

Tag 1: Der Daten-Lifecycle und: Woher kommen Daten?

Session 2: Datengenerierung und was dabei schief gehen kann

Simon Munzert
Hertie School

1. Schlechte Stichproben: Repräsentativität liegt im Auge des Betrachters
2. Schlechte Analytik: Signifikanz ist nicht alles, was zählt
3. Schlechte Schlussfolgerung: Korrelation impliziert keine Kausalität

Schlechte Stichproben: Repräsentativität liegt im Auge des Betrachters

IFLSCIENCE! ≡

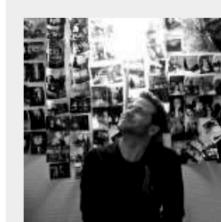
Survey Finds Most Americans Think That They Have Above Average Intelligence

DON'T SCOFF: YOU'RE NOT INVULNERABLE TO SELF-DELUSION. PR IMAGE FACTORY/SHUTTERSTOCK

A new US-based nationally representative survey has found that 65 percent of respondents (70 percent in men, 60 percent in women) agree with this rather telling statement: "I am more intelligent than the average person." Hopefully this doesn't require a rudimentary lesson in statistics to explain why this simply isn't possible.

Now, this is amusing, but let's not all pile in on the [American public](#). While this [PLOS ONE](#) systematic study is certainly noteworthy, it's not for the finding that many people overestimate their intellectual capabilities.

Instead, it's important because similar research conducted in the US half a century earlier found [much the same thing](#). Although the researchers caution about generalizing their findings, it's a good bet the same pattern can be found in other countries around the world too.



By Robin Andrews

31 JUL 2018, 14:55

Source [Robin Andrews, IFLScience](#)

80 percent of EU citizens want to scrap daylight savings: report

Some 4.6M EU citizens participated in European Commission survey.

By MAXIME SCHLEE | 8/29/18, 2:32 PM CET | Updated 8/29/18, 4:11 PM CET

A vast majority of EU citizens want to scrap daylight savings rules and stop changing their clocks twice a year, German media reported Wednesday.

Some 80 percent of respondents of a public consultation launched by the European Commission last month said they would support abolishing daylight savings, according to [Westfalenpost](#).

The Commission launched the consultation as part of its review of the EU [summer time directive](#). It has not provided details on its outcome, but has [said](#) some 4.6 million EU citizens participated.

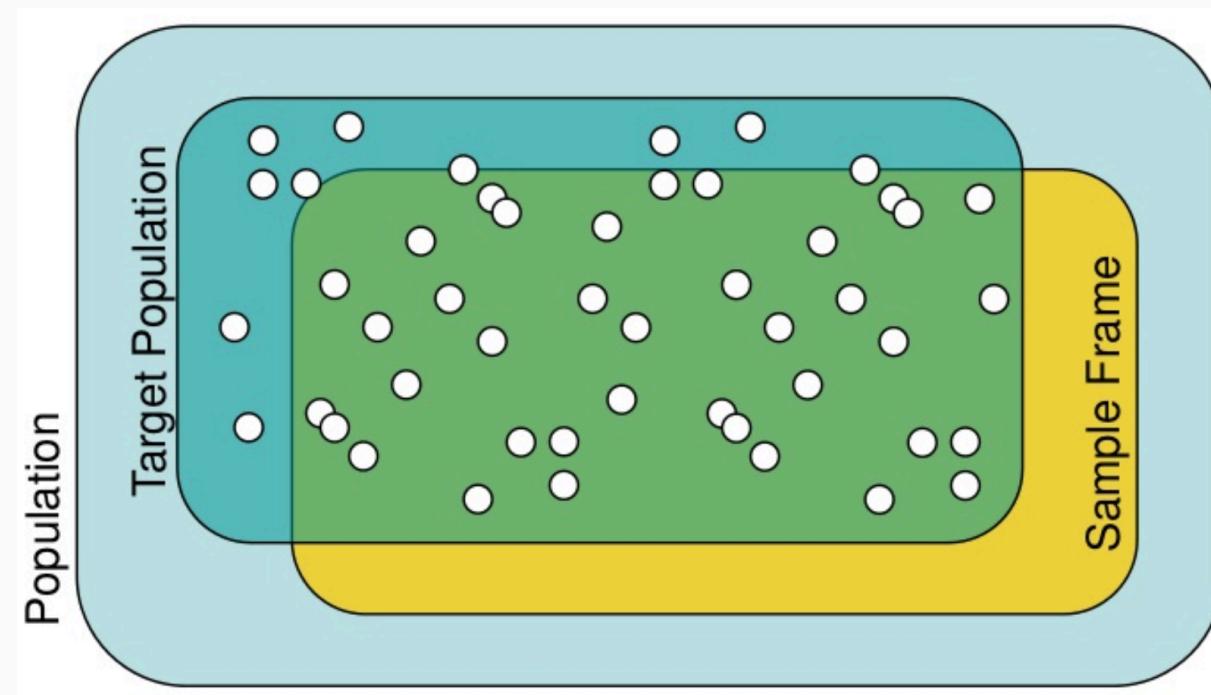
Source Maxime Schlee, Politico

Eine volkstümliche Definition von Repräsentativität

Eine Stichprobe (oder Daten im Allgemeinen) ist „repräsentativ“, wenn **die aus der Stichprobe gezogenen Schlüsse verallgemeinert werden können** für die Grundgesamtheit von Interesse.

Eine formalere Definition

Eine Stichprobe ist repräsentativ, wenn sie so gezogen wird, dass sie **statistisch nicht von der interessierenden Grundgesamtheit unterscheidbar** ist.



Warum „Repräsentativität“ ein problematischer Begriff ist

1. Ob eine Stichprobe repräsentativ ist, hängt von Ihrem Interesse ab.
2. Man kann eine Stichprobe nicht a priori als „repräsentativ“ bezeichnen.
3. Die Beurteilung der Repräsentativität einer Stichprobe erfordert starke Annahmen über Ihr Wissen über die Grundgesamtheit und Ihre Messungen der Merkmale, die „repräsentativ“ sein sollten.

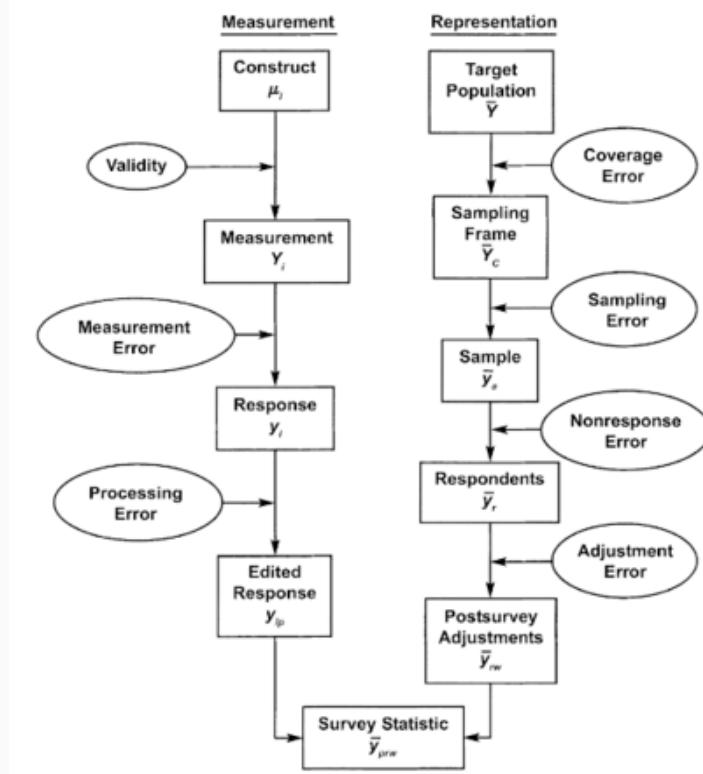
Inferenz in der Umfrageforschung

- Etwas über die Verteilung von Merkmalen in einer Grundgesamtheit erfahren
- Sammeln von Informationen aus einer Teilmenge der Grundgesamtheit

Zwei Arten von Fehlern

- **Messfehler:** was man misst, ist nicht das, was man messen will
- Fehler der **Repräsentation:** Die Gruppe, die Sie beobachten, ist nicht verallgemeinerbar auf die interessierende Population

Gesamtfehler der Umfrage Framework



Source Groves et al. 2009, Survey Methodology

Überrepräsentation und falsche Angaben in Wahlumfragen

- Die Zahlen aus Umfragen nach der Wahl überschätzen die Wahlbeteiligung oft erheblich.
- Zwei unterschiedliche Phänomene sind für diese Diskrepanz verantwortlich:
 1. Überrepräsentation der tatsächlichen Wähler
 2. Falsche Angaben zur Wahlbeteiligung durch Nichtwähler unter den Umfrageteilnehmern.
- Studien zur Validierung der Wahlbeteiligung helfen, das Problem auf individueller Ebene zu identifizieren.
- Eine Verzerrung der Wahlbeteiligung kann sich auch auf Analysen nachgelagerter Variablen (z.B. Wahlverhalten) auswirken.

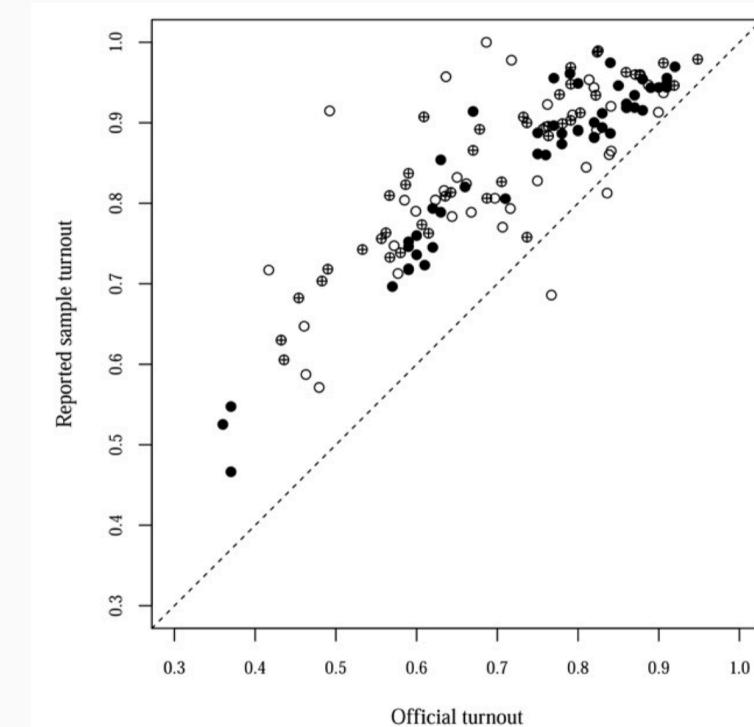


Fig. 1. Reported sample turnout rates from 130 postelection surveys versus official turnout. Data are taken from Modules 1–3 of the *Comparative Study of Electoral Systems* (CSES), and from a collection of election surveys for which vote validation studies (VVS) are available. For more detailed information, see the [Appendix](#) to this paper.

Source [Selb and Munzert 2013, Electoral Studies](#)

Was bedeutet das für Sie?

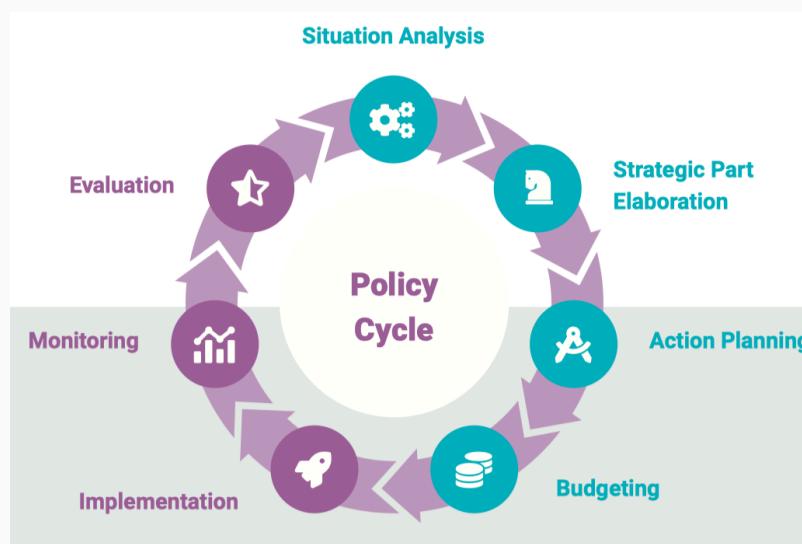
- Nehmen Sie die angegebene „Repräsentativität“ nicht für bare Münze.
- Der Stichprobenumfang allein garantiert keine Repräsentativität.
- Lassen Sie sich nicht von „großen Datenmengen“ täuschen (sie sind standardmäßig nicht repräsentativ).
- Lassen Sie sich nicht von „Zufallsstichproben“ täuschen (sie sind standardmäßig nicht repräsentativ).
- Wahrscheinlichkeitsstichproben sind kein Allheilmittel, da Menschen immer noch eine Selbstselektion in/aus Stichproben vornehmen.
- Schlechte Stichproben sind nicht auf Erhebungen beschränkt (denken Sie z.B. an Daten aus sozialen Medien, die Auswahl von Fällen für eine medizinische Studie oder die Auswahl von Ländern

Suchen Sie stattdessen nach den folgenden Punkten:

1. **Transparenz** über das Stichprobenverfahren.
2. **Einschätzung** der Repräsentativität der Stichprobe.
3. **Validierung** der Stichprobe anhand externer Benchmarks.
4. **Gesunder Menschenverstand** (ist es sinnvoll, den Repräsentanten anzurufen?)

Eckpunkte für die Diskussion

1. Zu welchem Zeitpunkt des Politikzyklus könnte welche Art der Konsultation sinnvoll sein?
2. Welche Vor- und Nachteile haben die verschiedenen Konsultationsarten für die Beobachtung und Bewertung?



Source Policy Planning, Monitoring and Evaluation Handbook, Government of Georgia

3. Consultation Methods.....
 - 3.1. Physical Consultations
 - (1) Surveys/Polls.....
 - (2) In-Depth Interviews.....
 - (3) Focus Group.....
 - (4) Public Meetings and Workshops.....
 - (5) Conferences, Forums.....
 - (6) Leaflet – Information Brochure.....
 - (7) Consultation Days, Exhibitions and Roadshow
- 3.2. Online Methods.....
 - (1) Online Consultation – Public Commenting
 - (2) Social Media.....
 - (3) Internet Forum, Commenting
 - (4) Online Polls and Surveys

Source Annex 11: Guideline for Public Consultations, Government of Georgia

Schlechte Analytik: Signifikanz ist nicht alles, was zählt

Statistische Signifikanz in der Praxis

- Konventionell sollten Fehler vom Typ I um jeden Preis vermieden werden.
- Ein Ergebnis gilt als statistisch signifikant, wenn es sehr unwahrscheinlich ist, dass es unter einer wahren Nullhypothese aufgetreten wäre.
- Ein Signifikanzniveau α gibt die Wahrscheinlichkeit eines Fehlers vom Typ I an. Üblicherweise wird es auf 5% festgelegt.

Gewisse Probleme

- Nur weil ein Effekt signifikant ist, heißt das nicht, dass er substantiell bedeutsam (groß) ist.
- Es gibt einen Anreiz für Forscher, statistisch signifikante Ergebnisse zu produzieren → "publication bias"
- Statistische Signifikanz ist (auch) eine Funktion der Stichprobengröße. Es ist **trivial, mit großen Daten signifikante Ergebnisse zu erzielen.**
- Leider ist es auch oft **trivial, mit kleinen Daten signifikante Ergebnisse zu erzielen**, wenn man in Bezug auf seine Hypothesen flexibel ist.

Lassen Sie sich nicht von übertriebenen Ausdrücken der Signifikanz

Theorie School

Glauben.

„Die folgende Liste stammt aus begutachteten Zeitschriftenartikeln, in denen (a) die Autoren sich selbst den Schwellenwert von 0,05 für die Signifikanz gesetzt haben, (b) diesen Schwellenwert für p nicht erreicht haben und (c) ihn so beschrieben haben, dass er interessanter erscheint.“ [Matthew Hankins, Probable Error](#)

(knapp) nicht statistisch signifikant ($p=0,052$), ein kaum nachweisbarer statistisch signifikanter Unterschied ($p=0,073$), ein grenzwertig signifikanter Trend ($p=0,09$), ein gewisser Trend zur Signifikanz ($p=0,08$), eine klare Tendenz zur Signifikanz ($p=0,052$), ein klarer Trend ($p<0,09$), ein klarer, starker Trend ($p=0,09$), ..., sehr knapp an der Grenze der statistischen Signifikanz ($p=0,051$), sehr knapp an der Signifikanz vorbei ($p<0,06$), sehr knapp signifikant ($p=0,0656$), sehr knapp nicht signifikant ($p=0,10$), sehr knapp signifikant ($p<0,1$), praktisch signifikant ($p=0,059$), schwach signifikant ($p>0,10$), abgeschwächt signifikant ($p=0,06$), schwach nicht signifikant ($p=0,07$), schwach signifikant ($p=0,11$), schwach statistisch signifikant ($p=0,0557$), nahezu signifikant ($p=0,11$)

Das Problem

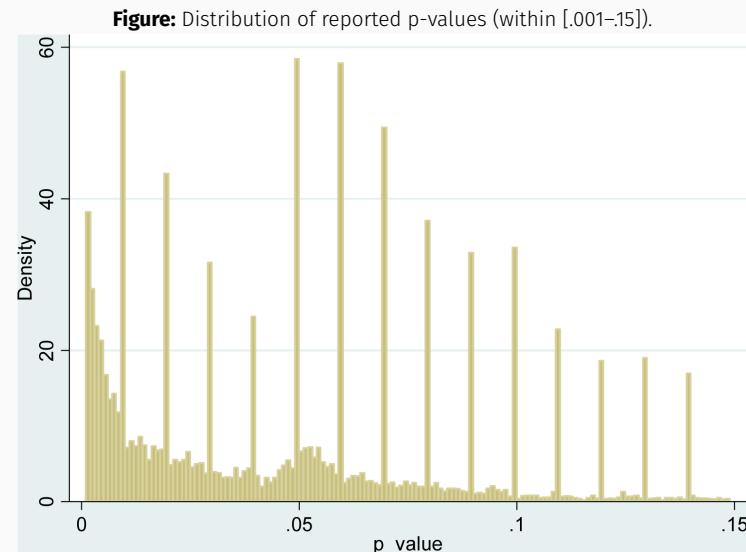
Die Dominanz der statistischen Signifikanz als Entscheidungskriterium bei wissenschaftlichen Veröffentlichungen macht die p-Werte zu einem wichtigen Zielkriterium bei der statistischen Analyse. Kleine p-Werte werden häufiger berichtet, als man erwarten würde!

Die Symptome

- **Daten-Fishing:** Testen vieler Hypothesen bis zur Signifikanz
- **p-Hacking:** Optimieren der Analyse (z.B. Hinzufügen/Entfernen von Kontrollen, Transformieren von Variablen, Ändern von Modellen) bis zur Signifikanz
- **HARKing:** Hypothesizing **a**fter the **r**esults are **k**nown
(Hypothesenbildung nachdem die Ergebnisse bekannt sind)

Das Gegenmittel?

Sonderausgabe in **The American Statistician, 2019**: "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$ "



"Der Datensatz besteht aus über 135'000 Datensätzen. Die Daten wurden mittels computergestützter Suche aus (...) fünf Journals of Experimental Psychology im Zeitraum von Januar 1996 bis März 2008 gewonnen."

Daten-Fishing and p-Hacking: Übung

Übung

- Schau dir <https://projects.fivethirtyeight.com/p-hacking/> an.
- Verbringen Sie fünf Minuten damit, sich zu wissenschaftlichem Ruhm zu hauen
- Mehr Hintergrundinformationen [hier](#).

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans

Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- Presidents
- Governors
- Senators
- Representatives

How do you want to measure economic performance?

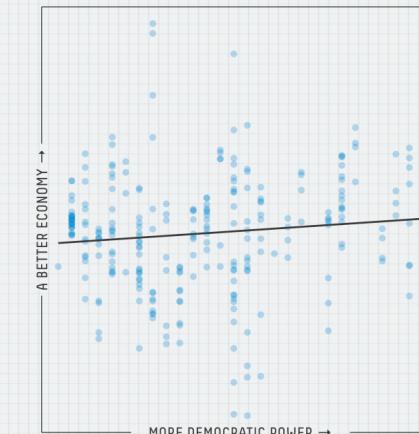
- Employment
- Inflation
- GDP
- Stock prices

Other options

- Factor in power
Weight more powerful positions more heavily
- Exclude recessions
Don't include economic recessions

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



Result: Almost

Your 0.06 p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Die Folgen der Vorregistrierung

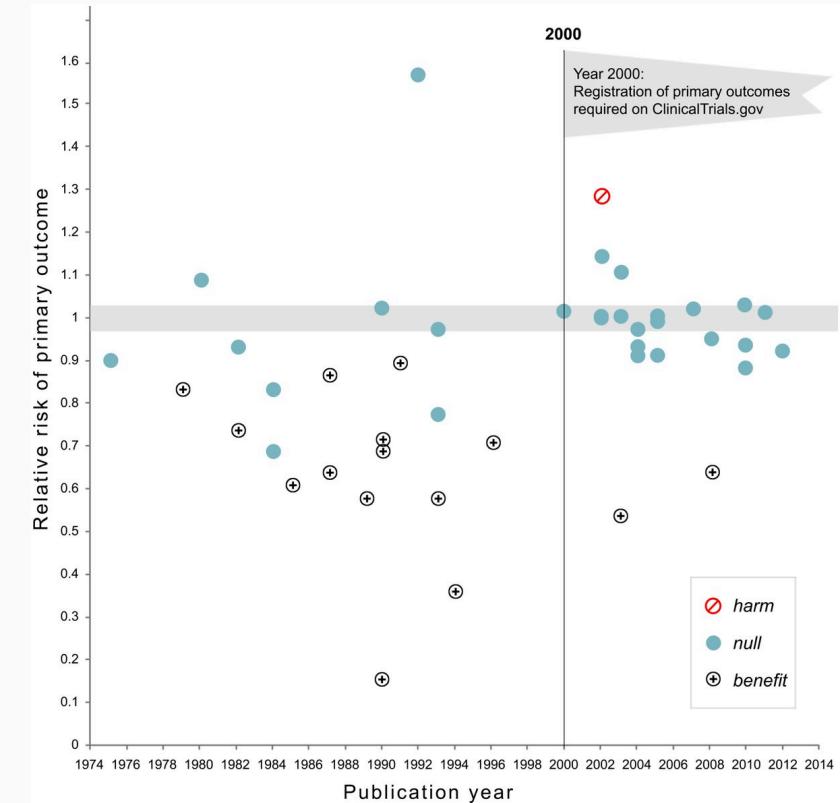
Die Idee

- Vorregistrierung bedeutet, dass Sie Ihre Studie (Hypothesen, Methoden, Analysen) registrieren (z.B. indem Sie sie online stellen), bevor sie durchgeführt wird
- In den letzten Jahren hat sich die Forschungspraxis stark verändert; siehe das [Open Science Framework \(OSF\) Registry](#) und das [American Economic Association \(AEA\) RCT Registry](#)

Warum das wichtig ist

"17 von 30 Studien (57\%), die vor dem Jahr 2000 veröffentlicht wurden, zeigten einen signifikanten Nutzen der Intervention für das primäre Ergebnis im Vergleich zu nur 2 der 25 (8\%) Studien, die nach 2000 veröffentlicht wurden." (siehe Grafik)

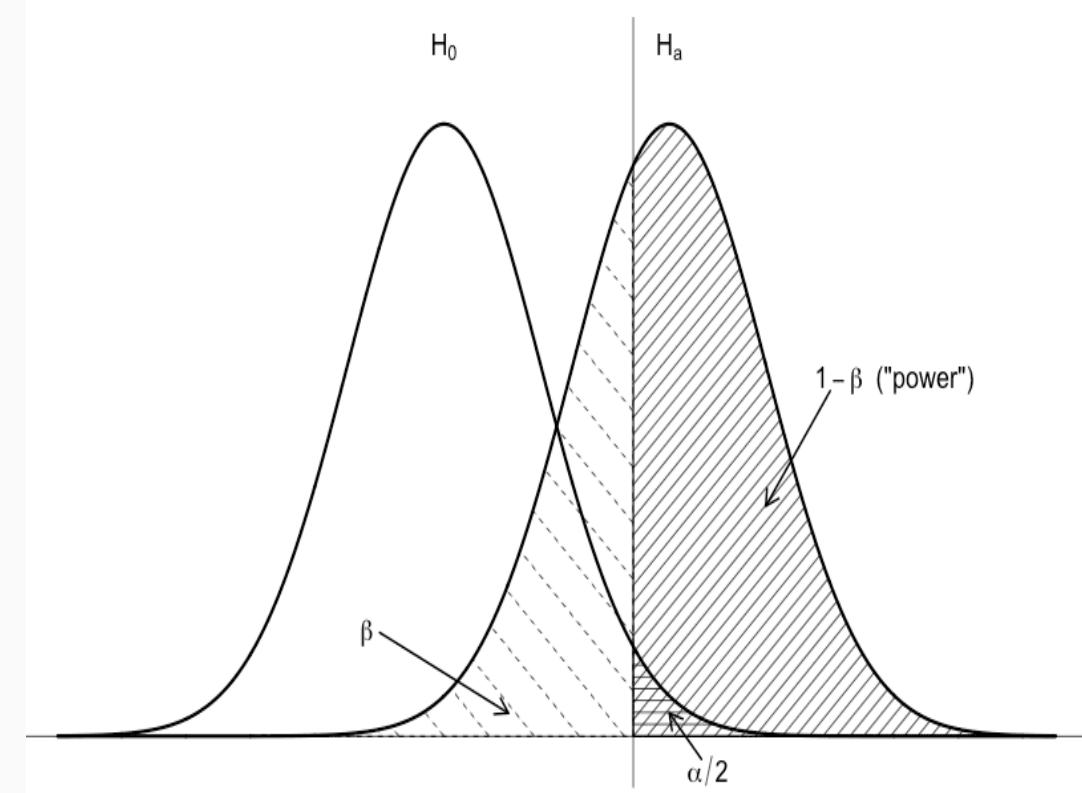
Figure: Relatives Risiko, einen Nutzen oder Schaden der Behandlung zu zeigen, nach Jahr der Veröffentlichung für große NHLBI-Studien zu Arzneimitteln und Nahrungsergänzungsmitteln.



Source Kaplan et al. 2015; PLOS ONE

Das Konzept

- Wir sind daran gewöhnt, uns vor falsch-positiven Ergebnissen (Signifikanztests!) zu schützen, aber auch falsch-negative Ergebnisse können schaden - insb., wenn die Durchführung einer Studie sehr kostspielig war
- Die statistische Aussagekraft ist die **Wahrscheinlichkeit, die Nullhypothese korrekt zurückzuweisen, wenn sie falsch ist.**
- Die Fähigkeit, Signal und Rauschen zu unterscheiden, wobei das Signal der interessierende Behandlungseffekt ist
- $P(\text{"Es gibt einen Effekt und ich sehe ihn"})$:
Aussagekraft = 1 - Fehler vom Typ II
- Je höher die statistische Aussagekraft eines Experiments ist, desto geringer ist die Wahrscheinlichkeit eines Fehlers vom Typ II.



Motivation

- Ist Ihre Stichprobe groß genug, um einen Effekt einer bestimmten Größe aufzudecken? Führen Sie eine Power-Analyse durch (idealerweise vor der Datenerhebung)!
- Bei der **Power-Analyse** wird die Wahrscheinlichkeit ermittelt, mit der ein Effekt einer bestimmten Größe bei einem bestimmten Stichprobenumfang entdeckt wird.
- Wenn Sie es sich leisten können, passen Sie den Stichprobenumfang und/oder das Design auf der Grundlage Ihrer Power-Berechnungen an

Berechnung

- Mehrere Teststärke-Formeln für verschiedene Versuchs- (und Beobachtungs-) Designs
- Die Formeln können umgestellt werden, um z.B. N zu bestimmen.
- Es gibt viele handelsübliche Potenzrechner, z. B. [hier](#) (Erläuterung [hier](#))
- In der Praxis erfordert die Durchführung von Teststärke-Analysen

Formeln

Teststärkeberechnung für zwei Gruppen Differenz-im-Mittelwert-Test mit gleichen Varianzen und Gruppengrößen:

$$\text{power} = \Phi\left(\frac{|\mu_t - \mu_c| \sqrt{N}}{2\sigma}\right) - \Phi^{-1}(1 - \frac{\alpha}{2})$$

- Φ ist die CDF der Normalverteilung → Teststärke unter der Annahme, dass sie der Normalverteilung folgt
- $\mu_{t,c}$ ist das durchschnittliche Ergebnis in der Behandlungs-/Kontrollgruppe → Effekt
- σ ist die Standardabweichung der Ergebnisse → Störanfälligkeit
- α ist das gewählte Signifikanzniveau, häufig 0,05 nach

Teststärke Analyse: Beispiel

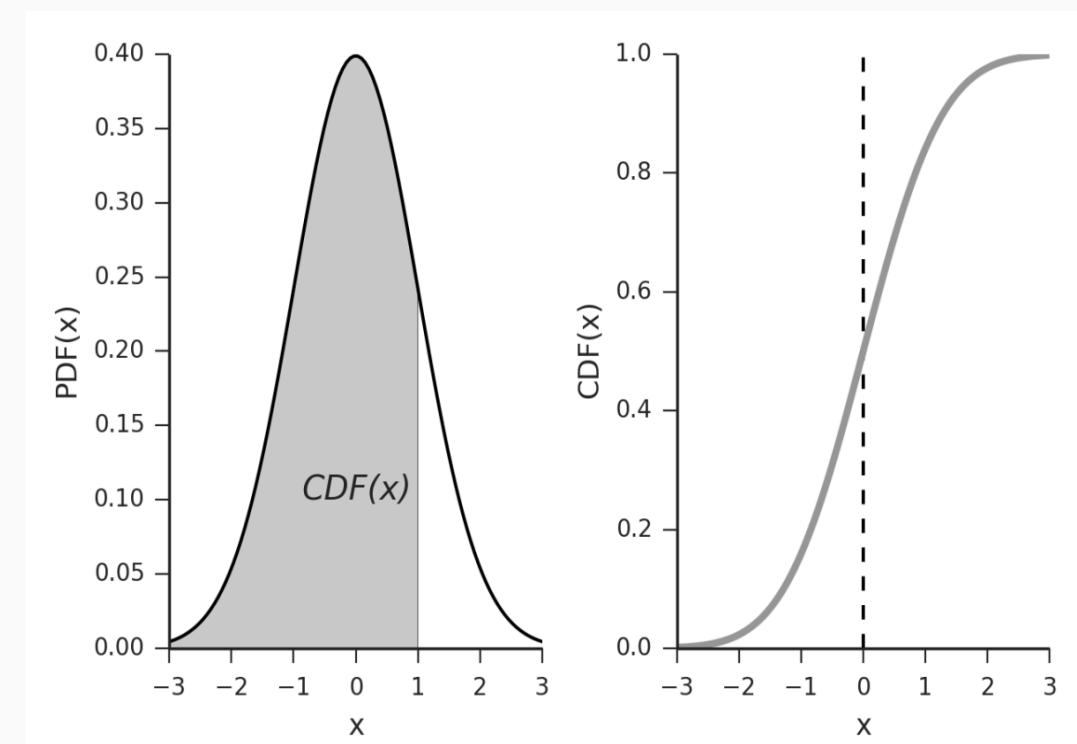
Beispiel

- Sie möchten einen Anstieg der Wahlbeteiligung um 5 % aufgrund einer neuen Kampagne feststellen
- Sie haben eine Stichprobe von 500 Wählern
- Gehen Sie von einer Standardabweichung der Wahlbeteiligung von 20% aus.
- Nehmen Sie ein Signifikanzniveau von 0,05 an.
- Wie hoch ist die Aussagekraft Ihrer Studie?
- Verwenden Sie z.B. [EGAP-Rechner](#)

Berechnung

- $\mu_t - \mu_c = 0.05$
- $\sigma = 0.2$
- $N = 500$
- $\alpha = 0.05$
- Power = ?

Probability density function (PDF) vs. Cumulative density function (CDF)

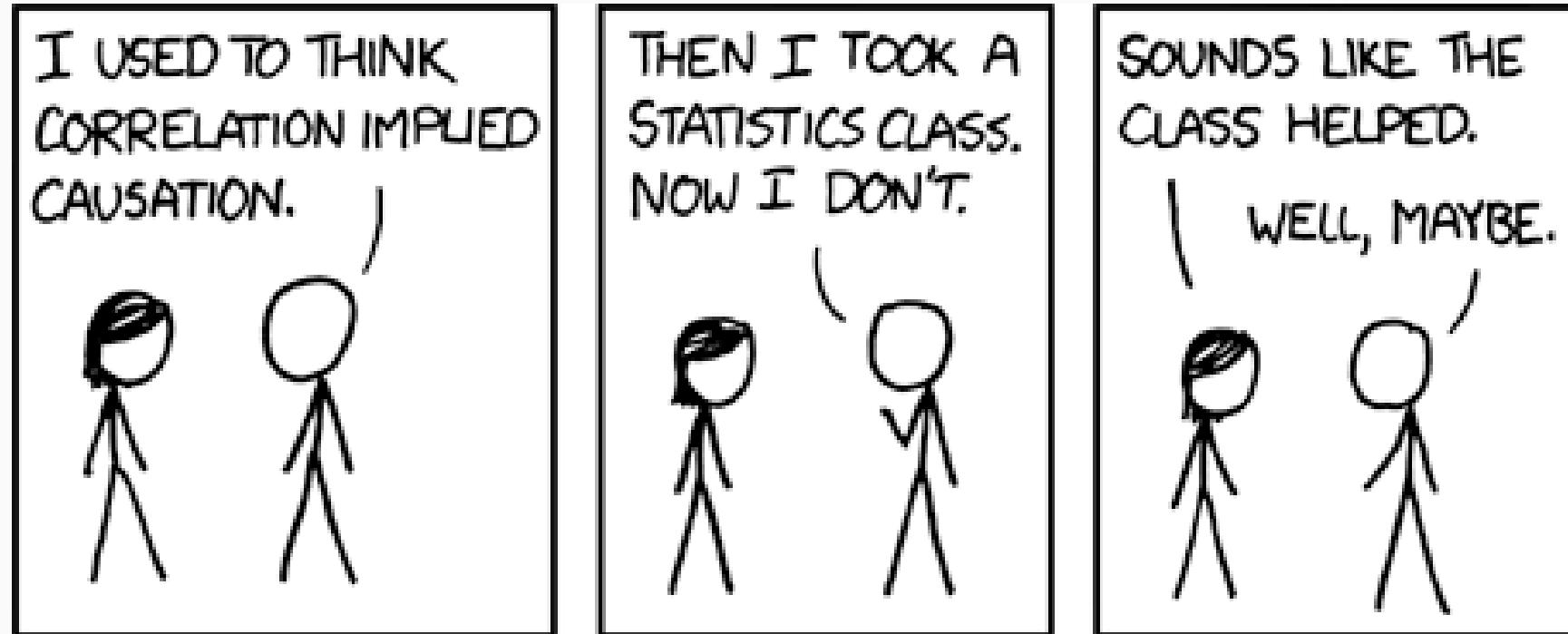


Checkliste

- **Prüfen Sie die Evidenz.** Ist sie wirklich statistisch signifikant?
- **Betrachten Sie die Theorie und die Evidenz.** Ist sie plausibel?
- **Schauen Sie sich das Design an.** Können Sie irgendwelche größeren Fehler entdecken?
- **Betrachten Sie die Effektgröße.** Ist sie aussagekräftig? Ist sie zu gut, um wahr zu sein?
- **Schauen Sie sich die Stichprobengröße an.** Ist sie angemessen groß? Ist die Studie gut getestet?
- **Prüfen Sie, ob die Studie vorregistriert wurde.** Erkennen Sie Ad-hoc-Hypothesen.
- **Vertrauen Sie keiner einzelnen Studie.** Achten Sie auf Meta-Analysen!

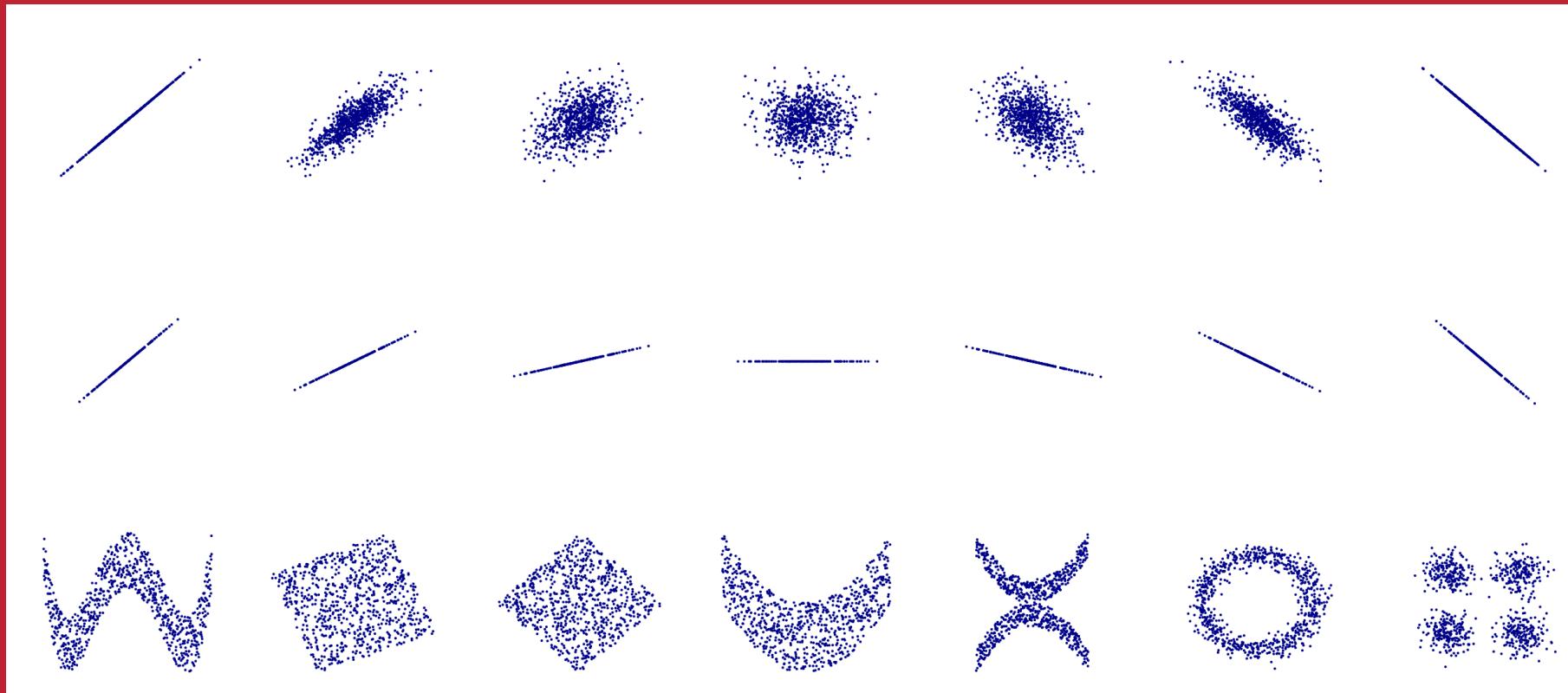


Schlechte Inferenz: Korrelation impliziert keine Kausalität

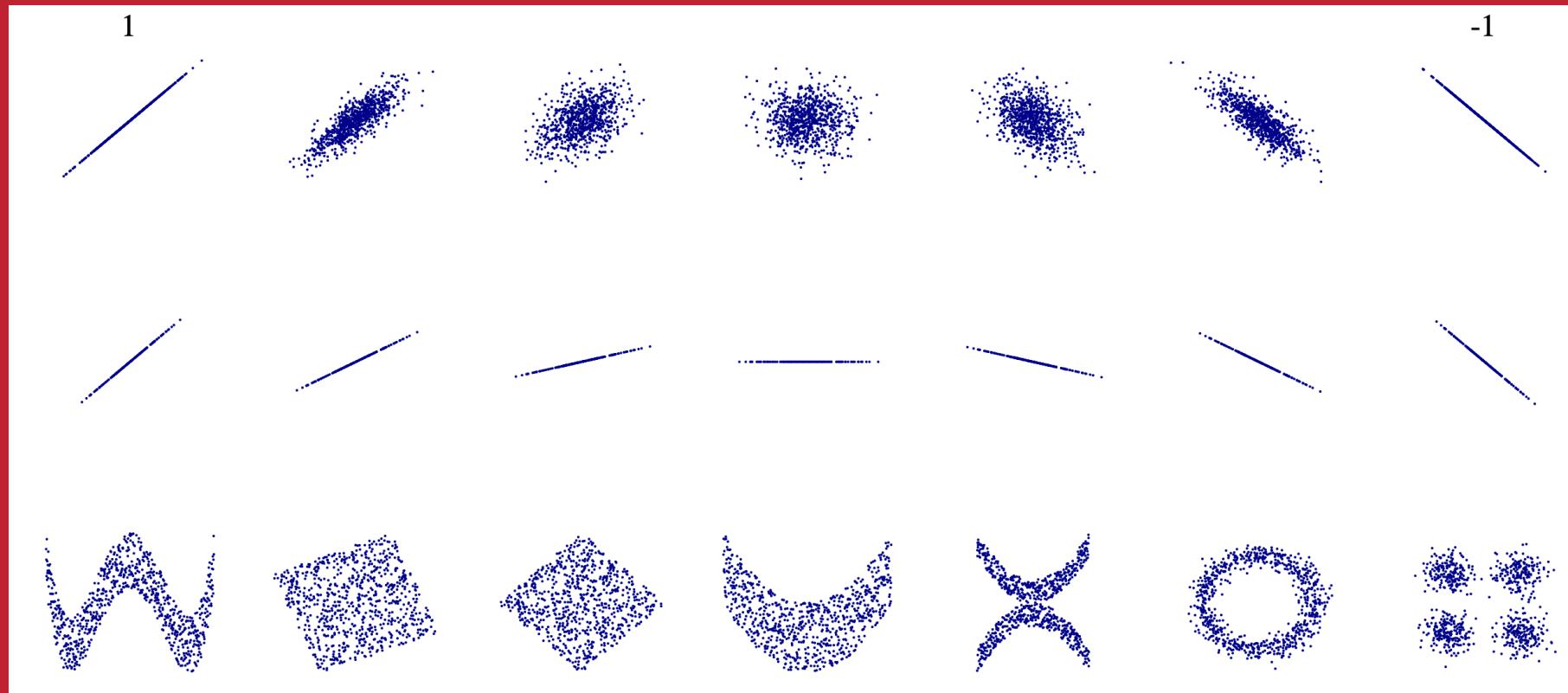


Source XKCD 552

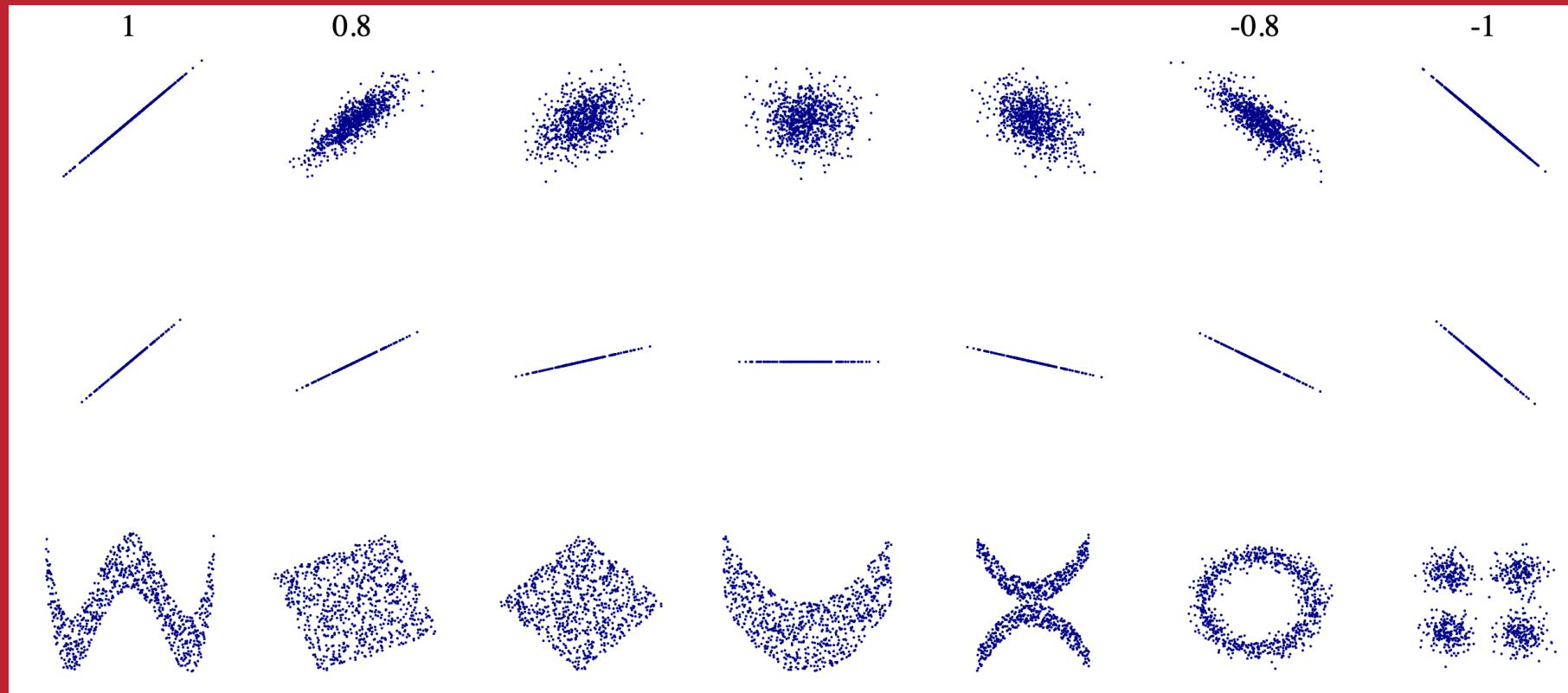
Raten Sie die Korrelationen!



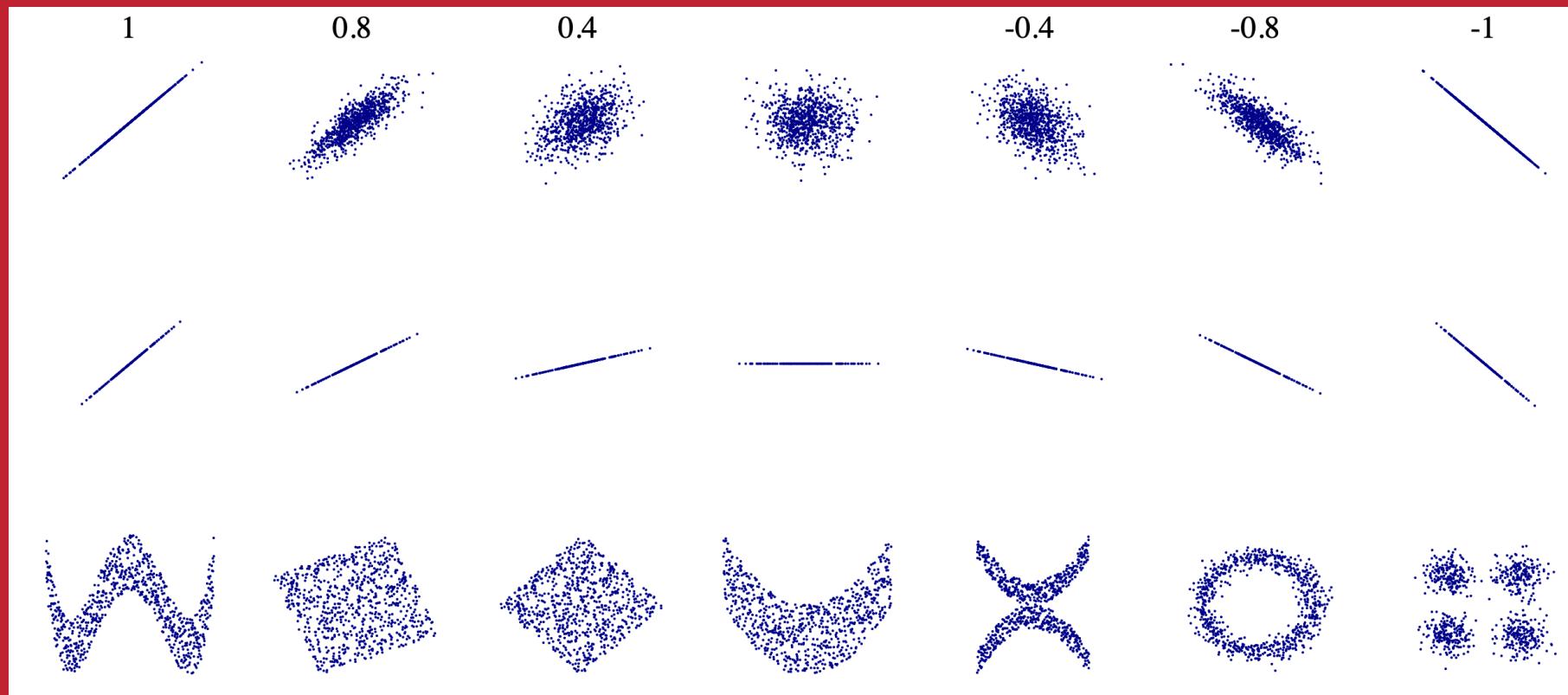
Raten Sie die Korrelationen!



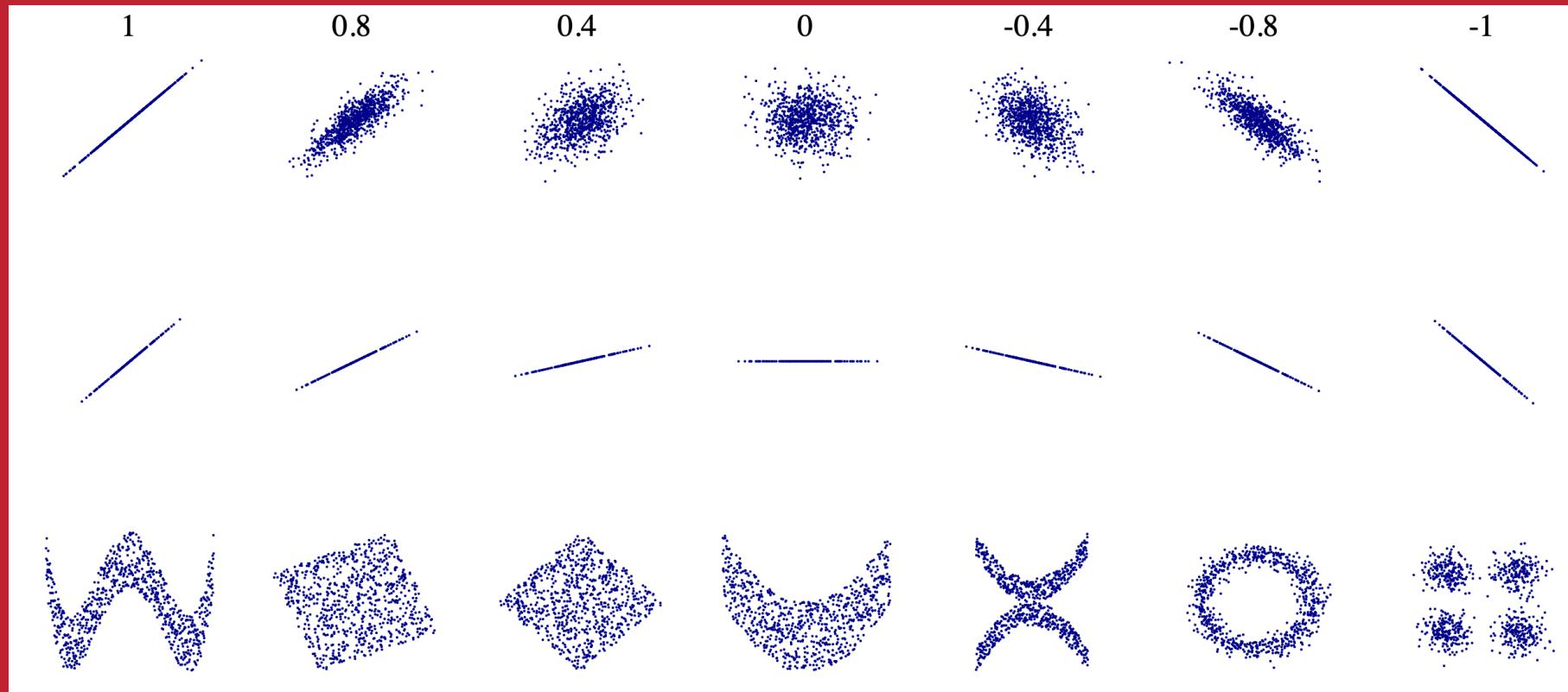
Raten Sie die Korrelationen!



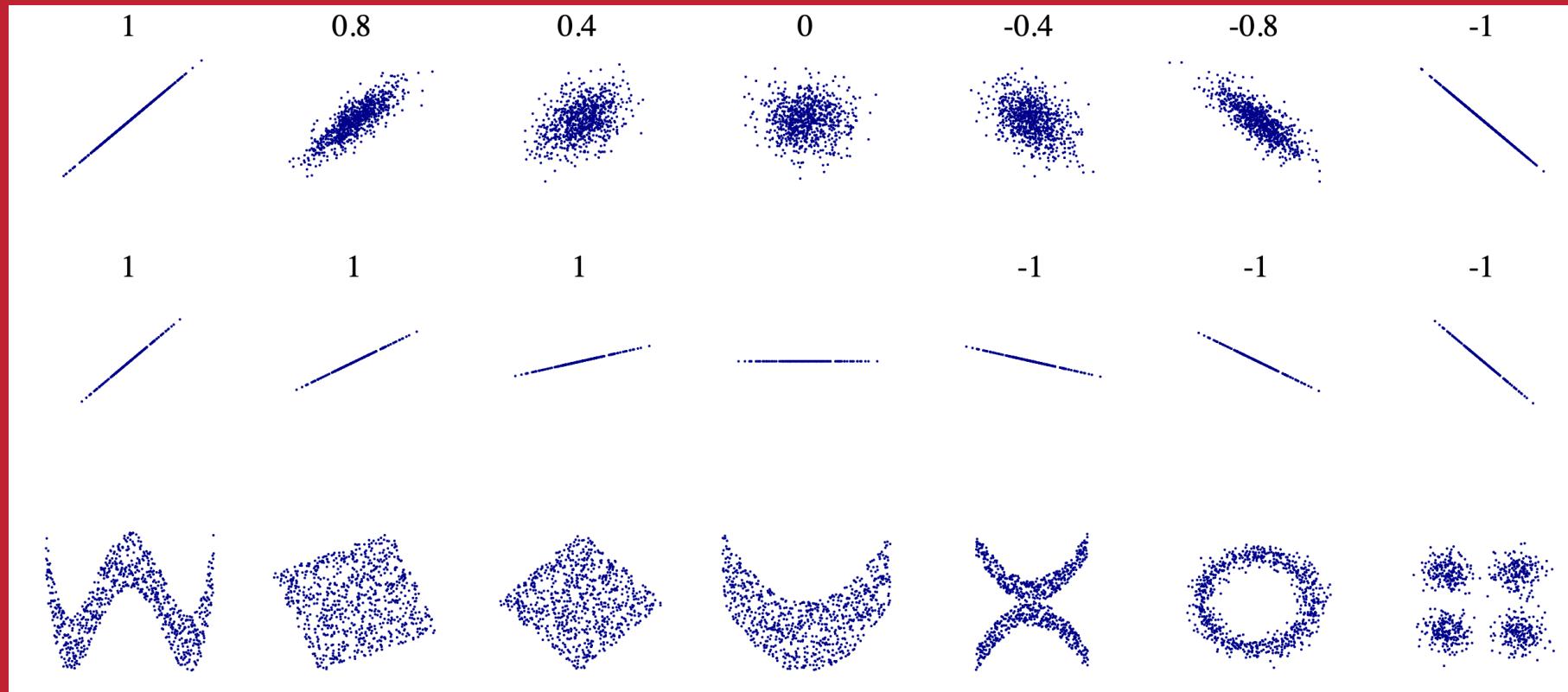
Raten Sie die Korrelationen!



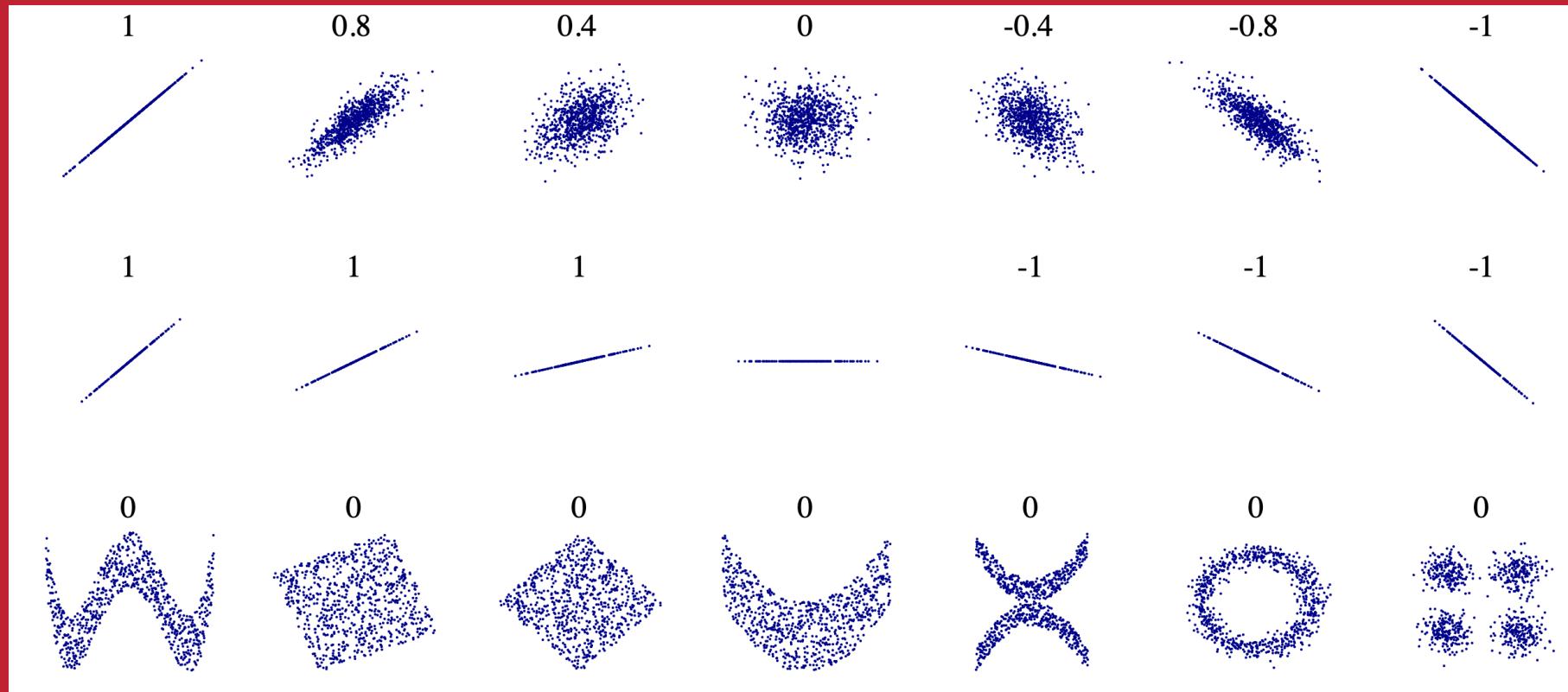
Raten Sie die Korrelationen!



Raten Sie die Korrelationen!

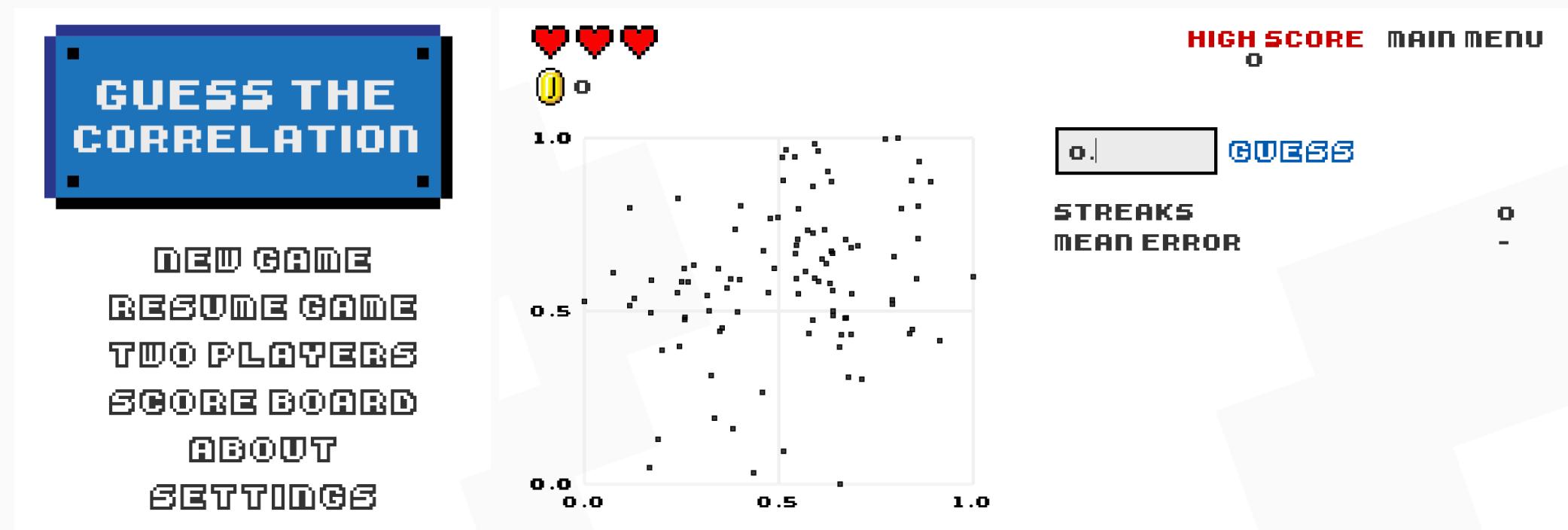


Raten Sie die Korrelationen!



Korrelationen besser erraten

Siehe <http://guessthecorrelation.com/>



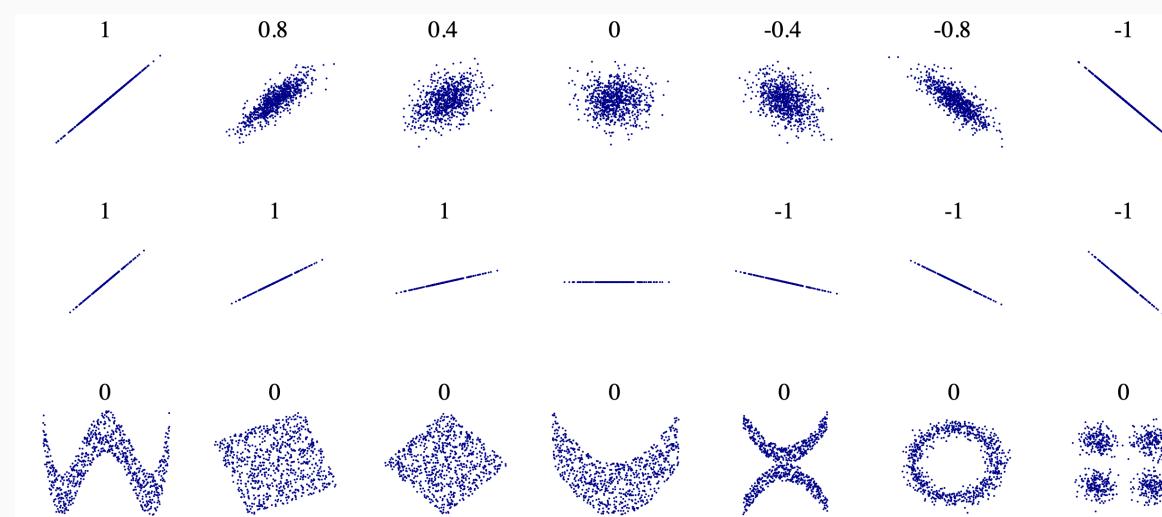
Der Pearson-Korrelationskoeffizient

- Misst die **Stärke und Richtung** einer **linearen Beziehung** zwischen zwei Variablen
- Der Bereich reicht von -1 bis 1
- Formel zur Berechnung:

$$r_{xy} = \frac{\text{covariation of X and Y}}{\text{separate variation of X and Y}} = \frac{Cov(x,y)}{s_x s_y} = \sum_i \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

Pearson's r ist...

- positiv, wenn die Variablen A und B gemeinsam steigen
- negativ, wenn A[B] steigt und B[A] sinkt
- 1, wenn A und B gemeinsam steigen -1, wenn A steigt und B sinkt
- 0, wenn A und B nicht kovariieren



Korrelation

Zwei Variablen sind **korreliert**, wenn die Kenntnis des Wertes der einen Variablen Aufschluss über den wahrscheinlichen Wert der anderen gibt.

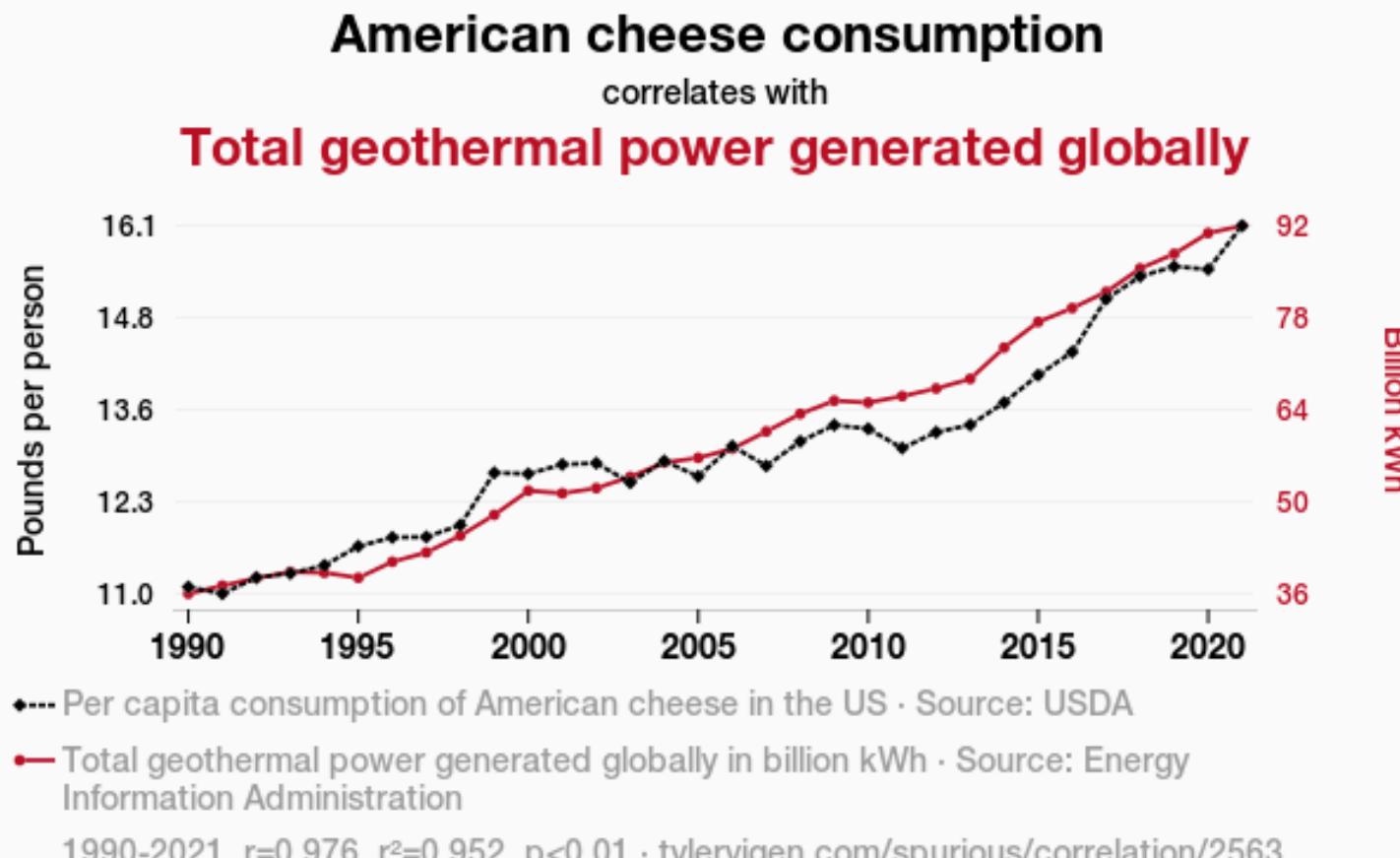
Kausalität

Zwei Ereignisse sind **kausal miteinander verbunden**, wenn das Auftreten des einen Ereignisses eine Folge des Auftretens des anderen ist.

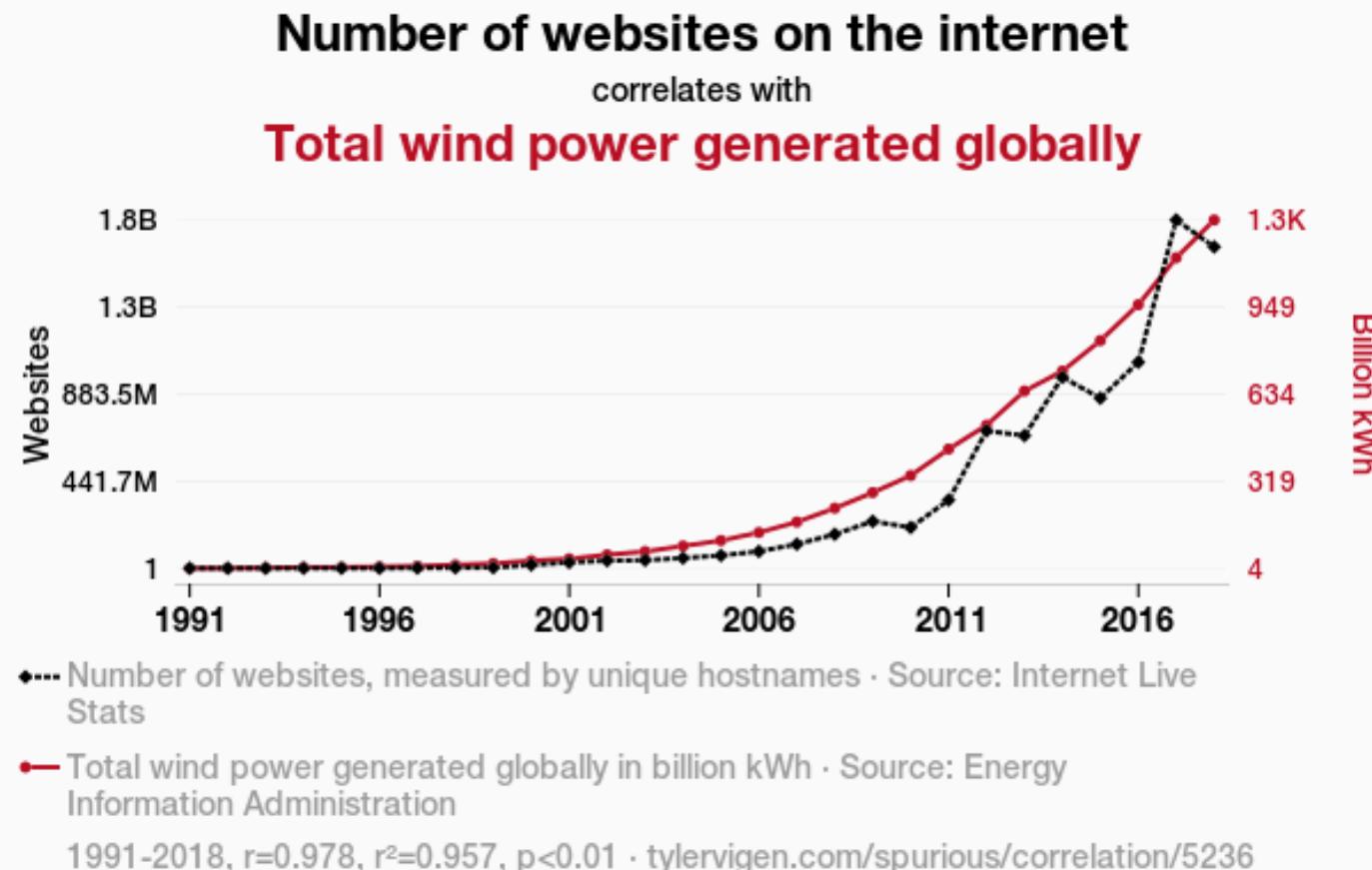
Der kausale Trugschluss

Zwei Variablen, die miteinander korreliert sind, stehen nicht notwendigerweise in einer Ursache-Wirkungs-Beziehung.

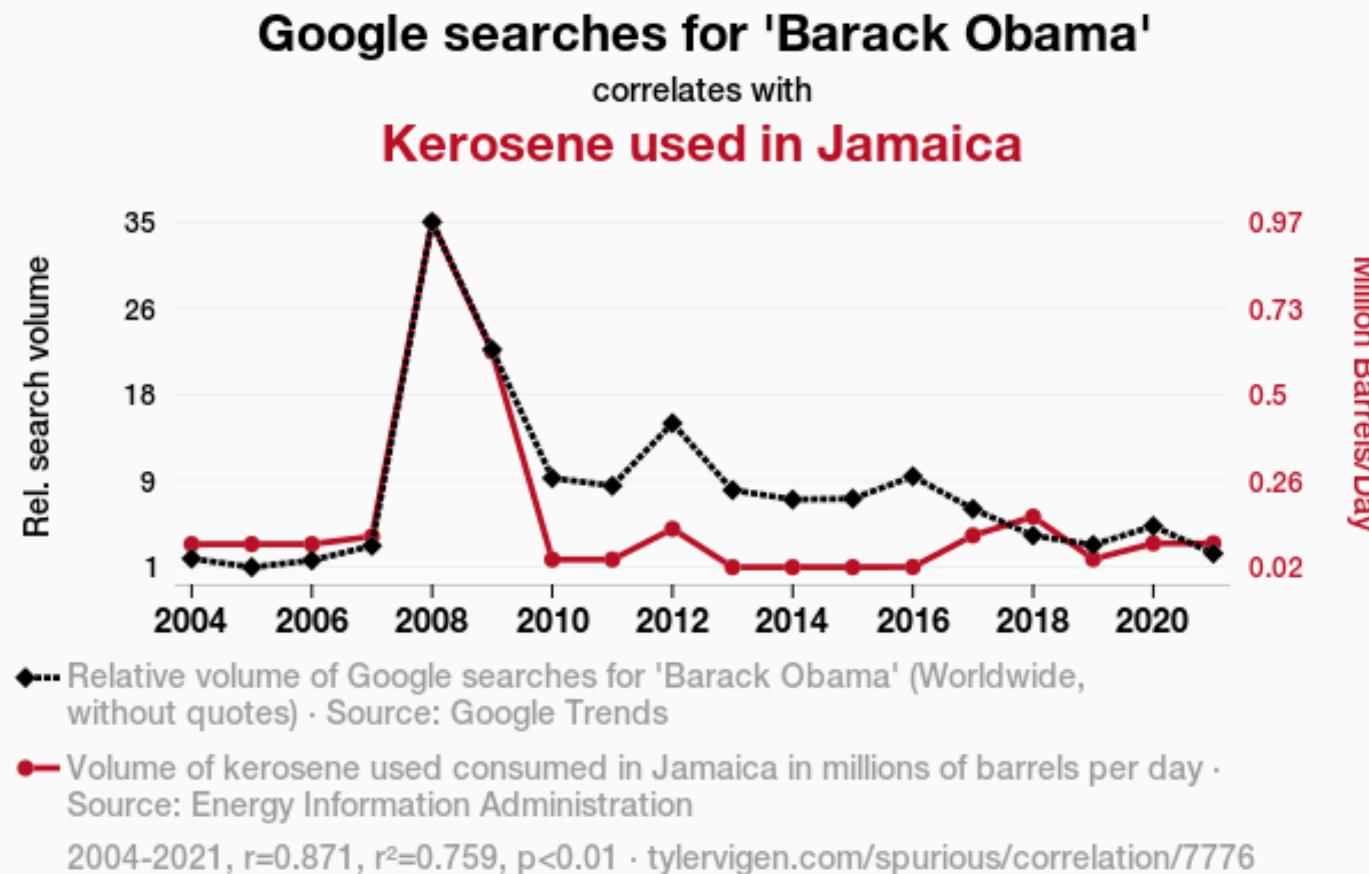
- cum hoc ergo propter hoc („damit, also deswegen“)
- post hoc ergo propter hoc („nach diesem, also wegen diesem“)



Source Tyler Vigen, <http://tylervigen.com/spurious-correlations>



Source Tyler Vigen, <http://tylervigen.com/spurious-correlations>



Source Tyler Vigen, <http://tylervigen.com/spurious-correlations>

Was bedeutet es also, wenn die Variablen A und B korreliert sind?

Hertie School

Einige mögliche Erklärungen

- A verursacht B (direkte Kausalität)
- B verursacht A (umgekehrte Kausalität)
- A und B sind Folgen einer gemeinsamen Ursache (Confounding)
- A verursacht C, das B verursacht (Mediation)
- A und B verursachen beide C, das bedingt wird (collider bias)
- Es gibt keinen Zusammenhang zwischen A und B, die empirische Korrelation ist ein Zufall (Scheinkorelation)

Wie kann man zwischen diesen unterscheiden?

- **Experimentelle Designs** (randomisierte kontrollierte Studien)
- **Natürliche Experimente** (Quasi-Experimente)
- **Observationsdaten** (sorgfältige Analyse von Störfaktoren und Kollisionen)
- **Gesunder Menschenverstand** (ist es sinnvoll, dass A Ursache für B ist?)
- **Rivalisierende Erklärungen** (können wir andere Erklärungen ausschließen?)

Was bedeutet das für Sie?

- Lassen Sie sich nicht von **hoher Korrelation oder großer Effektgröße** täuschen.
- Lassen Sie sich nicht von der **statistischen Signifikanz** täuschen.
- Lassen Sie sich nicht von **großer erklärter Varianz (R-Quadrat)** täuschen.

Stellen Sie sich stattdessen die folgenden Fragen:

1. Ist es wirklich sinnvoll, dass A Ursache für B ist?
2. Gibt es Beweise, die andere Erklärungen ausschließen?
3. Stützen sich die Beweise auf einen sauberen Versuchsplan?
4. Beruhen die Beweise auf einem natürlichen Experiment mit einer überzeugenden Geschichte?
5. Wenn es keine (Quasi-)Experimente gibt, beruht der Nachweis auf einer sorgfältigen Analyse von Beobachtungsdaten, wobei plausible Störfaktoren und Kollider berücksichtigt werden?