

Statistik erklärt: P-Wert, Korrelation & Induktion

Autor: Wolfram Klatt

Veröffentlicht am: 10.08.2026

Dokument generiert über Gesundheitsmythen.de

ZUSAMMENFASSUNG (ABSTRACT)

Die Statistik ist das Werkzeug der Wissenschaft zur Quantifizierung von Unsicherheit. Sie unterscheidet Signale von statistischem Rauschen durch Hypothesentests, p-Werte und Konfidenzintervalle. Ein fehlendes Verständnis begünstigt Irrtümer wie den Verwechseln von Korrelation und Kausalität oder das Hereinfallen auf p-Hacking.

Statistik: Die Mathematik der Erkenntnis im Rauschen der Daten

Die Welt, in der wir leben, ist komplex, dynamisch und voller natürlicher Variabilität. Kein Mensch gleicht biologisch exakt dem anderen, kein Medikament wirkt bei jedem Patienten absolut identisch, und Körperfunktionen wie der Blutdruck oder der Blutzuckerspiegel schwanken von Minute zu Minute. Wie können wir in einem derart vielschichtigen, von individuellen Faktoren geprägten Umfeld überhaupt belastbare, medizinische Erkenntnisse gewinnen? Die Antwort darauf liefert eine Disziplin, die mitunter als abstrakt empfunden wird, in Wahrheit aber das verlässlichste Werkzeug der menschlichen Vernunft ist: die Statistik.

1. Definition und Grundlagen: Das Ordnen des Chaos

Statistik ist weitaus mehr als das bloße Sammeln von Zahlen in Tabellen oder das Zeichnen von Diagrammen. Im wissenschaftlichen Sinne ist die Statistik die mathematische Lehre vom systematischen Umgang mit quantitativen Daten und, was noch wichtiger ist, mit der *Unsicherheit*. Sie ist die formale Methode, um aus Beobachtungen verlässliche Informationen abzuleiten, bedeutsame Muster von reinem Zufall zu unterscheiden und objektive Schlussfolgerungen über die Realität zu ziehen.

Man unterteilt die Disziplin grundlegend in zwei große Bereiche, die Hand in Hand arbeiten:

Die deskriptive (beschreibende) Statistik

Sie fasst vorhandene Datenmengen zusammen und macht sie übersichtlich. Wenn wir beispielsweise das Durchschnittsalter aller Patienten in einer Klinik berechnen oder die Streuung von Blutdruckwerten in einer Grafik visualisieren, beschreiben wir objektiv, was *in dieser konkreten Gruppe* vorliegt. Wichtige Kennzahlen sind hierbei der Mittelwert (Durchschnitt) und die Standardabweichung (ein Maß dafür, wie stark die einzelnen Werte um den Durchschnitt herum streuen).

Die induktive (schließende) Statistik

Die eigentliche Erkenntnisarbeit passiert in der induktiven Statistik. Hierbei nutzt man die Daten einer begrenzten Stichprobe (z. B. 1.000 Testpersonen), um Rückschlüsse auf eine viel größere Grundgesamtheit zu ziehen (z. B. alle Menschen, die an einer bestimmten Krankheit leiden). Da wir in der Medizin praktisch nie alle betroffenen Menschen gleichzeitig testen können, bleibt stets eine Restunsicherheit. Die induktive Statistik erlaubt es uns, genau diese Unsicherheit mathematisch exakt zu quantifizieren. Sie beantwortet die essenzielle Frage: "Wie wahrscheinlich ist es, dass das Ergebnis, das wir in unserer Stichprobe beobachtet haben, durch reinen Zufall zustande kam?"

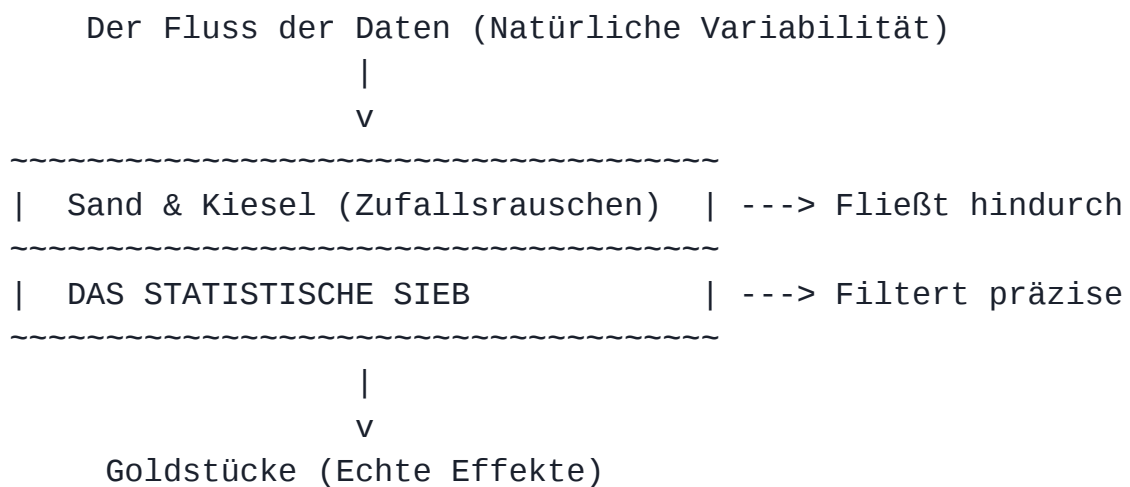
2. Die Metapher vom Sieb im Fluss: Signale im Rauschen finden

Um die Funktion der Statistik intuitiv zu verstehen, hilft ein gedankliches Bild:

Stellen Sie sich vor, Sie stehen in einem wilden Fluss. Dieser Fluss steht symbolisch für die natürliche Variabilität des Lebens und all die Schwankungen unserer Körperfunktionen. Sie suchen nach kleinen Goldstücken im Wasser – diese repräsentieren den echten medizinischen Effekt eines neuen Medikaments. Das Wasser ist jedoch trüb, und unzählige Sandkörner und Kieselsteine strömen hindurch. Dieser Sand ist das sogenannte "statistische Rauschen", also zufällige Schwankungen, die keine tiefere Bedeutung haben.

Ohne ein geeignetes Werkzeug werden Sie in das trübe Wasser greifen und vielleicht hin und wieder einen glänzenden Kieselstein herausfischen, den Sie im ersten Moment fälschlicherweise für Gold halten. In der Forschung nennt man dies ein falsch-positives Ergebnis.

Die Statistik fungiert in diesem Bild als ein hochpräzises, feinmaschiges Goldschürfsieb. Es ist exakt auf die Eigenschaften der gesuchten Goldstücke kalibriert.



Das statistische Sieb lässt den wertlosen Sand passieren und fängt zuverlässig die echten Erkenntnisse auf. Je feiner und besser die statistischen Methoden angewendet werden, desto verlässlicher trennen wir das echte Signal vom Rauschen der Daten.

3. Der naturwissenschaftliche Mechanismus:

Hypothesen auf dem Prüfstand

Um zu verstehen, wie evidenzbasierte Medizin heute funktioniert, muss man sich die Werkzeuge der schließenden Statistik genauer ansehen. Das Herzstück der medizinischen Forschung ist das **Testen von Hypothesen**.

Nehmen wir an, ein Forschungsteam hat einen neuen Wirkstoff zur Senkung des Blutdrucks entwickelt. Es möchte herausfinden, ob dieser Wirkstoff tatsächlich besser funktioniert als ein Scheinmedikament (Placebo).

In der Statistik beginnt man diesen Prozess stets mit der sogenannten **Nullhypothese (H0)**. Diese lautet aus Gründen der Vorsicht und Objektivität fast immer: *"Es gibt keinen echten Unterschied zwischen dem Wirkstoff und dem Placebo. Jegliche gemessene Abweichung ist das Resultat zufälliger Schwankungen."* Dem gegenüber steht die **Alternativhypothese (H1)**: *"Der Wirkstoff senkt den Blutdruck messbar besser als das Placebo."*

[Forschungsfrage formulieren]

|
v

+-----+
| Nullhypothese (H0) |
| "Kein echter Effekt" |
+-----+

+-----+
| Alternativhypothese (H1) |
| "Es gibt einen Effekt" |
+-----+

\
v

/
v

[Studiendurchführung & Datensammlung]

|
v

[p-Wert-Berechnung]

|

+-----+

|
v

$p < 0,05$

|
v

[H0 verwerfen: Effekt]

|
v

$p \geq 0,05$

|
v

[H0 beibehalten: Kein Nachweis]

Der p-Wert (Probability Value)

Die Wissenschaftler verabreichen Gruppe A den Wirkstoff und Gruppe B das Placebo. Am Ende der Studie stellen sie fest: Gruppe A hat im Durchschnitt einen etwas niedrigeren Blutdruck. Doch reicht das als Beweis? Blutdruck schwankt natürlicherweise stark. Vielleicht waren in Gruppe A zufällig Menschen, die an diesem Tag ohnehin entspannter waren.

Hier greift der berühmte p-Wert. Er gibt an, wie wahrscheinlich es ist, genau diese (oder noch extremere) Daten zu beobachten, *unter der strikten Annahme, dass die Nullhypothese wahr ist* (dass das Medikament also in Wirklichkeit gar nicht wirkt). Ist der p-Wert extrem klein – in der Medizin gilt meist ein Schwellenwert von $p < 0,05$ (kleiner als 5 Prozent) –, folgern die Statistiker: Es ist höchst unwahrscheinlich, dass wir einen derart großen Unterschied im Blutdruck rein zufällig messen würden, wenn das Medikament keine Wirkung hätte. Daraufhin wird die Nullhypothese verworfen. Das Ergebnis gilt als "statistisch signifikant".

Fehler erster und zweiter Art

Statistik bietet keine absolute Gewissheit, sondern berechnet Wahrscheinlichkeiten. Daher kann es zu zwei Arten von Irrtümern kommen:

- **Typ-I-Fehler (Alpha-Fehler):** Ein falsch-positives Ergebnis. Man behauptet, ein Medikament wirkt, obwohl der gemessene Unterschied nur reiner Zufall war. (Man schlägt Alarm, obwohl keine Gefahr droht).
- **Typ-II-Fehler (Beta-Fehler):** Ein falsch-negatives Ergebnis. Das Medikament wirkt tatsächlich, aber die Studie konnte den Effekt nicht nachweisen, vielleicht weil zu wenige Menschen teilgenommen haben. (Man übersieht einen tatsächlichen Effekt).

Die Grenze von 5 % beim p-Wert ist ein Kompromiss, um den Typ-I-Fehler so gering wie möglich zu halten, ohne den Typ-II-Fehler ausufern zu lassen.

Konfidenzintervalle und Stichprobengröße

Ein signifikantes Ergebnis allein sagt noch nichts darüber aus, *wie stark* die Wirkung ist. Hierfür nutzt man das **Konfidenzintervall**. Ein 95%-Konfidenzintervall bedeutet vereinfacht ausgedrückt: Wenn wir das Experiment 100 Mal exakt gleich wiederholen würden, läge der wahre Effekt in 95 Fällen innerhalb dieses berechneten Bereichs. Es vermittelt uns eine klare Vorstellung davon, mit welcher Schwankungsbreite der Effekt in der echten Welt zu erwarten ist.

Entscheidend für all diese Berechnungen ist die **Stichprobengröße (n)**. Je größer die Datenmenge, desto mehr gleichen sich zufällige Schwankungen aus. Das Gesetz der großen Zahlen besagt, dass bei genügend Teilnehmern das Rauschen verstummt und das wahre

biologische Signal deutlich zutage tritt.

4. Häufige Missverständnisse und menschliche Wahrnehmung

Dass statistische Konzepte im Alltag oft missverstanden werden, liegt nicht an mangelnder Intelligenz, sondern an der Art und Weise, wie unser menschliches Gehirn evolutionär geformt wurde. Wir sind darauf programmiert, in unserer Umwelt schnell Kausalitäten zu erkennen und Muster zu finden, selbst dort, wo keine sind.

Korrelation vs. Kausalität

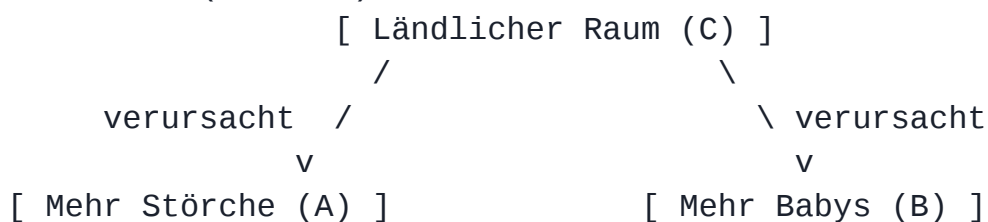
Das wohl häufigste Missverständnis lautet: *Korrelation bedeutet Kausalität*. Wenn zwei Entwicklungen parallel verlaufen (Korrelation), bedeutet das keineswegs zwingend, dass die eine Entwicklung die Ursache für die andere ist (Kausalität).

Das klassische Lehrbuchbeispiel ist der Zusammenhang zwischen Störchen und der Geburtenrate. In europäischen Regionen, in denen viele Störche nisten, gibt es oft auch eine höhere Anzahl an Neugeborenen. Bedeutet das, dass der Storch die Babys bringt? Natürlich nicht. Beide Faktoren werden von einer dritten, oft übersehenen Variable beeinflusst (dem sogenannten Confounder): dem ländlichen Raum.

Fehlschluss:

[Mehr Störche (A)] ----- verursacht? -----> [Mehr Babys (B)]

Die tatsächliche (kausale) Realität:



Der ländliche Raum bietet viel Lebensraum für Störche und beherbergt oft Familien, die traditionell mehr Kinder bekommen. Solche Scheinkorrelationen existieren in der Medizin und im Alltag zuhauf und erfordern sorgfältige statistische Bereinigung, um Fehlinterpretationen zu vermeiden.

Die Macht der Anekdote

Ein weiterer Faktor ist unsere Vorliebe für persönliche Geschichten. Wenn wir krank sind und leiden, suchen wir verständlicherweise nach Heilung, nach Sinn und nach Geschichten, die uns Hoffnung spenden. Es ist zutiefst menschlich, dem Erfahrungsbericht einer vertrauten Person ("Mir ging es schlecht, ich habe dieses neue Mittel ausprobiert, und am nächsten Tag war ich gesund!") eine enorme Bedeutung beizumessen.

Unser Gehirn reagiert stark auf solch emotionale Einzelerfahrungen (Anekdoten, $N=1$). Für die Wissenschaft jedoch muss eine Anekdote neutral betrachtet werden, da individuelle Besserungen durch den natürlichen Krankheitsverlauf, das Immunsystem oder den beruhigenden Placebo-Effekt zustande kommen können. Statistik fordert uns auf, einen Schritt zurückzutreten und den breiten Horizont zehntausender Datenpunkte zu betrachten – ein oft schwieriger, aber notwendiger Schritt zur objektiven Wahrheitsfindung.

5. Fehlinterpretationen und Science-Washing

Da wissenschaftliche Publikationen in unserer Gesellschaft ein hohes Ansehen genießen, versuchen mitunter auch Anbieter von Verfahren ohne gesicherten Wirkungsnachweis, ihren Produkten einen Anstrich von Wissenschaftlichkeit zu verleihen ("Science-Washing"). Oft geschieht dies gar nicht aus böser Absicht, sondern aus fehlender statistischer Methodenkenntnis. Wenn Studien und Statistiken fehlerhaft interpretiert werden, entstehen falsche Heilsversprechen.

Folgende statistische Verzerrungen kommen besonders häufig vor:

1. P-Hacking (Data Dredging) Bei einer Untersuchung werden sehr viele verschiedene Parameter gemessen (z.B. Blutdruck, Müdigkeit, Hautbild, Laune, Schlafqualität, Nierenwerte). Wenn man zwanzig Parameter testet, steigt die Wahrscheinlichkeit extrem an, dass allein durch Zufall ein Parameter den p-Wert von unter 0,05 erreicht – völlig unabhängig davon, ob das untersuchte Mittel wirklich wirkt. Werden dann nur die "erfolgreichen" Ergebnisse publiziert und die negativen verschwiegen, entsteht ein verzerrtes Bild. In der seriösen Forschung müssen Endpunkte daher *vorab* festgelegt werden.

2. Rosinenpickerei (Cherry Picking) Es existieren oft Dutzende Studien zu einem Thema. Wenn systematisch nur jene wenigen Untersuchungen hervorgehoben werden, die einen

positiven Effekt suggerieren, während die überwiegende Mehrheit der hochwertigen Studien ignoriert wird, entsteht ein falscher Gesamteindruck. Solide Evidenz basiert stets auf der Zusammenfassung *aller* verfügbaren Daten, beispielsweise in sogenannten Meta-Analysen.

3. Zu kleine Stichprobengrößen (Small n) Ergebnisse aus Beobachtungen mit nur 10 oder 15 Teilnehmern sind statistisch betrachtet kaum aussagekräftig (sie haben eine zu geringe "Power"). Bei so wenigen Personen können zufällige Schwankungen das Ergebnis enorm verfälschen. Belastbare medizinische Empfehlungen erfordern in der Regel deutlich größere Kohorten.

4. HARKing (Hypothesizing After the Results are Known) Hierbei wird zuerst in Daten gesucht, bis sich ein auffälliges Muster zeigt. Erst danach wird eine vermeintlich logische Erklärung dafür konstruiert und so getan, als hätte man dieses Ergebnis genau so vorhergesagt. Dies verkehrt den wissenschaftlichen Erkenntnisprozess ins Gegenteil.

Ein grundlegendes Verständnis dieser Mechanismen hilft immens dabei, Gesundheitsversprechen im Alltag besser einordnen zu können und wohlinformierte Entscheidungen für das eigene Wohlbefinden zu treffen.

6. Faszination Wissenschaft: Das Theorem des fortlaufenden Lernens

Die wahre Eleganz der Statistik liegt in ihrer Fähigkeit, unser Verständnis der Welt schrittweise zu verfeinern. Ein besonders faszinierendes Konzept hierbei ist die **Bayessche Statistik**, benannt nach dem englischen Mathematiker und Pfarrer Thomas Bayes.

Während klassische (frequentistische) Statistik oft harte "Ja/Nein"-Grenzen wie den p-Wert zieht, funktioniert die Bayessche Statistik eher wie ein offenes, lernendes Gehirn. Sie kombiniert unsere bestehende, bisherige Einschätzung (die "A-priori-Wahrscheinlichkeit") mit neuen Daten, um zu einer aktualisierten, besseren Einschätzung (der "A-posteriori-Wahrscheinlichkeit") zu gelangen.

Ein Beispiel aus der Diagnostik: Angenommen, eine extrem seltene Krankheit betrifft nur einen von 100.000 Menschen. Ein Screening-Test auf diese Krankheit ist zu 99 % zuverlässig. Wenn eine Person nun positiv getestet wird, ist die Wahrscheinlichkeit, dass sie die Krankheit *tatsächlich* hat, immer noch überraschend gering. Warum? Weil die Krankheit so extrem selten ist, dass selbst ein 1%iges Falsch-Positiv-Risiko bei Zehntausenden getesteten gesunden

Menschen mehr falsche Alarmerzeugt, als es echte kranke Menschen gibt. Die Bayessche Statistik bezieht dieses Vorwissen um die Seltenheit der Krankheit mathematisch korrekt ein.

Das Prinzip gilt auch für die Forschungsgemeinschaft: Wenn eine völlig neue These aufgestellt wird, die etabliertem medizinischem und physikalischem Wissen stark widerspricht, ist ihre anfängliche Wahrscheinlichkeit gering. Eine einzige, kleine Studie reicht dann nicht aus, um unser gesamtes medizinisches Weltbild umzuwerfen – "Außergewöhnliche Behauptungen erfordern außergewöhnliche Beweise". Werden jedoch im Laufe der Jahre immer mehr hochwertige Daten gesammelt, passt die Wissenschaft (wie das Theorem von Bayes) ihre Überzeugung fließend an die neue Realität an.

Statistik ist somit nicht bloß ein trockenes Regelwerk. Sie ist das universelle Instrument, mit dem die Menschheit gelernt hat, das leise Flüstern der Wahrheit im unaufhörlichen Lärm der Welt zu hören. Sie schützt Patienten vor unwirksamen Behandlungen, hilft uns, die Wirksamkeit von Therapien verlässlich zu sichern, und ermöglicht es uns, fundierte, angstfreie Entscheidungen für unsere eigene Gesundheit zu treffen. Es ist die Mathematik, die Mitgefühl und Fürsorge erst sicher und verlässlich macht.