Die Verwendung von p-Werten in den Agrarwissenschaften

Björn Christensen[1], Sören Christensen[2] und Tobias Sohr[2]

Abstract: Weil Datenanalyse in den Agrarwissenschaften einen immer höheren Stellenwert gewinnt, wird auch hier der richtige Umgang mit dem p-Wert immer wichtiger. Der p-Wert ist eine der meistgenutzten statistischen Größen, um Signifikanz festzustellen und daraus Entscheidungsregeln abzuleiten. Gleichzeitig wird die richtige Interpretation und die Aussagekraft des p-Wertes bei vielen Anwendungen in der Statistik lebhaft diskutiert. Im Folgenden werden wichtige Aspekte dieser Diskussion dargestellt, erläutert und anhand eines aktuellen Beispiels illustriert.

Keywords: Angewandte Statistik, p-Wert

1 Einleitung

Im Zuge der digitalen Transformation gewinnt die statistische Auswertung großer Datensätze immer größere Bedeutung in den Agrarwissenschaften. In den letzten Jahren entstanden vielfältige neue technische Möglichkeiten zur Datenerfassung und -auswertung. Die dadurch gestiegene Verfügbarkeit an Informationen schafft neben der Wissenschaft sogar für einzelne landwirtschaftliche Betriebe und andere Unternehmen des Agrarsektors die Möglichkeit, durch Datenanalyse Wettbewerbsvorteile zu erlangen. Infolgedessen ist die Notwendigkeit, datenbasiert Entscheidungen zu treffen, zunehmend Bestandteil sowohl der Agrarwissenschaften als auch der agrarökonomischen Praxis.

Insofern ist das Verständnis statistischer Methoden in den Agrarwissenschaften hilfreich. Eine besondere Rolle nimmt dabei der p-Wert ein. Einerseits ist der p-Wert eine der meistgenutzten Größen, um in der Statistik Signifikanz festzustellen, andererseits werden Interpretation und Aussagekraft desselben in der Statistik lebhaft diskutiert. Verwiesen sei dabei auf die Stellungnahme der American Statistical Association (ASA) zur Interpretation des p-Wertes [ASA16] und zum korrekten Umgang mit diesem. In diesem Beitrag wird diese Diskussion zusammengefasst und anhand des Beispiels aus [LH16] exemplarisch für die Agrarwissenschaften erläutert.

[1] Fachhochschule Kiel, Institut für Statistik und Operations Research, Sokratesplatz 2, 24149 Kiel, Bjoern.Christensen@fh-kiel.de

[2] Universität Hamburg, Bereich Mathematische Statistik und Stochastische Prozesse, Bundesstraße 55, 20146 Hamburg, soeren.christensen@uni-hamburg.de, tobias.sohr@uni-hamburg.de

2 Kernpunkte der ASA-Stellungnahme

2.1 Eigenschaften des p-Wertes

Der p-Wert wird zur Untersuchung der Plausibilität von Hypothesen in statistischen Modellen genutzt. Unter einer solchen Hypothese ergibt sich, wie eine Testgröße (oft z.B. das Stichprobenmittel) theoretisch verteilt ist, wenn die Hypothese wahr wäre. Der p-Wert ist dann die Wahrscheinlichkeit, dass die Kenngröße bei wahrer Hypothese mindestens so extreme Werte annimmt, wie man sie in den Daten beobachtet hat. Aus dieser Definition ergeben sich einige Implikationen für den richtigen Umgang mit dem p-Wert:

- Der p-Wert kann ein Indikator dafür sein, wie kompatibel der analysierte Datensatz mit der vorher aufgestellten Hypothese ist.
- Er ist jedoch nicht die Wahrscheinlichkeit, dass die aufgestellte Hypothese wahr ist. Ebenso wenig „beweist" ein hoher p-Wert die Gültigkeit der aufgestellten Hypothese.
- Es können bei demselben Datensatz sehr viele Hypothesen zu einem hohen p-Wert führen, sodass ein hoher p-Wert nicht als Beleg für die Richtigkeit einer Hypothese angesehen werden kann. Ein hoher p-Wert kann auch bei falscher Hypothese eintreten, etwa durch Zufall oder weil eine unpassende Testgröße gewählt wurde. Beispielsweise benötigt man für einen t-Test eine Stichprobe normalverteilter Ergebnisse. Wenn das nicht gewährleistet ist, sind Schlüsse aus einem hohen oder niedrigen p-Wert beim t-Test nicht zulässig.
- Auch ist bei einem niedrigen p-Wert die aufgestellte Hypothese keinesfalls zwangsläufig falsch. Ein p-Wert von 0,05 bedeutet gerade, dass, führt man dasselbe Zufallsexperiment 20-mal durch, im Mittel eines der Experimente ein Ergebnis mindestens genauso weit weg vom erwarteten Ergebnis liefert.
- Die oft angewandte Entscheidungsregel, bei $p \leq 0{,}05$ die Hypothese abzulehnen und andernfalls anzunehmen, ist demzufolge kein Naturgesetz.
- Der p-Wert erlaubt keinerlei Schlussfolgerungen auf die Größe des gemessenen Effektes, genauso wenig zu dessen Wichtigkeit oder dessen ökonomischem Nutzen. Einfach ausgedrückt kann man, misst man nur genau genug, jeden beliebig kleinen Effekt messen. Das heißt in Termen der Statistik ausgedrückt: Ist die Stichprobengröße hinreichend groß, liefert schon eine nur marginal von der Hypothese abweichende Testgröße einen sehr geringen p-Wert. Dann mag diese Abweichung zwar signifikant sein, aber nicht unbedingt relevant.

2.2 Der richtige Umgang mit dem p-Wert

Aus den genannten Eigenschaften des p-Wertes lassen sich einige Grundsätze zum Umgang mit demselben ableiten.

- Der p-Wert sollte nicht als einzige Entscheidungsgrundlage genutzt werden, da z.B. die Tatsache, dass der p-Wert für eine bestimmte Testgröße nahe 0,05 liegt, isoliert betrachtet nur geringe Aussagekraft besitzt. Stattdessen sollten weitere Testgrößen und Methoden betrachtet werden. Vor allem aber sollte der größere Rahmen der Datenerhebung nicht aus den Augen verloren werden.
- In der Datenanalyse allgemein sowie insbesondere beim Umgang mit p-Werten ist eine genaue Betrachtung der Methodik der Datenerfassung extrem wichtig. Testet man beispielsweise 20 Düngemittel gegen die Nullhypothese, die Mittel seien wirkungslos, indem man einfach die Abweichung der Ergebnisse von null betrachtet, wird mit hoher Wahrscheinlichkeit eines einen p-Wert $\leq 0{,}05$ aufweisen, auch wenn die Mittel in Wahrheit wirkungslos sind und die Ergebnisse sich nur wegen Zufallseinflüssen unterscheiden. Insofern würde „cherry-picking“, also die isolierte Betrachtung nur dieses einen Ausreißers, zu vollkommen falschen Schlussfolgerungen führen. Ein damit eng verwandtes Problem ist der sog. „Publikations-Bias“, der entsteht, wenn (ungewöhnliche) Ergebnisse mit einem niedrigen p-Wert veröffentlicht werden, nicht aber jene aus den selben Daten stammenden mit keinerlei Auffälligkeiten. Letztere sind jedoch (wie oben beispielhaft erklärt) unbedingt nötig, um die Ergebnisse richtig zu deuten.
- Der p-Wert ist keine isolierte Teststatistik, sondern gehört immer zu einem Test. Genauer gesagt gibt er an, wie sich das Ergebnis einer berechneten Testgröße mit den Annahmen über diese verträgt. Je nachdem, was man testet, bekommt man also möglicherweise sehr unterschiedliche p-Werte. Insofern ist es wichtig, für die aus der Analyse benötigten Rückschlüsse zunächst einen passenden Test und sinnvolle Annahmen zu finden und die p-Werte dann unter Berücksichtigung dieser zu interpretieren.
- Es ist nicht zulässig, aus p-Werten Rückschlüsse auf die Gründe oder die Höhe von gemessenen Effekten zu ziehen. Ein niedrigerer p-Wert bedeutet nicht automatisch einen stärkeren Effekt. Der p-Wert kann lediglich ein Indikator dafür sein, dass ein Effekt existiert. Die Stärke des Effekts muss ergänzend und vor allem inhaltlich bewertet werden. Auch die Frage, warum ein Effekt existiert, kann durch den p-Wert nicht erklärt werden.

3 Ein praxisnahes Beispiel: Erzeugerpreise von Weizen in Norddeutschland

In [LH16] wird der Erlös von 204 norddeutschen Betrieben durch den Verkauf von Win-

terweizen über den Zeitraum von 2003 bis 2014 analysiert: Unter anderem wird untersucht, ob die Daten Rückschlüsse darauf zulassen, dass einzelne Betriebe ihre Produkte signifikant besser vermarkten als der Durchschnitt. Der mittlere Verkaufspreis aller Betriebe in diesem Zeitraum lag bei 15,64 Euro pro dt. Die jeweils mittleren Verkaufspreise der Betriebe liegen zwischen 13,36 und 17,66 Euro pro dt. Nun stellt sich die Frage, ob einige Betriebe wirklich „besser" in der Vermarktung agieren oder ob dessen höherer Erlös „Glück", also zufallsbedingt ist. Es wäre jetzt genau das oben erwähnte „Cherry-Picking" und damit ein Fehler, einfach das Ergebnis einzelner Betriebe gesondert zu betrachten. Denn bei 204 betrachteten Betrieben ist es vollkommen natürlich, wenn einige davon teilweise weit vom Mittelwert abweichende Ergebnisse erzielen. Vielmehr muss man zuerst sinnvolle Tests zur Hypothese auswählen, in diesem Beispiel einen t-Test und einen Wilkoxon-Test. Bei diesen dann kann man überprüfen, ob wirklich ca. 5% der Betriebe bei den Tests ein Ergebnis mit einem p-Wert unter 0,05 erzielen, denn das wäre zu erwarten, wäre die Hypothese wahr. Allerdings erzielen in dem Beispiel bei dem t-Test 11% und bei dem Wilkoxon-Test sogar 11,7% der Betriebe einen p-Wert von unter 0,05. Deswegen wird die Hypothese, die Unterschiede der Betriebserlöse seien rein zufälliger Natur, abgelehnt. Anhand dieses Beispiels ist folglich zweierlei erkennbar. Erstens ist es wichtig, beim Interpretieren von p-Werten das Design der Studie zu berücksichtigen und nie einzelne Daten isoliert zu betrachten. Zweitens können unterschiedliche Tests derselben Hypothese zu Unterschiedlichen p-Werten und sogar zu unterschiedlichen Ergebnissen und in der Konsequenz zu unterschiedlichen Entscheidungen führen.

4 Fazit

P-Werte sind keine eigenständigen Teststatistiken, sondern immer nur ein Mittel zur Interpretation solcher. Insofern hängen sie stark von den jeweiligen Annahmen und den jeweiligen Tests ab. Demzufolge sollten sie nie isoliert betrachtet werden und geben allenfalls Hinweise auf statistische Zusammenhänge vor dem Hintergrund des gewählten Testdesigns, nicht jedoch auf kausale Aussagen.

Literaturverzeichnis

[ASA16] American Statistical Association.: ASA Statement on Statistical Significance and P-Values. The American Statistician, 2016, Vol. 70, Nr. 2, 129-133.

[LH16] Loy, J.; Holzer, P.: Messung des Vermarktungserfolges. Intelligente Systeme Stand der Technik und neue Möglichkeiten, Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016, 113-116

.