

Tag 2: Big Data, Good Data? Die kritische Nutzung von Daten

Session 6: Informierter Konsum von Daten und Statistiken

Simon Munzert
Hertie School

1. Deskriptive Statistik verstehen
2. Wahrscheinlichkeit verstehen
3. Statistische Effekte verstehen
4. Statistische Signifikanz verstehen

Deskriptiven Statistik verstehen

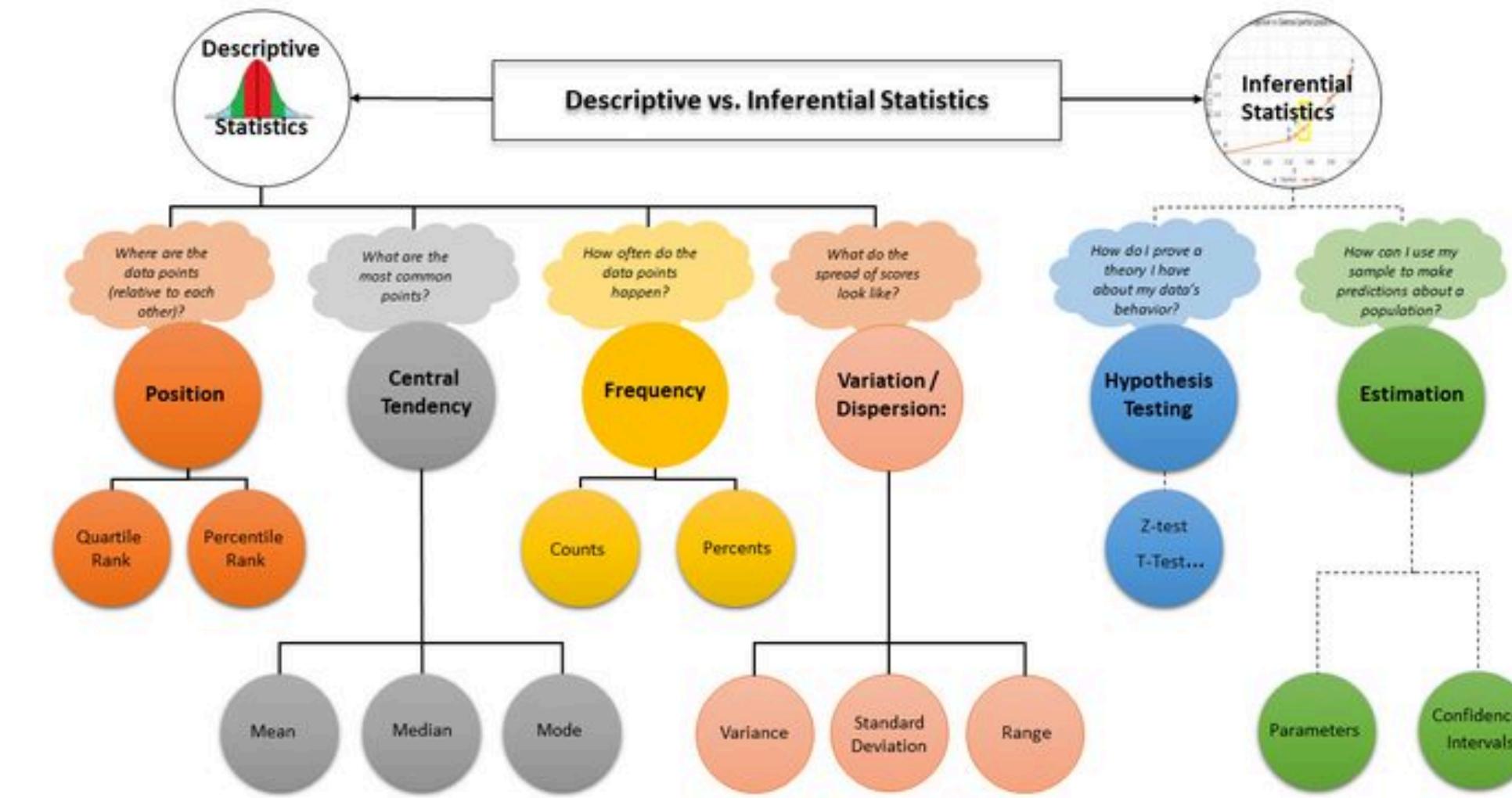
Descriptive Statistik

- Zusammenfassen und Beschreiben von Merkmalen einer Stichprobe oder Population
- Kann numerisch und visuell kommuniziert werden
- Unterschiedliche Skalen (Messniveaus) erfordern unterschiedliche deskriptive Statistiken
- Eine gute Beschreibung kann schwierig sein, wenn die Datenerfassung oder Messung komplex ist.

Inferenzstatistik

- Schlussfolgerungen über eine Grundgesamtheit auf der Grundlage einer Stichprobe
- Es können Rückschlüsse auf Mittelwerte, Proportionen, Beziehungen usw. gezogen werden
- Kann numerisch und visuell kommuniziert werden
- Gute Beschreibung ist die Grundlage für gute Schlussfolgerungen

Deskriptive vs. Inferenzstatistik



Drei populäre Maße für zentrale Tendenz

- **(Arithmetischer) Mittelwert:** Der **Durchschnitt** aller Werte in einem Datensatz
- **Median:** Der **mittlere** Wert eines Datensatzes
- **Modus:** Der **häufigste** Wert in einem Datensatz

Warum „zentrale Tendenz“? Beschreibt die Tendenz von quantitativen Daten, sich um einen zentralen Wert zu gruppieren.

Probieren Sie es aus

Ermitteln Sie den Modus, den Median und den Mittelwert der folgenden Werte:

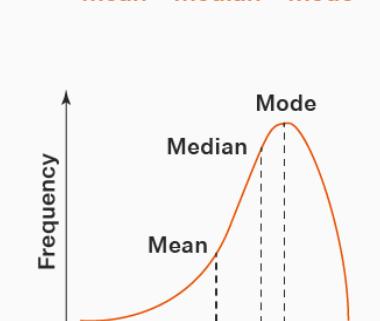
8, 2, 4, 2, 18, 6, 2

Welches Maß ist zu verwenden?

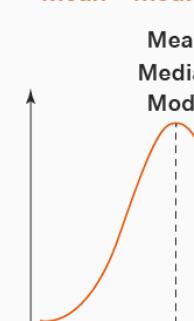
- **Mittelwert:** Empfindlich für Outlier, aber mit einer intuitiven Bedeutung
- **Median:** Robust gegenüber Outliern, aber etwas weniger intuitiv
- **Modus:** Nützlich für kategoriale Daten, kann aber bei kontinuierlichen Daten irreführend sein

Verzerrte Verteilungen

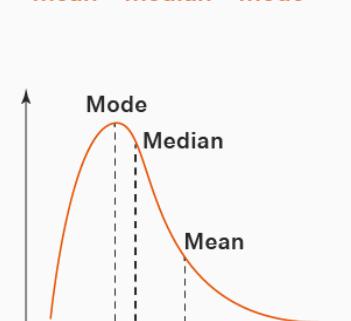
mean < median < mode



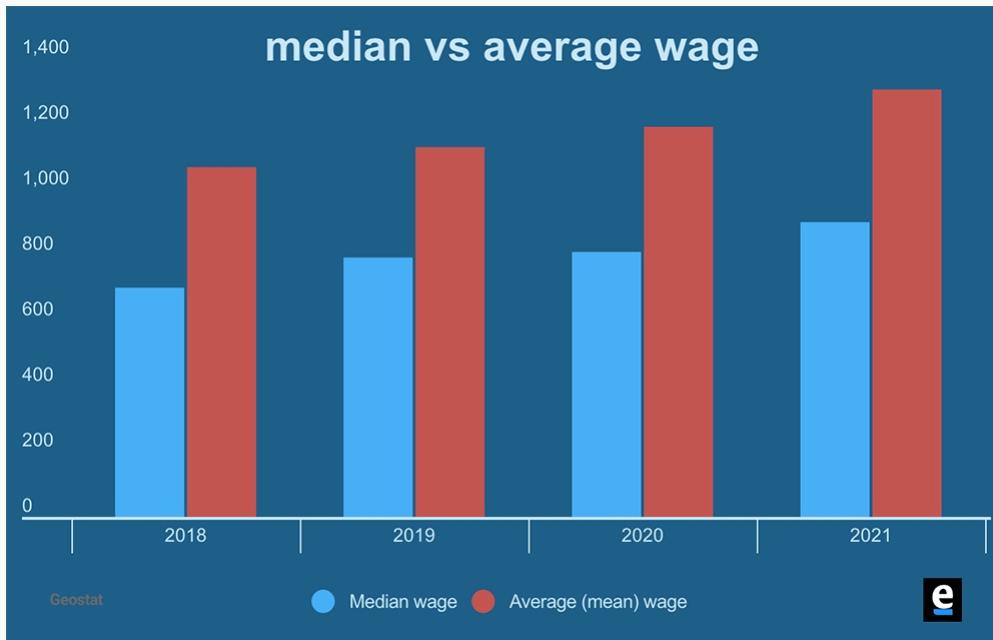
mean = median = mode



mean > median > mode



Maße für zentrale Tendenz: Beispiele



(Measured in Lari)

Source eurasianet, 2022

Average vs median income

Median and mean income between 2012 and 2014 in selected OECD countries in USD; weighted by the currencies' respective [purchasing power \(PPP\)](#).

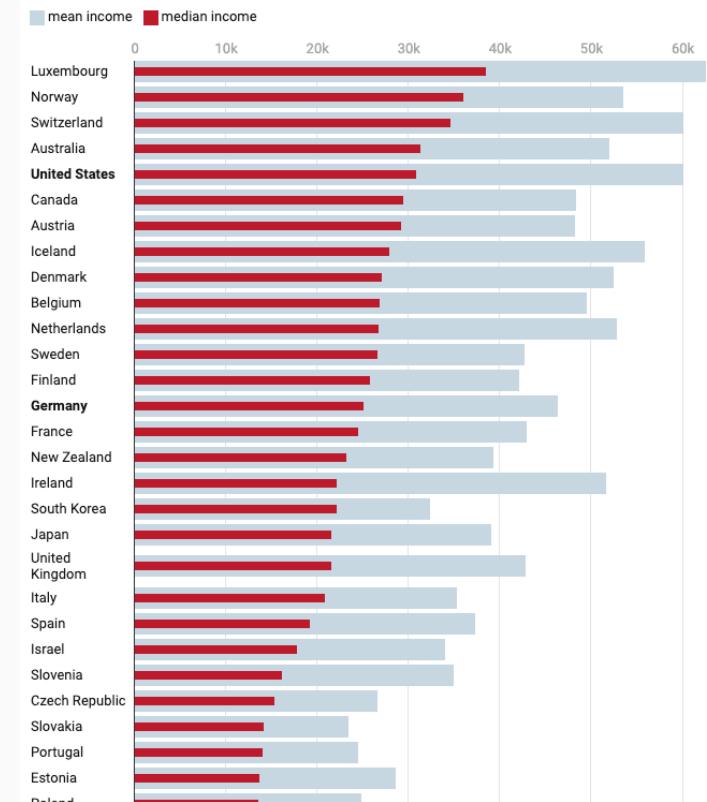


Chart: Lisa Charlotte Rost, Datawrapper • Source: [OECD](#) • [Get the data](#) • Created with Datawrapper

Source Lisa Muth, Datawrapper

Warum brauchen wir Streuungsmaße?

- Die zentrale Tendenz allein ist nicht aussagekräftig
- „Wie weit sind unsere Daten gestreut/ausgedehnt?“

Drei gängige Streuungsmaße

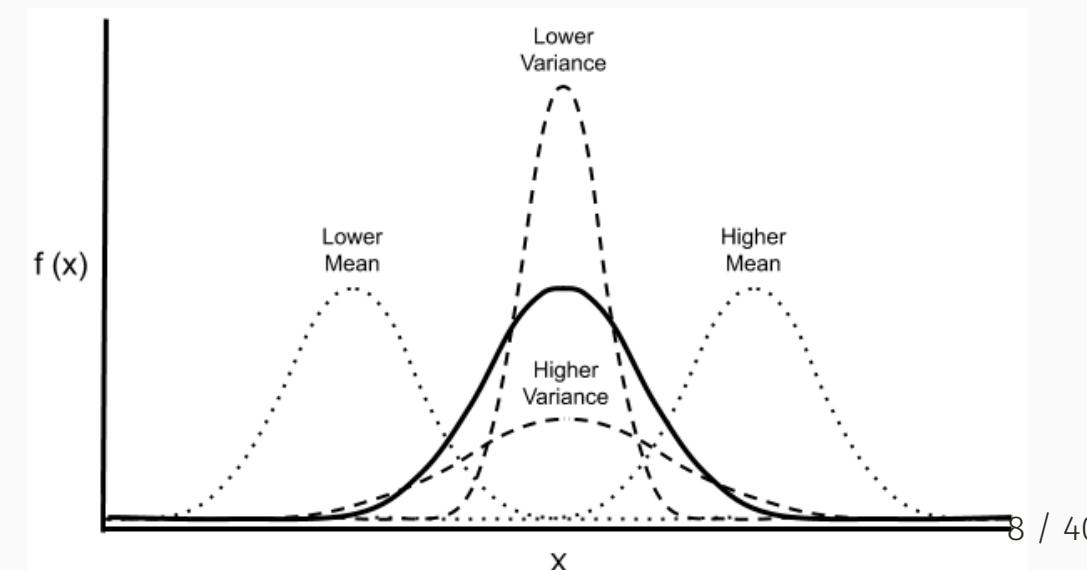
- **Bereich:** Die Differenz zwischen dem höchsten und dem niedrigsten Wert in einem Datensatz
- **Varianz:** Der Durchschnitt der quadrierten Differenzen vom Mittelwert
- **Standardabweichung:** Die Quadratwurzel aus der Varianz

Formel zur Berechnung der Varianz: $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

Siehe [hier](#) für interaktive Intuition.

Warum ist das so wichtig?

- Die meisten Menschen sind in der Tat nicht „durchschnittlich“. Variation kann eine Quelle für Einblicke in die zugrunde liegenden Prozesse sein
- Schlüsselmaß für nachgelagerte Statistiken, z.B. der Standardfehler als Schätzung der Stichprobenvariabilität (Unsicherheit einer Schätzung)



Wie können alle der folgenden Punkte wahr sein?¹

1. 80 % der 100 prominentesten georgischen TikTokers sind männlich.
2. Weibliche georgische TikToker haben im Durchschnitt 500 Follower, männliche nur 300.
3. Es gibt ungefähr gleich viele männliche und weibliche georgische TikTokers.



¹ „Wahr“ im Sinne von ‚theoretisch wahr‘. Die Zahlen sind alle erfunden.

How can all of the following be true?¹

1. At a Georgian university, the acceptance rate to each of four departments is higher for females than for males.

2. Aggregated over the departments, the acceptance rate is higher for males.

Wie können alle der folgenden Punkte wahr sein?¹

3. An einer georgischen Universität ist die Zulassungsquote in jedem der vier Fachbereiche für Frauen höher als für Männer.

¹Gleicher Hinweis wie im vorherigen Beispiel.

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	825	62%	108	82%
B	585	63%	560	63%	25	68%
C	792	34%	417	33%	375	35%
D	714	6%	373	6%	341	7%
Total	3024	39%	2175	47%	849	31%

Legend:

 greater percentage of successful applicants than the other gender

 greater number of applicants than the other gender

bold - the two 'most applied for' departments for each gender

Paradox erklärt

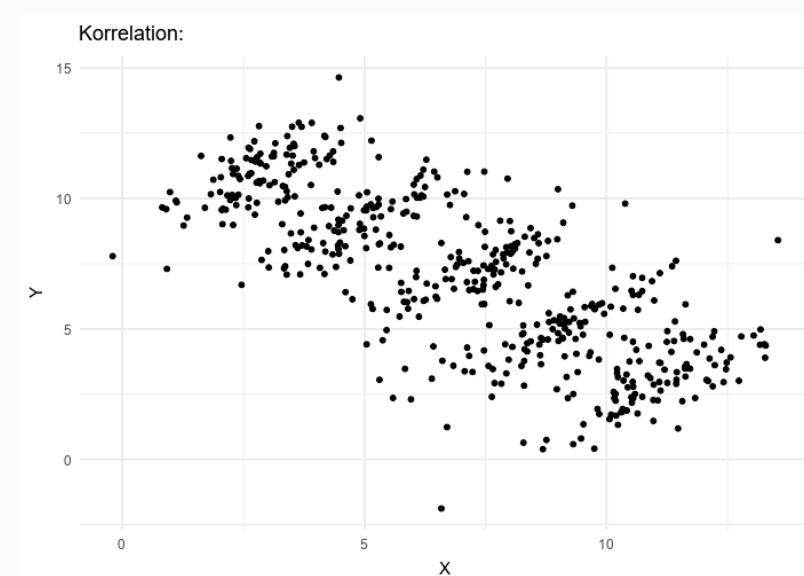
- Einige Abteilungen (C+D) waren wettbewerbsfähiger als andere, und dort bewarben sich mehr Frauen.
- Die allgemeinen Zulassungsquoten und die Zulassungsquoten innerhalb eines Fachbereichs haben unterschiedliche Ausgangspunkte!
- Es handelt sich nicht wirklich um ein „Paradoxon“, sondern um ein verwirrendes Problem: Wenn wir die gruppierende Variable anpassen/kontrollieren, ändert sich die Beziehung zwischen den Variablen.

Das Phänomen, verallgemeinert

- Ein Trend erscheint in verschiedenen Datengruppen, verschwindet aber oder kehrt sich um, wenn diese Gruppen kombiniert werden.
- Dies kann auch bei Korrelationen auftreten (positive vs. negative Korrelation innerhalb vs. zwischen Gruppen)

Relevanz für die Politikgestaltung

- Analyse von Mustern auf verschiedenen Ebenen (z.B. Region vs. Bund, Schulen vs. Schulbezirke)
- Wenn gruppeninterne Muster nicht berücksichtigt werden, könnten die politischen Schlussfolgerungen irreführend sein.



Source Wikipedia, "Simpson's paradox"

Wahrscheinlichkeit verstehen

Was sind Wahrscheinlichkeiten?¹

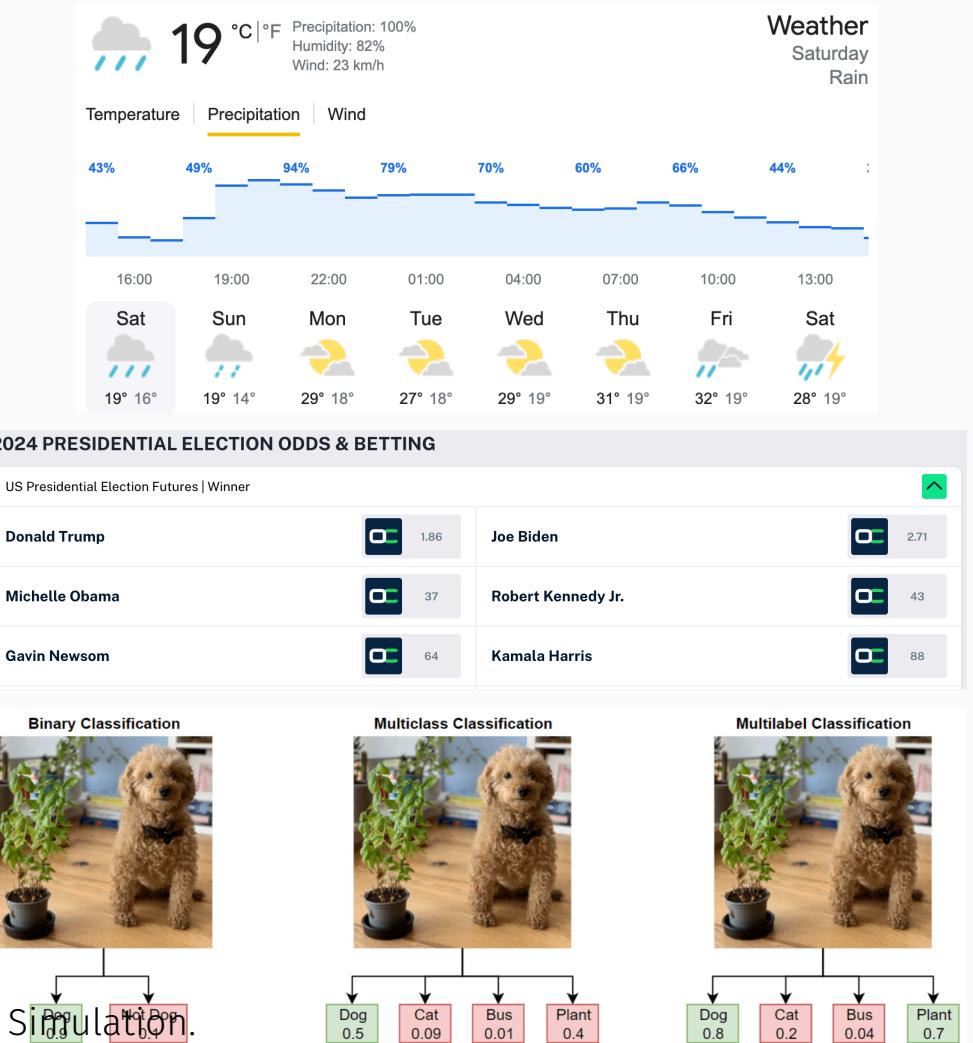
- Wahrscheinlichkeiten quantifizieren die Möglichkeit eines Ereignisses
- Wahrscheinlichkeiten liegen zwischen 0 und 1 (oder 0 und 100%)
- Wahrscheinlichkeiten können verbal oder numerisch kommuniziert werden

Relevanz von Wahrscheinlichkeiten für politische Entscheidungen

Wahrscheinlichkeiten sind ...

- ... das Kernstück der Risikobewertung und Entscheidungsfindung
- ... werden zur Quantifizierung von Unsicherheit
- ... zur Bewertung der Wirksamkeit von Maßnahmen

¹See [hier](#) für eine gute Einführung in die Wahrscheinlichkeitsrechnung und Simulation.



Marginalwahrscheinlichkeit

- Die Wahrscheinlichkeit des Eintretens eines Ereignisses: $p(A)$
- Unbedingte Wahrscheinlichkeit ist nicht von einem anderen Ereignis abhängig
- Beispiel: $p(\text{Würfeln einer } 5) = 1/6$

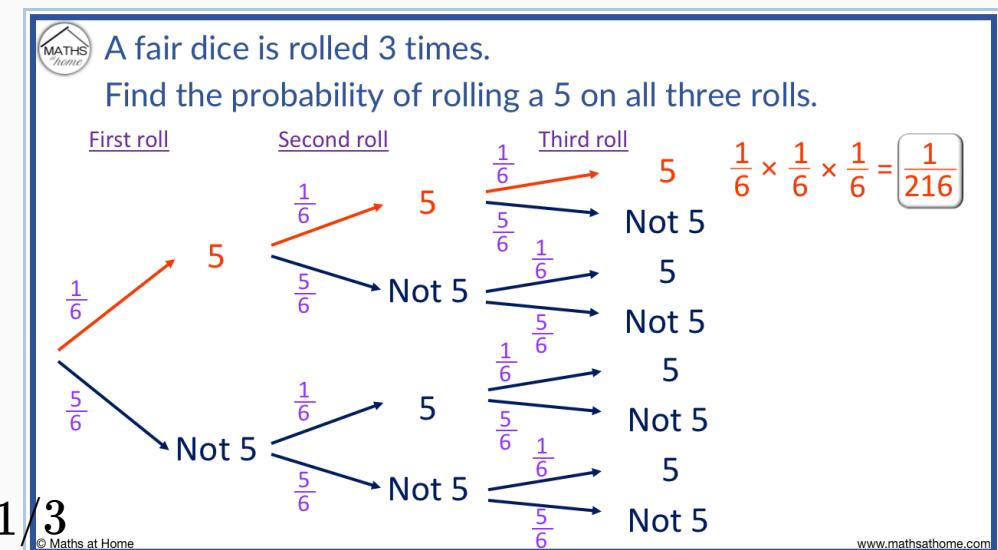
Bedingte Wahrscheinlichkeit

- Die Wahrscheinlichkeit, dass das Ereignis A eintritt, wenn das Ereignis B eintritt: $p(A|B)$
- Wichtig: Die Randwahrscheinlichkeit von B spielt hier keine Rolle!
- Beispiel:

$$p(\text{Würfeln einer } 5 | \text{Würfeln einer ungeraden Zahl}) = 1/3$$

Gemeinsame Wahrscheinlichkeit

- Die Wahrscheinlichkeit, dass Ereignis A und Ereignis B eintreten: $p(A \text{ und } B) = p(A \cap B)$
- Beispiel:
 $p(\text{Würfeln einer } 5 \text{ und einer geraden Zahl}) = 0$



Bedingte Wahrscheinlichkeiten

$$P(A) = 0.500 \text{ or } 50.0\%$$

$$P(B) = 0.300 \text{ or } 30.0\%$$

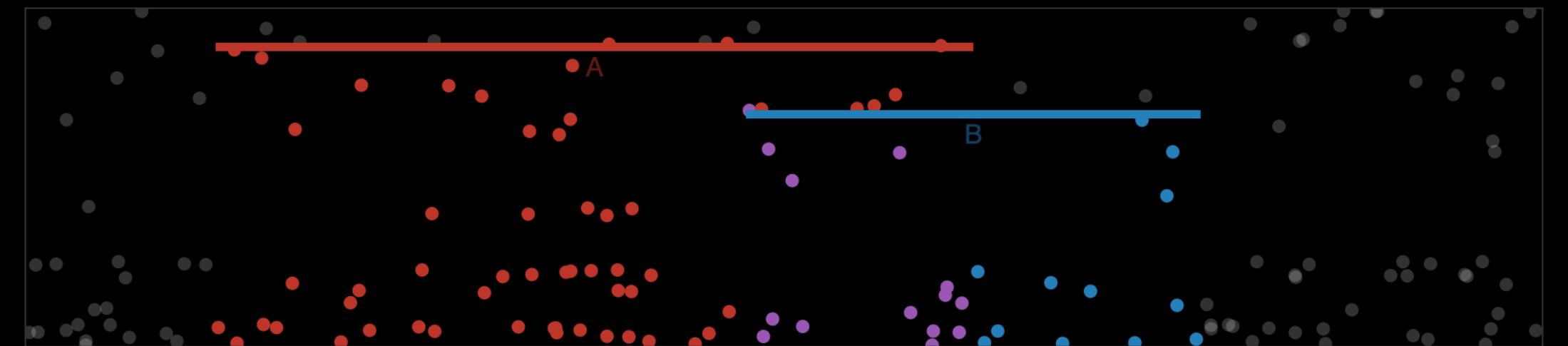
$$P(A \cap B) = 0.150 \text{ or } 15.0\%$$

$$P(B|A) = 0.300 \text{ or } 30.0\%$$

If we have a ball and we know it hit the red shelf, there's a 30.0% chance it also hit the blue shelf.

$$P(A|B) = 0.500 \text{ or } 50.0\%$$

If we have a ball and we know it hit the blue shelf, there's a 50.0% chance it also hit the red shelf.



Source Victor Powell, setosa.io (Siehe zur interaktiven Simulation)

Wie können alle der folgenden Punkte wahr sein?

1. Ein Impfstoff ist hochwirksam beim Schutz gegen eine Krankheit.
2. Die meisten Menschen, die die Krankheit bekommen, sind geimpft worden.



Source [Hakan Nural, Unsplash](#)

Wie können alle der folgenden Punkte wahr sein?

1. Ein Impfstoff schützt hochwirksam vor einer Krankheit.
2. Die meisten Menschen, die die Krankheit bekommen, sind geimpft worden.

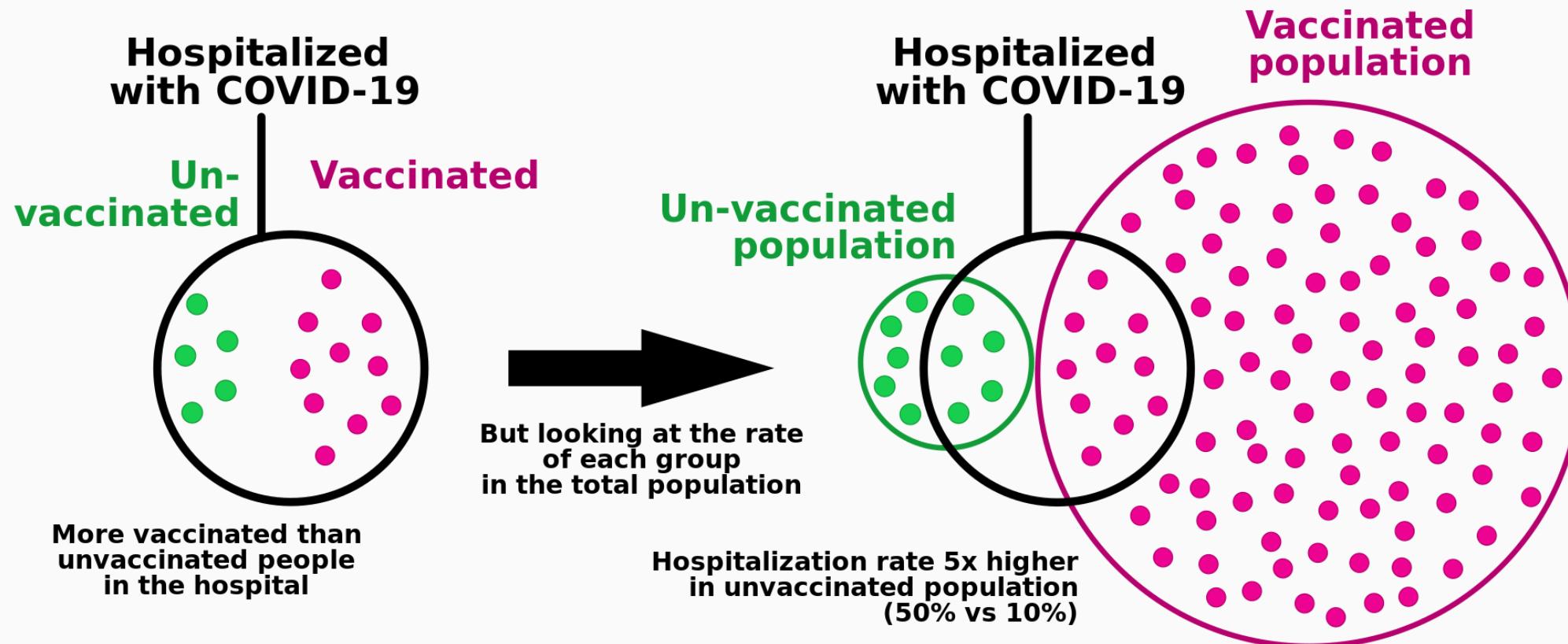
Irrtum der Basisrate

- Dies ist ein klassischer Fall des „Base Rate Fallacy“ oder „Prosecutor's fallacy“.
- Wenn die Impfrate $P(\text{geimpft})$ in der Bevölkerung hoch ist, besteht für geimpfte Personen einfach eine viel größere Chance, ins Krankenhaus zu kommen, als für ungeimpfte Personen.



Source [Hakan Nural, Unsplash](#)

Das Irrtum der Basisrate illustriert



Source Marc Rumilly

Politik für Ereignisse mit geringer Wahrscheinlichkeit

- Politikmaßnahmen sind darauf ausgerichtet, Ereignisse mit geringer Wahrscheinlichkeit zu verhindern, die extrem kostspielig sind, wenn sie eintreten.
- Beispiele: Terroranschläge, Krieg, Naturkatastrophen
- Die Vorhersage solcher Ereignisse ist von Natur aus schwierig.
- KI-gestützte Erkennungssysteme versprechen jedoch hohe Erkennungsraten. Aber selbst bei einer sehr hohen Genauigkeit kann die Anzahl der falsch-positiven Meldungen prohibitiv hoch sein.

Leitfaden

- Berücksichtigen Sie bei der Interpretation von Wahrscheinlichkeiten immer die **Basisrate**.

Sieben Sie vorsichtig, wenn Sie **bedingte**

Beispiel: Terrorist Identifikation

- In einer Stadt mit 1 Mio. Einwohnern gibt es 100 Terroristen und 999.900 Nicht-Terroristen:
 $p(\text{Terrorist}) = 0,0001$
- Überwachungssystem auf der Grundlage von Gesichtserkennungssoftware mit zwei Fehlerquoten von 1%:
 1. Falsch-negativ-Rate:
 $p(\text{kein Alarm}|\text{Terrorist}) = 0,01$
 2. Falsch-positiv-Rate:
 $p(\text{Alarm}|\text{kein Terrorist}) = 0.01$

Was bedeutet das, wenn wir einen Alarm erhalten?¹

$$p(\text{Terrorist}|\text{Alarm}) = \frac{p(\text{Alarm}|\text{Terrorist})p(\text{Terrorist})}{p(\text{Alarm})}$$
$$= \frac{0.99 * 0.0001}{0.01} = 0.01$$

¹Getting $p(\text{Alarm}) = p(\text{Alarm}|\text{Terrorist}) * p(\text{Terrorist}) + p(\text{Alarm}|\text{Nicht-Terrorist}) * p(\text{Nicht-Terrorist}) = p(\text{Terrorist}|\text{Alarm}) = 0.99 * 0.0001 + 0.01 * 0.9999 = 0.01$

Kommunikation von Wahrscheinlichkeiten mit verbalen Ausdrücken

Utrecht University
Erasmus School of Management



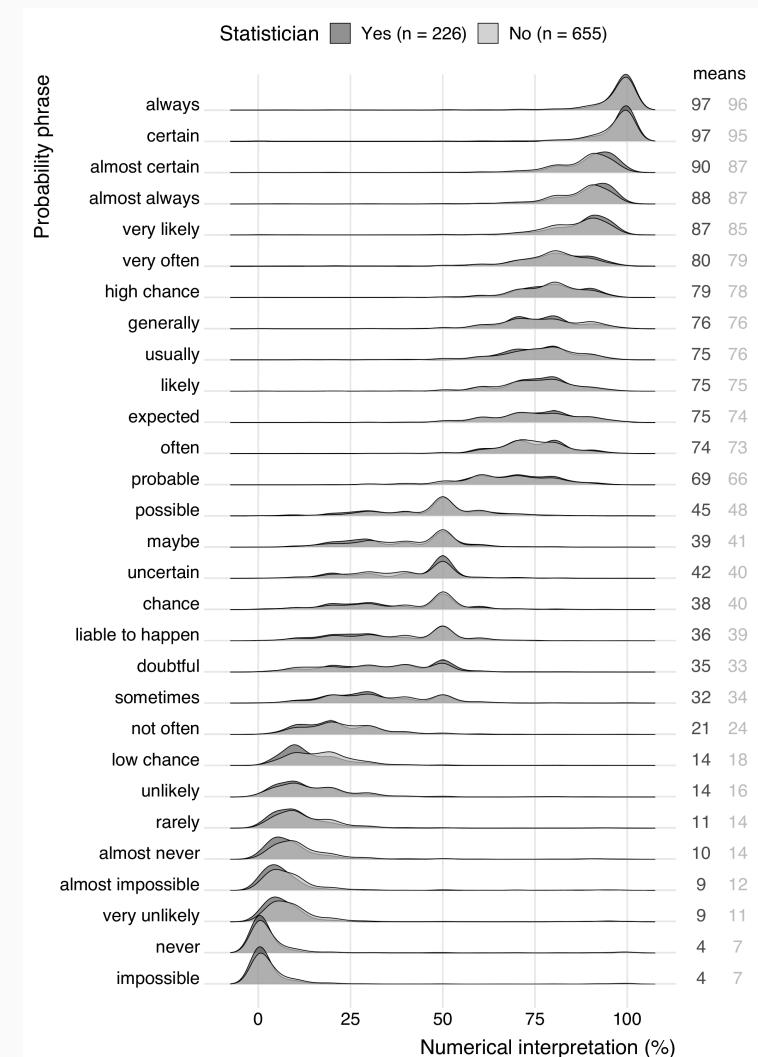
Variability in the interpretation of probability phrases used in Dutch news articles — a risk for miscommunication

Sanne Willems, Casper Albers and Ionica Smeets

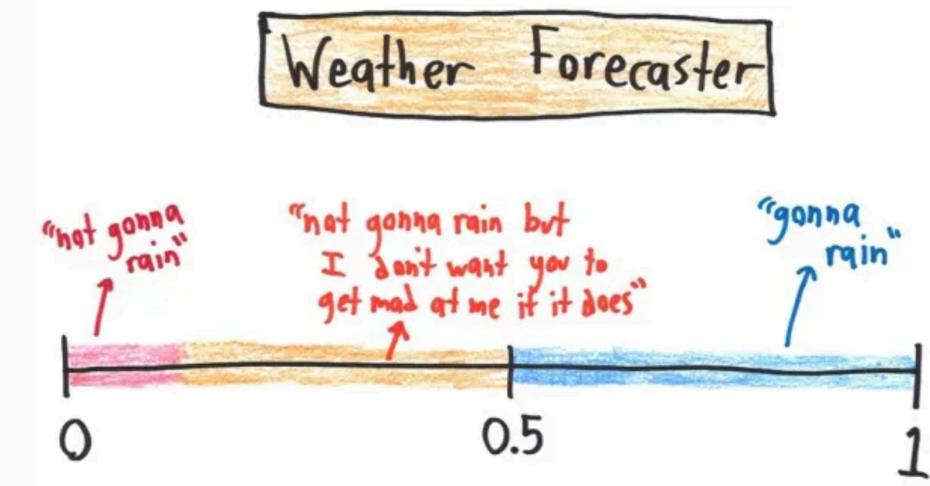
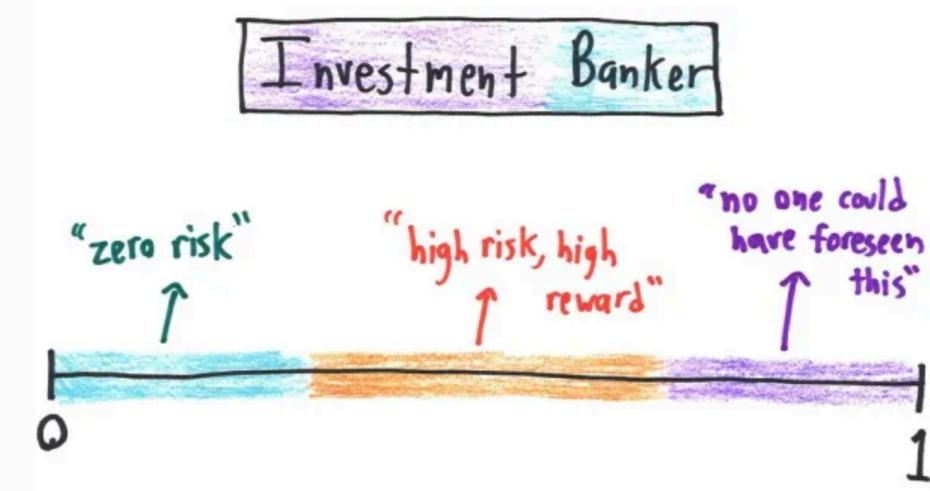
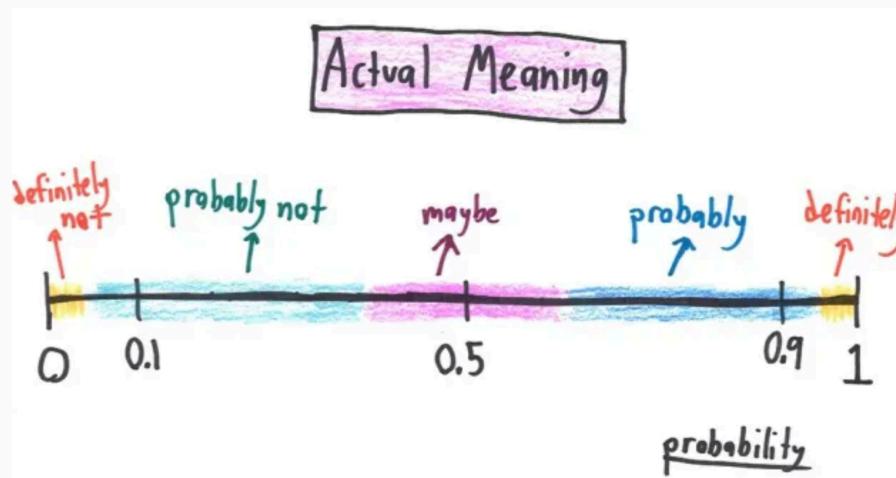
Abstract Verbal probability phrases are often used in science communication to express estimated risks in words instead of numbers. In this study we look at how laypeople and statisticians interpret Dutch probability phrases that are regularly used in news articles. We found that there is a large variability in interpretations, even if the phrases are given in a neutral context. Also, statisticians do not agree on the interpretation of the phrases. We conclude that science communicators should be careful in using verbal probability expressions.

Keywords Risk communication; Science and media; Science writing

Source [Willems et al. 2020](#)



Was bedeutet die Wahrscheinlichkeit für verschiedene Berufe? Hertie School



Statischste Effekte verstehen

Was sind statistische Effekte?

- Statistische Effekte quantifizieren einen Unterschied oder eine Beziehung zwischen Variablen.

Beispielfragen zu Effekten

- Wie groß ist der marginale Effekt von Bildung auf das Einkommen?
- Wie stark steigt die Wahrscheinlichkeit, wählen zu gehen, mit dem Alter?

Statistischer Effekt \neq kausaler Effekt!

- Bei statistischen Effekten geht es um statistische Beziehungen zwischen Variablen, nicht um kausale Beziehungen.
- Nur weil Ihr Modell Ihnen beispielsweise sagt, dass ein zusätzliches Einkommen von 100 \$ pro Monat mit einem zusätzlichen Jahr Bildung verbunden ist,

Effektgrößen

- Effektgrößen sind quantitative Maße für die **Stärke einer Beziehung**.
- Effektgrößen drücken das Ausmaß eines Unterschieds oder einer Beziehung auf standardisierte Weise aus.

Beispiele

- Ein (nicht standardisierter oder standardisierter) **Gruppenmittelwertunterschied**
- Der **Korrelationskoeffizient** ist eine Effektgröße für die Beziehung zwischen zwei kontinuierlichen Variablen (siehe später!), z.B. $r = 0,1 \rightarrow$ schwach, $r = 0,5 \rightarrow$ mittel, $r = 0,9 \rightarrow$ stark
- Der **Regressionskoeffizient** drückt die vorhergesagte marginale Änderung eines Ergebnisses im Verhältnis zu einer Einheitsänderung des Prädiktors aus (möglicherweise in Abhängigkeit

Beispiel: Regressionseffekt

Hourly wage	
Education	0.505*** (0.051)
Female	-2.275*** (0.279)
Nonwhite	-0.119 (0.460)
Intercept	0.650 (0.681)
N	526
R ²	0.259
Adjusted R ²	0.255

Einige Hinweise zur Verwendung von Effektgrößen

Einige Fragen, die Sie sich stellen sollten

- Was bedeutet die Effektgröße **substanzial**? Was bedeutet z.B. ein „Effekt von 0,87“?

	Prominence	Influence
Senate	0.906*** (0.060)	1.483*** (0.067)
Sessions served	0.163*** (0.016)	0.292*** (0.017)
Party (Independent)	0.701* (0.368)	1.059** (0.412)
Party (Republican)	0.035 (0.047)	-0.080 (0.052)
Office: Governor	0.266* (0.158)	0.450** (0.177)
Office: Lt. Governor	-0.031 (0.257)	0.089 (0.288)
Office: US Secretary	0.551** (0.262)	0.372 (0.294)
Position: House Speaker	1.896*** (0.385)	2.670*** (0.431)
Position: Majority / Minority Leader	0.185 (0.308)	0.711** (0.345)
Position: Whip	0.231 (0.233)	0.848*** (0.261)
Position: Deputy Whip	0.698*** (0.234)	0.462* (0.262)
Position: Party Chairman	-0.115 (0.215)	-0.255 (0.241)
(Intercept)	1.648*** (0.050)	1.527*** (0.057)
N	492	492
R-squared	0.493	0.694
Adj. R-squared	0.481	0.687
Residual Std. Error (df = 479)	0.505	0.565
F Statistic (df = 12; 479)	38.890***	90.715***

*** p < .01; ** p < .05; * p < .1

Einige Hinweise zur Verwendung von Effektgrößen

Einige Fragen, die Sie sich stellen sollten

1. Was bedeutet die Effektgröße **substanzial**? Was bedeutet z.B. ein „Effekt von 0,87“?
2. Ist die Effektgröße **plausibel**? Wie steht sie im Vergleich zu Ihrer Intuition und anderen Effekten in der Literatur?

Political Science Research and Methods (2020), page 1 of 7
doi:10.1017/psrm.2019.63



RESEARCH NOTE

Longevity returns to political office

Sebastian Barfort¹, Robert Klemmensen² and Erik Gahner Larsen^{3*}

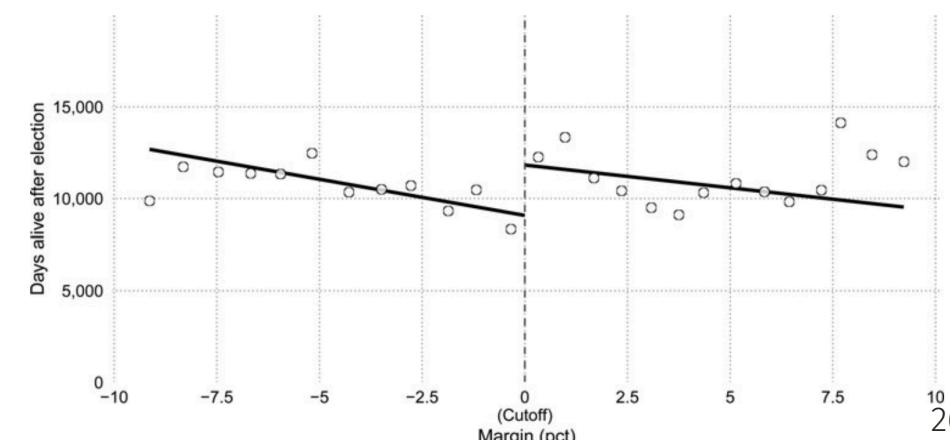
¹Independent Researcher, Copenhagen, Denmark, ²Political Science, University of Southern Denmark, Odense, Denmark and ³School of Politics and International Relations, University of Kent, Canterbury, Kent, United Kingdom of Great Britain and Northern Ireland

*Corresponding author. E-mail: E.G.Larsen@kent.ac.uk

(Received 18 April 2019; revised 19 September 2019; accepted 13 November 2019)

Abstract

Does political office cause worse or better longevity prospects? Two perspectives in the literature offer contradicting answers. First, increased income, social status, and political connections obtained through holding office can increase longevity. Second, increased stress and working hours associated with holding office can have detrimental effects on longevity. To provide causal evidence, we exploit a regression discontinuity design with unique data on the longevity of candidates for US gubernatorial office. The results show that politicians winning a close election live 5–10 years longer than candidates who lose.



Einige Fragen, die Sie sich stellen sollten

1. Was bedeutet die Effektgröße **substanzial**? Was bedeutet z.B. ein „Effekt von 0,87“?
2. Ist die Effektgröße **plausibel**? Wie steht sie im Vergleich zu Ihrer Intuition und anderen Effekten in der Literatur?
3. Wie **präzise** wird der Effekt geschätzt?

Statistische Signifikanz verstehen

Attractive names sustain increased vegetable intake in schools

Brian Wansink ^{a,*}, David R. Just ^b, Collin R. Payne ^c, Matthew Z. Klinger ^d

^a Department of Applied Economics and Management at Cornell University, 15 Warren Hall, Ithaca, NY 14853-7801, USA

^b Department of Applied Economics and Management at Cornell University, 16 Warren Hall, Ithaca, NY 14853-7801, USA

^c New Mexico State University, College of Business, MSC 5280, PO Box 30001, Las Cruces, NM 88003-8001, USA

^d Half Hollow Hills High School East, 50 Vanderbilt Parkway, Dix Hills, NY 11746, USA

ABSTRACT

Objective: This study will determine if the selective use of attractive names can be a sustainable, scalable means to increase the selection of vegetables in school lunchrooms.

Methods: Study 1 paired an attractive name with carrots in five elementary schools ($n=147$) and measured selection and consumption over a week compared to controls. Study 2 tracked food sales of vegetables in two elementary schools ($n=1017$) that were systematically attractively named or not named over a two-month period. Both studies were conducted in New York in 2011.

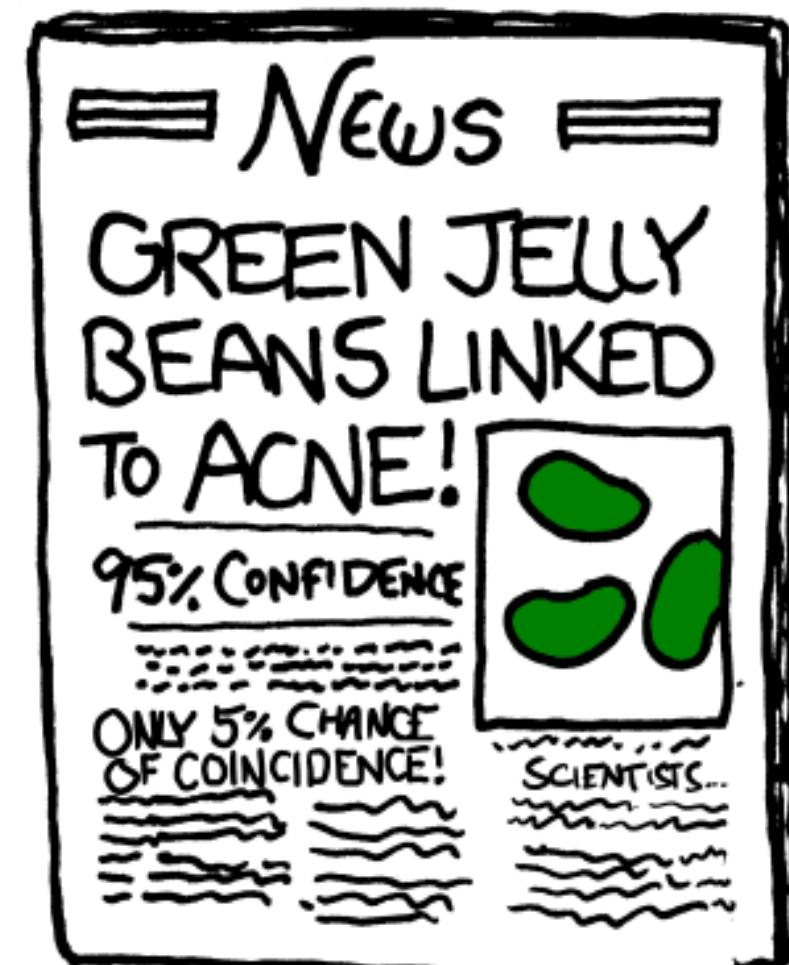
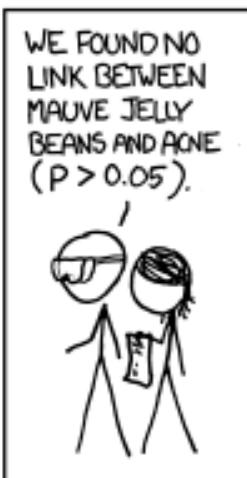
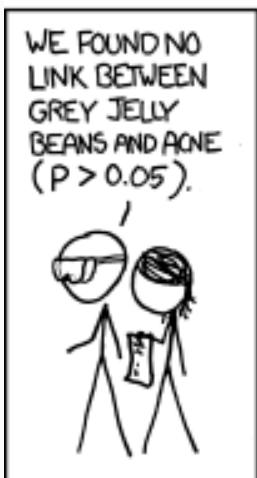
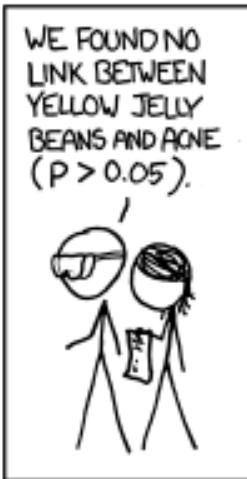
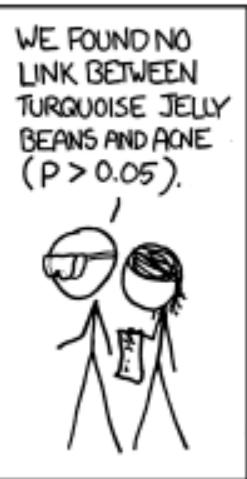
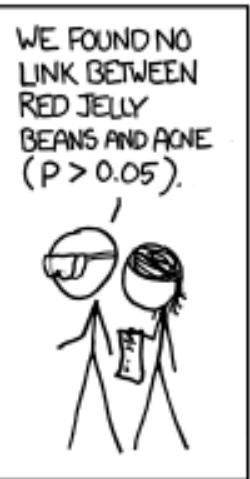
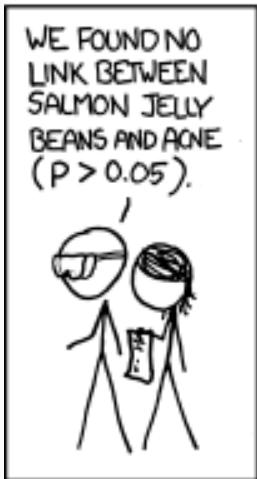
Results: Study 1 found that elementary students ate twice the percentage of their carrots if attractively named as "X-ray Vision Carrots," than if un-named or generically named as the "Food of the Day." Study 2 found that elementary school students were 16% more likely to persistently choose more hot vegetable dishes ($p<0.001$) when they were given fun or attractive names.

Discussion: Attractive names effectively and persistently increased healthy food consumption in elementary schools. The scalability of this is underscored by the success of Study 2, which was implemented and executed for negligible cost by a high school student volunteer.



Source Wansink et al.,
Retraction Watch

"Statistische Signifikanz" überall



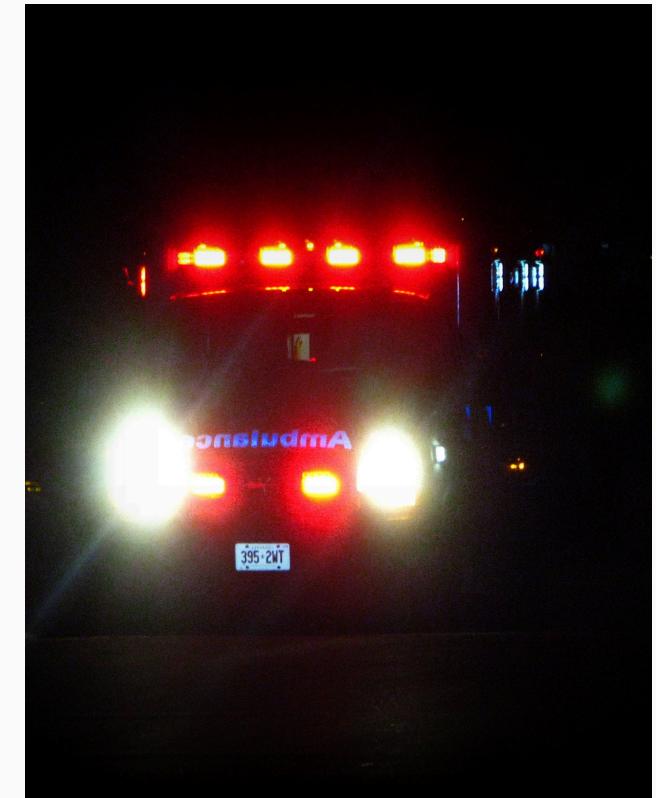
Beispiel statistischer Fehler, bei denen es um Leben und Tod geht

Sie sind Rettungssanitäter und nähern sich dem Ort eines Autounfalls. Ein Opfer liegt regungslos auf der Straße und Sie müssen einschätzen, ob das Opfer tot oder lebendig ist, und das Opfer entsprechend behandeln.

Ausgehend von diesen Informationen, **welcher Fehler führt zum größten Fehlschlag?**

Hypothesen

- **Nullhypothese:** Das Opfer ist am Leben.
- **Alternative Hypothese:** Das Opfer ist nicht am Leben.



Source [jeffalltogether, StackExchange.com](#)

Hypothesen

- **Nullhypothese:** Das Opfer ist am Leben.
- **Alternative Hypothese:** Das Opfer ist nicht am Leben.

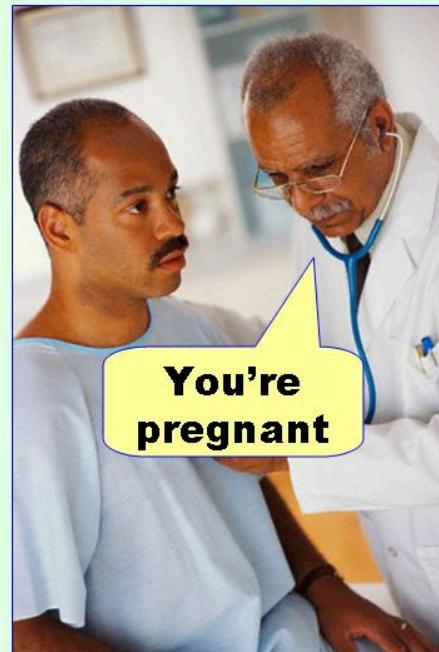
Fehlertypen

- **Typ-I-Fehler:** Sie weisen die Null zurück, obwohl die Null tatsächlich wahr ist („falsch positiv“).
- **Typ II-Fehler:** Sie verwerfen die Null nicht, wenn die Null tatsächlich falsch ist. („falsch negativ“)

Kosten

- **Typ I-Fehler:** Sie erklären das Opfer für tot, obwohl es in Wirklichkeit noch lebt. Sie erhalten keinen Krankenwagen, der sie zur lebensrettenden medizinischen Behandlung ins Krankenhaus bringt. → **Extrem kostspieliger Fehler**
- **Typ II-Fehler:** Sie erklären das Opfer für lebendig, obwohl es in Wirklichkeit tot ist. Sie schicken eine tote Person fälschlicherweise mit einem Krankenwagen ins Krankenhaus → **Nicht so kostspieliger Fehler**

Type I error
(false positive)



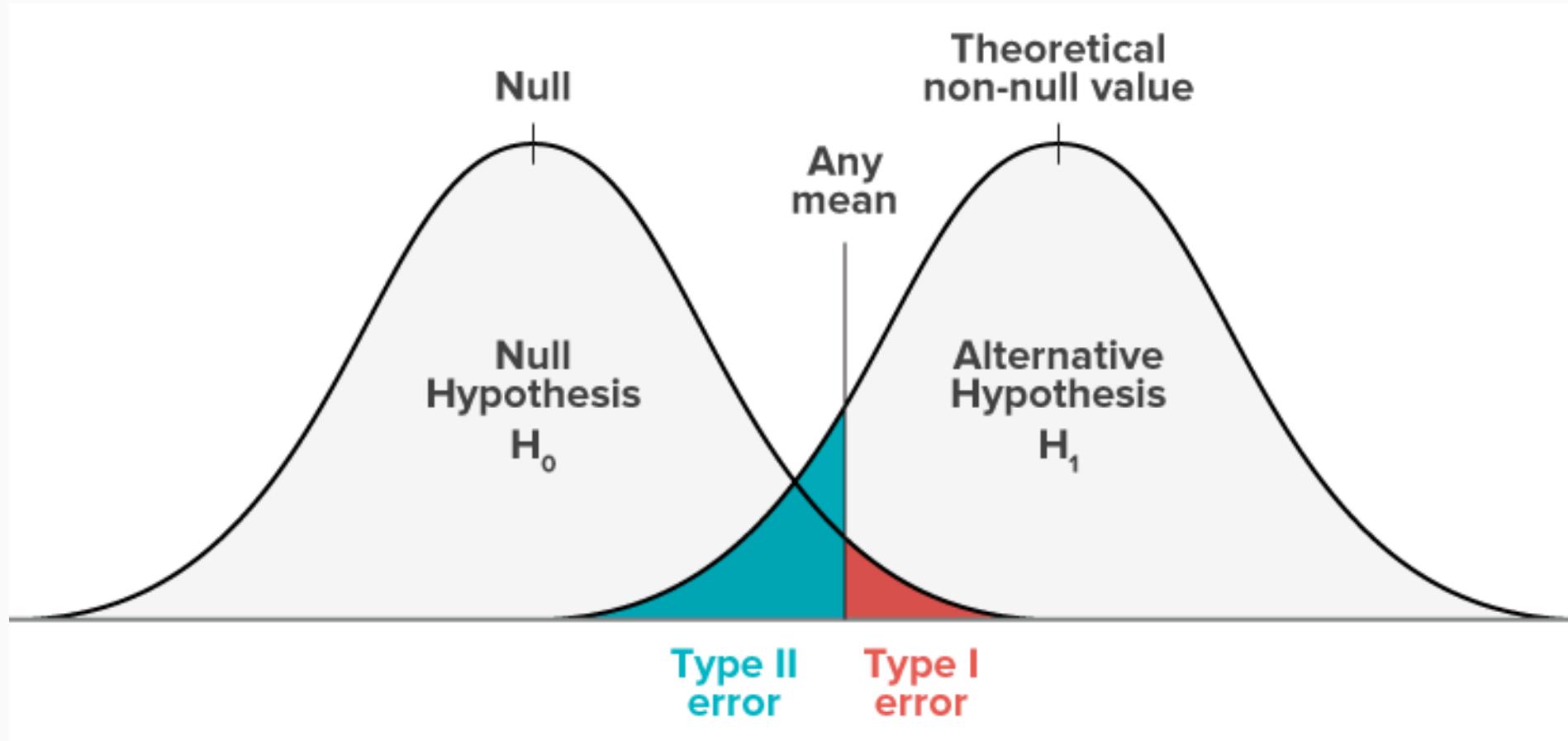
Type II error
(false negative)



Fehlertypen im Hypothesentesten

		The Truth (Based on Entire Population)	
		Nothing Is There (H_0 Is True)	Something Is There (H_0 Is False)
Your Conclusion (Based on Your Sample)	I Don't See Anything (Nonsignificant)	Right!	Wrong (Type II Error)
	I See Something (Significant)	Wrong (Type I Error)	Right!

Fehlertypen im Hypothesentesten



Statistische Signifikanz vs. praktische Signifikanz

- Sie sind nicht dasselbe.
- Bei der statistischen Signifikanz geht es um die Wahrscheinlichkeit der Beobachtung der Daten unter Berücksichtigung der Nullhypothese.
- Bei der praktischen Signifikanz geht es um die Bedeutung des Ergebnisses in der realen Welt.

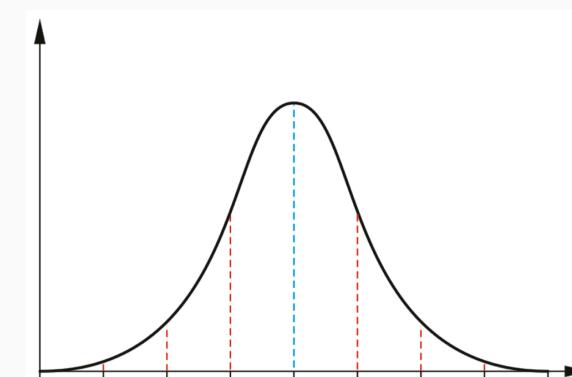
Vom Hypothesentest zur statistischen Signifikanz

Der dreistufige Ansatz:

1. Formulierung von Null- und Alternativhypotesen.
2. Berechnen Sie eine Teststatistik. Zum Beispiel die Effektgröße in einer Regression geteilt durch den Standardfehler.
3. Vergleichen Sie die Teststatistik mit einem kritischen Wert; berechnen Sie einen p-Wert.

Der p-Wert

- Der p-Wert ist die Wahrscheinlichkeit, dass ein Ergebnis beobachtet wird, das mindestens so extrem ist wie das beobachtete Ergebnis, wenn die Nullhypothese wahr wäre.
- Der p-Wert wird mit einem Schwellenwert (z.B. 0,05) verglichen, um zu entscheiden, ob die Nullhypothese verworfen werden soll.
- Wichtig ist, dass der p-Wert nicht die Wahrscheinlichkeit angibt, dass die Nullhypothese wahr oder falsch ist!



Statistische Signifikanz im Auge behalten

	Prominence	Influence
Senate	0.906*** (0.060)	1.483*** (0.067)
Sessions served	0.163*** (0.016)	0.292*** (0.017)
Party (Independent)	0.701* (0.368)	1.059** (0.412)
Party (Republican)	0.035 (0.047)	-0.080 (0.052)
Office: Governor	0.266* (0.158)	0.450** (0.177)
Office: Lt. Governor	-0.031 (0.257)	0.089 (0.288)
Office: US Secretary	0.551** (0.262)	0.372 (0.294)
Position: House Speaker	1.896*** (0.385)	2.670*** (0.431)
Position: Majority/Minority Leader	0.185 (0.308)	0.711** (0.345)
Position: Whip	0.231 (0.233)	0.848*** (0.261)
Position: Deputy Whip	0.698*** (0.234)	0.462* (0.262)
Position: Party Chairman	-0.115 (0.215)	-0.255 (0.241)
(Intercept)	1.648*** (0.050)	1.527*** (0.057)
N	492	492
R-squared	0.493	0.694
Adj. R-squared	0.481	0.687
Residual Std. Error (df = 479)	0.505	0.565
F Statistic (df = 12; 479)	38.890***	90.715***

*** p < .01; ** p < .05; * p < .1

Table 1. Summary statistics

Variable	South (1)	North (2)	Difference in means (3)	Adjusted difference in means (4)	P value (5)
Panel 1: Air pollution exposure at China's Disease Surveillance Points					
TSPs, $\mu\text{g}/\text{m}^3$	354.7	551.6	196.8***	199.5***	<0.001/0.002
SO ₂ , $\mu\text{g}/\text{m}^3$	91.2	94.5	3.4	-3.1	0.812/0.903
NO _x , $\mu\text{g}/\text{m}^3$	37.9	50.2	12.3***	-4.3	<0.001/0.468
Panel 2: Climate at the Disease Surveillance Points					
Heating degree days	2,876	6,220	3,344***	482	<0.001/0.262
Cooling degree days	2,050	1,141	-910***	-183	<0.001/0.371
Panel 3: Demographic features of China's Disease Surveillance Points					
Years of education	7.23	7.57	0.34	-0.65	0.187/0.171
Share in manufacturing	0.14	0.11	-0.03	-0.15***	0.202/0.002
Share minority	0.11	0.05	-0.05	0.04	0.132/0.443
Share urban	0.42	0.42	0.00	-0.20*	0.999/0.088
Share tap water	0.50	0.51	0.02	-0.32**	0.821/0.035
Rural, poor	0.21	0.23	0.01	-0.33*	0.879/0.09
Rural, average income	0.34	0.33	0.00	0.24	0.979/0.308
Rural, high income	0.21	0.19	-0.02	0.27	0.772/0.141
Urban site	0.24	0.25	0.01	-0.19	0.859/0.241
Predicted life expectancy	74.0	75.5	1.54***	-0.24	<0.001/0.811
Actual life expectancy	74.0	75.5	1.55	-5.04**	0.158/0.044

The sample ($n = 125$) is restricted to DSP locations within 150 km of an air quality monitoring station. TSP ($\mu\text{g}/\text{m}^3$) in the years 1981–2000 before the DSP period is used to calculate city-specific averages. Degree days are the deviation of each day's average temperature from 65°F, averaged over the years 1981–2000 before the DSP period. The results in column (4) are adjusted for a cubic in degrees of latitude north of the Huai River boundary. Predicted life expectancy is calculated by OLS using all of the demographic and meteorological covariates shown. All results are weighted by the population at the DSP location. One DSP location is excluded due to invalid mortality data. *Significant at 10%, **significant at 5%, ***significant at 1%. Sources: China Disease Surveillance Points (1991–2000), *China Environment Yearbook* (1981–2000), and World Meteorological Association (1980–2000).

Statistische Signifikanz im Auge behalten

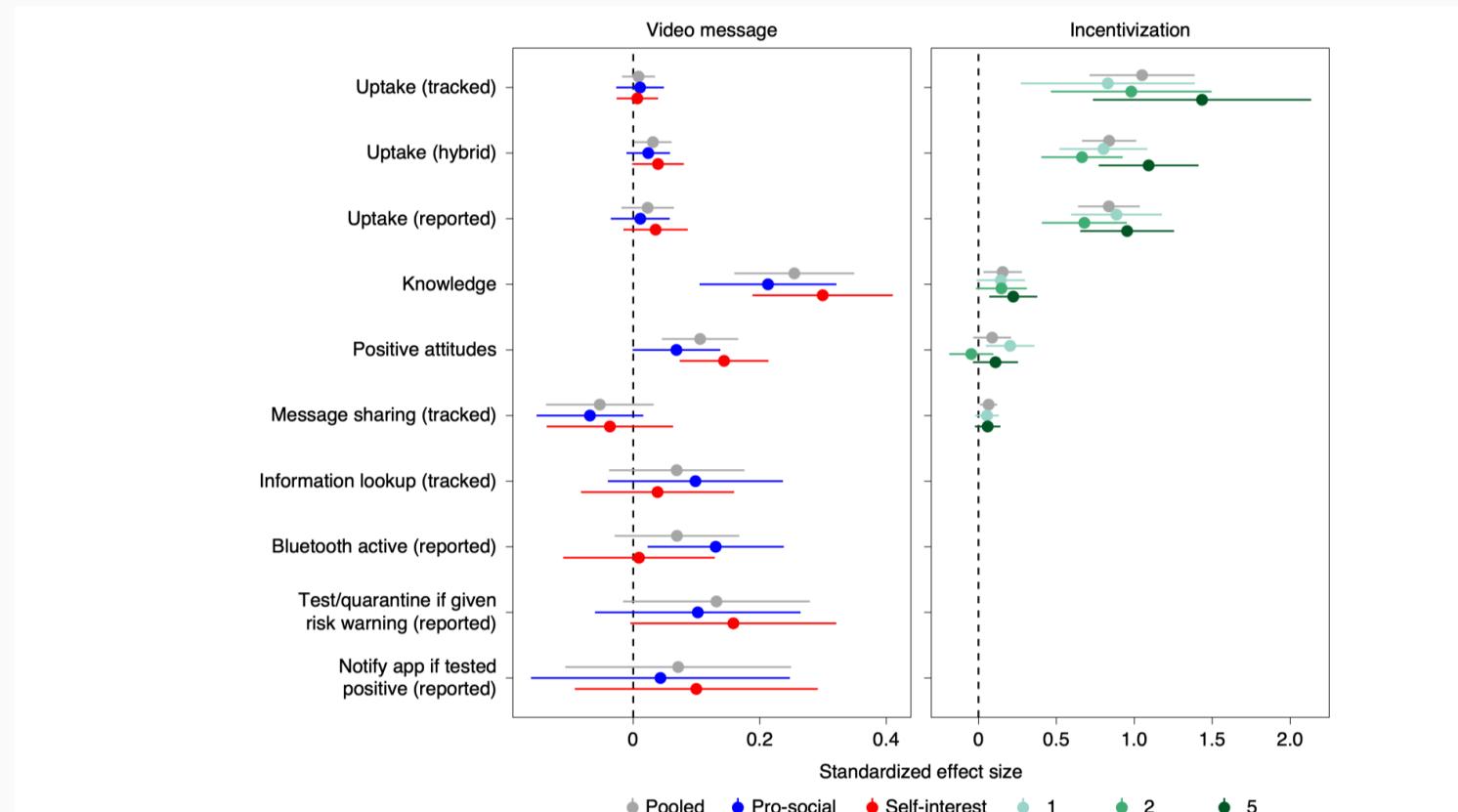


Fig. 3 | Effect of message and incentive treatments on uptake, knowledge, attitudes and behaviour. Each plot shows standardized ITT estimates with 95% CIs from fully saturated ordinary least squares regression models fit using the pre-registered LASSO covariate selection procedure. The video message sample comprises $n=2,044, 1,356$ and $1,337$ respondents for estimation of the pooled, pro-social and self-interest treatment effects, respectively. The incentive sample comprises $n=1,015, 513, 516$ and 494 respondents for estimation of the pooled, €1, €2 and €5 treatment effects, respectively.

Kontroversen um statistische Signifikanz¹

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
<http://dx.doi.org/10.1080/00031305.2016.1154108>

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Sechs Prinzipien

1. *p*-Werte können angeben, wie unvereinbar die Daten mit einem bestimmten statistischen Modell sind.
2. *p*-Werte messen nicht die Wahrscheinlichkeit, dass die untersuchte Hypothese wahr ist, oder die Wahrscheinlichkeit, dass die Daten allein durch Zufall entstanden sind.
3. Wissenschaftliche Schlussfolgerungen und geschäftliche oder politische Entscheidungen sollten nicht nur darauf beruhen, ob ein *p*-Wert einen bestimmten Schwellenwert überschreitet.
4. Eine korrekte Schlussfolgerung erfordert eine vollständige Berichterstattung und Transparenz.
5. Ein *p*-Wert sagt nichts über die Größe eines Effekts oder die Bedeutung eines Ergebnisses aus.
6. Ein *p*-Wert an sich ist kein guter Maßstab für die Evidenz eines Modells oder einer Hypothese.

¹See also [here](#) for a nice primer to this controversy.

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland¹ · Stephen J. Senn² · Kenneth J. Rothman³ · John B. Carlin⁴ · Charles Poole⁵ · Steven N. Goodman⁶ · Douglas G. Altman⁷

Received: 9 April 2016/Accepted: 9 April 2016/Published online: 21 May 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific

Editor's note This article has been published online as supplementary material with an article of Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process and purpose. *The American Statistician* 2016.

literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P* values they produce) can lead to small *P* values even if the declared test hypothesis is correct, and can lead to large *P* values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P* values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

Statistiken verwenden: gelernte Lektionen

Lügen, verdammt Lügen und Statistiken

- Politische Debatten drehen sich fast zwangsläufig auch um Statistiken
- Strategischer Anreiz, Beweise zu seinen Gunsten zu beeinflussen
- Statistische Fallstricke: Nicht alles, was logisch klingt, ist statistisch fundiert
- Ein grundlegendes Verständnis der grundlegenden Konzepte der Statistik ist der Schlüssel zu einem kritischen Verbraucher statistischer Informationen
- Einige populäre Irrtümer und Fehler treten immer wieder auf - lernen Sie, sie zu erkennen!

