1.2 Signifikanz (p-Wert)

Um welches Problem geht es in diesem Unterkapitel?

Im Rahmen einer wissenschaftlichen Studie können in der Regel nicht alle Patienten untersucht werden. Das führt zu dem Problem, dass von einer Auswahl (Stichprobe) auf die Grundgesamtheit (Population) geschlossen werden muss. Wie aber ist es möglich, von einigen wenigen etwas über alle auszusagen? Derartige Aussagen sind grundsätzlich mit gewissen Unsicherheiten behaftet. Wenn eine Therapie bei den Patienten der Stichprobe wirkt, ist sie dann auch bei allen Patienten wirksam? Und wenn ja, wie können wir das wissen, wenn wir gar nicht alle untersucht haben? Der p-Wert (von lat. probabilitas: Wahrscheinlichkeit) liefert Informationen über die Wahrscheinlichkeit, mit der man sich irrt, wenn man annimmt, ein in der Stichprobe gefundener Unterschied sei auch in der Population vorhanden.

Wenn Sie dieses Unterkapitel gelesen haben, können Sie folgende Fragen beantworten:

Wie können Aussagen über die Population ausgehend von einer Stichprobe getroffen werden?
Was ist der p-Wert und in welchem Zusammenhang steht er mit statistischen Hypothesen?
Welche Fehlentscheidungen sind beim Schließen von der Stichprobe auf die Population möglich?
Warum ist es problematisch, im Rahmen einer Studie sehr viele statistische Tests zu rechnen?
Was sind ein- und zweiseitige Hypothesen und worin liegen mögliche Gefahren, wenn man eine zweiseitige Hypothese im Nachhinein zu einer einseitigen umformuliert?

Schlüsselbegriffe

Signifikanz, Signifikanzniveau, wissenschaftliche Hypothesen, Alltagshypothesen

Das Schließen von einer Stichprobe auf die Grundgesamtheit

Frau Alma Tiener züchtet seit dreißig Jahren Dalmatiner. Durch sorgfältige Auswahl ist es ihr gelungen, einen neuen Typus zu erschaffen: schwarze Dal-

matiner mit weißen Punkten (die sogenannten „Höllendalmatiner"). Im Laufe der Zeit stellt sich jedoch heraus, dass es bei den Höllendalmatinern öfter zu Taubheit kommt als bei den herkömmlichen, schwarzgepunkteten Dalmatinern. (Taubheit ist allgemein ein Problem bei Tieren dieser Hunderasse.) Deshalb entschließt sich Frau Tiener, bei den folgenden Würfen die Sache genauer zu untersuchen. Sie besitzt jeweils fünf Zuchthündinnen und ermittelt bei jedem Wurf den Anteil der tauben Hunde. Bei den herkömmlichen Dalmatinern waren unter den insgesamt 30 Welpen 4 auffällige (4 von 30, das sind 13,3 %), bei den Höllendalmatinern fand sie 6 auffällige Welpen unter 26 (6 von 26, das sind 23,1 %). Frau Tiener verallgemeinert deshalb auf alle künftigen Würfe und sagt: „Meiner Erfahrung nach ist der Anteil tauber Hunde unter den Höllendalmatinern signifikant höher als unter den herkömmlichen Dalmatinern."

Die Bedeutung des Wortes „signifikant"

Wir wollen nun erkunden, was es mit dem Wort „signifikant" auf sich hat. Ist es gerechtfertigt, dass Frau Tiener von einem signifikanten Ergebnis spricht, weil in der einen Gruppe 13,3 % und in der anderen 23,1 % der Tiere auffällig sind und sie es deshalb als erwiesen erachtet, dass dieser Unterschied auch in der Population vorhanden sei? „Signifikant" bedeutet so viel wie „überzufällig", das heißt, ein in der Stichprobe gefundener Unterschied wird nicht mehr auf bloßen Zufall zurückgeführt, sondern es ist davon auszugehen, dass der gefundene Unterschied auch „in Wirklichkeit", also in der Population, besteht.

Führen wir das Gedankenexperiment weiter und nehmen an, dass es schon sehr viele Züchter von Höllendalmatinern gibt. Von allen registrierten Dalmatinern und Höllendalmatinern werden je 30 Hunde entnommen, sodass wir nun zwei Stichproben mit jeweils 30 Tieren haben (siehe Abbildung 1.5).

Wir wissen bereits, dass es sehr viele Möglichkeiten gibt, solche Stichproben zu entnehmen. Das Ziel ist, aufgrund der beiden konkreten Stichproben etwas über die Populationen (das sind die Populationen der normalen und der Höllendalmatiner) auszusagen. Und wir haben auch schon festgestellt, dass die Anteile auffälliger Hunde in Stichproben aus ein und derselben Population eben je nach Stichprobe selbst variieren werden, je nachdem, welche Individuen „zufällig" in die Stichprobe kommen.

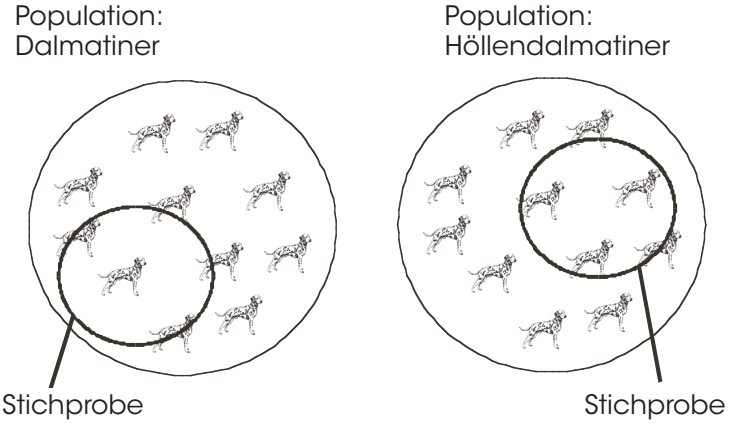

Abb. 1.5: Stichproben der Dalmatiner und Höllendalmatiner

Für das Verständnis des p-Wertes sind nun folgende zwei Gedankengänge bedeutsam:

Gedankengang 1: Selbst wenn es „in Wirklichkeit", also in den beiden Populationen der normalen und der Höllendalmatiner, keinen Unterschied in der Häufigkeit tauber Hunde gibt, wird deren Anteil in den beiden Stichproben ziemlich sicher unterschiedlich sein!

Die Gretchenfrage lautet daher: Ist der unterschiedliche Anteil nur zufällig – zufällig bedingt durch die Zufälligkeit, mit der die jeweils 30 Hunde ausgewählt wurden – oder ist dieser zu erwartende Unterschied dadurch zustande gekommen, dass es in der Population der Höllendalmatiner tatsächlich mehr taube Tiere gibt?

Gedankengang 2: Wir haben bereits festgestellt, dass Aussagen über die Population aufgrund von Stichproben stets mit Unsicherheit behaftet sind, wir können nichts mit absoluter Sicherheit sagen. Sicherheit wäre nur dann gegeben, wenn wir die gesamte Population untersuchen würden.

Das bedeutet leider: Selbst wenn wir ein signifikantes Ergebnis erhalten – das heißt, selbst wenn wir aufgrund entsprechender Stichprobendaten schlussfolgern, dass es in der Population der Höllendalmatiner mehr taube Tiere als in der Population der normalen Dalmatiner gibt –, könnte es sein, dass wir uns irren. Mit anderen Worten: Es könnte sein, dass sich die beiden

Populationen trotz der Signifikanz, die sich aus den Stichproben ergibt, nicht unterscheiden. Aber, und das ist das Gute: So ein Irrtum wäre ziemlich unwahrscheinlich.

Und jetzt zum p-Wert: Wenn der p-Wert (zur Erinnerung: p steht für probabilitas: Wahrscheinlichkeit) eine bestimmte Höhe nicht übersteigt (zumeist: kleiner oder gleich 0,05 [= 5 %] ist), dürfen wir schlussfolgern, dass es in der Population der Höllendalmatiner wirklich mehr taube Tiere gibt. Natürlich könnte diese Schlussfolgerung falsch sein, aber es ist sehr unwahrscheinlich, dass sie falsch ist.

Wie groß darf der p-Wert sein, damit das Ergebnis signifikant ist? Diese Frage kann nicht einheitlich beantwortet werden, etabliert haben sich die Grenzen 5 %, 1 % und 0,1 % (bzw., was äquivalent ist: $p \leq 0{,}05$, $p \leq 0{,}01$ und $p \leq 0{,}001$). Das heißt: Wenn der p-Wert höchstens 0,05 bzw. 0,01 oder 0,001 beträgt, dann haben wir ein signifikantes Ergebnis auf dem 5 %-, dem 1 %- bzw. dem 0,1 %-Niveau. Die Signifikanzniveaus sind übrigens völlig willkürlich gewählt, was berechtigterweise oft kritisiert wird. Es kann durchaus sein, dass in einer Studie ein einziger Patient den Ausschlag gibt zwischen nicht signifikant und signifikant (neben weiteren möglichen Einflüssen, die das Zünglein an der Waage spielen können).

Der p-Wert selbst ist übrigens stets das Resultat eines statistischen Tests (wie beispielsweise des t-Tests, einer Korrelationsrechnung oder einer Varianzanalyse), also eines Rechenvorganges, der nach bestimmten Regeln abläuft.

Doch sehen wir uns an, wie solch ein p-Wert in Publikationen üblicherweise präsentiert wird.

Aus einer Studie

"For all statistical tests a 5 % significance level was considered acceptable and used throughout the analysis ... Scores for these primary efficacy variables decreased significantly (p = 0.000 ...) for both groups by the end of the treatment."

Agublia, E., Cacacchia, M., Cassano, G. B., Faravelli, C., Ferrari, G., Giordano, P., Pancheri, P., Ravizza, L., Trabucchi, M., Bolino, F., Scarpato, A., Berardi, D., Provenzano, G., Brugnoll, R. & Rozzini, R. (1993). Double-blind study of the efficacy and safety of sertraline versus fluoxetine in major depression. International Clinical Psychopharmacology, 8, 197–202.

In dieser Studie wurden Sertraline und Fluoxetine hinsichtlich ihrer Wirksamkeit zur Behandlung von Depression verglichen. Als Grenze, bis zu der ein Ergebnis als signifikant akzeptiert wurde, sind 5 % angegeben. Als „primary efficacy variables" wurden die Testwerte von zwei psychologischen Tests zur Diagnostik von Depression verwendet (es handelte sich um die Tests HAMD und CGI). Berechnet wurde u. a., ob sich die Testwerte von HAMD und CGI über die Behandlungsdauer veränderten. (Dazu mussten die Patienten der beiden Gruppen „Sertraline" und „Fluoxetine" diese Tests mehrmals bearbeiten.) Die Autoren hatten somit für alle Patienten dieser Gruppen einen Anfangs- und einen Endwert, wodurch die Differenz zwischen diesen beiden Zeitpunkten berechnet werden konnte. Aufgrund des angeführten p-Wertes von $p = 0{,}000$ war der in den Stichproben (Stichprobe 1: Patienten, die Sertraline bekamen; Stichprobe 2: Patienten, die Fluoxetine bekamen) gefundene Unterschied in der Veränderung der Testwerte von HAMD und CGI zwischen diesen beiden Gruppen signifikant, das heißt: Die Patienten der Sertraline- und Fluoxetine-Gruppen wiesen eine signifikant unterschiedliche Veränderung der HAMD- und CGI-Werte zwischen den beiden Messzeitpunkten auf. Dies bedeutet: Die ermittelten Stichprobenunterschiede bestehen mit hoher Wahrscheinlichkeit auch in der Population! Wäre der p-Wert größer als 5 % (bzw. 0,05), zum Beispiel $p = 0{,}12$ (bzw. 12 %), wäre das Ergebnis nicht signifikant und die Schlussfolgerung würde lauten: Der gefundene Unterschied in der Veränderung der HAMD- und CGI-Werte zwischen den Gruppen „Sertraline" und „Fluoxetine" müsste auf den Zufall zurückgeführt werden, wäre also durch zufällige Stichprobenschwankungen zu erklären.

Wir haben bisher festgestellt:

- Der p-Wert gibt Auskunft darüber, ob ein Resultat signifikant ist oder nicht.
- Damit wir ein Ergebnis als signifikant bezeichnen, darf der p-Wert eine bestimmte Höhe nicht überschreiten, wobei es unterschiedliche Grenzen gibt: nicht höher als 5 % oder 1 % oder 0,1 %. (Diese Grenzen heißen Signifikanzniveau und es wird aufgrund inhaltlicher Überlegungen vor jeder statistischen Auswertung festgelegt, welche dieser drei Grenzen gelten soll.)
- Der p-Wert informiert uns über die Wahrscheinlichkeit, dass wir uns irren, wenn wir eine bestimmte Schlussfolgerung über die Population treffen.

Wenn wir die letzte Feststellung mit anderen Worten ausdrücken und sagen, dass wir eine richtige Schlussfolgerung über die Population treffen, so müssen wir uns noch Gedanken darüber machen, was mit „richtiger Schlussfolgerung“ gemeint ist.

Das wissenschaftliche Vorgehen zeichnet sich u. a. dadurch aus, dass sehr genaue Hypothesen darüber aufgestellt werden, was man erwartet oder nicht erwartet. Im Alltag verwenden wir ebenfalls Hypothesen, die uns aber sehr oft gar nicht bewusst sind. Wenn wir etwa eine Person aufgrund der Art, wie sie sich kleidet, in eine „Schublade“ stecken, so denken wir meistens nicht explizit darüber nach. Es geschieht einfach. Unsere impliziten Hypothesen könnten lauten:

- Wer Markenkleidung trägt, ist eitel.
- Wer Markenkleidung trägt, hat viel Geld.
- Wer sich schlampig kleidet, dem ist nicht zu trauen.

Wissenschaftliche Hypothesen müssen im Gegensatz zu unseren Alltagshypothesen gewissen Ansprüchen genügen, und sie müssen empirisch überprüfbar sein. Ein statistischer Test hat die Aufgabe, auf der Basis von Datenmaterial diese Hypothesen zu bestätigen oder zu verwerfen. Im eben genannten Alltagsbeispiel ist die Kleidung sozusagen das Datenmaterial, von dem wir ausgehen, und wir benutzen implizit unsere Werte, Einstellungen, Erfahrungen etc., um unsere Alltagshypothesen zu überprüfen. Und meistens neigen wir dazu, das, was wir glauben, zu bestätigen. Das hat viel mit Vorurteilen, mit vorgefassten Meinungen zu tun.

Im Idealfall geht der empirisch arbeitende Wissenschaftler ohne Vorurteile und feste Erwartungen an seine Studie, obwohl er natürlich auch Annahmen trifft, die er zu bestätigen sucht. Die Logik des statistischen Tests beinhaltet allerdings eine Art Pessimismus. Was heißt das? Um dies zu verstehen, müssen wir unseren Blick auf zwei wichtige Begriffe richten: die Nullhypothese und die Alternativhypothese.