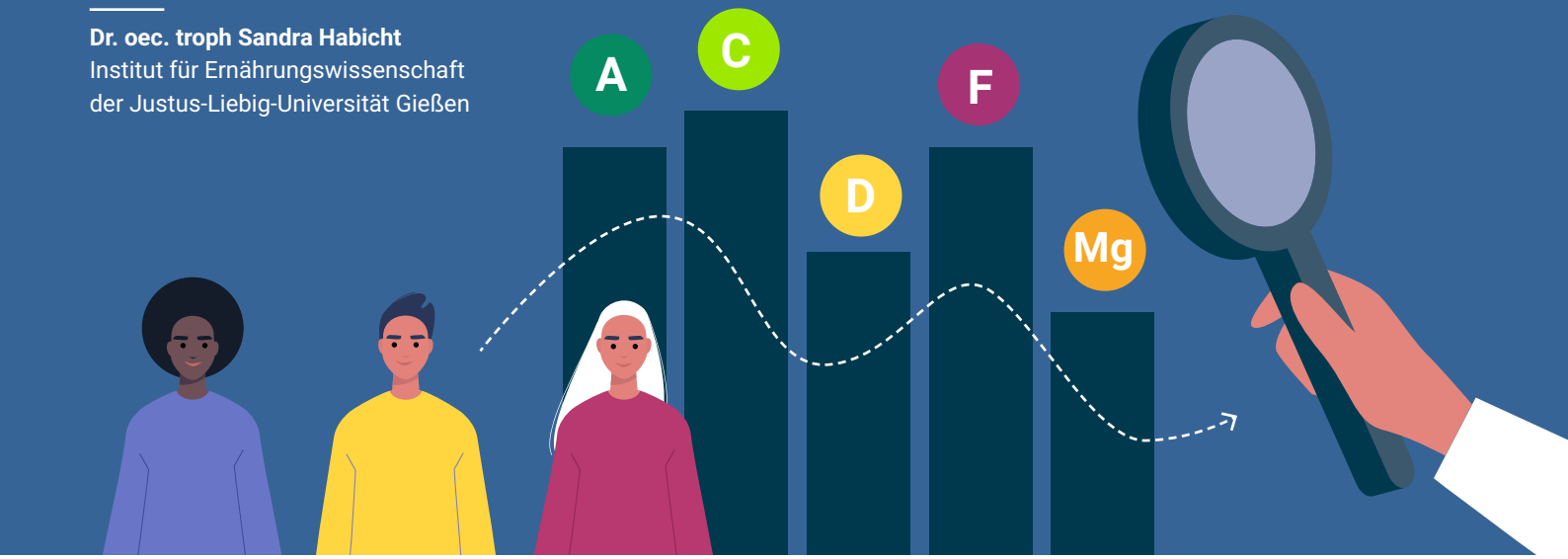
FOKUS WISSENSCHAFT

ERNÄHRUNGSFORSCHUNG INTERPRETIEREN

Dr. oec. troph Sandra Habicht
Institut für Ernährungswissenschaft
der Justus-Liebig-Universität Gießen



Deskriptive und analytische Statistik

Was sagen mir statistische Begriffe über die Studie und ihre Ergebnisse?

Ernährungsbezogene Fragestellungen in der Forschung und entsprechende Studiendesigns sind in den letzten Jahrzehnten immer komplexer geworden. Je komplexer die Fragestellungen sind, desto aufwendiger ist möglicherweise die statistische Auswertung der erhobenen Daten. Damit einher geht auch eine entsprechend ausdifferenziertere Darstellung der Studien und ihrer Ergebnisse, was an diejenigen, die Publikationen korrekt lesen und verstehen wollen, immer höhere Ansprüche stellt. Sich mit Auswertungsmethoden und den wichtigsten Begrifflichkeiten aus der Statistik auszukennen, ist daher wichtig. In der ersten Ausgabe von FOKUS WISSENSCHAFT wurden verschiedene Studientypen der Ernährungsforschung vorgestellt. Dieser zweite Teil erläutert häufig verwendete Begriffe aus der Statistik sowie Darstellungen, damit Lesende von Ernährungsstudien diese richtig einordnen können.

Repräsentativ: Für was und für was nicht?

Das übergeordnete Ziel aller statistischen Analysen ist es, Erkenntnisse, die mithilfe der jeweiligen Studie und den/der dafür ausgewählten Probandinnen und Probanden/Stichprobe gewonnen wurden, auf die allgemeine Bevölkerung (die Grundgesamtheit), eine bestimmte Risikopopulation oder eine Patientinnen-/Patientengruppe mit einer bestimmten Erkrankung zu übertragen. Um auf die Gesamtbevölkerung schließen zu können, werden repräsentative Stichproben benötigt; das heißt, die untersuchte Stichprobe entspricht bezogen auf relevante Merkmale wie Alter, Geschlecht oder Einkommen der Gesamtbevölkerung. Mit repräsentativen Studien können beispielsweise Daten oder Einstellungen zu Ernährung und Gesundheit oder zum Konsum von Nahrungsergänzungsmitteln erhoben werden. Zur Erforschung von Erkrankungsursachen sowie Präventionsmaßnahmen oder Handlungsmöglichkeiten werden häufig keine Stichproben benötigt, die repräsentativ für die gesamte

Inhalt

→ Mehr auf Seite 2

Bevölkerung sind. Die Erkenntnisse lassen sich jedoch auf Personen mit ähnlichen Merkmalen wie die der Studienpopulation übertragen. Deshalb ist es wichtig, dass mit den Studienergebnissen auch die Merkmale des Studienkollektivs dargestellt werden. Zum einen werden sie als Ein- und Ausschlusskriterien definiert und beschrieben, zum anderen wird die Stichprobe mithilfe der sogenannten deskriptiven Statistik beispielsweise anhand des Durchschnittsalters oder der Geschlechtsverteilung der Probandinnen und Probanden charakterisiert. Daran anknüpfende statistische Methoden sollen die eigentliche wissenschaftliche Fragestellung der Studie beantworten und umfassen unter anderem die Beschreibung von Zusammenhängen zwischen bestimmten Parametern oder Vergleiche von Gruppen.

Deskriptive Statistik: Mehr als nur die Beschreibung der (Studien-)Population

Zur deskriptiven Statistik gehören Häufigkeiten, aber auch berechnete Größen wie Mittelwert, Varianz, Standardabweichung, Median und Interquartilsabstand. Mittelwert und Median sind dabei Maßzahlen, mit denen Aussagen über den Durchschnitt bzw. die „Mitte" der Stichprobe getroffen werden können. Um ein Bild von der Verteilung der Daten zu erhalten, werden Mittelwert und Median mit den jeweils passenden Maßen für die Streuung der Daten angegeben. Die Streuung gibt an, wie stark die einzelnen Werte der Stichprobe von der Mitte abweichen.

Der **Mittelwert** bildet den Durchschnitt der Werte einer Population/Probandinnen- und Probandengruppe zu einem Zeitpunkt. Passende Streumaße für den Mittelwert sind die Varianz oder die Standardabweichung, wobei die Standardabweichung als die Wurzel aus der Varianz berechnet wird. Die Standardabweichung wird in Studien häufiger angegeben. Mittelwert und Standardabweichung (bzw. Varianz) sind gut geeignet, Daten abzubilden, die normalverteilt sind. **Normalverteilte Daten** sind zum Beispiel die Körpergröße von Erwachsenen, weil es jeweils wenig Kleine und Große, aber viele mit mittleren Größen gibt (*siehe Abbildung 1*). Im Vergleich von Gruppen oder Zeitpunkten werden Mittelwert und Standardabweichung graphisch häufig als Balkendiagramm mit Fehlerbalken dargestellt (*siehe Abbildung 2, S. 3*).

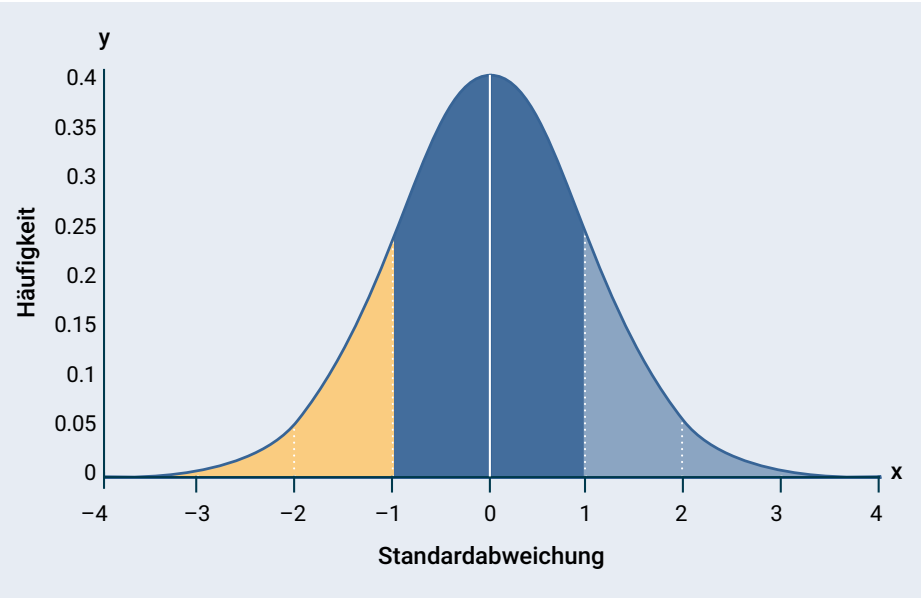


Abbildung 1:
Schematische Darstellung normalverteilter Daten als Gaußsche Kurve.

Nicht immer aber sind die Daten einer Stichprobe normalverteilt. Dies kann unterschiedliche Ursachen haben. Bei Daten, die nicht normalverteilt sind, wird der **Median** als Maß für die Mitte angegeben und stellt den Wert dar, unter und über welchem jeweils 50 Prozent der Werte einer bestimmten Population/Probandinnen- und Probandengruppe zu einem Zeitpunkt liegen. Der Median ist also der Wert in der Mitte der Datenreihe und ist rechnerisch das Gleiche wie die 50. Perzentile einer Stichprobe (*siehe auch S. 6*).

Der **Interquartilsabstand** (*engl. interquartile range, IQR*) ist in Kombination mit dem Median ein gutes Maß für die Streuung. Dieser liegt zwischen der 25. und der 75. Perzentile (auch als Q25 und Q75 bezeichnet) und bildet die mittleren 50 Prozent der Werte einer Stichprobe ab.

Median und Interquartilsabstand bilden auch die Grundlage für die graphische Darstellung der Daten als **Boxplot** *(siehe Abbildung 2, rechts)*. Werte, die mehr als die 1,5-fache Länge des Interquartilsabstandes von diesem nach oben oder unten abweichen, werden als mögliche Ausreißer betrachtet. Als zusätzliches Maß zur Verteilung der Daten werden beim Boxplot deshalb häufig der niedrigste bzw. höchste Wert der Datenreihe, der nicht mehr als die 1,5-fache Länge des Interquartilsabstandes von diesem abweicht, als sogenannte Whisker *(engl. für Schnurrhaar)* abgebildet.

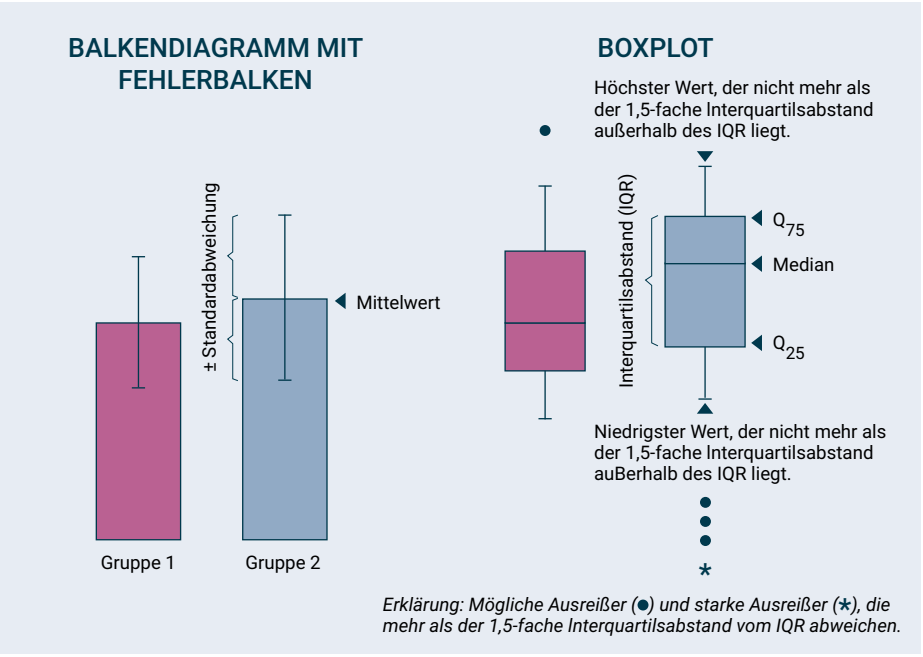


Abbildung 2: Schematische Darstellung von Balkendiagramm und Boxplot und welche statistischen Größen jeweils abgelesen werden können.

Deskriptive und analytische Statistik in Beobachtungsstudien

Neben Mittelwert, Median und Streuungsmaßen spielen Häufigkeiten, Prävalenzen und Inzidenzen in deskriptiven Beobachtungsstudien (auch Querschnittsstudien oder Prävalenzstudien) eine große Rolle.

Die **Prävalenz** ist eine Kennzahl, mit der z. B. Erkrankungshäufigkeiten, aber auch das Vorliegen von bestimmten Lebensstilfaktoren wie der Versorgungsstatus oder Vegetarismus in der Stichprobe oder der Gesamtpopulation dargestellt werden können. Sie wird häufig in Prozent angegeben. Die Prävalenz bezeichnet den Anteil der Stichprobe mit einem gesuchten Merkmal oder einer bestimmten Erkrankung an der beobachteten Gesamtstichprobe.

Zu den deskriptiven Beobachtungsstudien gehören im weiteren Sinne auch repräsentative Längsschnittstudien. Diese haben eine lange Beobachtungszeit der Probandinnen und Probanden, um zu erheben, wie viele der zu einer definierten Risikopopulation gehörenden und zunächst gesunden Personen eine zu untersuchende Erkrankung entwickeln. Risikopopulation bedeutet in diesem Fall beispielsweise, dass nur Frauen hinsichtlich der Entwicklung von Gebärmutterhalskrebs untersucht werden und keine Männer, da diese nicht erkranken können. Die Neuerkrankungsrate wird als **Inzidenz** bezeichnet und unterscheidet sich von der Prävalenz. Es werden außerdem die kumulative Inzidenz und Inzidenzrate unterschieden, welche voneinander abweichen können (*siehe Infokasten unten*).

Die **kumulative Inzidenz** (*auch Inzidenzrisiko*) ist der Anteil an Personen, der in einem definierten Zeitraum eine definierte Erkrankung entwickelt. Berechnet wird die kumulative Inzidenz als Neuerkrankungsrate innerhalb eines Zeitraums im Verhältnis zur untersuchten Gesamtstichprobe (Studien- oder Risikopopulation) und wird in Bezug auf den Beobachtungszeitraum angegeben.

Die **Inzidenzrate** (*auch Inzidenzdichte*) stellt eine Alternative zur kumulativen Inzidenz dar. Es wird dabei nicht das Verhältnis der Neuerkrankungen zur gesamten Stichprobe, die zu Beginn eingeschlossen war, berechnet, sondern die Neuerkrankungsrate im ***Verhältnis zur Personenzeit unter Risiko***. Dazu werden üblicherweise die Jahre der gesunden Personen unter Beobachtung addiert; Drop-outs (ausgeschiedene Probandinnen/Probanden) oder Erkrankte haben eine entsprechend geringere Personenzeit unter Risiko *(siehe Abbildung 3, S. 5)*.

Infokasten

Prävalenz	=	Alle Erkrankten zu einem Zeitpunkt / Gesamte Stichprobe	(x 100)
Kumulative Inzidenz/ Inzidenzrisiko	=	Neuerkrankungen innerhalb eines Zeitraums / Gesamte Stichprobe	(x 100)
Inzidenzdichte/ Inzidenzrate	=	Neuerkrankungen innerhalb eines Zeitraums / Personenzeit unter Risiko	(x 100)

Als weiterführende Statistik können Häufigkeiten von Erkrankungen, Versorgungstatus oder andere Gesundheitsdaten mit weiteren Merkmalen der Personen oder des Lebensstils in Verbindung gebracht werden. Derlei **Korrelationen** können auch in deskriptiven Studien erste Assoziationen von zwei oder mehreren Parametern miteinander ermitteln. Beispielsweise können Breitengrad des Wohnortes und Häufigkeit von Osteoporose in einer Bevölkerung miteinander korreliert werden. Diese Korrelation wäre positiv, wenn beide Parameter gemeinsam steigen oder sinken. In unserem Beispiel würde dies bedeuten, dass je höher der Breitengrad des Wohnorts ist, desto eher steigt die Wahrscheinlichkeit einer Osteoporose-Erkrankung. Eine negative Korrelation bedeutet eine inverse Beziehung: je weniger, desto mehr bzw. je mehr, desto weniger. Neben dem Vorzeichen der Beziehung ist auch die Stärke des Zusammenhangs bzw. der Assoziation bedeutend. Der **Korrelationskoeffizient** kann Werte zwischen - 1 und + 1 annehmen. Ein Wert von + oder - 1 bedeutet, dass eine Variable bzw. ein Parameter zu 100 Prozent durch eine andere Variable erklärt werden kann. Ein Wert von 0 bedeutet demnach, dass kein Zusammenhang zwischen den beiden Parametern besteht. Je nach Forschungsdisziplin oder Fragestellung wird ein schwacher, mittlerer oder starker Zusammenhang beispielsweise ab einem Korrelationskoeffizienten von +/-0,3, +/-0,5 oder +/-0,7 definiert. In der Ernährungsforschung sind Korrelationskoeffizienten von + 1 oder - 1 relativ selten, da die Ernährung häufig nur einer von vielen Einflussfaktoren auf z. B. Erkrankungen ist.

Zwei Parameter, die miteinander korrelieren, also häufig bei Personen gleichzeitig auftreten, sind miteinander assoziiert, müssen aber nicht kausal zusammenhängen. Möglicherweise gibt es gemeinsame weitere Merkmale, die diese Assoziation bedingt haben. Wie mehrere Eigenschaften oder Variablen von Personen miteinander zusammenhängen und wie sich verschiedene Variablen gegenseitig bedingen, kann mit unterschiedlichen **multivariaten statistischen Testverfahren** untersucht werden. Beispielsweise kann der Vitamin-D-Status mit der Häufigkeit von Typ-2-Diabetes mellitus negativ korrelieren. Ein guter Vitamin-D-Status kann aber wiederum mit mehr Bewegung im Freien oder der Einnahme von Multivitamin-Supplementen zusammenhängen. Welchen Anteil am Schutz vor Typ-2-Diabetes mellitus der Vitamin-D-Status oder die anderen möglichen Variablen haben, kann so annähernd beleuchtet werden. Stellt sich eine Variable als starker Einflussfaktor heraus, obwohl andere Variablen in der Analyse berücksichtigt wurden, wird oft von einem sogenannten **unabhängigen Risikofaktor** oder **Präventionsfaktor** gesprochen. Das bedeutet z. B., dass die multivariate Analyse hat zeigen können, dass die Supplementierung von Vitamin D unabhängig von der Bewegung im Freien einen starken Einfluss auf das Risiko von Typ-2-Diabetes mellitus haben kann. Damit wäre dann die zusätzliche Einnahme des Vitamins ein Präventionsfaktor.

Wie können Risiko- oder Präventionsfaktoren definiert werden?

Um eine kausale Beziehung zwischen einem Lebensstil- oder Ernährungsfaktor oder dem Versorgungsstatus einerseits und einem Erkrankungsrisiko andererseits zu erforschen, können analytische Beobachtungsstudien durchgeführt werden: Kohorten-Studien bzw. Fall-Kontroll-Studien.

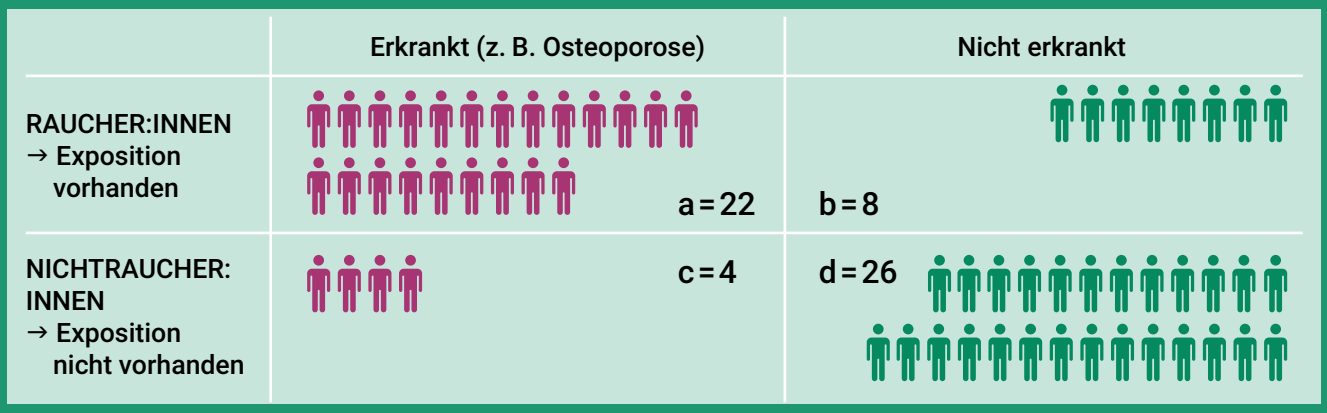
Kohorten-Studien untersuchen das Auftreten von Versorgungsproblemen bzw. Erkrankungen im Langzeitverlauf. Während allerdings in den oben beschriebenen repräsentativen/epidemiologischen Studien die Inzidenz einer Bevölkerung oder der Grundgesamtheit ein mögliches relevantes Ergebnis darstellt, ist das Ziel von Kohorten-Studien ein anderes:

In einer Kohorten-Studie werden mindestens zwei zunächst gesunde Kohorten (Probandinnen- und Probandengruppen) miteinander verglichen, die sich in mindestens einem Merkmal des Lebensstils oder des Versorgungsstatus unterscheiden (Expositionsfaktor). Wie sich die Inzidenzen einer zu untersuchenden Erkrankung zwischen den beiden Kohorten im Studienverlauf unterscheiden, ist hier das eigentliche Ergebnis.

Um **Risiko- oder Präventionsfaktoren** zu identifizieren, ist es also nicht so sehr relevant, wie viele Personen z. B. mit niedrigem Vitamin-D-Status im Verlauf der Beobachtung eine Osteoporose entwickeln, sondern wie hoch beispielsweise das Risiko verglichen mit Personen mit einem guten Status ist. Das Risiko je Kohorte ist rechnerisch die (kumulative) Inzidenz je Kohorte am Ende einer definierten Beobachtungszeit. Da bei einer Kohorten-Studie nicht nur ein Merkmal untersucht wird, sondern zwei Merkmale, bietet sich die Darstellung als **Kreuztabelle** (*auch 4-Feldertafel genannt*) an. In einer solchen Tabelle werden Häufigkeiten zweier Merkmalsausprägungen (erkrankt/nicht erkrankt und exponiert/nicht exponiert) in einer Stichprobe/Studienpopulation dargestellt, wobei mit Exposition das Vorhandensein eines bestimmten Merkmals wie Rauchen, Mikronährstoffstatus oder Vegetarismus gemeint ist. Die Kreuztabelle kann bei Kohorten-Studien für die Berechnung des Relativen Risikos helfen oder kann bei Fall-Kontroll-Studien analog der Berechnung der Odds Ratio dienen (*siehe Abbildung 3, S. 5)*.

Das **Relative Risiko** (*engl. Relative Risk, RR)* wird aus Daten einer Kohorten-Studie berechnet und ist die Inzidenz der Kohorte mit dem zu untersuchenden Merkmal/Expositionsfaktor (z. B. niedriger Vitamin-D-Status) geteilt durch die Inzidenz der Kohorte ohne dieses Merkmal.

Für alle drei Verhältnisse gilt: Wenn die OR, aber auch das RR oder die HR, gleich 1 oder nahe 1 ist und/oder das 95 %-CI die 1 umschließt, scheint der Expositionsfaktor in diesem Studiensetting keinen Einfluss auf das Erkrankungsrisiko gehabt zu haben.



Relatives Risiko (RR) = $\frac{\text{Wahrscheinlichkeit unter Exposition zu erkranken}}{\text{Wahrscheinlichkeit ohne Exposition zu erkranken}} = \frac{a : (a + b)}{c : (c + d)} = \frac{22 : (22 + 8)}{4 : (4 + 26)} = 5{,}5$

Hazard Ratio (HR) = Die HR wird analog zum RR berechnet, bezieht jedoch den zeitlichen Verlauf mit ein.

Odds Ratio (OR) = $\frac{\text{Anzahl Erkrankter unter Exposition} \times \text{nicht Erkrankte ohne Exposition}}{\text{Nicht Erkrankte unter Exposition} \times \text{Anzahl Erkrankter ohne Exposition}} = \frac{a \times d}{b \times c} = \frac{22 \times 6}{8 \times 4} = 17{,}9$

Ein RR/eine HR/eine OR von 1 bedeutet: keinen Unterschied
> 1 bedeutet: Risiko/Chance unter Exposition größer (Exposition = Risikofaktor)
< 1 bedeutet: Risiko/Chance unter Exposition kleiner (Exposition = Präventionsfaktor)

Abbildung 3: Beispiel einer 4-Feldertafel mit Formeln für RR und OR bzw. Interpretation von RR, HR und OR.

Werte können zwischen 0 und unendlich liegen. Werte > 1 bedeuten ein höheres Risiko; das Merkmal ist ein Risikofaktor. Zwischen 0 und 1 liegt das RR, wenn unter der Exposition eine Erkrankung weniger häufig auftritt, sodass die Exposition demnach als Präventionsfaktor bezeichnet werden kann. Bei einem RR von 0,5 sind unter Exposition nur halb so viele erkrankt, bei einem RR von 2 wären es doppelt so viele bzw. bei einem RR von 1,5 wären es 50 Prozent mehr. Das RR wird häufig in Kombination mit einem **p-Wert** und einem 95 %-**Konfidenzintervall** (*engl. Confidence Interval, CI*) berichtet. Das CI wird immer als Spanne angegeben. Ein bestimmter Expositionsfaktor ist ein statistisch signifikanter Risiko- oder Präventionsfaktor, wenn der p-Wert kleiner ist als ein zuvor festgelegter Wert, meist kleiner 0,05. Dies ist der Fall, wenn der Bereich des 95 %-Konfidenzintervalls nicht die 1 umschließt. Bei einem RR von 0,6 und einem 95 %-CI von [0,5; 0,87] bzw. bei einem RR von 1,7 und einem 95 %-CI von [1,2; 2,1] wäre das Ergebnis signifikant. Die Bezeichnung **statistisch signifikant** taucht in der Statistik und beim Berichten von Ergebnissen sehr häufig auf. Konfidenzintervall und Signifikanz hängen eng miteinander zusammen. Das Konfidenzintervall, meist das 95 %-CI, wird je nach Parameter (*Mean of Difference oder Odds Ratio/Hazard Ration/Relative Risk*) mit einem bestimmten Verfahren geschätzt und soll mit einer Wahrscheinlichkeit von 95 Prozent den wahren Wert der Grundgesamtheit enthalten. Alternative Bezeichnungen für das CI sind **Vertrauensintervall, Vertrauensbereich** oder **Erwartungsbereich**. Als Signifikanzniveau wird häufig ein p-Wert von < 0,05 gewählt. Statistisch signifikant bedeutet dann, dass sich die Daten oder Entwicklungen von Gruppen oder Zeitpunkten so deutlich unterscheiden, dass es sich mit einer Wahrscheinlichkeit von 95 Prozent nicht um einen reinen Zufall handelt, sondern auf die untersuchten Unterschiede zurückgeführt werden kann.

Neben dem RR wird häufig die sogenannte **Hazard Ratio (HR)** in Kohorten-Studien angegeben. Beide unterscheiden sich zunächst scheinbar kaum in der Interpretation. Der Hauptunterschied liegt darin, dass das RR am Ende der geplanten Untersuchungszeit die Inzidenzen erfasst, wohingegen das HR das Ereignis Krankheit/Mortalität im zeitlichen Verlauf über einen langen Studienzeitraum, zum Beispiel jährlich, abbildet.

Berichtet werden können ein RR oder eine HR als ein höheres/niedrigeres Risiko unter Exposition zu erkranken. Davon unterscheidet sich die **Odds Ratio (OR)**. Diese wird nicht bei Kohorten-Studien, sondern bei Fall-Kontroll-Studien errechnet. Die Studienteilnehmer:innen werden nur zu einem Erhebungszeitpunkt befragt und ggf. untersucht und unterteilen sich schon zu Beginn in Erkrankte und nicht Erkrankte. Interessant ist hier, wie viele Erkrankte bzw. Gesunde in der Vergangenheit einem Expositionsfaktor ausgesetzt gewesen sind. Auch die OR kann mithilfe einer 4-Feldertafel berechnet werden. Falls der Expositionsfaktor vor einer Erkrankung schützen kann, wie eine Vitamin-D-Einnahme vor Osteoporose, würde die OR analog zum RR und HR zwischen 0 und 1 liegen. Falls die Exposition beispielsweise Rauchen oder ein erniedrigter Mikronährstoffstatus (Vitamin D/Calcium) wäre, ist es wahrscheinlich, dass die OR > 1 ist und somit ein Hinweis für einen Risikofaktor für die zu untersuchende Erkrankung darstellt. Im Vergleich zum RR oder HR müsste der Zusammenhang

bei der OR so formuliert werden, dass jemand mit einer bestimmten Erkrankung eine größere Chance hatte, dem entsprechenden Risiko-/Präventionsfaktor ausgesetzt gewesen zu sein. Auch das OR wird gemeinsam mit dem p-Wert und/oder dem 95 %-CI angegeben.

Neben Risikofaktoren können auch Referenzbereiche mit dem niedrigsten Krankheits- oder Sterberisiko ermittelt werden für Kriterien wie beispielsweise den Blutdruck oder den BMI.

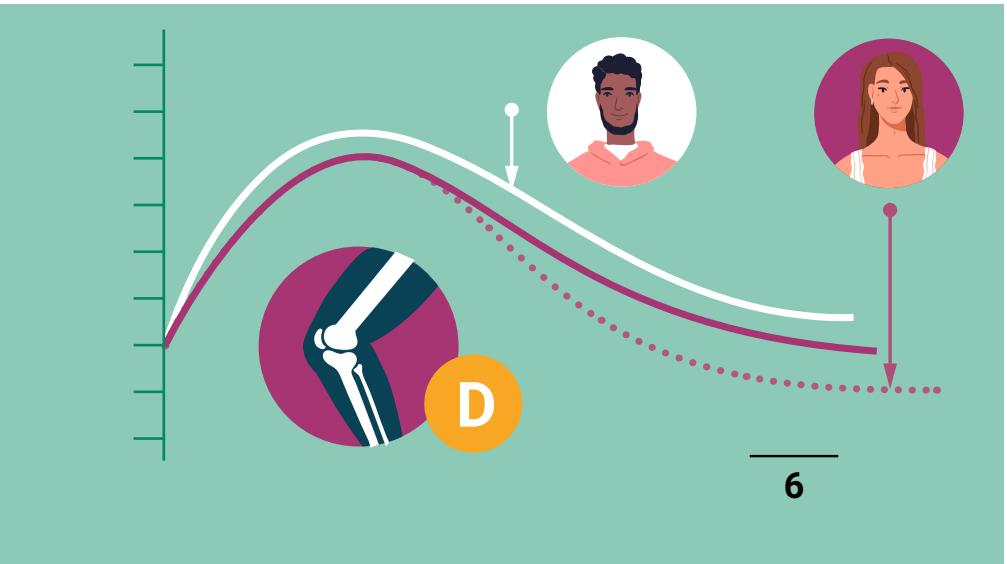
Perzentilen und z-Scores: Werte außerhalb der Norm ermitteln

Aus den Daten einer Stichprobe können *Perzentile* berechnet werden, die anzeigen, wie viel Prozent der Personen einer Stichprobe einen bestimmten Wert oder einen kleineren Wert als diesen aufweisen. Oder andersherum interpretiert: Welchen Wert erreichen mindestens 5 bzw. 10 Prozent der Stichprobe.

Hintergrund ist, dass Parameter wie Längenwachstum, Knochendichte, Muskelkraft, Vitamin-D-Aufnahme oder ähnliches auch bei Gesunden sehr unterschiedlich sein können. Es ist jedoch davon auszugehen, dass Werte, die sehr stark von den meisten anderen Personen desselben Alters/Geschlechts/o. a. abweichen, einen Handlungsbedarf anzeigen können. Meist ist die Abweichung von der sogenannten Norm definiert als < 10. bzw. > 90. Perzentile. Beispiel: Ein Kind, das kleiner ist als 90 Prozent der Kinder im selben Alter, hat eine Größe entsprechend der 10. Perzentile. Werte zwischen der 10. und der 90. Perzentile werden meist als normal bewertet, Werte kleiner der 5. Perzentile oder über der 95. Perzentile werden als kritische Abweichung von der Norm betrachtet.

Ähnlich können **z-Scores** interpretiert werden. Ein z-Score sagt aus, um ein Wievielfaches der Standardabweichung einer Stichprobe der Messwert einer Person vom Mittelwert der Stichprobe abweicht. Da der z-Score nach oben oder unten vom Mittelwert abweichen kann, kann der z-Score positive oder negative Werte annehmen. Eine starke Abweichung ist je nach Parameter bei einem abweichenden Wert von mehr als eins, zwei oder drei Standardabweichungen vom Mittelwert der Stichprobe vorhanden.

In großen Studien mit einer entweder nach Gesundheit oder Repräsentativität ausgewählten Stichprobe können Perzentilen mit dem Ziel berechnet werden, als Referenz oder Diagnosekriterium für die Allgemeinbevölkerung zu dienen. Beispiele hierfür sind Perzentilenkurven für das Längenwachstum von Heranwachsenden oder z-Scores der Knochendichte als Diagnosekriterium für Osteoporose.



Den Erfolg einer Intervention durch statistische Vergleiche messen

In Interventionsstudien gibt es bestenfalls neben der Gruppe, welche die zu untersuchende Intervention erhält (beispielsweise die Einnahme eines bestimmten Mikronährstoffs oder Pflanzenextraktes), eine Kontrollgruppe mit oder ohne Placebobehandlung, die miteinander verglichen werden sollen.

Zur Erfassung der Unterschiede von zwei oder mehreren Messzeitpunkten und/oder Behandlungen (Intervention vs. Kontrolle) eignen sich unterschiedliche Testverfahren. Häufig werden Mittelwerte verglichen.

Ein **Mittelwertvergleich** von zwei Gruppen kann je Messzeitpunkt mittels **T-Test** berechnet werden. Da es sich um unterschiedliche Personen in den Gruppen handelt, würde ein T-Test für unverbundene/unabhängige Stichproben ausgewählt werden. Unverbunden (oder unabhängig) heißt, dass unterschiedliche Personen gruppenweise miteinander verglichen werden.

Bei mehr als zwei Gruppen würde statt des T-Tests beispielsweise je Erhebungs- oder Messzeitpunkt eine **ANOVA** (*engl. Analysis of Variance, Varianzanalyse*) berechnet werden. Die ANOVA korrigiert für multiples Testen, weil mehrere Vergleiche durchgeführt werden, und bezieht die Varianzen aller Gruppen in die Tests mit ein. Bei einer ANOVA können die Mittelwerte aller Gruppen paarweise miteinander verglichen und p-Werte für alle einzelnen Vergleiche berechnet werden. Es kann eine Aussage getroffen werden, ob sich die Werte eines Parameters zwischen den Gruppen zum jeweiligen Zeitpunkt statistisch signifikant unterscheiden.

Es können auch Mittelwertvergleiche von zwei oder mehreren Zeitpunkten berechnet werden. Wenn Daten derselben Personen vorher und nachher gemessen und verglichen werden sollen, würde ein T-Test für verbundene/abhängige Stichproben gewählt werden.

Verbunden (oder abhängig) heißt, dass dieselben Personen vor und nach der Intervention vermessen, untersucht oder befragt wurden.

Manche Ergebnisse werden als Mittelwert und Standardabweichung der Differenzen dargestellt: Je Probandin und Proband wird zunächst die Differenz aus zwei Zeitpunkten (z. B. Nachher minus Vorher) ermittelt und aus diesen Differenzen der **Mittelwert der Differenzen** (*engl. Mean of differences*) aller Probandinnen und Probanden einer Gruppe berechnet. Der Mittelwert der Differenzen kann dann positiv sein, wenn sich ein Parameter innerhalb einer Gruppe häufiger oder stärker erhöht hat und sich weniger häufig oder weniger stark verringert hat. Der Mittelwert der Differenzen wird negativ, wenn sich die Werte der einzelnen Probandinnen und Probanden eher verringern. Bei zwei oder mehr Gruppen kann mit einem Vergleich dieser Mittelwerte eine Aussage getroffen werden, ob sich die Entwicklungen zwischen den Gruppen unterscheiden. Die mittleren Differenzen werden häufig mit dem 95 %-CI angegeben. Bei der Betrachtung einer Gruppe als Vorher-Nachher-Vergleich sollte das 95 %-CI die 0 nicht enthalten, also eindeutig positiv oder negativ sein, damit das Ergebnis als statistisch signifikant bezeichnet werden kann. Dieses Verfahren entspricht dem Mittelwertvergleich mit verbundenen Stichproben.

Um neben der Signifikanz auch die praktische oder klinische Relevanz annähernd widerzuspiegeln, kann die **Effektstärke** (*engl. effect size*) berechnet werden. Für die Effektstärke gibt es verschiedene Verfahren, beispielsweise **Cohen´s d**, welcher zu Daten passt, die mit dem T-Test analysiert wurden. Der Effekt ist umso stärker, je höher der Wert ist. Ein Wert < 0,5 gilt als kleiner Effekt, ein Wert > 0,8 als starker Effekt; Werte zwischen 0,5 und 0,8 beschreiben eine mittlere Effektstärke. Je nachdem, wie viele Gruppen es gibt oder ob die Daten normalverteilt sind oder nicht, gibt es verschiedene Effektstärkemaße, die unterschiedlich interpretiert werden.

Allgemein werden Mittelwertvergleiche wie T-Test oder ANOVA bei normalverteilten Daten durchgeführt; für nicht-normalverteilte Daten werden verschiedene nicht-parametrische Testverfahren für Gruppen- und Behandlungsunterschiede gewählt.

Die Effektstärke und ob ein Mittelwertunterschied signifikant sein kann, hängt von der Streuung der Daten und von der Stichprobengröße ab. Es ist deshalb ein Vorteil, wenn sich Werte innerhalb der Gruppe nicht stark unterscheiden und wenn die Gruppen möglichst groß sind.

Auswertung mit Prinzip: Die Qualität der Daten beurteilen

Studien beinhalten verschiedene Informationen, die Aufschluss über die Qualität der Daten geben und darüber, wie mit möglichen Schwierigkeiten in der Analyse umgegangen wurde.

Die meisten Studien geben an, ob es im Vorfeld eine Fallzahlberechnung gegeben hat und ob diese gewünschte Fallzahl abzüglich der Drop-outs erreicht werden konnte. Es wird angegeben, wie viele Personen im Rahmen eines Screenings voruntersucht und anhand der Ein- und Ausschlusskriterien in die Studie eingeschlossen werden konnten. Die Anzahl der Teilnehmenden, die im Nachhinein ausgeschieden sind oder die die Studie mit allen Untersuchungen abgeschlossen haben, wird ebenfalls angegeben.

Je nach Fragestellung und Studiendesign findet die Auswertung nach einem von zwei Prinzipen statt: **Intention-To-Treat** oder **Per-Protocol**. Bei der Intention-to-Treat-Analyse werden auch Daten von Personen eingeschlossen, für die keine kompletten Datensätze vorliegen und teilweise Untersuchungen und Werte fehlen. Bei einer Per-Protocol-Analyse werden nur Probandinnen und Probanden mit vollständigen Datensätzen aller geplanten Untersuchungen eingeschlossen. Bei Interventionsstudien wird dies graphisch als Consort Diagramm dargestellt. Auch **Ausreißer** (*engl. Outlier*) können zum Ausschluss aus der Datenanalyse führen und müssen beschrieben und kenntlich gemacht werden. Ausreißer sind Messwerte mit einer starken Abweichung von der Stichprobe oder der Behandlungsgruppe. Da es sich nicht immer um unrealistische oder unplausible Daten handelt, sollten diese nur ausgeschlossen werden, wenn für die Werte bestimmte Bedingungen gelten. Beispielsweise können Ausreißer definiert sein als Werte, die mehr als der 1,5-fache Interquartilsabstand außerhalb der Grenzen des Interquartilsabstands liegen (*siehe Abbildung 2, S. 3 oben*). Dabei sind die jeweiligen Grenzen die 75. und die 25. Perzentile, und die Länge des Interquartilsabstands stellt die Differenz zwischen der 75. und der 25. Perzentile dar. Eine Alternative zur Identifizierung möglicher Ausreißer sind Werte, die mehr als drei Standardabweichungen vom Mittelwert abweichen.

Die Auswahl der untersuchten Kriterien muss zur Fragestellung passen und sollte zuverlässig (reliabel) und valide (der Wahrheit entsprechend) bestimmt werden können. Bei Angaben zum Qualitätsmanagement werden deshalb auch Hinweise zur Vermeidung von **Messfehlern** bei anthropometrischen Messungen, Fragebögen oder Laboranalysen gemacht und wird erläutert, wie zuverlässig diese Methoden sind.

Daneben ist auch die Auswahl der untersuchten Parameter entscheidend – insbesondere ist wichtig, ob die gemessenen Parameter die Fragestellung ausreichend beantworten können. Bei der Validität der Daten werden die **interne** und die **externe Validität** unterschieden. Die interne Validität kann durch den Ausschluss von **Störfaktoren** oder **Störvariablen** (*engl. Confounder*) hergestellt oder verbessert werden. Die externe Validität ist gegeben, wenn die Studienergebnisse nicht nur für die Studienpopulation gelten, sondern auch auf die Allgemeinheit übertragen werden können. Die externe Validität ist umso höher, je besser die Studienpopulation ausgewählt wurde, also kein Sampling Bias (*auch Selection Bias*) vorliegt. Ein **Bias** ist eine Verzerrung, die beispielsweise zu einem **systematischen Fehler** führen kann, wenn die Methode der Probandinnen-/Probandenauswahl oder Gruppeneinteilung bzw. die Erhebungsmethode entscheidende Einflussfaktoren auf die Fragestellung nicht beachtet hat. Ein Sampling Bias kann beispielsweise vorliegen, wenn bei Studierenden einer Hochschule erfasst werden soll, wie viele junge Menschen Nahrungsergänzungsmittel einnehmen, weil Studierende sich sehr wahrscheinlich von nicht Studierenden derselben Altersgruppe in ihren Gewohnheiten unterscheiden.

Daten statistisch zusammenfassen: Meta-Analysen

Meta-Analysen gehören zu den Übersichtsarbeiten, die Ergebnisse von Studien mit ähnlichen Fragestellungen zusammentragen und für eine übergeordnete Statistik nutzen. Eine Meta-Analyse kann Beobachtungsstudien oder Interventionsstudien zusammenfassen.

Je nachdem, welcher Studientyp betrachtet wird, wird das Hauptergebnis als Verhältnis (z. B. OR, RR oder HR) bzw. als Differenz dargestellt (z. B. Intervention – Kontrolle). In einem Forrest-Plot (*siehe Abbildung 4*) wird das Ergebnis einer jeden Studie als ein Kästchen dargestellt. Die Größe des Kästchens symbolisiert oft die Anzahl der Studienteilnehmer:innen in dieser Studie. Zusätzlich wird eine waagerechte Linie abgebildet, die seltener die Varianz, häufiger das 95 %-CI darstellt. Wenn das Ergebnis ein Verhältnis ist (OR oder RR), gilt: Das Ergebnis ist signifikant, wenn das 95 %-CI die 1 nicht umschließt. Bei einigen Meta-Analysen, die Interventionsstudien einschließen, können Unterschiede zwischen Gruppen oder Zeitpunkten als Verhältnis oder stattdessen als Differenzen abgebildet sein; dann wäre entscheidend, dass das 95 %-CI die 0 nicht umschließt. Zusätzlich zu den einzelnen Studienergebnissen wird eine gewichtete Auswertung für alle Studien durchgeführt. Gewichtet heißt, dass das Ergebnis einer Studie mit vielen Teilnehmer:innen stärker in das Endergebnis der Meta-Analyse einfließt als das Ergebnis von kleinen Studien.

Die Studien, die für die Meta-Analyse verwendet werden, sollten möglichst ähnliche Parameter und Studiendesigns sowie Behandlungsdauern und Merkmale von Versuchsteilnehmer:innen haben. Deshalb werden in einer Meta-Analyse Angaben zur Heterogenität gemacht.

Die **Heterogenität** ist insbesondere dann groß, wenn unterschiedliche Studien zu einer Fragestellung unterschiedliche Ergebnisse aufweisen. Die Ergebnisse können unterschiedlich deutlich sein oder sogar in unterschiedliche Richtungen voneinander abweichen. Ob Unterschiede im Design dies erklären können, sollte deshalb bei der Interpretation beachtet werden. Falls unterschiedliche Designs die Ursache sein können, kann eine Subgruppen-Analyse, also eine separate Analyse der Studien, die methodisch zusammenpassen, ergänzt werden.

Als ein mögliches Problem bei Meta-Analysen gilt ein *Publication Bias*. Dieser bedeutet, dass hauptsächlich Studien, die die Hypothese über einen Behandlungserfolg bestätigen, veröffentlicht werden und Studien mit unerwarteten Ergebnissen eher nicht öffentlich kommuniziert werden.

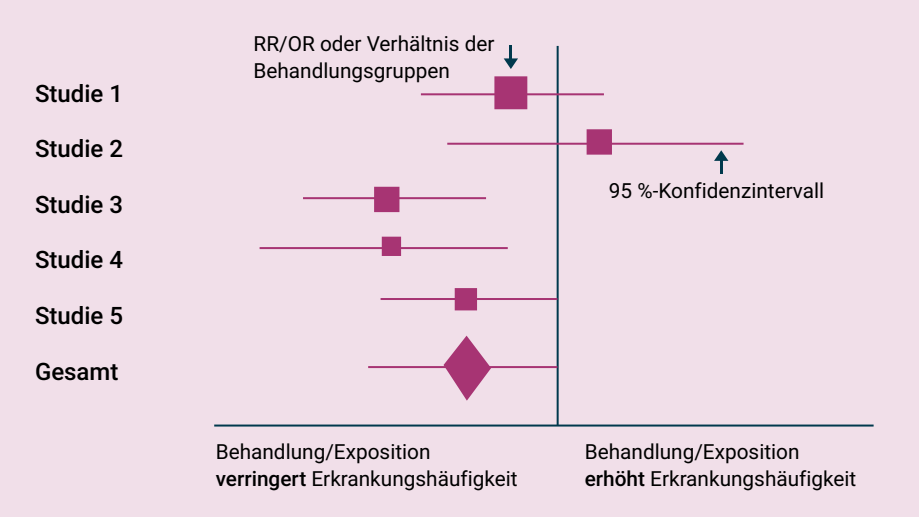


Abbildung 4: Schematische Darstellung eines Forrest Plots, wie es häufig in Meta-Analysen dargestellt wird.

Ein **Funnel Plot** (*auch Trichter Plot; siehe Abbildung 5*) kann darüber Auskunft geben. Hierbei werden die Studieneffekte mit der Studiengröße als Streudiagramm dargestellt, wobei davon ausgegangen wird, dass größere Studien eher mittlere Effekte zeigen und kleinere Studien anfälliger sind für zufällige Abweichungen in die eine oder andere Richtung. Wenn die Streuung der Punkte symmetrisch ist und die Form einem Trichter ähnelt, ist ein Publication Bias unwahrscheinlich.

Meta-Analysen helfen, ein Gesamtbild über mehrere Studien zu geben. Mehrere Studien, bestenfalls Beobachtungs- und Interventionsstudien in Kombination mit Grundlagen aus der Tier- und Zellforschung, können helfen, die aktuelle Studienlage einzuordnen. Fachgesellschaften versuchen so, evidenzbasierte Empfehlungen und Referenzen zu formulieren.

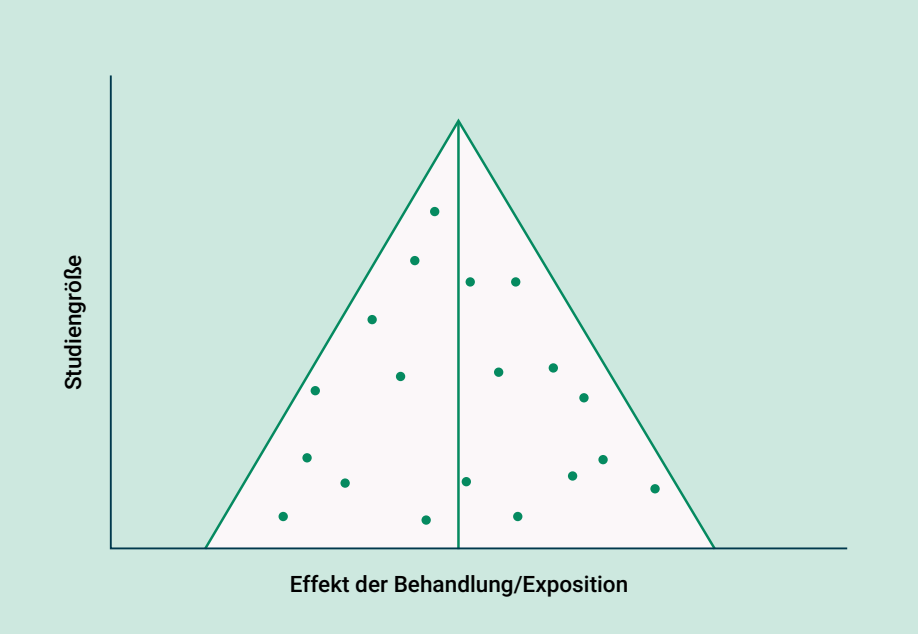


Abbildung 5: Schematische Darstellung eines Funnel Plots, der sich zur Identifikation eines Publication Bias eignet.

Studien kritisch und kundig lesen und ihre Ergebnisse korrekt wiedergeben – alles andere kann jede:r

Der Einfluss von Ernährungsfaktoren auf die Gesundheit ist ein komplexes Zusammenspiel zwischen verschiedenen Parametern und individuellen Unterschieden, die mit Studien und statistischen Methoden immer nur näherungsweise erfasst werden können. Umso wichtiger ist es, dass sich Leser:innen von Studien und alle, die über Studien kommunizieren, nicht nur mit den verschiedenen Designs auskennen, sondern auch statistische Auswertungen verstehen und „lesen" können. Die Ergebnisse einer Studie einzuordnen bedeutet auch, die Qualität der Studie und der statistischen Analyse kritisch zu hinterfragen und zu prüfen, ob sich Ergebnisse durch andere Studien bestätigen lassen und die postulierte Wirkung auch klinische Relevanz hat. Das heißt für alle, die Studienergebnisse veröffentlichen und weitergeben, dass sie Merkmale der untersuchten Population, die gemessenen Parameter und Grundzüge des Designs und der Statistik mit angeben, um Irrtümer und Fehlinterpretationen zu vermeiden.

Literaturempfehlungen

Krickhahn T, Poß D (Autoren): Statistik kompakt für Dummies Wiley-VCH GmbH; ISBN: 3527711546; EAN: 9783527711543

Rauch G, Kruppa J, Grittner U, Neumann K, Herrmann C (Autoren): Medizinische Statistik für Dummies. Wiley-VCH GmbH; ISBN: 3527715843; EAN: 9783527715848

The EQUATOR Network | Enhancing the Quality and Transparency of health Research. https://www.equator-network.org (letzter Zugriff: 08.08.2022)

AK NEM
Arbeitskreis Nahrungsergänzungsmittel im Lebensmittelverband Deutschland e.V.

Herausgeber
Arbeitskreis Nahrungsergänzungsmittel (AK NEM) im Lebensmittelverband Deutschland e. V.
Postfach 06 02 50, 10052 Berlin
Claire-Waldoff-Straße 7, 10117 Berlin
Telefon: +49 30 206143-0
aknem@lebensmittelverband.de

Autorin/Kontakt:
Dr. oec. troph Sandra Habicht
Institut für Ernährungswissenschaft
der Justus-Liebig-Universität Gießen
Wilhelmstraße 20, 35392 Gießen
sandra.d.habicht@ernaehrung.uni-giessen.de

Gestaltung: Ariane Skibbe, DFY Berlin

www.nahrungsergaenzungsmittel.org