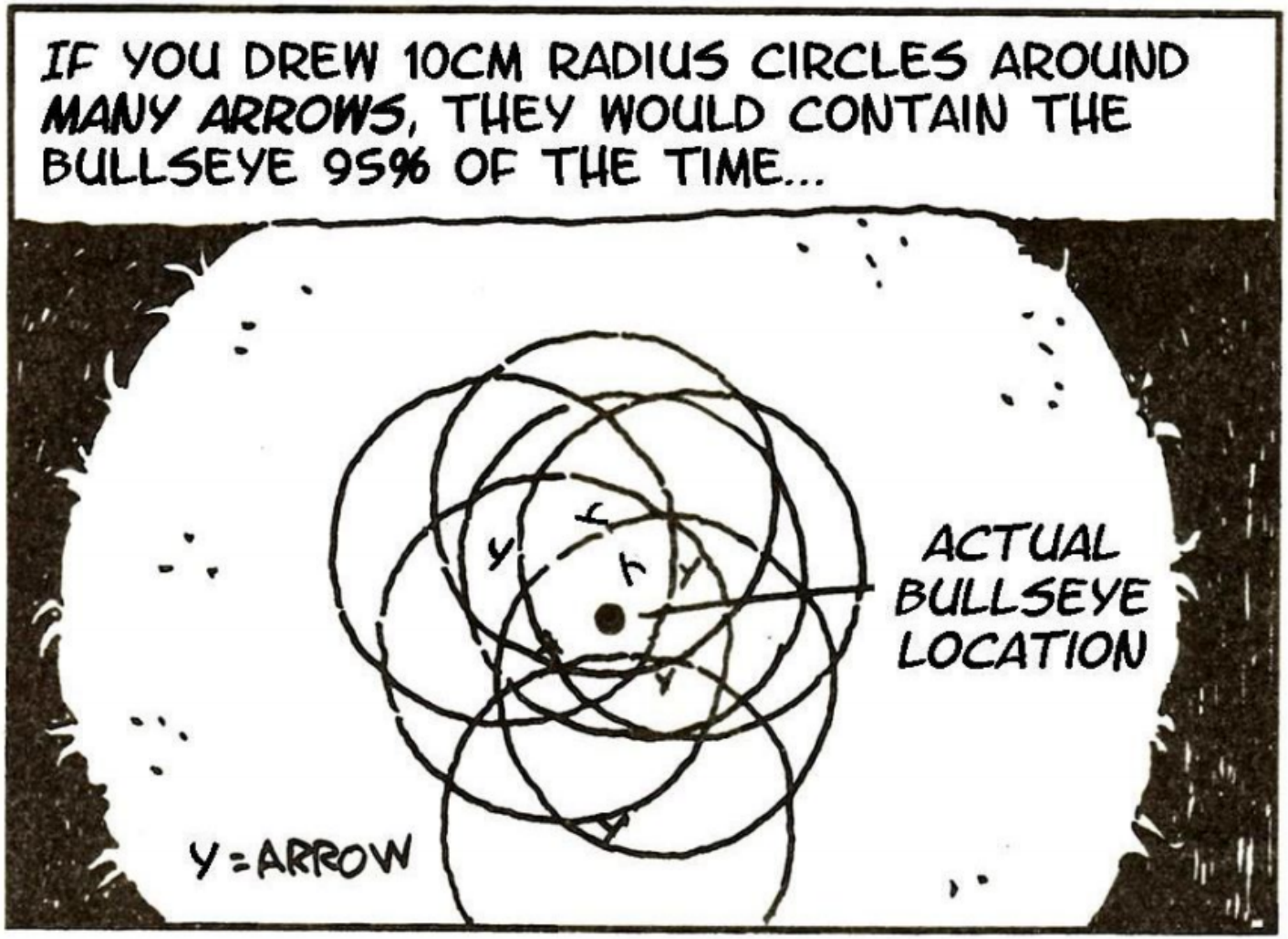
From **“The Cartoon Guide to Statistics”, by Gonick & Smith**



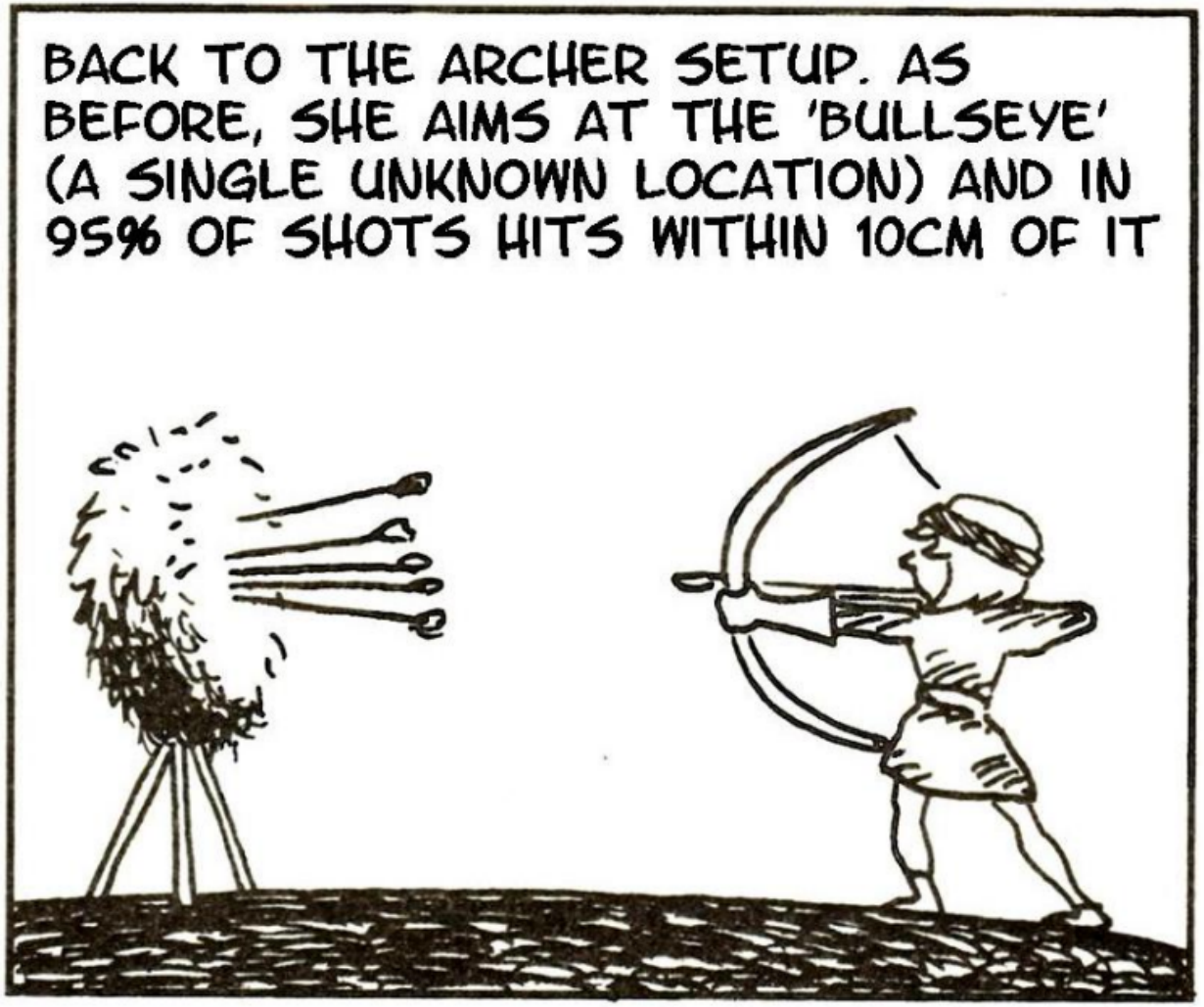
From “The Cartoon Guide to Statistics”, by Gonick & Smith



From **"The Cartoon Guide to Statistics", by Gonick & Smith**

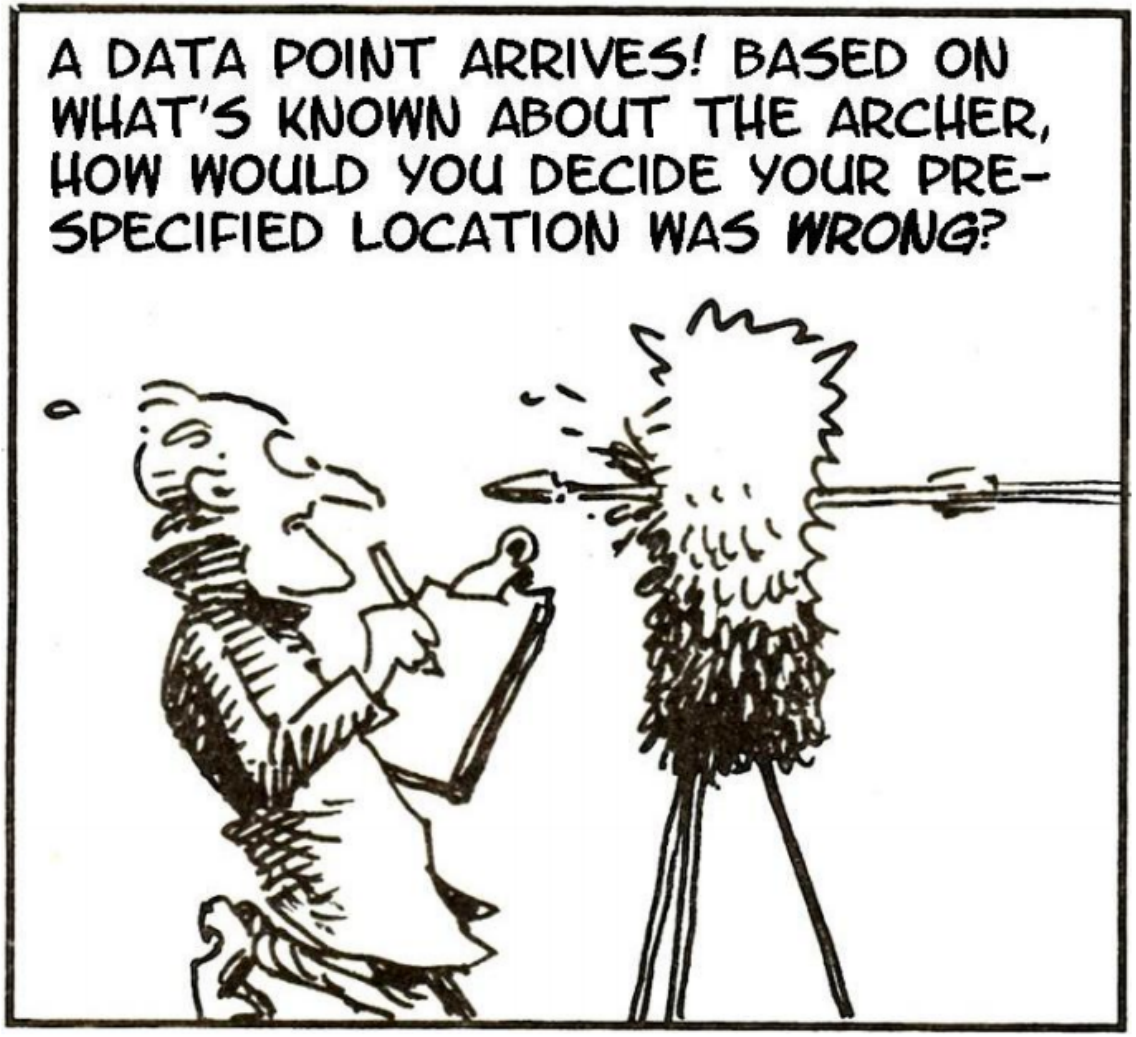


From **"The Cartoon Guide to Statistics", by Gonick & Smith**

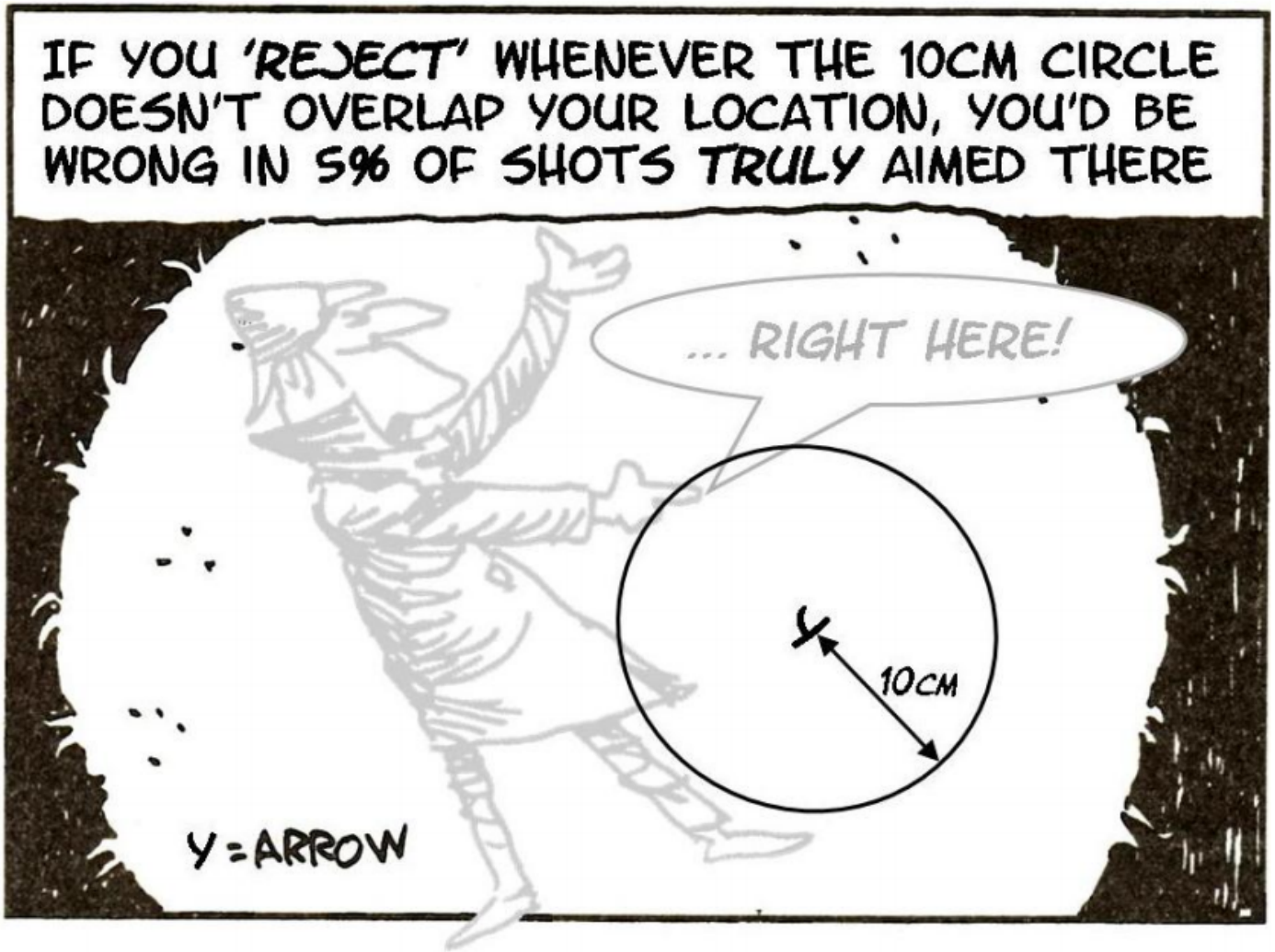


From **"The Cartoon Guide to Statistics", by Gonick & Smith**

From **"The Cartoon Guide to Statistics", by Gonick & Smith**



From **"The Cartoon Guide to Statistics", by Gonick & Smith**



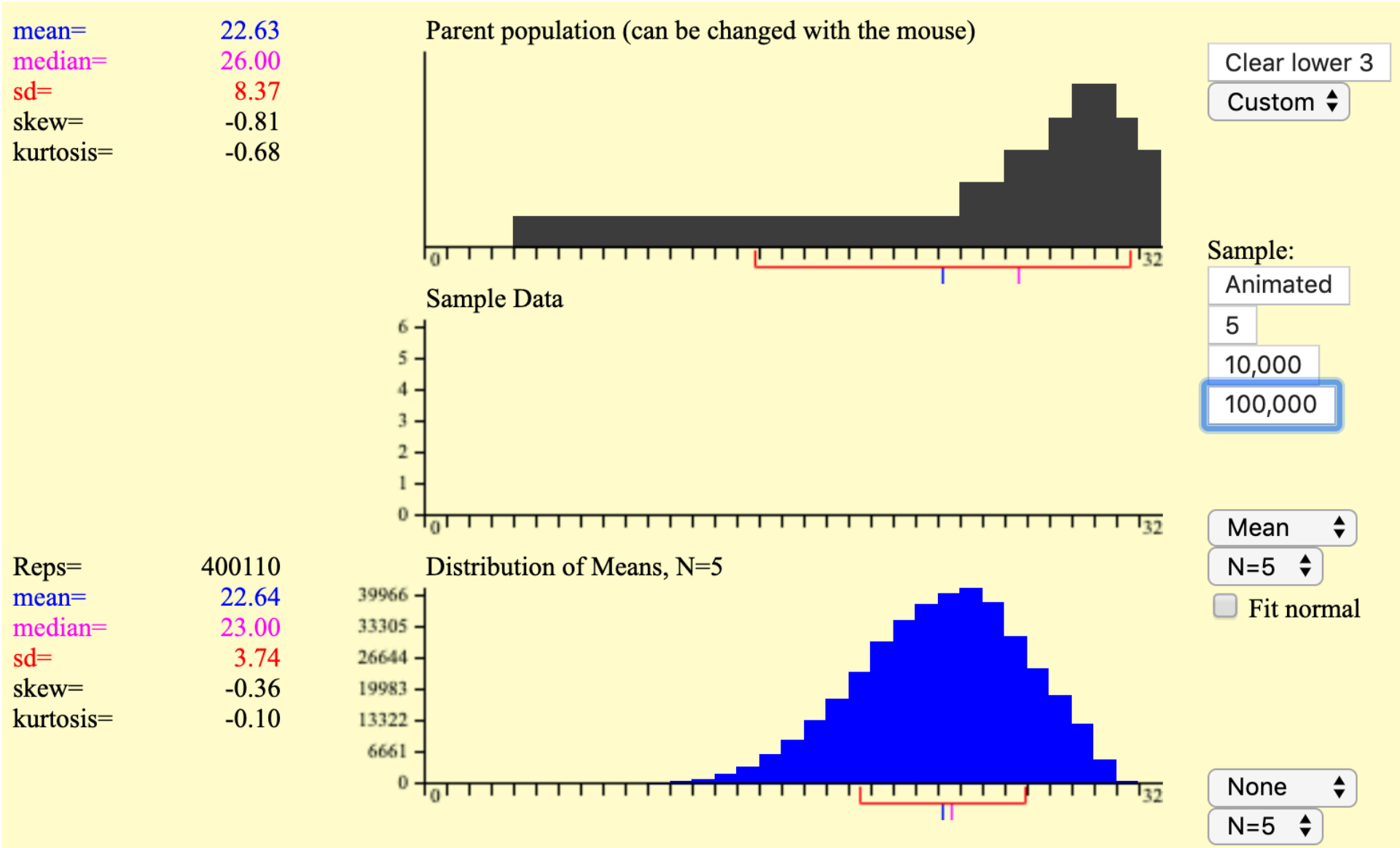
Frequentist inference

- Make better inferences by **repeated sampling**
- Outputs:
  - **Point estimates:** "best guess" of an unknown population parameter (e.g., mean), given the sample
  - **Interval estimates:** range containing the true parameter value with the probability at the stated confidence level (confidence intervals)

Inferential statistics

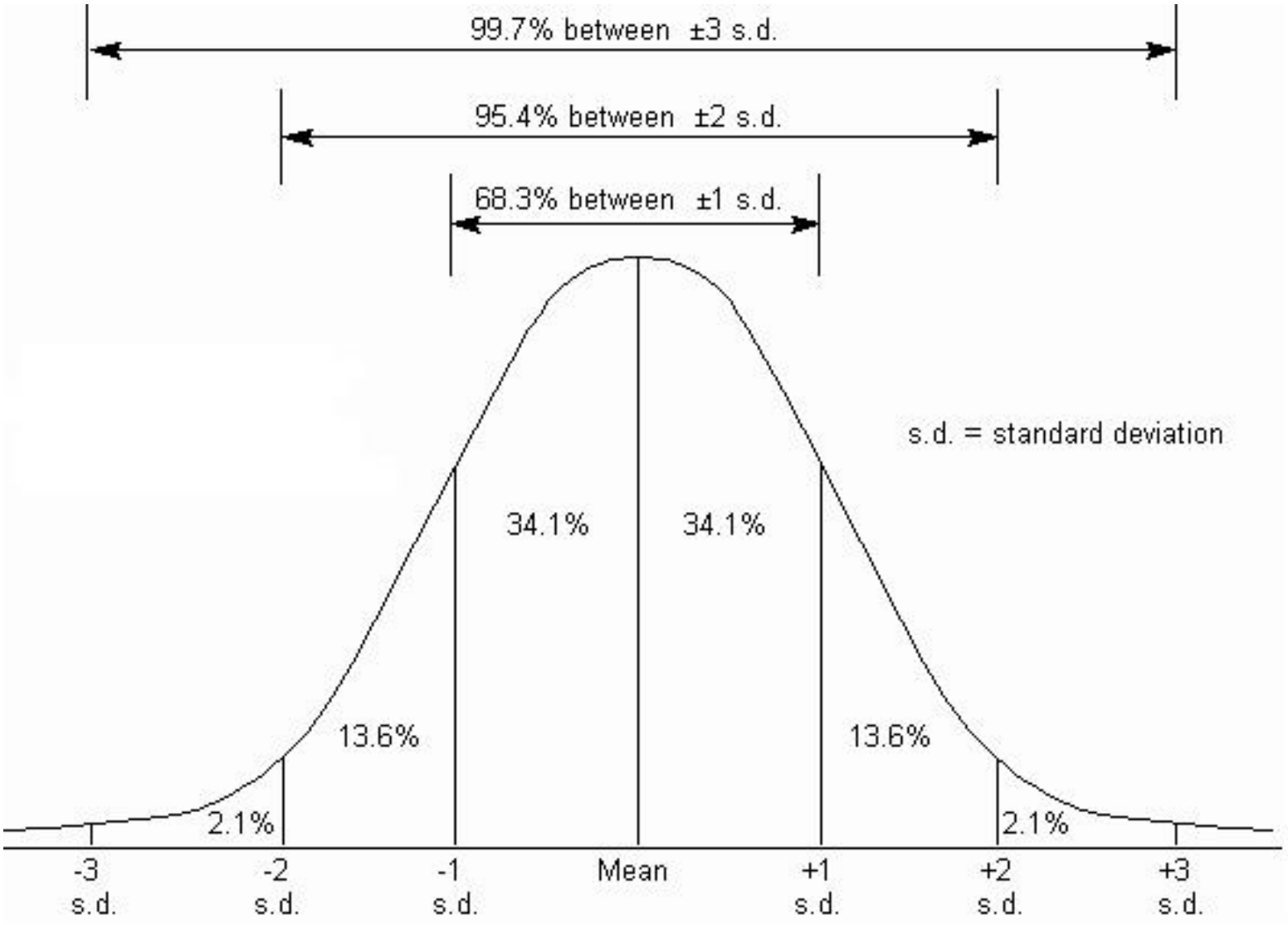
- Inferential statistics allow you to make inferences to the population from which you are sampling
- Underpinned by the Central Limit Theorem
  - Sampling distribution of the mean follows a **normal distribution** with a **mean** (μ) and a **variance** of (σ^2/N, where N is the sample size)…
  - …no matter what the original distribution was

Sampling from a non-normal distribution

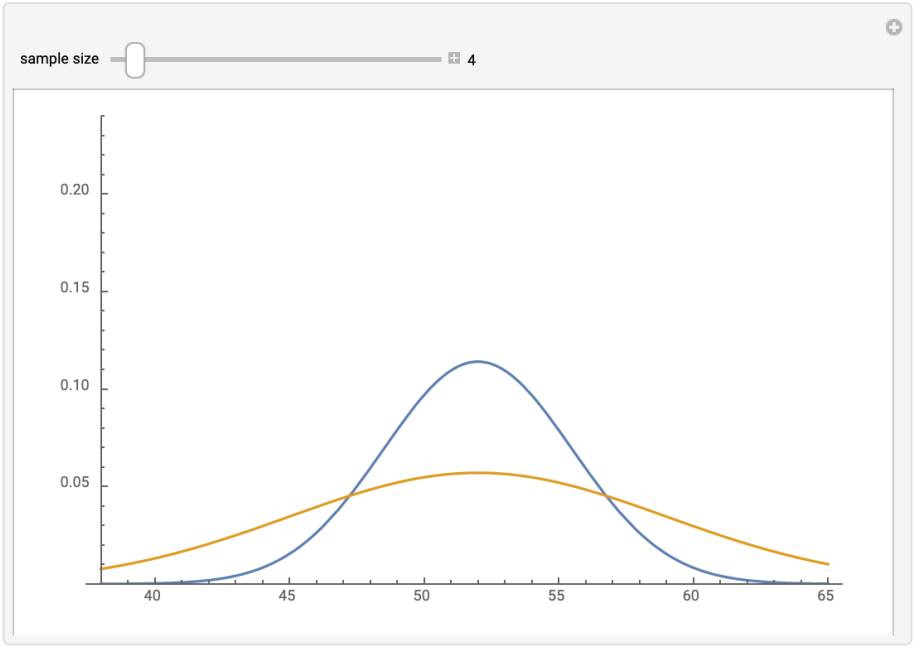


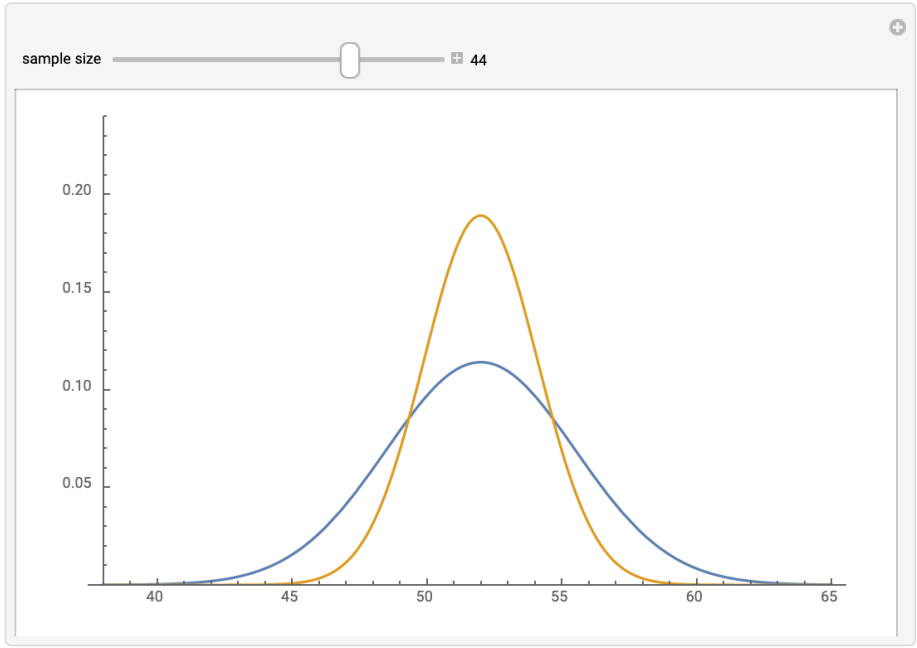
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Normal distribution

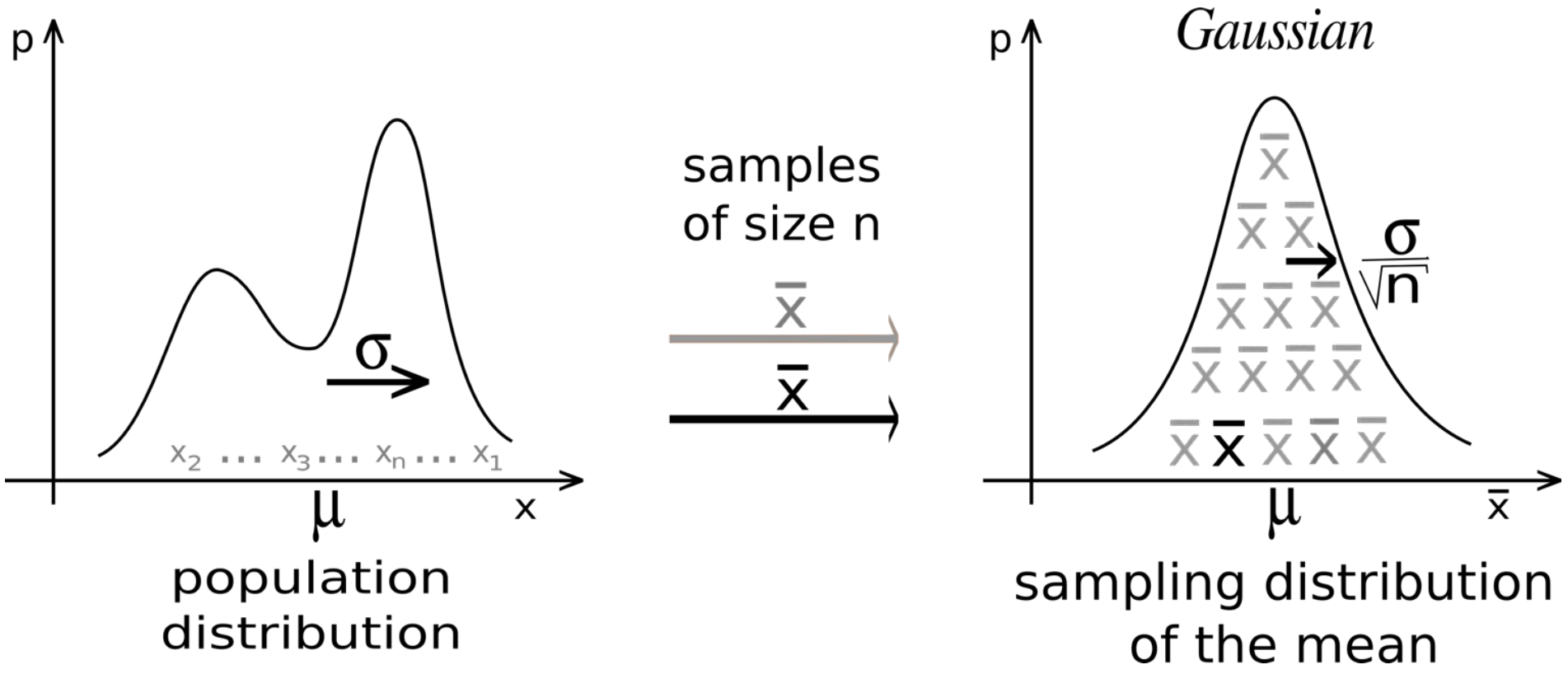


Effect of sample size on normal distribution





https://demonstrations.wolfram.com/DistributionOfNormalMeansWithDifferentSampleSizes/



https://en.wikipedia.org/wiki/Central_limit_theorem

Some crucial frequentist concepts

- **Central Limit Theorem**: states that given a distribution with a mean μ and variance σ^2, the **sampling distribution of the mean** approaches a normal distribution with a mean (μ) and a variance σ^2/N as N, the sample size
- The very fundamental aspect about the central limit theorem is that **no matter what the shape of the original distribution**, the **sampling distribution of the mean approaches a normal distribution**.

Interactive demo on sampling distribution:
http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Null hypothesis significance testing (NHST)

- Commonly-used process for determining whether an effect measured in a **sample** is likely to exist in the **population**, within certain bounds
  - Does a drug have an effect on the symptom severity
- Helps determine whether there is statistical support for the effect observed in the sample to also be in the population.
  - The bigger the sample size, the more likely this is

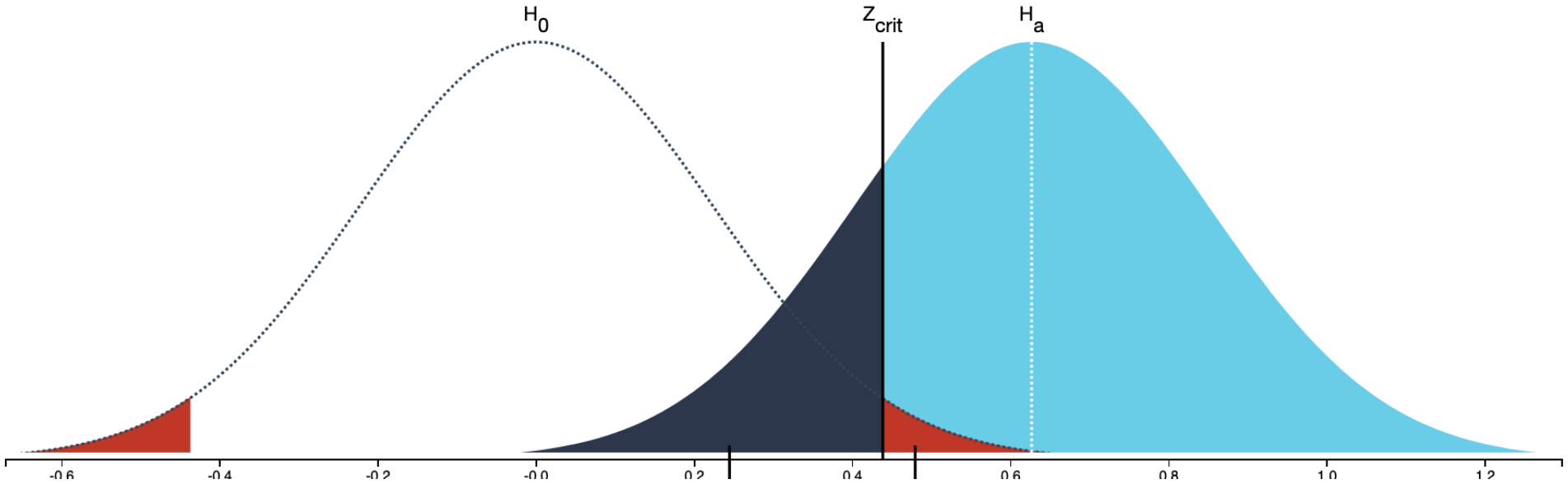
Null hypothesis significance testing (NHST

- **Hypothesis** (**H_1**; confusingly called the "Alternative Hypothesis"): *there is an effect*
  - treatment with drug **affects** the severity of the symptoms
- **Null hypothesis (H_0):** *there is not an effect*
  - treatment with drug reduces **has no effect** on the severity of the symptoms
- See if you can **reject** the null hypothesis or **fail to reject it**.
- Often boils down to the *difference* in means between two samples

p-values and significance

- Significance level is the probability that rejection of the null hypothesis is true
  - Conventionally 5% (0.05) is considered “significant”
- The p-value is the probability that obtaining the observed results (or results more extreme) if H_0 is true
  - So need small values to reject the null hypothesis
- If a mean difference is statistically significant, then the effect is likely to be present in the population

Is the difference between the means statistically significant?



https://rpsychologist.com/d3/NHST/

nature International weekly journal of science

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

ScienceNews
MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Archive

Search Science News...

Explore

LATEST | MOST VIEWED

TELEVISION
Hollywood-made science documentary series comes to TV
BY TINA HESMAN SAEY OCTOBER 20, 2015

SCIENCE VISUALIZED
'Whalecopter' drone swoops in for a shot and a shower
BY SUSAN MILIUS OCTOBER 20, 2015

SCIENCE TICKER
Climate change could shift New

Context

SCIENCE PAST AND PRESENT
TOM SIEGFRIED

CONTEXT NUMBERS

P value ban: small step for a journal, giant leap for science

Editors reject flawed system of null hypothesis testing

BY TOM SIEGFRIED 3:18PM, MARCH 17, 2015

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

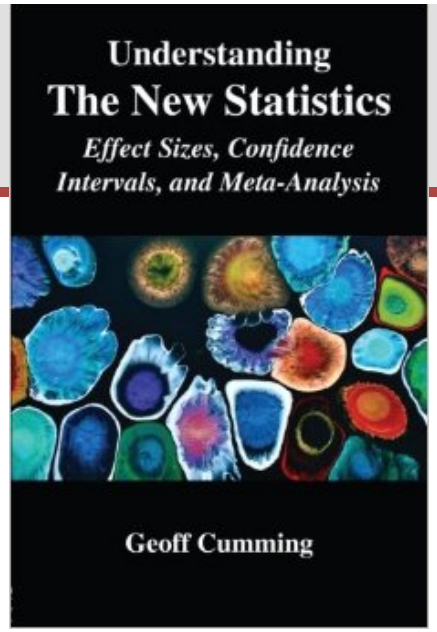
Editorial

David Trafimow and Michael Marks
New Mexico State University

The Basic and Applied Social Psychology (BASP) 2014 Editorial **emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it** (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. **From now on, BASP is banning the NHSTP.**

…. Some might view the NHSTP ban as indicating that it will be easier to publish in BASP, or that less rigorous manuscripts will be acceptable. This is not so. On the contrary, we believe that the **$p < .05$ bar is too easy to pass** and sometimes serves as an excuse for lower quality research.

The “new” statistics



- Are not really new
- ***Avoid binary decision-making*** as in NHST
- Replace ***estimation*** with testing
- Use effect sizes (ESs) and confidence intervals
- From Cumming, 2014:

“… Suppose you read in the news that “support for Proposition X is 53%, in a poll with an error margin of 2%.” Most readers immediately understand that the 53% came from a sample and, assuming that the poll was competent, conclude that 53% is a fair estimate of support in the population. The 2% suggests the largest likely error. Reporting a result in such a way, or as 53 ± 2%, or as 53% with a 95% CI of [51, 55], is natural and informative. It is more informative than stating that support is “statistically significantly greater than 50%, $p < .01$.” The 53% is our point estimate, and the CI our interval estimate, whose length indicates precision of estimation. “…

Cumming, G. (2013). The new statistics why and how. *Psychological science*.

Effect Sizes

- Measure of the strength of an effect
  - mean
  - differences between means
  - correlations
  - (p-values don’t indicate effect size)
- Their interpretation requires **informed judgment in context**
  - Is the size of the effect important?

Cohen's *d*

- A standardised effect size
- Standardized difference between sample means

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$$

sample means

standard deviation

- .. where, s is the pooled standard deviation

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

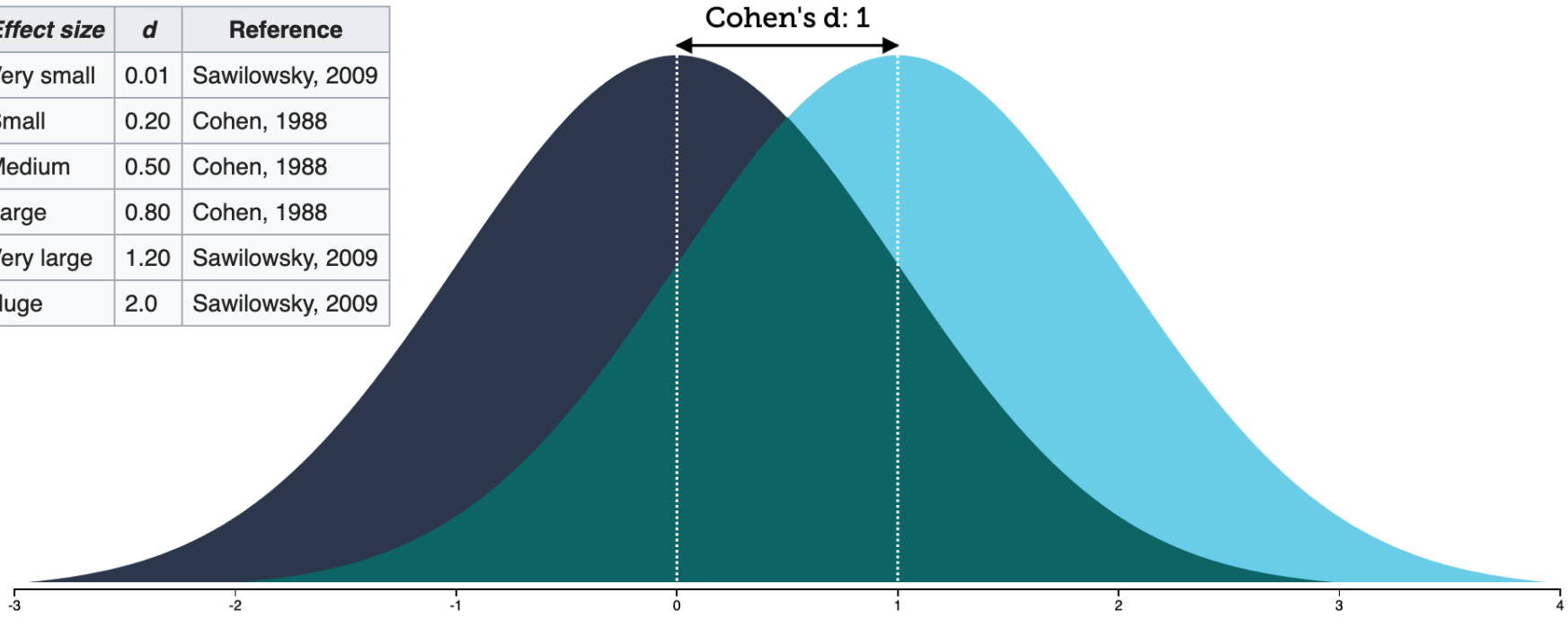
- .. and variance for the groups can be computed as:

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n1} (x_{1,i} - \bar{x}_1)^2,$$

d (Jacob Cohen (1988). Statistical Power Analysis for the Behavioral Sciences)

Cohen's *d*

Effect size	*d*	Reference
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009



https://en.wikipedia.org/wiki/Effect_size#Cohen's_d

https://rpsychologist.com/d3/cohend/

Interpretation

Cohen's U_3

% Overlap

Probability of Superiority

Number Needed to Treat[1]

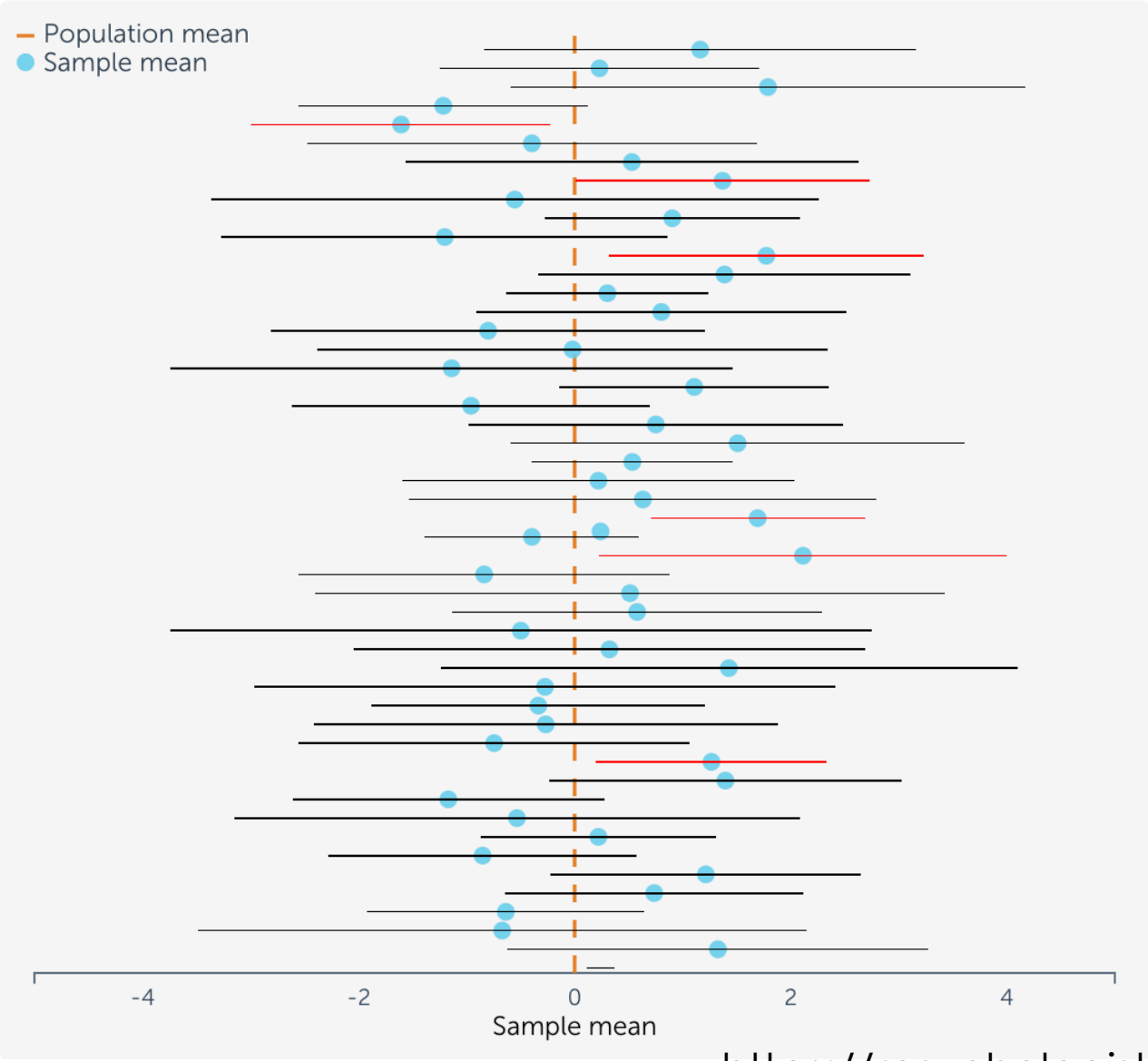
A Common Language Explanation

With a Cohen's *d* of 1, 84 % of the treatment group will be above the mean of the control group (Cohen's U_3), 62 % of the two groups will overlap, and there is a 76 % chance that a person picked at random from the treatment group will have a higher score than a person picked at random from the control group (probability of superiority). Moreover, in order to have one more favorable outcome in the treatment group compared to the control group we need to treat 2.8 people. This means that if 100 people go through the treatment, 36.3 more people will have a favorable outcome compared to if they had received the control treatment[1].

Confidence interval (CI)

- An interval estimate of a population parameter.
  - observed interval (calculated from observations)
  - For the 90% case, CIs from repeated sampling should include the parameter, 90% of the time

Confidence interval (CI)



https://rpsychologist.com/d3/CI/

Suggestions from Cumming (2013)

- **Formulate research questions in estimation** terms…
  - E.g. How large is the effect? To what extent is the effect?
- …**rather than in dichotomous** terms.
  - E.g. Is there a difference? Is this treatment better?
- Identify the effects size that **best answers the research question**.
  - If the question asks about the difference in amount, use the different between two means as the effect size
- See the article.

From:Cumming, G. (2013). The new statistics why and how. *Psychological science*.

But the debate will continue …

The "new statistics" are built on fundamentally flawed foundations

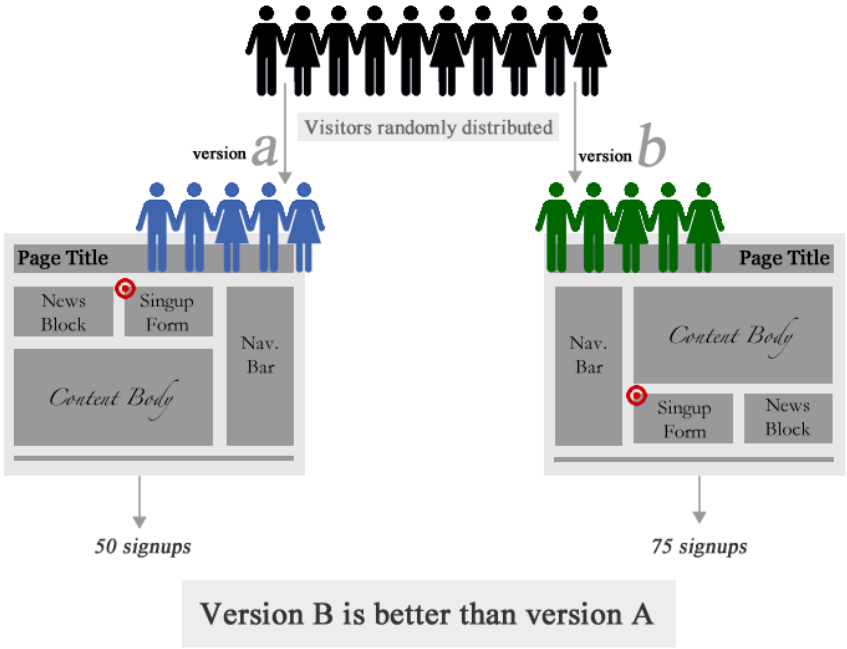
Michael D. Lee

University of California, Irvine

…. Methodologically, it is wrong to rely on flawed orthodox statistical theory. We should all just be Bayesian.

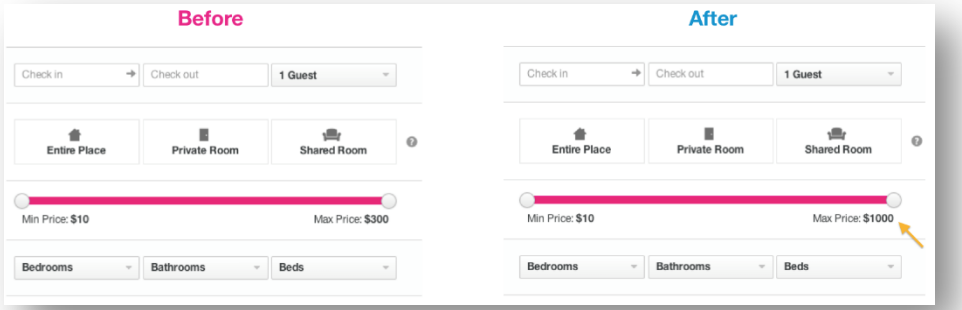
A/B testing

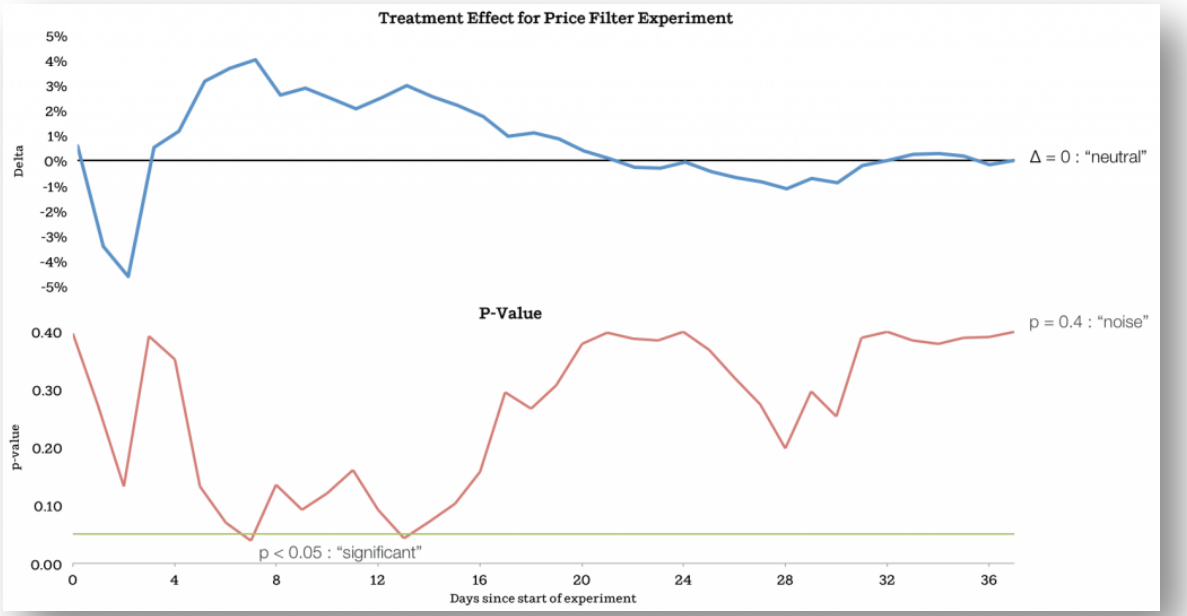
- Commonly use in technology firms:
  - Introduce one design to a group A.
  - And another to group B.
  - Check for differences.

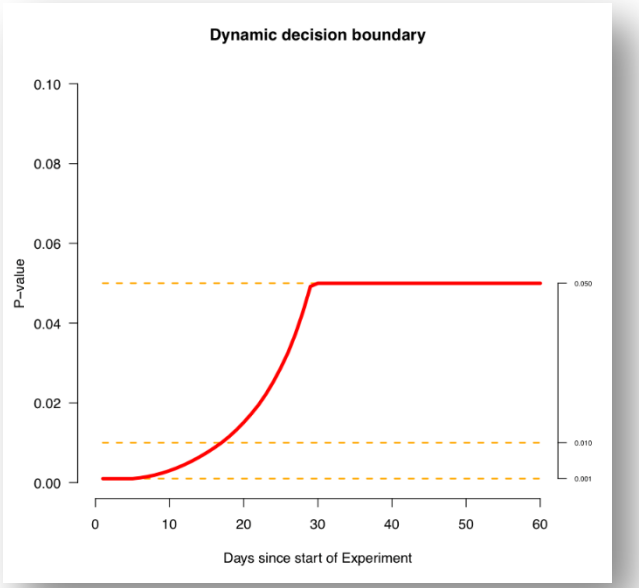


http://www.smashingmagazine.com/2010/06/the-ultimate-guide-to-a-b-testing/

An example from Airbnb (http://nerds.airbnb.com/experiments-at-airbnb/)







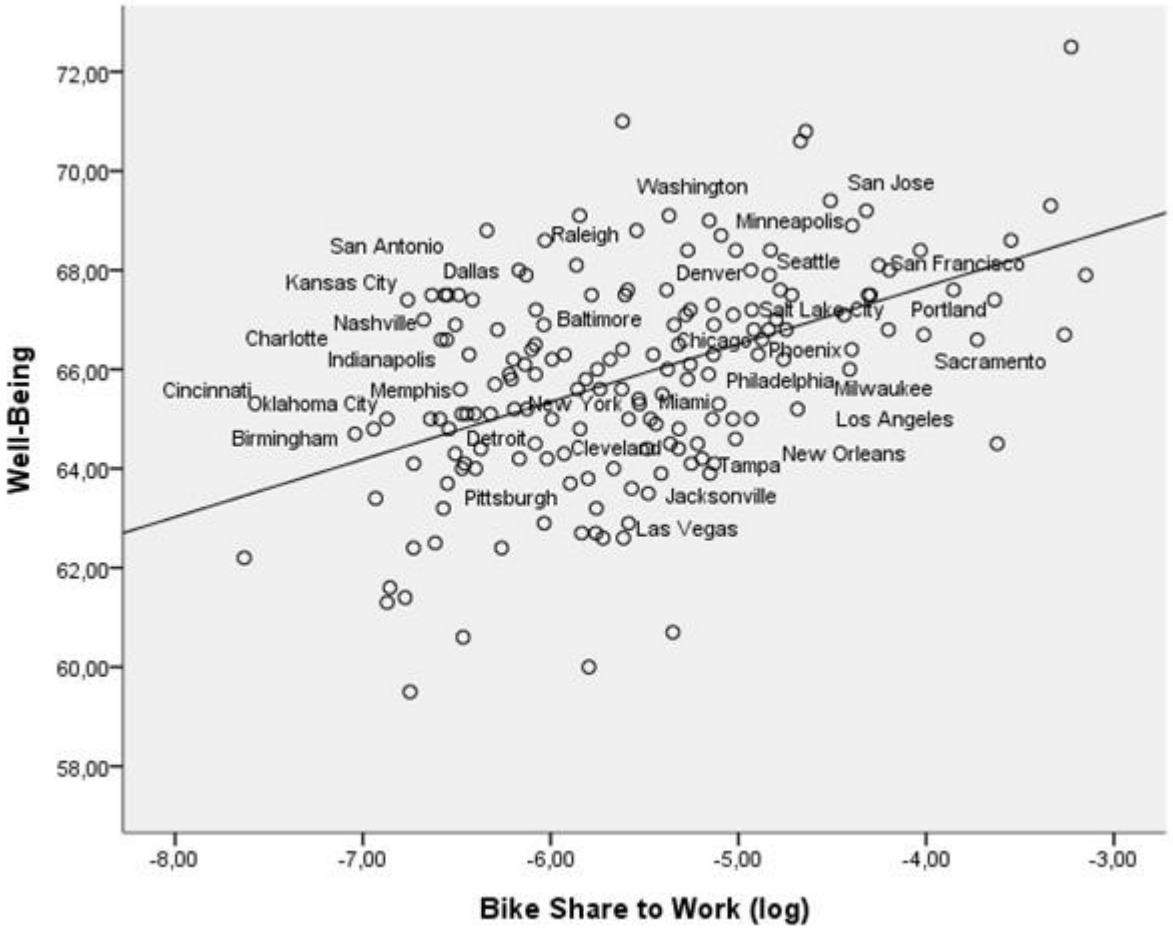
Use data to identify **relationships** among variables and use these relationships to build **models** and predictions

Correlation (a.k.a. dependence)

Finding the relationship between two quantitative variables **without** being able to infer **causal relationships**

Correlation is a statistical technique used to determine the **degree** to which two variables are **related**

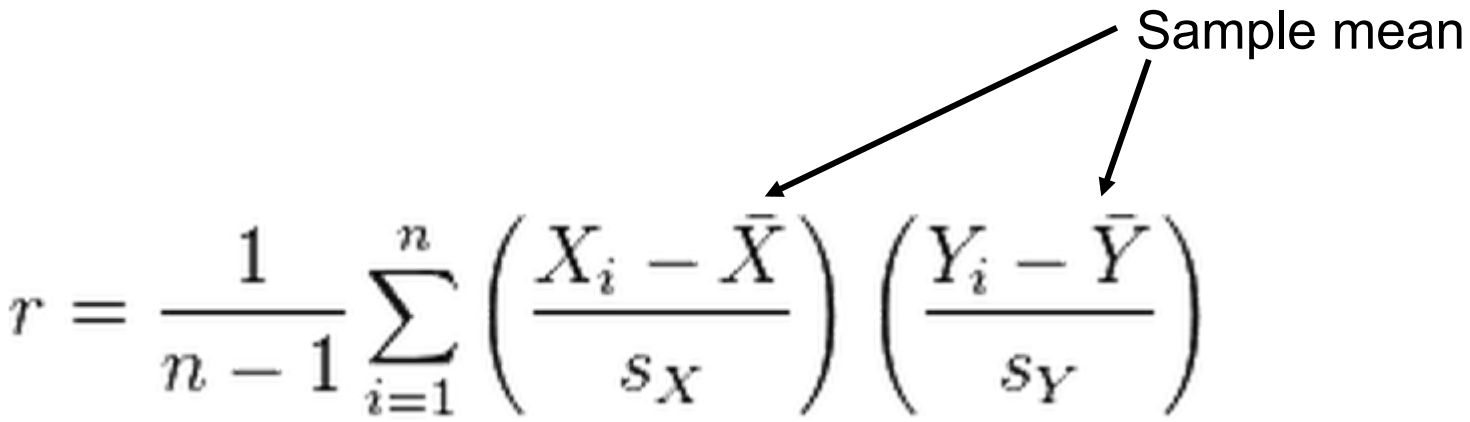
Example – Well-being vs. cycling



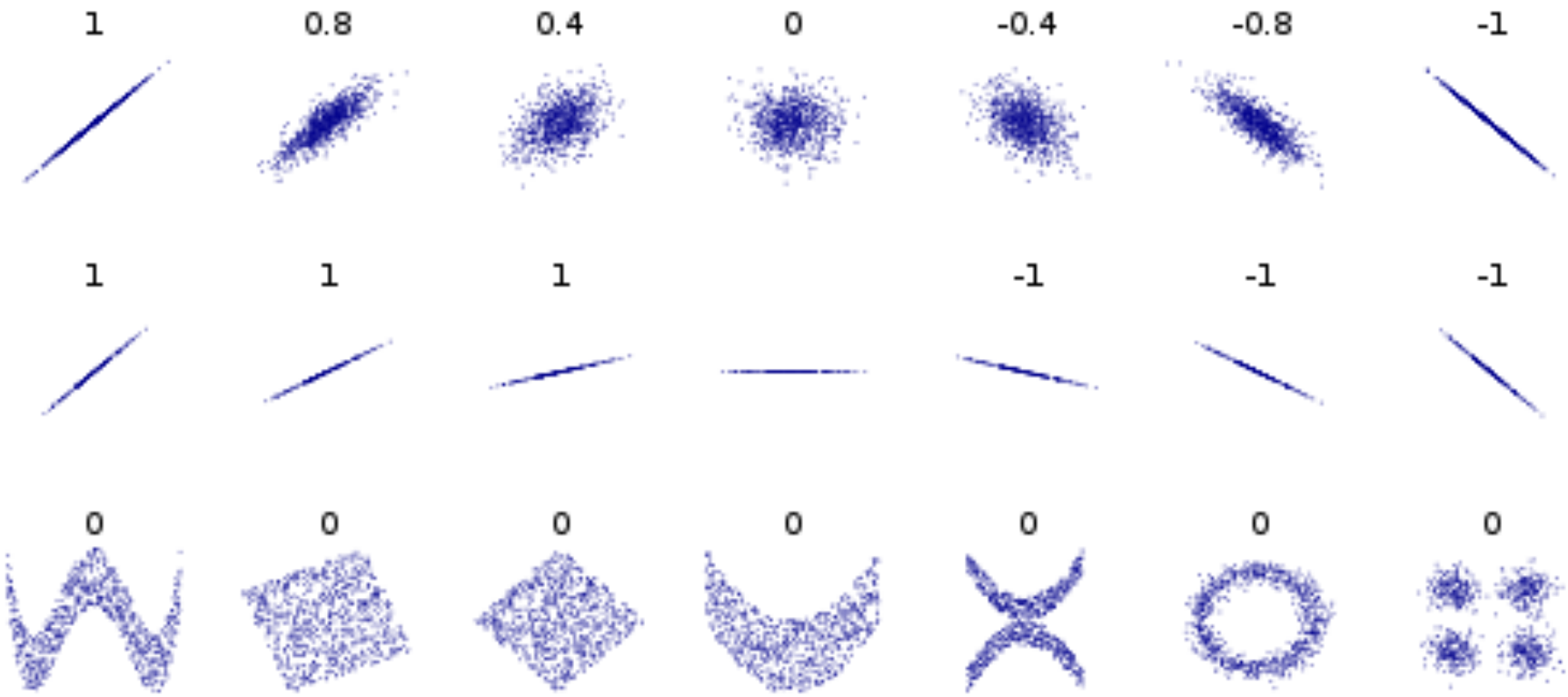
http://www.theatlantic.com/national/archive/2011/06/americas-top-cities-for-bike-commuting-happier-too/240265/

Pearson's Correlation

- Correlation coefficient $\boldsymbol{\rho}$ (or r) in [-1, 1], where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation



Some levels of correlation



Pearson's Correlation

- Suitable for linear relations (assumption)
- Outliers is an issue
- Might need to test for significance, use a t-test
- A rule of thumb, r =

  +.70 or higher Very strong positive relationship
  +.40 to +.69 Strong positive relationship
  +.30 to +.39 Moderate positive relationship
  +.20 to +.29 weak positive relationship
  +.01 to +.19 No or negligible relationship
  -.01 to -.19 No or negligible relationship
  -.20 to -.29 weak negative relationship
  -.30 to -.39 Moderate negative relationship
  -.40 to -.69 Strong negative relationship
  -.70 or higher Very strong negative relationship

Spearman's rank correlation

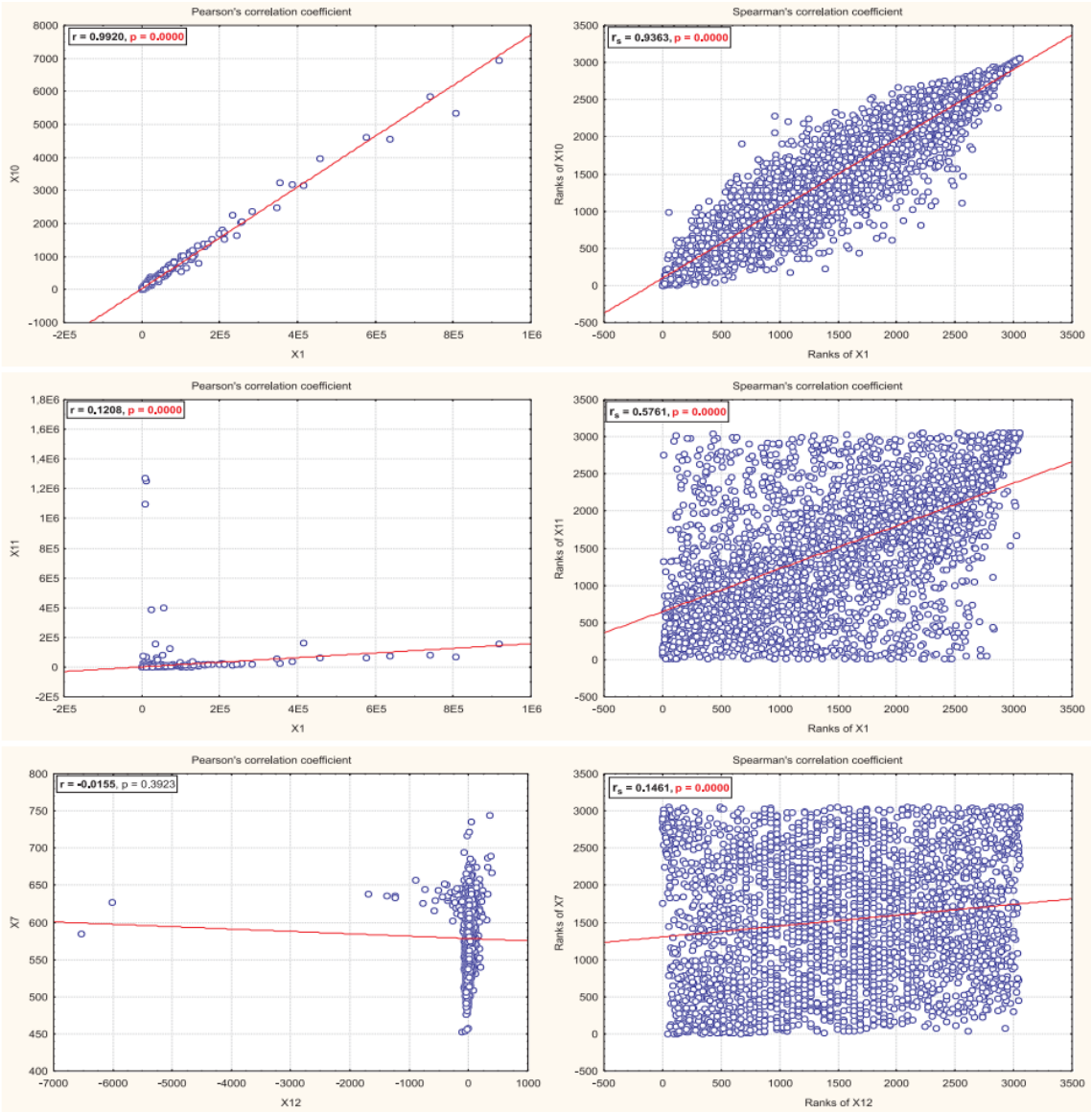
- Nonparametric measure (more resistant to outliers)
- Rank based, captures **monotonic** relations
- Linearity not assumed!
- Pearson corr. between **ranked variables**
- Works with categorical data

Distance in ranks

$$d_i = x_i - y_i$$

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Pearson vs. Spearman



Hauke, Jan, and Tomasz Kossowski. "Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data." *Quaestiones Geographicae* (2011)

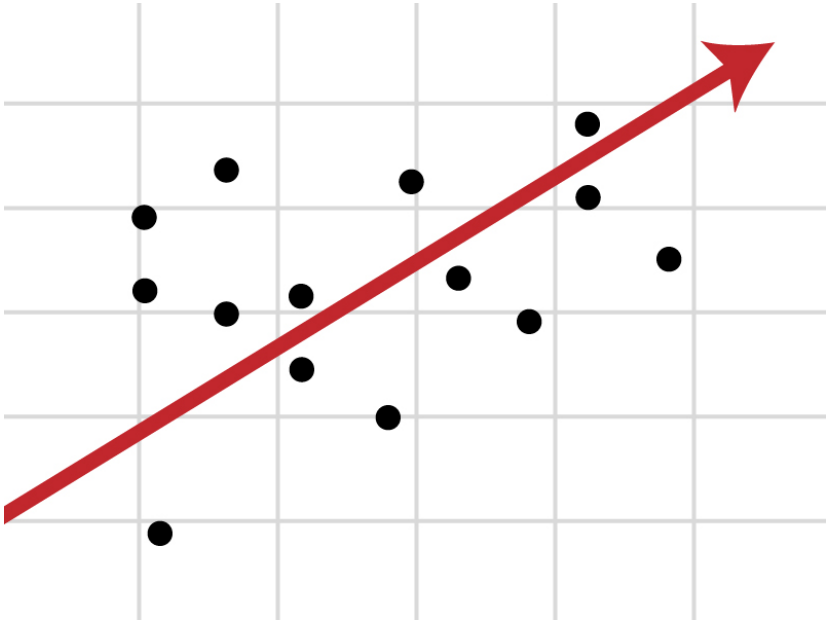
Modelling: some concepts

- Dependent variable
  - Output or effect
  - Phenomena being tested
- Independent variable
  - a.k.a., predictor, regressor, controlled
  - Input or **cause**

$$y=f(x)$$

Regression analysis

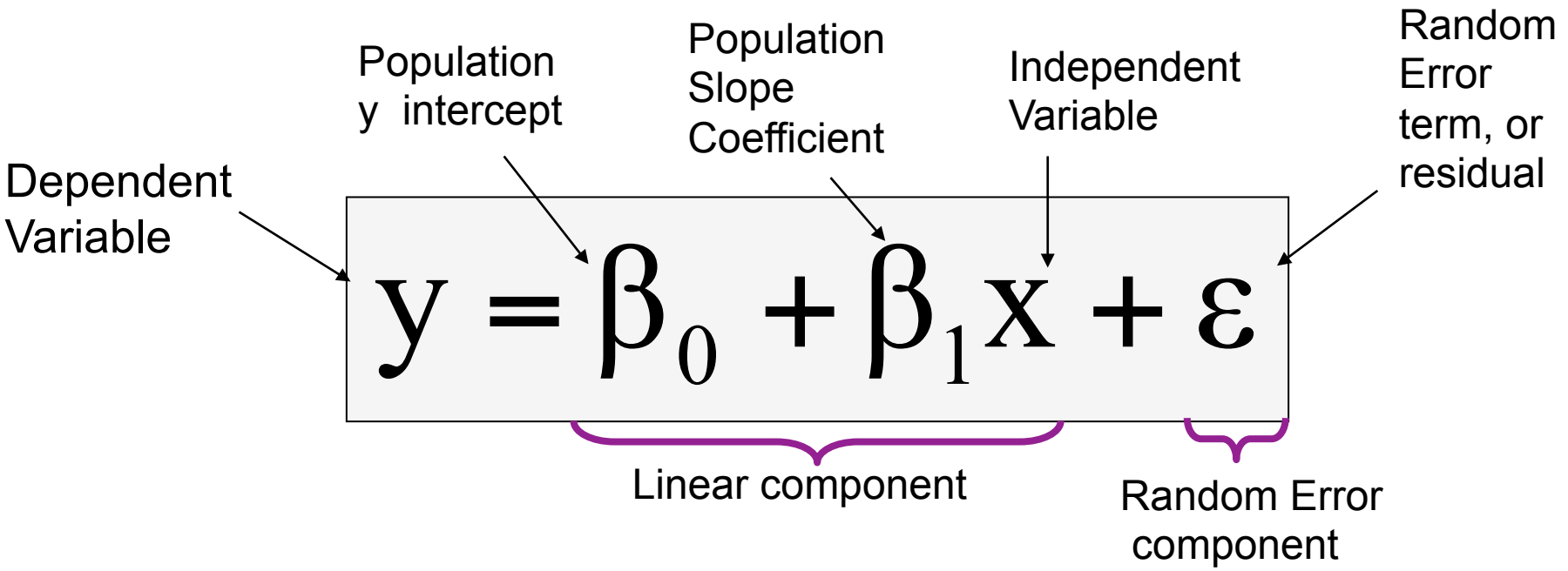
- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable



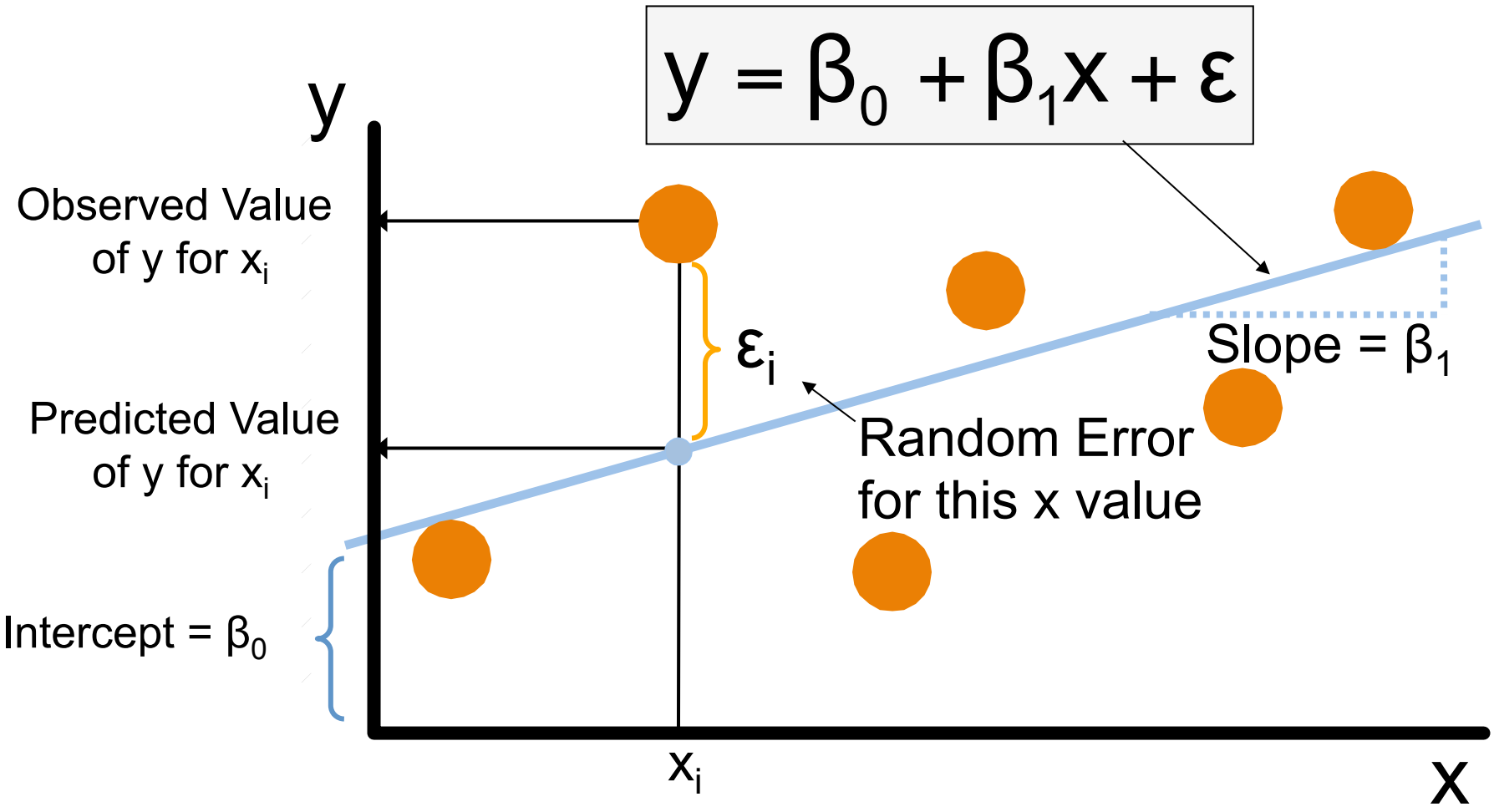
Simple linear regression

- Only one **independent variable**, *x*
- Relationship between *x* and *y* is described by a **linear function**
- Changes in *y* are assumed to be **caused by** changes in *x*

Linear regression model

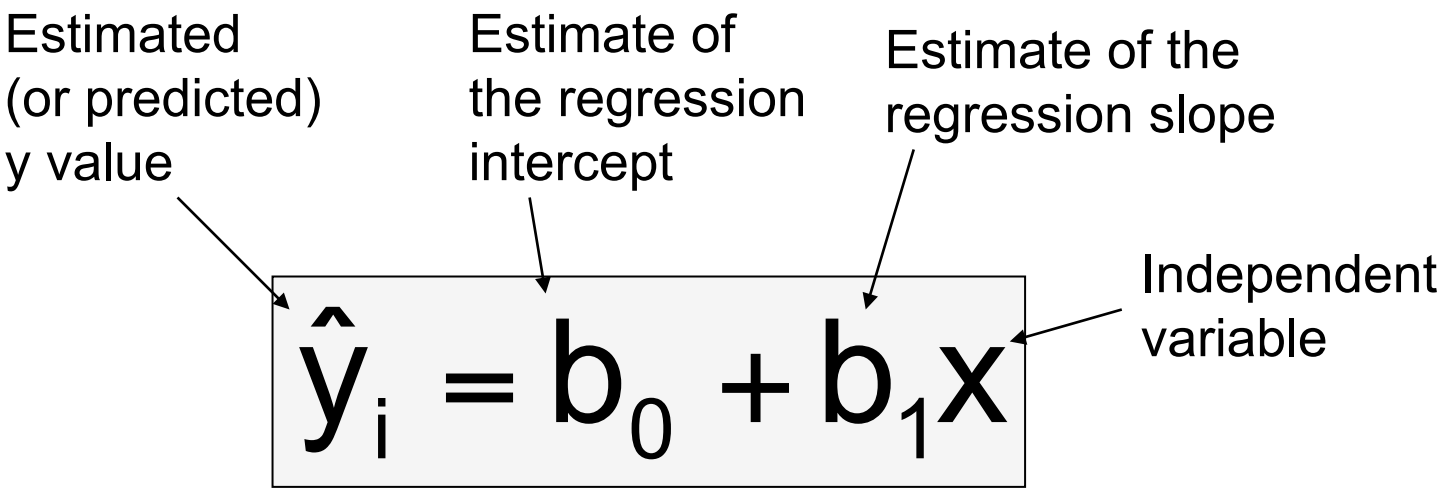


Linear regression model



Estimated Regression Model

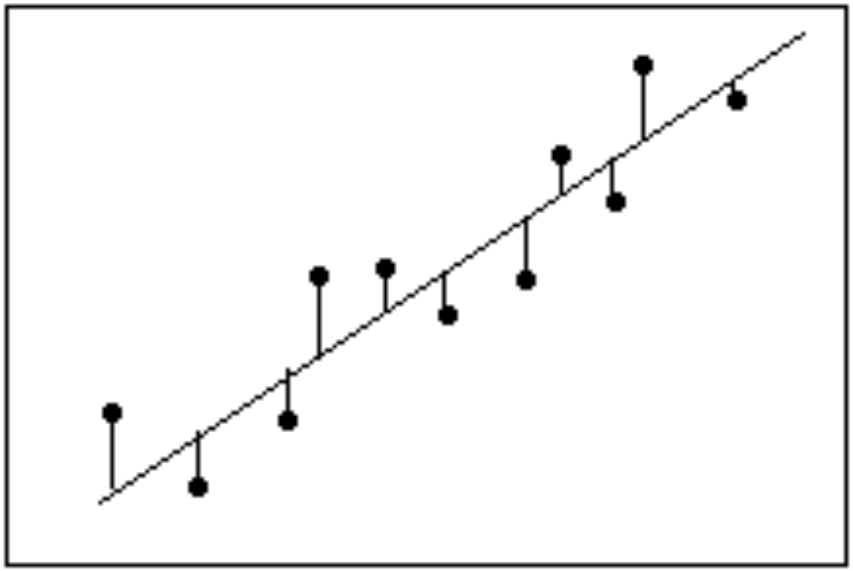
The sample regression line provides an **estimate** of the population regression line



The individual random error terms e_i have a mean of zero

Ordinary least squares (OLS)

- Method for estimating the unknown parameters in a linear regression model.
- Trying to find the best fit by minimizing modelling error



Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that **minimize the sum of the squared residuals**

$$\sum e^2 = \sum (y - \hat{y})^2$$

$$= \sum (y - (b_0 + b_1 x))^2$$

Interpretation

$$\hat{y}_i = b_0 + b_1 x$$

- b_0 is the estimated average value of y when the value of x is zero
- b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Example: Housing Prices

A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

A random sample of 10 houses is selected

- Dependent variable (y) = house price in $1000s
- Independent variable (x) = square feet

Sample Data for House Price Model

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Resulting regression model -- looking at b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977 \text{ x (square feet)}$$

- b_0 is the estimated average value of Y when the value of X is zero (if x = 0 is in the range of observed x values)
  - Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Resulting regression model -- looking at b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,\text{x (square feet)}$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, b_1 = .10977 tells us that the average value of a house increases by .10977($1000) = **$109.77**, on average, for each **additional one square foot** of size

Evaluating the model – looking at variations

$$SST = SSE + SSR$$

Total sum of Squares	Sum of Squares Error	Sum of Squares Regression
$SST = \sum (y - \bar{y})^2$	$SSE = \sum (y - \hat{y})^2$	$SSR = \sum (\hat{y} - \bar{y})^2$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

Explained and Unexplained Variation

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean y
- SSE = error sum of squares
  - Variation attributable to **factors other than** the relationship between x and y
- SSR = regression sum of squares
  - **Explained variation** attributable to the relationship between x and y

$$SST = \sum (y - \bar{y})^2 \qquad SSE = \sum (y - \hat{y})^2 \qquad SSR = \sum (\hat{y} - \bar{y})^2$$

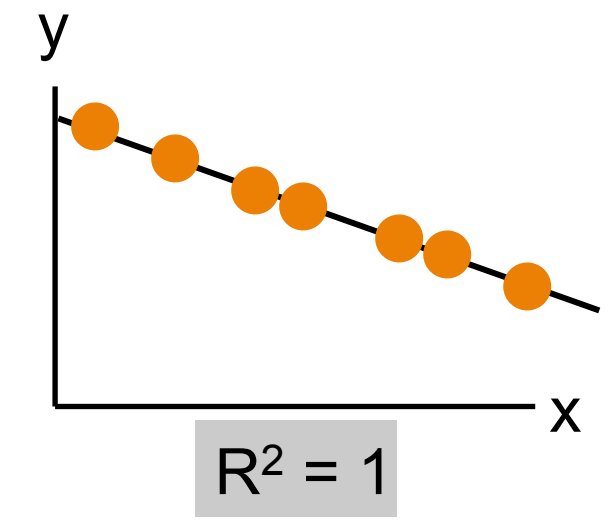
$$\text{SST} = \text{SSE} + \text{SSR}$$

Coefficient of Determination, R^2

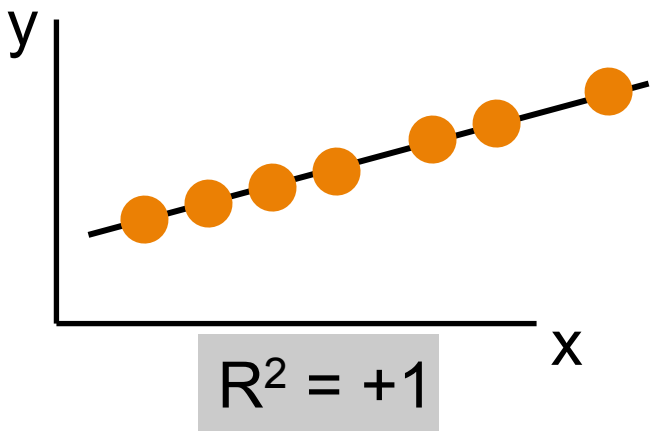
- Indicates how well data fits regression model
- portion of the total variation in the dependent variable that is explained by variation in the independent variable

$$R^2 = \frac{SSR}{SST} = \frac{sum \text{ of squares explained by regression}}{total \text{ sum of squares}}$$

Approximate R^2 Values



$R^2 = 1$



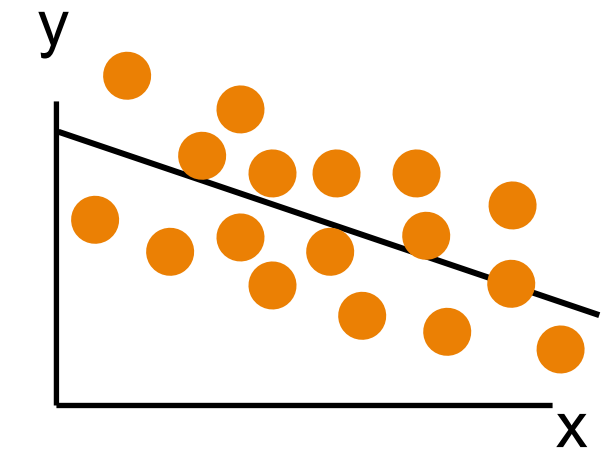
$R^2 = +1$

$R^2 = 1$

Perfect linear relationship between x and y:

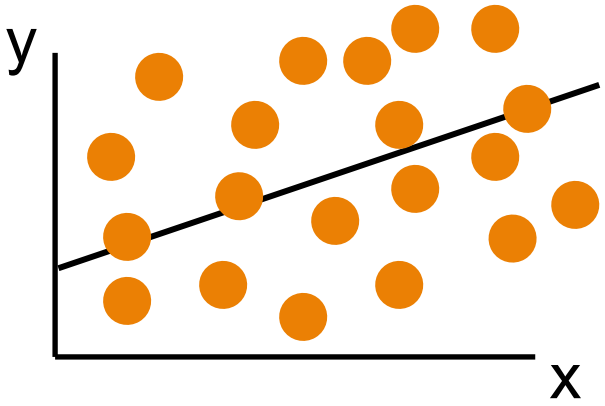
100% of the variation in y is explained by variation in x

Approximate R^2 Values



$0 < R^2 < 1$

Weaker linear relationship between x and y:



Some but not all of the variation in y is explained by variation in x

What if more variables?

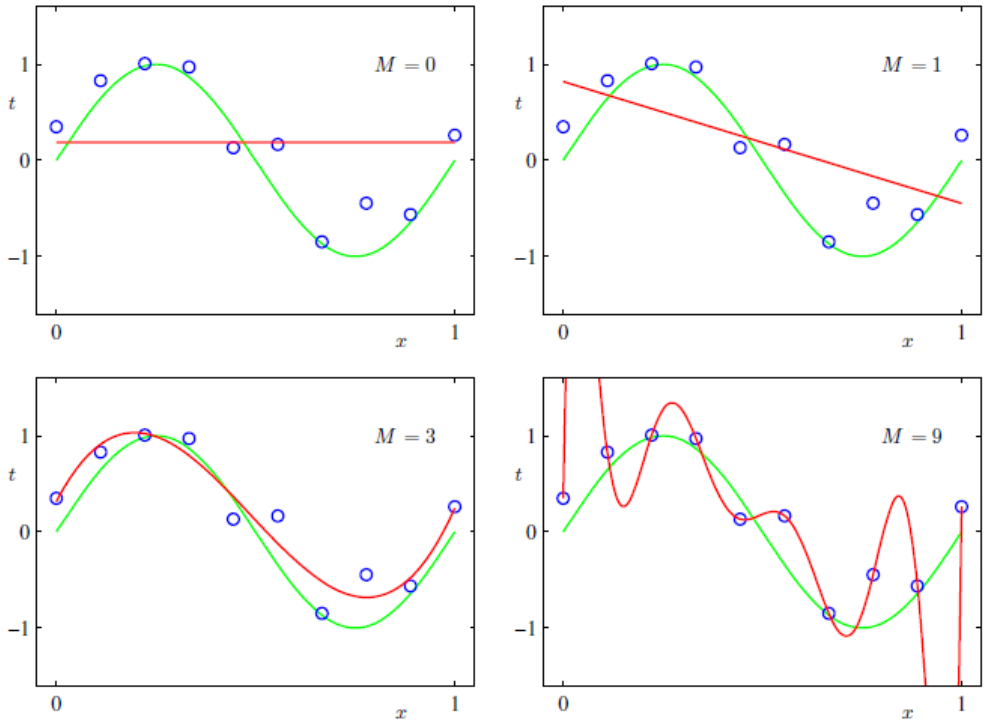
- **Multiple linear regression**
  - If you want to relate several independent variables to a dependent variable
  - e.g., House price ⇔ house size, house age

$$Y = a + b_1X_1 + b_2X_2 + ... + b_pX_p$$

Polynomial Regression

- Model the relation with *n*th degree polynomial
- Be careful not to over-fit, simple can be better (or not)

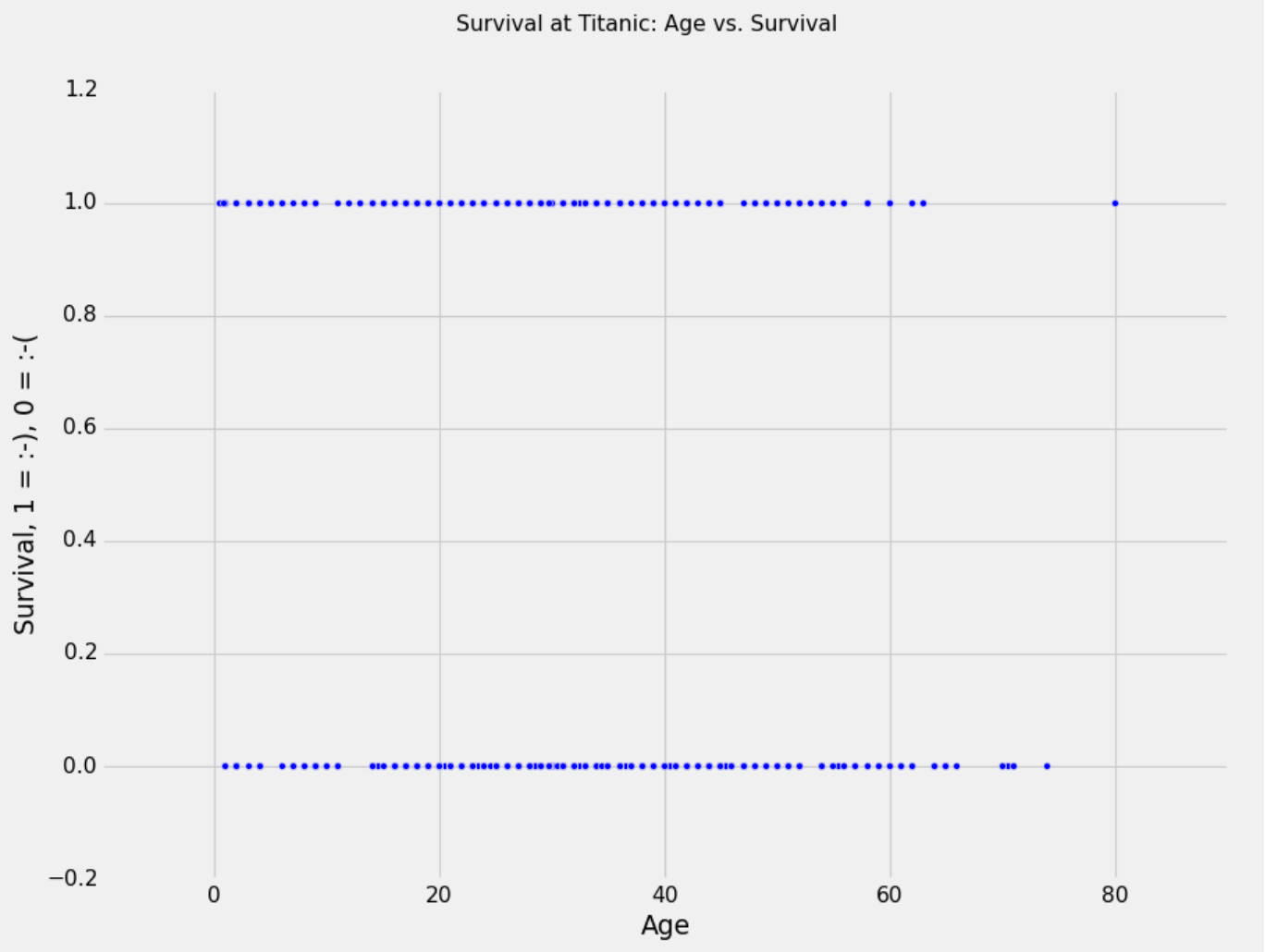
$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \varepsilon.$$



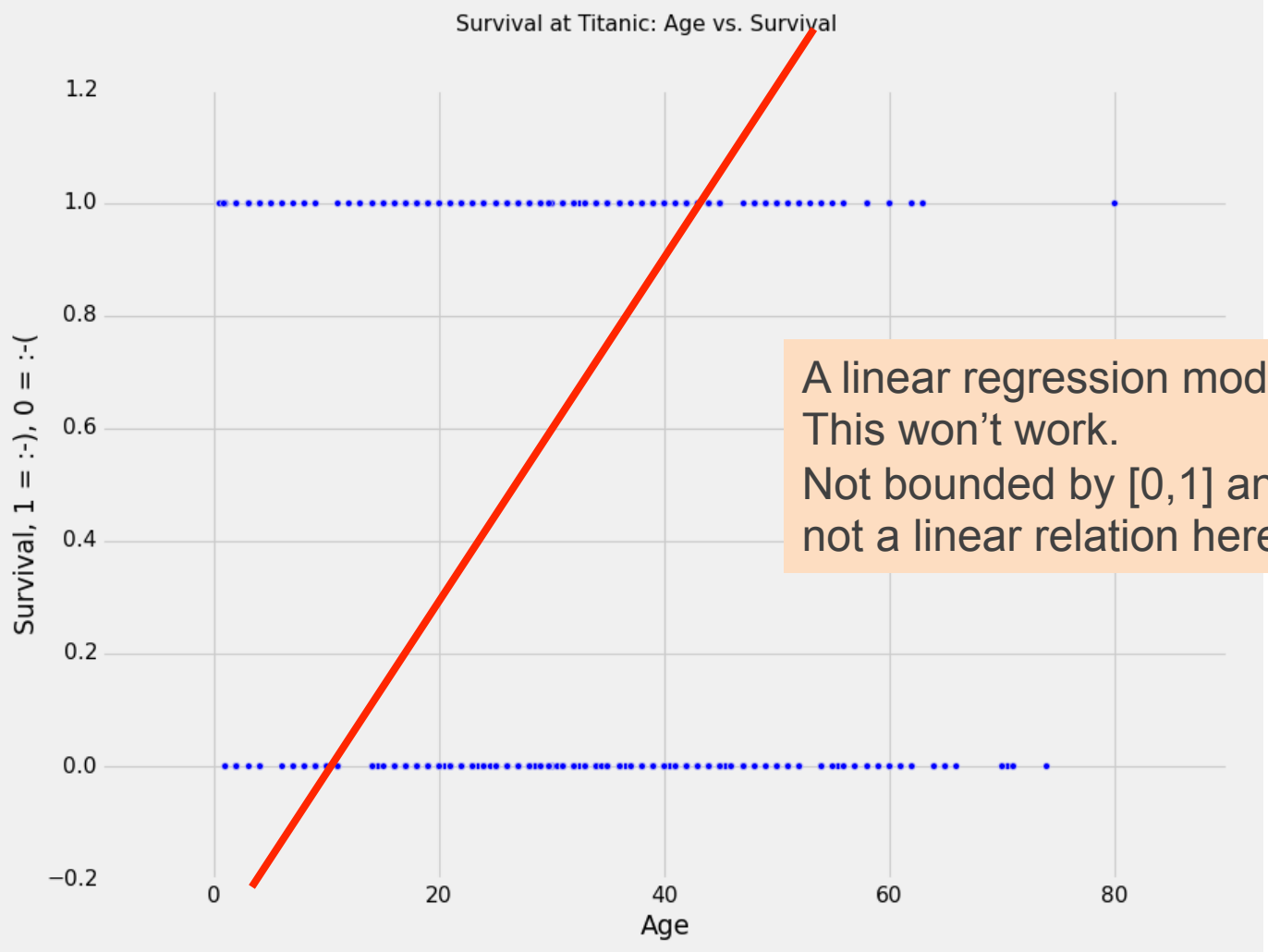
Logistic regression

- Models the relationship between a set of variables and a variable with a set of **limited outcome** (e.g., binary)
  - e.g., trying to model whether one will have a particular disease or not given age, gender, blood readings, etc.
- Simply, we have a binary output variable Y, and we want to model the conditional probability $Pr(Y = 1|X = x)$ as a function of x, i.e., we want a model of what values of x lead to an outcome of 1.

Ex, Titanic Survival



Ex, Titanic Survival: want to estimate survival



A linear regression model?
This won't work.
Not bounded by [0,1] and not a linear relation here.

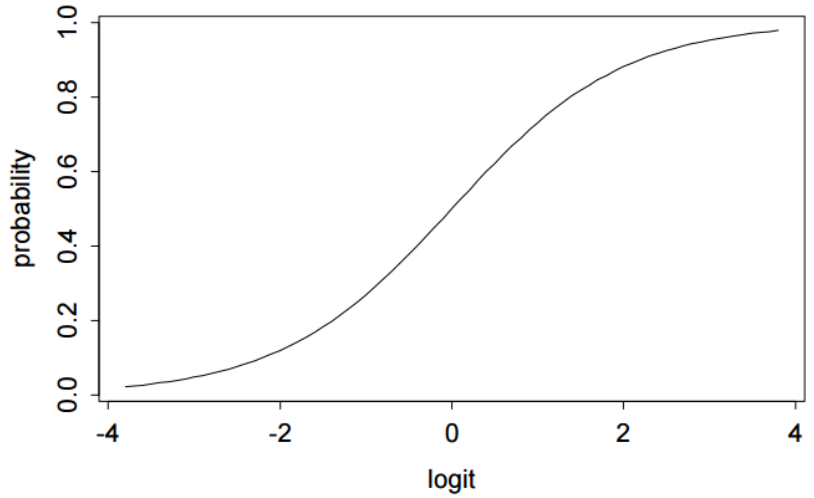
Logistic regression

- Estimate with a logistic function instea$\text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i}$

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta$$

- And solve for p(x):

$$p(x; b, w) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}} = \frac{1}{1 + e^{-(\beta_0 + x \cdot \beta)}}$$



- The logit of the probability $\pi\downarrow i$, rather than the probability itself, follows a linear model
- You end up with a classifier:

Y = 1 when p ≥ 0.5 and Y = 0 when p < 0.5

Other types of regression (that we won't cover)

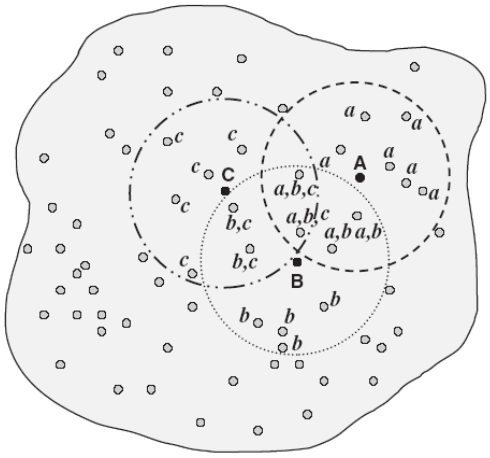
- **Multinomial Logistic Regression** – if you have more than one category in your dependent var.
- **Ridge Regression** – if you have multi-colinearity within your data, i.e., independent variables are not that independent
- **Lasso Regression –** similar to Ridge but this has the nice property that some variables in your model get a 0 weighting, easier to interpret.

Other types of dependence analysis

- Time-series data
  - Data that has a temporal aspect
  - **Cross-correlation**: Find the correlation between two time series as a **function of time difference** between them

[Chatfield, Chris. *The analysis of time series: an introduction*. CRC press, 2013.]

- Spatial data
  - Relations might vary geographically
  - Spatial auto-correlation
  - Geographically weighted regression



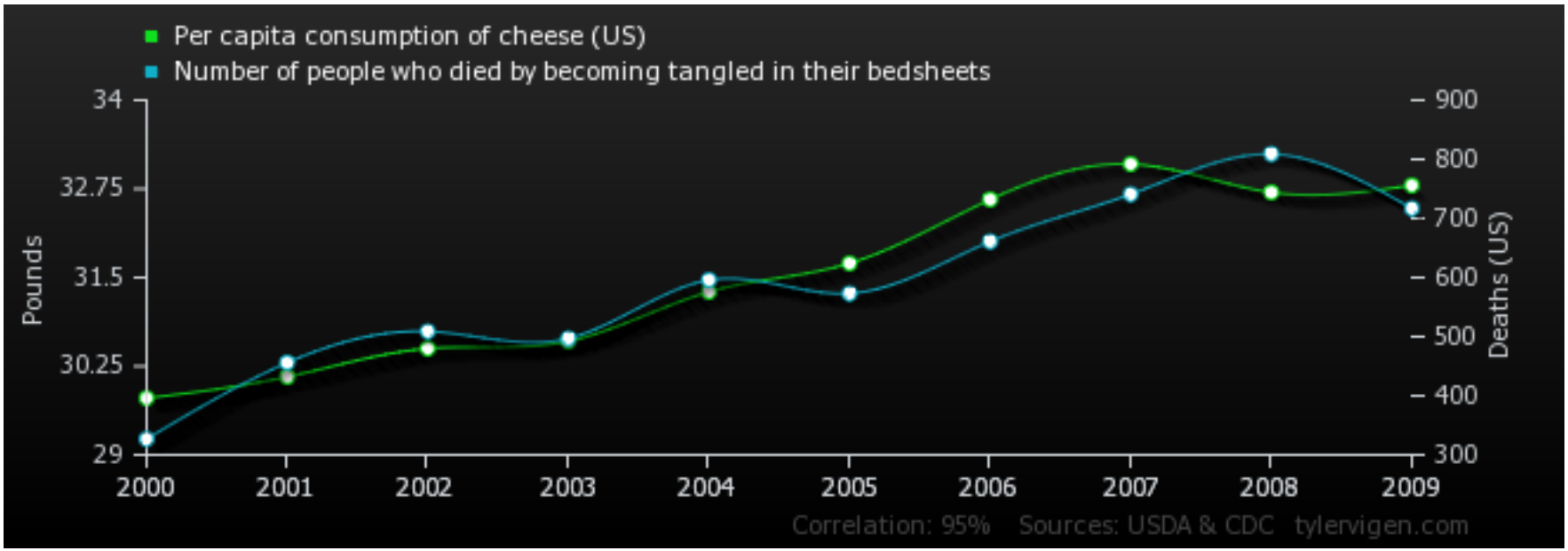
[Lloyd, Christopher D. *Local models for spatial analysis*. CRC Press, 2010.]

How to choose which method?

- Not always a clear answer
- Use visualise to understand your data first
- Try several alternatives
  - Look at errors, residuals

Causality vs. correlation

- Correlation do not always indicate causality



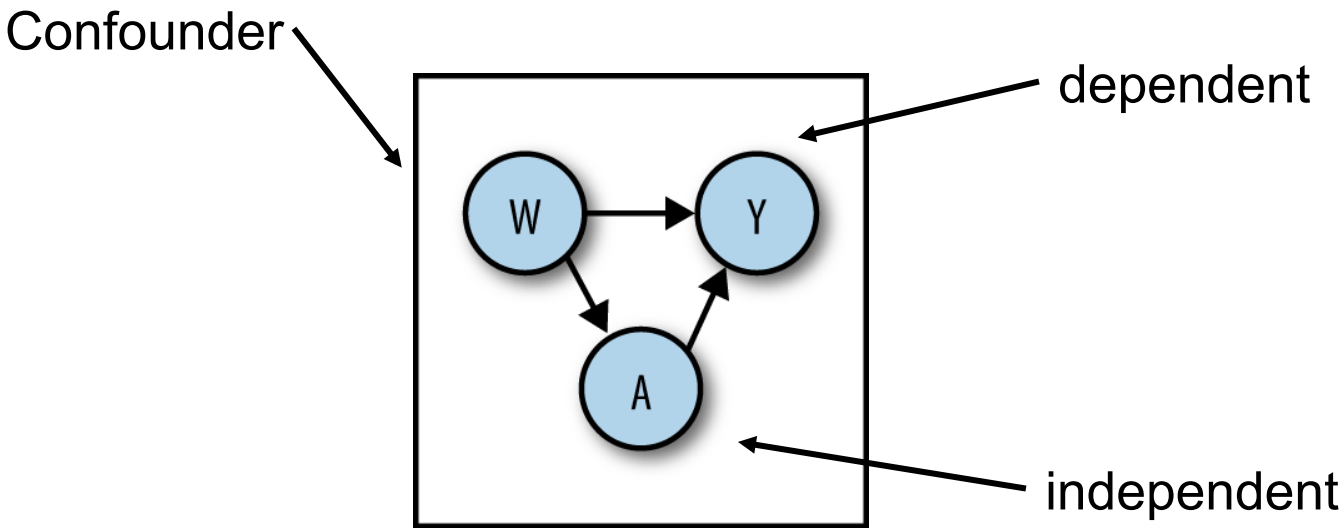
http://www.tylervigen.com/

spurious correlations

Discover a new correlation

Confounders

- Variable that correlates with both the dependent variable and the independent variable
- Might generate **spurious relations**
- Ex: ice-cream consumption vs. death by drowning



OK Cupid Example

- Researched-focused online dating company
- “beautiful” is likely a confounder

Figure 11-2. OK Cupid’s attempt to demonstrate that using the word “beautiful” in an email hurts your chances of getting a response

from *Doing Data Science*, p.277

http://blog.okcupid.com/index.php/online-dating-advice-exactly-what-to-say-in-a-first-message/

Simpson's Paradox

A condition where a trend appears in different groups of data but disappears or reverses when these groups are combined

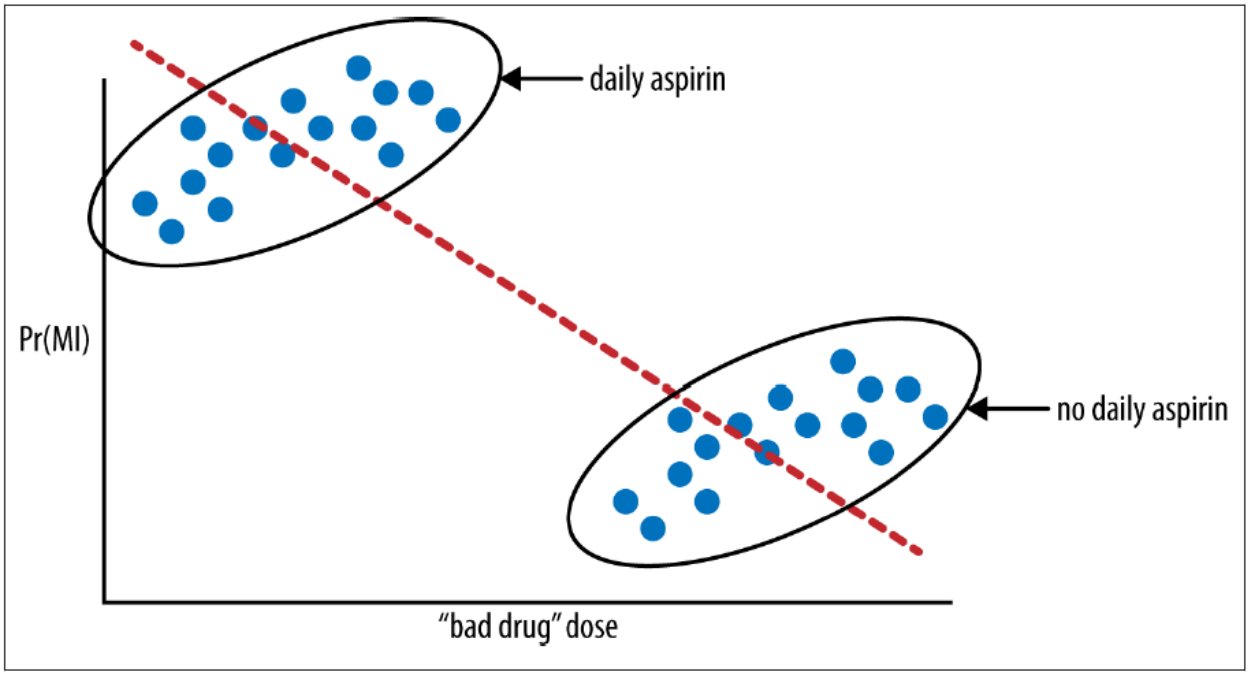


Figure 11-4. Probability of having a heart attack as a function of the size of the dose of a bad drug and whether or not the patient also took aspirin

Always check group characteristics and within-group variations and compare !!

Making causal statements

- Ask correct (sensible) questions first
  - Form: What is the effect of *x* on *y*?
- Best practice: “Randomized Clinical Trials”
  - randomly allocate one or other of the different phenomena being investigated
  - Hard to achieve in data science applications
- Often: “Observational Studies”
  - You get the data rather than curate it
  - Try avoiding biases

Labs & Next Week

- Practical exercises on:
  - Correlation
  - Linear regression
  - Pointers to Logistic Regression
- Next time (in 2 weeks):
  - Moving into high dimensional worlds
  - How to deal with many columns??
    - Feature selection
    - Dimension reduction
    - …