Simulationsbasiert Signifikanztests verstehen

MICHAEL RÖßNER, MÜNCHEN; KARIN BINDER, MÜNCHEN & STEFAN UFER, MÜNCHEN

***Zusammenfassung**: Data Literacy umfasst auch Kompetenzen zur Interpretation von Daten und Analysen, beispielsweise mithilfe von Signifikanztests. Die Bedeutung von Signifikanztests unterliegt jedoch vielen Fehlvorstellungen. Im Rahmen eines Kurses für begabte Schüler:innen wurden inferenzstatistische Methoden (simulationsbasiert mit R) eingeführt. Der Nachtest zeigt: Während die Schüler:innen nach dem Workshop qualitativ einschätzen können, was es bedeutet, wenn der p-Wert das Signifikanzniveau unterschreitet, unterliegen noch viele dem sogenannten Relevanzfehlschluss. Abschließend werden Vorschläge für die Weiterentwicklung des Workshops vorgestellt.*

***Abstract:** Data Literacy also includes skills for interpreting data, for example with the help of significance tests. However, there are many misconceptions about the meaning of significance tests. As part of a course for gifted students, inferential statistical methods were taught (simulation-based with R). The post-test shows: While the students can qualitatively assess after the workshop what it means when the p-value falls below the significance level, many are still subject to the so-called relevance fallacy. Finally, suggestions for the further development of the workshop are presented.*

1. Einleitung

„Der Anteil der Knieverletzungen steigt mit diesen Schuhen signifikant“, titelte die Süddeutsche Ende September 2023 in einem Interview mit einem Professor für Biomechanik (Knuth, 2023). Doch was bedeutet dieser signifikante Anstieg? Wer über ausreichende Data Literacy im Bereich der Signifikanztests verfügt, weiß, dass es sich dabei nicht um einen großen Anstieg handeln muss, weil statistische Signifikanz etwas anderes beschreibt als eine Effektstärke und daher nicht – wie das Alltagswort „signifikant“ – mit dem Wort „bedeutend“ oder „bezeichnend“ übersetzt werden sollte, auch wenn das der eigentlichen Wortherkunft aus dem Lateinischen entspräche (Gagnier & Morgenstern, 2017). Zahlreiche Studien belegen die Schwierigkeiten und typischen Fehlvorstellungen, die sogar Wissenschaftler:innen oder Statistik-Dozent:innen im Verständnis von Signifikanztests und der Interpretation von p-Werten haben (Oakes, 1986, Haller & Krauss, 2002, Herrera-Bennett, Heene, Lakens & Ufer, Preprint).

Im Rahmen eines einwöchigen Kurses für begabte Schüler:innen wurden an der LMU München in Vorlesungen und Übungen genau die inferenzstatistischen Methoden (wie z. B. t-Test, Chi-Quadrat-Test) gelehrt, die seit September 2024 in Bayerischen Gymnasien im Rahmen des sogenannten Vertiefungskurses in der Jahrgangstufe 12 im Modul „Statistik“ unterrichtet werden (ISB, 2024). Ein Fokus lag hierbei auf einem simulationsbasierten Ansatz mithilfe der Software R, der ebenso auf die in der gymnasialen Oberstufe sonst häufig unterrichteten Binomialtests angewendet werden könnte. Abschließend wurde geprüft, inwiefern die Schüler:innen nach der Einführung den aus der Literatur bekannten Fehlkonzepten von Signifikanztests unterliegen.

Im vorliegenden Beitrag werden zunächst die klassischen Fehlkonzepte in Bezug auf Signifikanztest und p-Werte zusammengefasst. Anschließend wird ein Training zu den Inhalten des Vertiefungskurses „Statistik“ vorgestellt, das auf den Erwerb konzeptuellen Wissens zu Signifikanztests und p-Werten abzielt (z. B. Zusammenhang zwischen Samplingvarianz und Stichprobengröße), statt auf prozedurales Wissen (wie z. B. der konkreten Berechnung von Teststatistikwerten, der Bestimmung von Annahme- und Ablehnungsbereichen). Die didaktische Schwerpunktsetzung des entwickelten Trainings lag auf einer simulationsbasierten Einführung von Signifikanztests (siehe auch Chance et al., 2022; Chandrakantha, 2020), also der Simulation von Stichprobenziehungen aus einer Nullpopulation (in der eben kein Unterschied, keine Abhängigkeit oder kein Zusammenhang vorliegt), um die Bedeutung des Teststatistikwerts einordnen zu können, den man bei der eigentlichen gezogenen Stichprobe erhalten hat. Damit blieben konkrete Formeln, Tabellen oder Herleitungen für Teststatistikverteilungen eher im Hintergrund, da vor allem die Berechnungen durch Simulationen mit dem Statistikprogramm R entlastet wurden. Das Training deckte die folgenden Inhaltsbereiche ab: Skalenniveaus, p-Werte, t-Test, Chi-Quadrat-Test, Korrelation, Regression, Simpson-Paradoxon. Ein weiterer Schwerpunkt des Trainings lag auf der Auseinandersetzung mit typischen Fehlvorstellungen in Anlehnung an den Aufbau negativen Wissens (Oser et al., 1999). Weiter wird eine empirische Studie vorgestellt, in der das Training mit einer Gruppe von über 40 Schüler:innen (Jahrgangsstufen 8-12) erprobt wurde. Die im Beitrag vorgestellten

Ergebnisse sollen Hinweise geben, inwiefern ein entsprechendes Training typische Fehlvorstellungen zu Signifikanztests bei Schüler:innen bereits beim Ersterwerb inferenzstatistischer Konzepte vermeiden kann. Ein analoger, simulationsbasierter Zugang ist auch für den Binomialtest anwendbar, auf den sich viele geltende gymnasiale Lehrpläne in Deutschland beschränken.

2. Theoretischer Hintergrund

2.1 Signifikanztests im Data Literacy Framework

Schüller et al. (2019) betonen im Data Literacy Framework die Bedeutung der Förderung der Datenkompetenz in unserer Gesellschaft. Der bisherige Fokus vieler Statistik-Lehrpläne an Schulen und Hochschulen ist jedoch nach wie vor die Vermittlung prozeduraler Fähigkeiten, wobei ein kompetenter Umgang mit Daten andere Zugänge benötigt (vgl. Ridgway, 2016). Beim Thema Signifikanztests wird dies besonders deutlich: Ein mechanisches Abarbeiten der Prozedur des Signifikanztests erscheint vor dem Hintergrund der dokumentierten Fehlvorstellungen zu Signifikanztests wertlos, wenn das dahinterliegende Konzept nicht mit vermittelt wird (siehe 2.2.).

Während Gal (2002) von „Statistical Literacy" spricht, um die Kompetenz zu beschreiben, richtige Schlussfolgerungen aus vorliegenden statistischen Daten zu ziehen und diese kritisch zu beurteilen, wird diese Sichtweise durch Gould (2017) zu einer Data Literacy erweitert, die auch Aspekte wie Datenerfassung und Datenerstellung beinhaltet. Aktuelle Rahmenmodelle zu Data Literacy, wie das Data Literacy Framework (Schüller et al., 2019) beinhalten in den Kompetenzfeldern auch den Aspekt der Datenethik, in dem sich die kritische Auseinandersetzung mit p-Werten, dem verbreiteten Umgang mit ihnen und den Ergebnissen von Signifikanztests verorten lässt.

Der Arbeitskreis Stochastik der GDM (2003) empfiehlt als Mindestziel für die Sekundarstufe II im Umgang mit Signifikanztests: „Die Schüler haben an Beispielen grundlegende Probleme statistischer Schlussweisen kennengelernt. Sie verstehen das prinzipielle Vorgehen bei einem Signifikanztest und können anhand eines Beispiels erklären, was eine signifikante Abweichung vom Erwartungswert ist. Insbesondere wissen sie, bei welchen Fragestellungen Signifikanztests ein angemessenes Werkzeug darstellen und bei welchen Problemen diese Tests keine brauchbaren Aussagen liefern. Sie sind imstande, Aussagen über Wahrscheinlichkeiten für Fehler 1. und 2. Art korrekt aufzustellen und zu interpretieren."

Nicht zuletzt gehört es – besonders im Kontext von „Fake News" – zur Kompetenz eines mündigen Bürgers, statistische Argumente anderer Personen kritisch hinsichtlich korrekter und falscher Interpretationen evaluieren zu können.

2.2 Empirische Befunde zur Schwierigkeit des Begriffs „signifikant"

Zahlreiche Studien belegen die Schwierigkeiten und typischen Fehlvorstellungen, die sich sogar bei Wissenschaftler:innen oder Statistik-Dozent:innen im Verständnis von Signifikanztests und der Interpretation von p-Werten zeigen (Oakes, 1986; Ioannidis, 2005; Haller & Krauss, 2002; Badenes-Ribera, Frias-Navarro, Lotti, Bonilla-Campos & Longobardi, 2016; Herrera-Bennett, Heene, Lakens & Ufer, Preprint). Eine unreflektierte Verwendung von Signifikanztests und entsprechende Fehlvorstellungen, was ein signifikantes Testergebnis bedeutet, werden auch in der wissenschaftlichen Praxis als eine mögliche Ursache für die sogenannte „Replikationskrise" gesehen, also für die Tatsache, dass sich viele Forschungsergebnisse nicht reproduzieren lassen (Hirschauer et al., 2016).

Ein Teil des Problems: Die heute eingesetzten Signifikanztests kombinieren Verfahren und Begriffe, die in den 1920er und 1930er Jahren in verschiedenen Strömungen innerhalb der Inferenzstatistik von Ronald A. Fisher (Fisher, 1925), sowie Jerzy Neyman und Egon Pearson entwickelt wurden (Neyman & Peason, 1933): p-Werte, Nullhypothese, Alternativhypothese, Fehler 1. Art, Fehler 2. Art. Jedoch entstammen diese Begriffe eigentlich aus (mindestens) zwei verschiedenen „Schulen". Ein typisches kombiniertes Verfahren prüft nun p-Werte gegen ein zuvor festgelegtes Signifikanzniveau α und ist teils zu einer Art „Ritual" verkommen, welches in der Wissenschaft schlimmstenfalls undurchdacht abgearbeitet wird (Gigerenzer, Krauss & Vitouch, 2004). Die Betonung von p-Werten ist in der Konzeption von Signifikanztests von Fisher (1925) entstanden, während die Unterscheidung der Fehler 1. und 2. Art sowie das Aufstellen einer alternativen Hypothese H1 von Neyman und Pearson (1933) entwickelt wurden – und zwar eigentlich, um sich von der von Fisher (1925) vorgeschlagenen Methode abzugrenzen. Während der p-Wert aus einer einzelnen Stichprobe gewonnen wird und als Maß für die Evidenz gegen

eine einschlägige Nullhypothese verwendet wird, entspricht der Gedanke eines zuvor festgelegtem Signifikanzniveaus α einer frequentistischen Sichtweise mit wiederholter Anwendung des Tests – und ohne die Verwendung eines p-Wertes (Passon & von der Twer, 2020). Problematisch wird es beispielsweise bei Verwechslungen von p-Werten und dem Fehler 1. Art, die oft aus einer Vermischung der beiden Schulen entstehen: Hier wird der p-Wert von Fisher (1925), der *nachher* aus den Daten berechnet wird, mit einem (nach Neyman und Pearson, 1933) *vor der Studie* festgelegten maximalen Risiko für einen Fehler 1. Art vermischt, um eine Entscheidung zu treffen. Ein guter Überblick über die verschiedenen Sichtweisen und die Entstehung einer Hybrid-Variante findet sich in Gigerenzer (2004) oder Schneider (2015).

Fehlinterpretationen in eingereichten wissenschaftlichen Artikeln haben sogar die American Statistical Association (ASA) dazu bewogen, ein offizielles Statement zu typischen Fehlern im Umgang mit Signifikanztests herauszugeben (Wasserstein & Lazar, 2016). Andere Journale verbieten sogar die empirische Evidenz anhand von p-Werten zu berichten (z. B. das Journal „Basic and Applied Social Psychology“, Trafimow & Marks, 2015). Als Lösungsansatz wird vorgeschlagen, wie beispielsweise bereits von der American Psychological Association empfohlen (APA, 2010), alternativ Konfidenzintervalle zu verwenden (Cumming, 2012), auch wenn diese wiederum mit entsprechenden Fehlvorstellungen verbunden sind (Hoekstra et al., 2014, Herrera-Bennett et al., Preprint).

Nicht zuletzt aufgrund der Replikationskrise haben sich zahlreiche Studien mit typischen Fehlvorstellungen im Zusammenhang mit Signifikanztests und p-Werten beschäftigt. Dabei sind verschiedene Klassen von typischen Fehlern konzeptualisiert worden und je nach Anzahl der herausgearbeiteten Fehlerkategorien wird in Publikationen von den *big five* (Kline, 2013), dem *dirty dozen* (Goodman, 2008) oder den *fantastischen Vier* (Passon & von der Twer, 2020) gesprochen.

Die richtige Interpretation eines p-Wertes lautet: Der p-Wert ist die Wahrscheinlichkeit dafür, die beobachteten oder sogar noch extremere Werte für eine Teststatistik zu erhalten (z. B. diesen oder einen noch extremeren Unterschied im Falle eines t-Tests im 2-Stichprobenfall), gegeben die Nullhypothese ist wahr (z. B. es gibt keinen Unterschied). Tabelle 1 zeigt sieben Fehlvorstellungen, die wir im Folgenden näher beschreiben. Manche Personen unterliegen dem Fehlschluss, dass Ergebnisse von Signifikanztests als *eindeutige Beweise* für bestimmte Aussagen zu werten sind. Die Unsicherheit der Aussagen, die durch das Ziehen von Stichproben zustande kommt, wird hierbei übersehen. Dieser Fehlschluss wurde beispielsweise von Oakes (1986) sowie Haller und Krauss (2002) beschrieben. Obwohl dieses Fehlkonzept durch eine statistische Grundbildung einfach zu beheben sein sollte, war der Prozentsatz der Personen, die diesem Fehlkonzept unterliegen, beachtlich: In Haller und Krauss (2002) glaubten beispielsweise 34% der Psychologiestudierenden, ein signifikantes Ergebnis beweise eindeutig, dass die Nullhypothese falsch ist.

Auch der *Inverse-Wahrscheinlichkeits-Fehlschluss* wurde in der Literatur vielfach beschrieben und untersucht (siehe z. B. Oakes 1986; Shaver, 1993; Kirk, 1996; Haller und Krauss, 2002; Herrera-Bennett et al., Preprint; Passon & van der Twer, 2020). Bei diesem Fehlschluss wird fälschlicherweise angenommen, es handelt sich beim p-Wert um die Wahrscheinlichkeit, dass die Nullhypothese wahr ist, unter Vorliegen der Daten, obwohl dies eine invertierte Sichtweise darstellt.

Der Replikations-*Fehlschluss* wurde ebenfalls bereits vielfach rezipiert oder in Studien untersucht (Carver, 1978; Oakes, 1986; Greenwald, 1996; Haller & Krauss, 2002; Passon & van der Twer, 2020). Beim Replikations-Fehlschluss unterliegen Personen der Fehlvorstellung, der p-Wert gebe die Wahrscheinlichkeit dafür an, dass das Ergebnis repliziert werden könne. Bei einem p-Wert von 3 % würde dies demnach bedeuten, dass man im Schnitt nur bei 3 von 100 entsprechenden Studien mit vergleichbaren Stichproben *kein* signifikantes Ergebnis erhalten würde. Eine ähnlich klingende Aussage über p-Werte ist aber durchaus zutreffend: Testen wir wahllos 100 verschiedene Hypothesen, die in der Realität alle nicht zutreffen, ist zu erwarten, dass bei einem Signifikanzniveau von 5 % ca. 5 Tests ein signifikantes Ergebnis liefern. Aus diesem Grund sollten bei der Durchführung mehrerer Signifikanztests auch Korrekturen vorgenommen werden, wie beispielsweise die Bonferroni-Korrektur (Sedgwick, 2012). Hirschauer et al. (2016) berichten außerdem noch von *Fehlschlüssen beim Überschreiten des Signifikanzniveaus*, die die asymmetrische Anlage von Signifikanztests nicht beachten.

Fehlvorstellung **Autoren**	**Erklärung der Fehlvorstellung**	**Antwort auf die Aussage**
Eindeutiger Beweis Oakes (1986), Haller & Krauss (2002)	Signifikante Ergebnisse werden als eindeutiger Beweis für das Zutreffen der Nullhypothese oder der Alternative gesehen.	Falsch, denn auch ein signifikanter p-Wert kann eine Hypothese nicht endgültig beweisen oder widerlegen.
Inverse-Wahrscheinlichkeits-Fehlschluss Passon & van der Twer (2020), Oakes (1986), Shaver (1993), Kirk (1996), Haller & Krauss (2002), Herrera-Bennett et al. (Preprint)	Der p-Wert gibt die Wahrscheinlichkeit für das Zutreffen der Nullhypothese (ggfs. unter Vorliegen der Daten) an.	Falsch, denn der p-Wert gibt die Wahrscheinlichkeit an, den gefundenen Wert der Teststatistik (oder einen noch extremeren) zu finden, wenn in Wahrheit die Nullhypothese gilt. Der p-Wert gibt also eine Wahrscheinlichkeit an, die das Zutreffen der Nullhypothese voraussetzt.
Replikations-Fehlschluss Carver (1978), Oakes (1986), Greenwald (1996), Haller & Krauss (2002), Passon & van der Twer (2020), Herrera-Bennett et al. (Preprint)	Der p-Wert sagt etwas über die Wahrscheinlichkeit aus, die signifikanten Ergebnisse replizieren zu können.	Falsch, denn der p-Wert lässt keinen Schluss auf Messwiederholungen in diesem Sinne zu.
Fehlschlüsse beim Überschreiten des Signifikanzniveaus Hirschauer, Mußhoff, Grüner, Frey, Theesfeld & Wagner (2016)	Ein p-Wert größer als das Signifikanzniveau bedeutet, dass die Studienergebnisse zeigen, dass die Nullhypothese gilt.	Falsch, denn die Tests sind asymmetrisch angelegt und ein großer p-Wert kann daher nicht als eine Bestätigung der Nullhypothese interpretiert werden.
Fehlschlüsse beim Unterschreiten des Signifikanzniveaus --	Ein p-Wert kleiner als das Signifikanzniveau bedeutet, dass die Studienergebnisse gegen die Hypothese der Autor:innen sprechen.	Falsch, denn ein kleiner p-Wert bedeutet, dass es sehr unwahrscheinlich ist, die vorliegenden Ergebnisse zu erhalten, wenn in Wahrheit die Nullhypothese gilt.
Relevanz-Fehlschluss Kirk (1996), Herrera-Bennett et al. (Preprint)	Interpretation von „statistischer Signifikanz“ im Sinne von (klinisch/wissenschaftlich) bedeutungsvoll bzw. relevant oder wichtig.	Falsch, denn Signifikanz macht keine Aussage über die Relevanz oder Größe eines Effektes.
Effektstärken-Fehlschluss Herrera-Bennett et al. (Preprint)	Es wird ein unmittelbarer Zusammenhang zwischen der Effektstärke und der Größe des p-Wertes angenommen, im Sinne von: Ein kleiner p-Wert ist automatisch mit einer großen Effektstärke assoziiert.	Falsch, denn der p-Wert macht keine Aussage über die Größe von Effekten und hängt insbesondere auch von der Stichprobengröße ab.

Tab. 1: Einige typische Fehlvorstellungen im Zusammenhang mit dem p-Wert

Würde man bei einem t-Test einen p-Wert über dem Signifikanzniveau erhalten, so heißt das nicht, dass man damit statistisch (signifikant) die Gleichheit der Gruppen A und B nachgewiesen hat, denn in diese Richtung ist der Test nicht ausgelegt (Sedlmeier & Gigerenzer, 1989). Ein solcher Schluss würde voraussetzen, dass die von der Stichprobengröße abhängige statistische Power des Tests, eine (festzulegende) minimale relevante Effektstärke zu identifizieren, hinreichend hoch ist.

In der vorliegenden Studie werden wir in analoger Weise *Fehler* untersuchen, die beim *Unterschreiten*

des Signifikanzniveaus zu erwarten sind und die das grundlegende Prinzip des Testens betreffen. Es soll darunter die Fehlvorstellung verstanden werden, ein kleiner p-Wert spreche für die Nullhypothese. In Wirklichkeit ist aber ja bei einem kleinen p-Wert gerade die Wahrscheinlichkeit klein, diesen oder einen noch extremeren Teststatistikwert rein zufällig zu erhalten, wenn die Nullhypothese richtig ist. Die Fehlvorstellung deutet darauf hin, dass basale Kenntnisse bezüglich der prinzipiellen Anlage von Signifikanztests fehlen.

Passon und van der Twer (2020) sowie Herrera-Bennett (Preprint) sprechen außerdem von einem *Relevanz-Fehlschluss*, wenn das Wort „statistisch signifikant" fälschlicherweise mit relevant bzw. (klinisch) bedeutsam oder wissenschaftlich bedeutsam gleichgesetzt wird. Inwiefern ein statistisch signifikantes Ergebnis auch aus wissenschaftlicher Perspektive als relevant erachtet werden muss, hängt jedoch von weiteren Faktoren ab.

Eine ebenfalls hartnäckige Fehlvorstellung ist der *Effektstärken-Fehlschluss* (Herrera-Bennett, Preprint), der sicherlich auch auf die umgangssprachliche Wortbedeutung von „signifikant" zurückzuführen ist. Hierbei wird zwischen der Effektstärke und der Größe des p-Wertes ein direkter Zusammenhang unterstellt, im Sinne von: Ein sehr kleiner p-Wert ist automatisch mit einer großen Effektstärke assoziiert. Jedoch muss beachtet werden, dass auch die Größe der gewählten Stichprobe eine entscheidende Rolle spielt, ob kleine, mittlere oder große Effekte signifikant werden oder nicht. Dies birgt auch die Gefahr unsauberen wissenschaftlichen Arbeitens, wenn man (bewusst oder unbewusst) die Begriffe „signifikantes Ergebnis" und „großer Effekt" verwechselt. Darüber hinaus sei bemerkt, dass die p-Werte unter der Annahme der Nullhypothese gleichverteilt sind. Ist die Nullhypothese also richtig, so erhalten wir mit gleicher Wahrscheinlichkeit einen p-Wert zwischen dem 13. oder 14. Perzentil oder dem 51. und 52. Perzentil. Wenn die Nullhypothese allerdings nicht richtig ist, sind die zu erwartenden p-Werte bei kleineren Werten konzentriert (Passon & von der Twer, 2020). Die Angabe einer Begriffsunterscheidung wie „signifikant", „hoch signifikant" oder „höchstsignifikant" kann auf eine solche Fehlvorstellung hindeuten (Schneider, 2015), auch wenn die Wahl des Signifikanzniveaus selbst durchaus eine bestimmte Kosten-Nutzen-Abwägung impliziert (siehe Sterner et al., 2024).

2.3 Das Modul „Statistik" im Vertiefungskurs der gymnasialen Oberstufe in Bayern

Da statistische Tests in zahlreichen Forschungsfeldern und wissenschaftlichen Journalen – und damit auch in der öffentlichen Kommunikation von und über Wissenschaft – aber nach wie vor von zentraler Bedeutung sind, hat die schulische Beschäftigung mit Signifikanztests eine wissenschaftspropädeutische wie auch allgemeinbildende Bedeutung. Das Konzept der Signifikanz wird von vielen Medien im Rahmen der Wissenschaftskommunikation aufgegriffen. Außerdem begegnen einem Großteil der gymnasialen Schüler:innen später die entsprechenden statistischen Inhalte in Studium oder Beruf, insbesondere auch außerhalb des MINT-Bereichs (Neumann et al., 2021). An aktuellen, beispielsweise an Hochschulen vorherrschenden didaktischen Konzepten kann kritisiert werden, dass sie zwar konkrete Rechenverfahren und die dazu notwendigen Teststatistiken und statistischen Verteilungen behandeln, konzeptuelles Wissen zu den von diesen Verteilungen beschriebenen Phänomenen (Freudenthal, 1983) – insbesondere die unter der Nullhypothese zu erwartende Samplingvarianz – aber vielfach vernachlässigen. Ebenso liegt der Fokus in den meisten gymnasialen Lehrplänen hauptsächlich auf einer schematischen Anwendung des Binomialtests, da hier die rechnerische Herleitung der Teststatistik auch mit den mathematischen Mitteln der Oberstufe möglich ist. An bayerischen Gymnasien werden nach dem neuen LehrplanPlus auch aus diesem Grund seit dem Schuljahr 2024/2025 erstmals p-Werte, Chi-Quadrat-Tests, t-Test, Skalenniveaus sowie die Themen Korrelation und Regression in einem Vertiefungskurs „Statistik" unterrichtet (ISB, 2024). Seit dem Schuljahr 2024/2025 werden in Bayern Vertiefungskurse in Deutsch und Mathematik angeboten. Da die Schüler:innen in Bayern nicht die Möglichkeit haben, diese beiden zentralen Fächer als Leistungsfach zu wählen, bieten diese Vertiefungskurse Schüler:innen mit besonderem Interesse in diesen Gebieten die Möglichkeit, einer intensiveren und gegebenenfalls auch studienvorbereitenden Auseinandersetzung in der Q12. Die Vertiefungskurse sind zweistündig zusätzlich zum normalen Mathematikunterricht und nicht abiturrelevant. Im Vertiefungskurs Mathematik wählt die Lehrkraft hierbei aus den fünf Modulen (1. Komplexe Zahlen, 2. Folgen und Reihen, 3. Matrizen, 4. Zahlentheorie und Kryptologie und 5. Statistik) drei Module aus, um diese in dem zweistündigen Fach ein Schuljahr lang zu unterrichten. Dabei werden im Lehrplan die

folgenden Kompetenzerwartungen im Modul „Statistik“ formuliert:

„Die Schülerinnen und Schüler …

> *…erläutern anhand von Beispielen die wesentlichen Eigenschaften der unterschiedlichen Skalenniveaus und unterscheiden dabei insbesondere nominalskalierte, ordinalskalierte und metrische (intervall- und verhältnisskalierte) Variablen. Sie stellen Zusammenhangs-, Unabhängigkeits- und Unterschiedshypothesen, auch zu gesellschaftsrelevanten Fragestellungen, bezüglich je zweier Variablen auf.*
>
> *…erläutern anhand von Streudiagrammen die Grundidee der linearen Regression und stellen Gleichungen von Regressionsgeraden auf, um Werte von Variablen im Sinne einer Vorhersage zu schätzen. Sie berechnen die Werte von Korrelationskoeffizienten und interpretieren diese als Maß für den Zusammenhang zweier Größen. Dass bei der Betrachtung dieses Zusammenhangs Fehlinterpretationen auftreten können (z. B. Simpson-Paradoxon), weisen sie anhand von geeigneten Beispielen graphisch und rechnerisch nach.*
>
> *…beschreiben das grundsätzliche Vorgehen beim Chi-Quadrat-Unabhängigkeitstest oder beim t-Test und bestimmen mithilfe des betrachteten Testverfahrens bei geeigneten Beispielen von Datensätzen p-Werte unter Verwendung einer Statistik-Software. Sie interpretieren ihre Ergebnisse im Sachzusammenhang.*
>
> *…prüfen mediale Darstellungen von Daten auf Korrektheit und analysieren den manipulativen Charakter fehlerhafter Darstellungen anhand typischer Beispiele.“ (ISB, 2024)*

Im Folgenden wird die didaktische Schwerpunktsetzung sowie der konkrete Aufbau eines Kurses zur Förderung des Verständnisses für Signifikanztests (und die soeben vorgestellten angrenzenden Gebiete) erläutert. Der Kurs erläutert Signifikanztests 1. über konzeptuelles Verständnis statt "Null Ritual", 2. simulationsbasiert, 3. mittels Statistik-Software und 4. mit einem Fokus auf dem Aufbau negativen Wissens (Oser et al. 1999).

2.4 Didaktische Schwerpunkte: Konzeptuelles Verständnis statt „Null Ritual“

Im schulischen Stochastikunterricht werden Signifikanztests oft formelbasiert unterrichtet mit einem Schwerpunkt auf prozeduralem Wissen im Sinne einer konkreten Berechnung von Annahme- und Ablehnungsbereichen. Ein Vorgehen, das auf einer undurchdachten Durchführung eines „Kochrezepts“ beruht, wurde in der Vergangenheit auch als „Null Ritual“ bezeichnet (Gigerenzer et al., 2004). Jedoch kann das reine Abarbeiten einer Prozedur zu einer oberflächlichen Verständnisweise führen, da die zugrunde liegenden Konzepte erst in den Vordergrund treten können, wenn rein rechnerische Aspekte des Signifikanztests in den Hintergrund treten dürfen, um konzeptuelles Wissen aufbauen zu können. Dies ist auch deshalb so wichtig, weil die händische Berechnung bei Signifikanztests (je nach Art des Tests) oft sehr zeitintensiv ist und viele Rechnungen oder spezielles Wissen beispielsweise zur Handhabung von Tafelwerken oder entsprechender Software erfordert. Liegt der Fokus hingegen auf dem zugrundeliegenden Konzept (z. B. indem die konkreten Berechnungen computergestützt ausgeführt werden), könnten typische Fehlvorstellungen, die beim Thema Signifikanztests sehr hartnäckig sind, möglicherweise vermieden werden. Zu kurz kommt in derartigen Ansätzen häufig die zentrale konzeptuelle Bedeutung des p-Wertes als Cut-Off in einer Sampling-Verteilung der Teststatistik unter der Nullhypothese. Dieses Konzept der Samplingvarianz erscheint zentral für das Verständnis von p-Werten. Biehler et al. (2023), Krauss und Wassner (2001), Meyfarth und Biehler (2006), Griese et al. (2020), Weber (2020) und Ufer (2022) empfehlen daher Signifikanztests in der gymnasialen Oberstufe mithilfe von p-Werten anstelle von Annahme- und Ablehnungsbereichen einzuführen. Das Kurskonzept nutzt eine Hybridvariante der Signifikanztests, bei der zunächst aber die Theorie von Fisher (1925) über die p-Werte und der Evidenz gegen die Nullhypothesen im Zentrum steht. Neben einer stärkeren Berücksichtigung konzeptuellen Wissens hat die Einführung von Signifikanztests über p-Werte auch den Vorteil, dass dieser Zugang authentisch im Sinne echter Anwendung statistischer Methoden ist (die ohne Tabellenbuch und händische Berechnungen auskommt; vgl. auch Krauss et al., eingereicht). Es wird jedoch als didaktische Reduktion eine scharfe Grenze bei α=5% eingeführt, unter der ein zugehöriger p-Wert als signifikant angesehen wird (vgl. auch Podworny, 2019).

Signifikanztests durch Simulationen verstehen

Um das Konzept der Samplingvarianz im Kontext von Signifikanztests zugänglich zu machen, müssen wiederholt Stichproben aus einer Population gezogen werden, die der Nullhypothese unterliegt, die wir aus diesem Grund im Rahmen des Kurses auch als

„Nullpopulation“ bezeichnet haben. Dies ist mangels entsprechender Populationen praktisch meist nicht möglich. Die Didaktik der Stochastik empfiehlt simulationsbasiertes Lernen, das hier konkrete Erfahrungen zu diesem Prozess ermöglichen und so zu einer verbesserten Beherrschung der Konzepte des Signifikanztests und des p-Wertes führen kann. Es werden hierfür auch konkrete Vorschläge zur Umsetzung unterbreitet (Mayfarth, 2008; Podworny, 2019; Chance et al., 2022; Chandrakantha, 2020). Dieses Prinzip wird auch im vorliegenden Kurskonzept genutzt, da der p-Wert im Sinne einer authentischen Anwendung von Statistik im Mittelpunkt stehen soll und p-Werte erst mithilfe entsprechender Simulationen für Lernende greifbar werden (vgl. Abschnitt 2.5). Diese Herangehensweise ermöglicht es den Lernenden, eine konkrete (approximative) Vorstellung zur Bedeutung des p-Wertes aufzubauen, indem sie die Simulation von Stichproben und die Beurteilung der Wahrscheinlichkeit von Ergebnissen unter der Annahme der Nullhypothese durchführen. Durch die direkte Erfahrung mit den simulierten Stichproben unter der Nullhypothese können die Lernenden zudem ein tieferes Verständnis dafür entwickeln, wie der p-Wert interpretiert werden kann und wie nicht.

Durchführung von Signifikanztests mithilfe entsprechender Software

In der gymnasialen Oberstufe werden Signifikanztests (wie z. B. Binomialtests) häufig mittels Tabellenbüchern und Taschenrechnern gelehrt, obwohl diese in authentischen Anwendungen mithilfe entsprechender statistischer Software durchgeführt werden (vgl. auch Krauss et al., eingereicht). Ein Lehransatz, der softwaregestützte Methoden verwendet, würde Schüler:innen eine authentischere Ausbildung in Datenanalyse bieten, weswegen wir im vorliegenden Kurskonzept Signifikanztests (simulationsbasiert) mithilfe von R erarbeiten und auch die Schüler:innen mit dieser Software arbeiten dürfen (vgl. auch Chandrakantha, 2020; Podworny, 2019 sowie Mayfarth, 2008).

Aufbau negativen Wissens

Bei einem Themenbereich, in dem sich typische Fehlvorstellungen so hartnäckig halten (selbst bei Personen mit entsprechenden grundlegenden bzw. sogar fortgeschrittenen Statistik-Kenntnissen), sollte 1. dieses Fehlerwissen einerseits als wesentliche Wissensfacette von Mathematiklehrkräften angesehen werden und 2. dieses negative Wissen auch mit Schüler:innen explizit thematisiert werden (vgl. auch Oser et al., 1999). Lernende haben das Konzept der Signifikanztests und die Bedeutung von p-Werten erst dann ausreichend verinnerlicht, wenn sie nicht nur artikulieren können, was ein signifikantes oder nicht-signifikantes Testergebnis bedeutet, sondern wenn sie auch abgrenzen können, welche Aussagen über signifikante bzw. nicht-signifikante Ergebnisse auf Fehlvorstellungen beruhen. Im vorgestellten Kurskonzept lag daher auch ein Schwerpunkt auf der Abgrenzung typischer Fehlvorstellungen.

2.5. Beschreibung des Kurskonzepts

Im Rahmen eines einwöchigen Kurses für begabte und interessierte Schüler:innen der Jahrgangsstufen 8-12 wurde das Thema „Einstieg in die Inferenzstatistik“ behandelt. Die Kursinhalte wurden in fünf 105-minütigen Vorlesungen erarbeitet und jeweils anschließend in einer zweistündigen Übungseinheit, getrennt nach Altersgruppen, vertieft.

Sämtliche statistische Berechnungen und Simulationen wurden mit der Statistiksoftware R durchgeführt. Um den Teilnehmenden den Installationsprozess der Software zu ersparen, wurde RStudio in einer kostenlosen Online-Umgebung (Posit-Cloud) verwendet, die alle für den Kurs relevanten Funktionen beinhaltet. Als Referenzpopulation diente ein nachgebildeter Datensatz deutscher Schüler:innen aus der PISA-Studie 2012 (N=4.545) mit 16 Variablen (ID, Geschlecht, sozio-ökonomischer Status, Migrationshintergrund, Schulart, Schulleistung in Mathematik, Deutsch und Naturwissenschaften zu je zwei Messzeitpunkten sowie Noten in Mathematik, Deutsch, Biologie, Chemie und Physik). Komplexere Operationen wie die Simulation von Stichprobenziehungen oder die Berechnung von p-Werten wurden dabei durch vordefinierte Befehle entlastet. Zudem wurden alle relevanten Code-Elemente in den Vorlesungen besprochen und anschließend über ein Online-Dokument zur Verfügung gestellt, sodass diese von den Teilnehmenden direkt während der Vorlesung kopiert, ausprobiert und entsprechend der jeweiligen Aufgabenstellungen in den Übungen leicht abgeändert werden konnten. Anstelle der üblichen R-Scripts wurden in der Regel R-Notebooks verwendet, bei denen eine Ausgabe aller Berechnungen nicht nur über die Konsole, sondern auch direkt im Notebook unter den jeweiligen Code-Abschnitten erscheint. Die R-Notebooks können zudem auch als übersichtliche HTML-Dateien abgespeichert werden und somit (auch zu späteren Zeitpunkten) unabhängig von RStudio geöffnet werden. Ein Überblick über die behandelten Vorlesungsinhalte wird in Tabelle 2 gegeben.

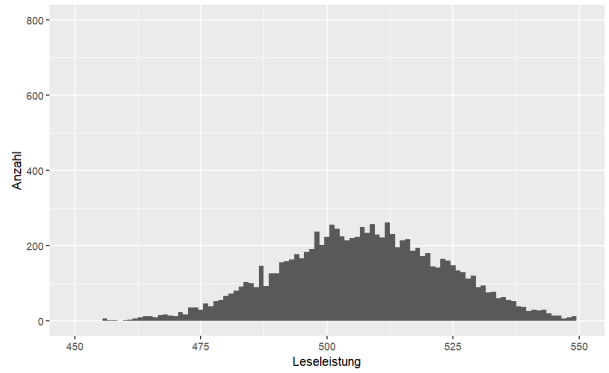
Tag 1	Tag 2	Tag 3	Tag 4	Tag 5
Statistische Grundbegriffe Lage- und Streuungsparameter Erwartungstreue Skalenniveaus Grundtypen von Forschungsfragen und Hypothesen Empirisches Gesetz der großen Zahlen/Samplingvarianz	t-Test (Mittelwert gegen festen Wert) Grundprinzip der Inferenzstatistik (simulationsbasiert)	Grundsätzliche Fehlertypen beim Testen von Hypothesen Chi-Quadrat Unabhängigkeitstest und zugehörige Effektstärke t-Test und adaptierte Effektstärke für den t-Test Kritik an Signifikanztests Interpretation des p-Wertes	Streudiagramme Korrelation und Test auf Korrelationskoeffizienten Regressionsanalyse und Interpretation von Regressionsgleichungen	Korrelation ≠ Kausalität Simpson-Paradoxon Will-Rogers-Phänomen HARKing/ data mining Publication bias

Tab. 2: Übersicht der Inhalte in den Vorlesungen und Übungen

In der ersten Vorlesungseinheit wurden wichtige statistische Grundbegriffe (z. B. Merkmal, Merkmalsausprägung), Lage- und Streuungsparameter und die Ordnung der Skalenniveaus behandelt. Anhand von Beispielen diskreter und metrischer Merkmale wurden analog zum neuen bayerischen Lehrplan drei Grundtypen inferenzstatistischer Hypothesen eingeführt (Abhängigkeitshypothesen, Unterschiedshypothesen, und Zusammenhangshypothesen). Für zwei kategoriale Merkmale werden Abhängigkeitshypothesen verwendet, für Gruppenunterschiede (kategoriales Merkmal) in einem metrischen Merkmal werden Unterschiedshypothesen formuliert, wohingegen Zusammenhangshypothesen zwei metrische Merkmale zueinander in Beziehung setzen. Abschließend wurde das empirische Gesetz der großen Zahlen im Kontext der Inferenzstatistik als abnehmende Samplingvarianz einer Teststatistik bei zunehmender Stichprobengröße illustriert. Der Begriff Samplingvarianz beschreibt dabei die qualitative Streuung der auftretenden Werte einer Teststatistik verschiedener, zufällig gezogener Stichproben aus derselben Population. Die Ausprägung der Samplingvarianz, das heißt wie stark die Werte der Teststatistik für Stichproben um den wahren Wert in der Population streuen, hängt dabei maßgeblich von der Größe der Stichprobe ab. Gemäß dem empirischen Gesetz der großen Zahlen werden größere Abweichungen von diesem Wert mit zunehmender Stichprobengröße unwahrscheinlicher beziehungsweise seltener (siehe Abb. 1).

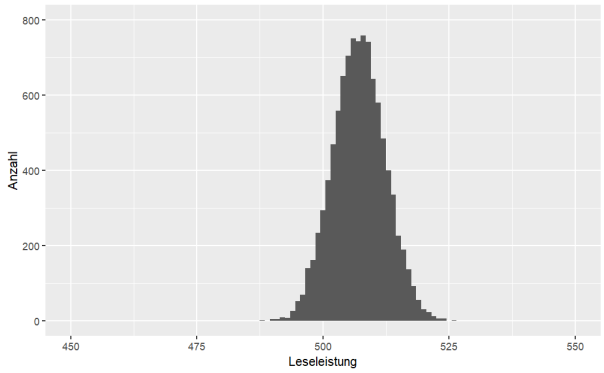
In der zweiten Vorlesungseinheit wurde das grundlegende Vorgehen der Inferenzstatistik präsentiert. Dieses besteht darin, abzuschätzen, ob der in einer Stichprobe gefundene Wert einer Teststatistik noch plausibel über die Samplingvarianz erklärbar ist, wenn man annimmt, dass die Nullhypothese gilt (siehe Abb. 2). Beispielsweise interessiert man sich für den Mittelwert μ_P eines Merkmals in einer Population P und es soll untersucht werden, ob sich dieser Mittelwert von einem bekannten Wert, beispielsweise dem Mittelwert einer anderen Gruppe, unterscheidet. Die Teststatistik ist damit das arithmetische Mittel der vorliegenden Merkmalsausprägungen. Als konkreten Beispielkontext könnte man sich ein bestimmtes Lehrkonzept für den Mathematikunterricht vorstellen („P-Learn“), das an einigen Gymnasien in Deutschland unterrichtet wird. Die Schüler:innen, die nach diesem Konzept unterrichtet werden, stellen die Population P dar. Nun möchte man wissen, ob sich die Mathematikleistung dieser Population P von der durchschnittlichen Mathematikleistung aller Schüler:innen der Referenzpopulation N (simulierter PISA-Datensatz) unterscheidet. Dieser Wert wurde anhand der vorliegenden Daten zum PISA-Test in der (simulierten) Referenzpopulation zu $\mu_N = 513$ bestimmt. Die Nullhypothese H_0 lautet also, dass kein Unterschied vorliegt und somit „$\mu_P = \mu_N$“ gilt. Um die Frage zu untersuchen, wird eine zufällige Stichprobe aus 30 P-Learn-Schüler:innen gezogen, welche den PISA-Test bearbeiten sollen.

a)



10.000 Stichproben, N=30

b)



10.000 Stichproben, N=300

Abb. 1: Veranschaulichung der Samplingvarianz. a) zeigt die Verteilung der Mittelwerte der Leseleistung von 10.000 Stichproben mit einer Stichprobengröße von je N=30 aus dem PISA-Datensatz. Bei b) wurde die Stichprobengröße auf N=300 erhöht. In beiden Fällen streuen die Werte um den Populationsmittelwert 507. Bei höherer Stichprobengröße nimmt jedoch die Breite der Verteilung ab und die Mittelwerte der Stichproben liegen näher am Populationsmittelwert.

Dabei wird eine durchschnittliche Mathematikleistung von $\hat{\mu}_P = 530$ ermittelt. Der Wert $\hat{\mu}_P$ ist damit ein *Schätzwert* für den wahren Mittelwert der Mathematikleistung μ_P in der Population P aller nach P-Learn unterrichteten Schüler:innen. Was bedeutet es nun, dass der Mittelwert der Stichprobe höher ist als der Mittelwert aller deutschen Schüler:innen der Referenzpopulation $\mu_N = 513$? Kann man guten Gewissens behaupten, dass P-Learn bessere Ergebnisse erzielt als herkömmlicher Unterricht? Von dieser Fragestellung, die zu einer einseitigen Testung führen würde, wurde übergegangen zu der Frage, ob es sich hierbei lediglich um eine „normale“ Schwankung handle, die im Rahmen der Samplingvarianz bei einer Stichprobe der Größe N=30 plausibel erscheint und zu erwarten ist. Die Überprüfung, welche Abweichungen typischerweise vorkommen (egal in welche Richtung), wenn der Mittelwert bei 513 Punkten liegt, ist mit einer zweiseitigen Testung verknüpft. Um diese Frage zu beantworten, hilft die Betrachtung einer *Nullpopulation*, das heißt, einer Menge von Schüler:innen, die im Mittel genau die Mathematikleistung μ_N erbringen und sich sonst nicht von der Population P unterscheiden (insbesondere die Standardabweichung der Mathematikleistung von Population P und Nullpopulation ist gleich). Den Begriff Nullpopulation wählen wir analog zum Begriff Nullmodell (siehe auch Podworny, 2019; S. 211), um hervorzuheben, dass es sich hierbei um eine Modellannahme handelt, in der es – je nach Zusammenhangs-, Unterschieds- oder Abhängigkeitshypothese – eben keinen Zusammenhang, keinen Unterschied oder keine Abhängigkeit zwischen den betrachteten Merkmalen gibt. Im Normalfall existiert diese Nullpopulation in der Realität nicht und muss simuliert werden. In unserem Beispiel können wir jedoch die konkrete Menge der Schüler:innen der Referenzpopulation als Nullpopulation wählen, da auf diese Menge die beschriebenen Anforderungen bereits zutreffen.

Nun werden aus dieser Nullpopulation sehr viele (z. B. 10.000) Stichproben der Größe N=30 gezogen, deren Mittelwerte der Mathematikleistung sich gemäß der Samplingvarianz um den Wert $\mu_N = 513$ verteilen. Jetzt kann der Anteil der Stichproben bestimmt werden, der mindestens 17 Punkte vom Mittelwert der Nullpopulation entfernt liegt, deren Mittelwert für die Mathematikleistung also mindestens 530 oder höchstens 496 beträgt. In diesem Fall sind das 3.308 Stichproben bzw. 33,08 %. Dieser Prozentsatz, genannt p-Wert, gibt die Wahrscheinlichkeit an, einen Wert der Teststatistik wie in der Population P (also den Mittelwert 530) zu finden oder einen Wert, der sogar noch mehr vom Mittelwert der Nullpopulation abweicht, wenn die Nullhypothese gilt, also der wahre Mittelwert $\mu_P = \mu_N = 513$ beträgt. In diesem Fall gehört der vorliegende Wert der Teststatistik 530 also „nur“ zu den 33,08 % der extremer abweichenden Werte, wenn in Wahrheit die Nullhypothese gilt.

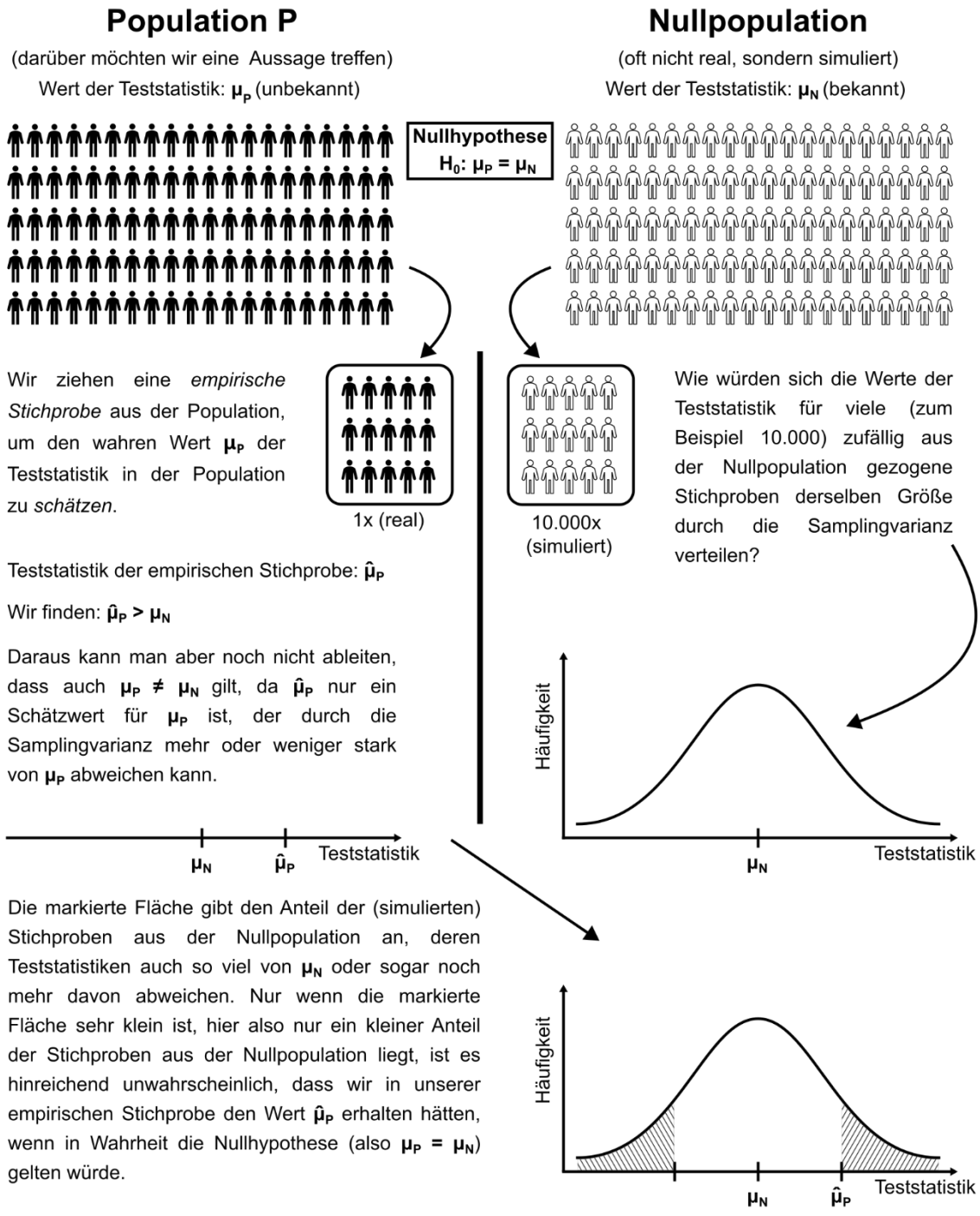


Abb. 2: Grundidee der Inferenzstatistik am Beispiel einer Unterschiedshypothese (simulationsbasiert)

Es scheint also durchaus wahrscheinlich zu sein, eine solche Abweichung auch dann zu erhalten, wenn die Nullhypothese richtig ist. Auch im Sinne einer einseitigen Testung kann man sich überlegen: Die Wahrscheinlichkeit, dass man rein zufällig eine Gruppe mit 30 Personen zieht, die so gut wie die P-Learn-Gruppe ist oder sogar besser, obwohl die 30 Personen eigentlich aus der Nullpopulation gezogen wurde, beträgt 16,54% und ist damit gar nicht so unwahrscheinlich (siehe Abb. 3). Im Sinne einer didaktischen Reduktion und in Anlehnung an Neyman und Pearson (1933) wurde eine feste Grenze von 5% als Signifikanzniveau vorgegeben (vgl. auch Podworny, 2019). Der in der Simulation erhaltene p-Wert (der eher aus der Sichtweise von Fisher, 1925 stammt), wird mit diesem vorher festgelegten Signifikanzniveau in Verbindung gebracht.

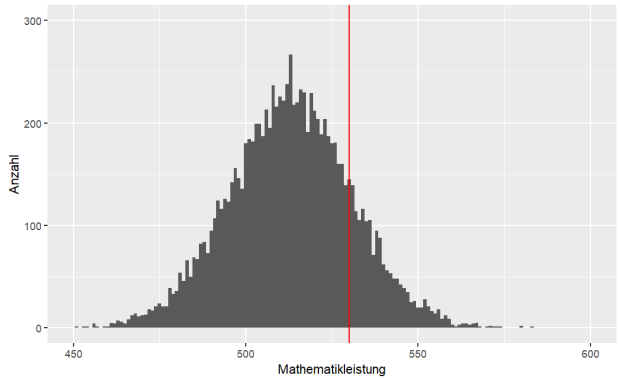


Abb. 3: Simulation von 10.000 Stichproben (N=30) aus der Nullpopulation. Die rote Linie gibt den Mittelwert $\hat{\mu}_P$ der Stichprobe an.

In unserem konkreten Fall konnten wir die Hypothese, dass nach P-Learn unterrichtete Schüler:innen abweichende Mathematikleistungen erbringen, statistisch damit nicht bekräftigen. Das bedeutet natürlich nicht, dass dadurch das Gegenteil bewiesen wurde (vgl. Fehlschluss bei Überschreiten des Signifikanzniveaus). Nur weil der Wert unserer Teststatistik „nicht völlig untypisch für die Nullpopulation" ist, können wir nicht schließen, dass unsere Stichprobe (sehr wahrscheinlich) aus der Nullpopulation kommt.

Bei einer simulationsbasierten Bestimmung eines p-Werts müsste man unendlich viele, unabhängige Stichproben aus einer Nullpopulation ziehen, um ein exaktes Ergebnis zu erhalten. In der Realität wird der p-Wert daher oft analytisch über theoretisch bestimmte Verteilungsfunktionen der Teststatistik (z. B. t-Verteilung, Chi-Quadrat-Verteilung) berechnet. Für die Binomialverteilung ist dies auch mit den mathematischen Kenntnissen der Oberstufe möglich.

In der dritten Vorlesungseinheit wurde das grundlegende Prinzip des Signifikanztestens wiederholt und herausgearbeitet, welche Arten von Fehlern beim Testen grundsätzlich unterschieden werden können. In den Übungen wurde zudem herausgearbeitet, wie sich die Wahrscheinlichkeiten der beiden möglichen Fehler bei einer festen Stichprobengröße gegenseitig beeinflussen.

Anschließend wurde zunächst ein informelles Vorgehen zur Überprüfung von Abhängigkeitshypothesen anhand einer Vierfeldertafel erarbeitet. Zur quantitativen Beschreibung der Abweichung zweier abhängiger Merkmale von deren Unabhängigkeit wurde dazu die χ^2-Statistik und der korrigierte Kontingenzkoeffizient als Maß für die Effektstärke eingeführt. Schließlich wurden simulationsbasierte χ^2-Unabhängigkeitstests anhand der simulierten Referenzpopulation durchgeführt. Diese simulierte

Referenzpopulation ermöglichte es damit auch, die jeweils „wahren" Werte der Teststatistik in der Population zum Vergleich zu bestimmen.

Anschließend wurde das Grundprinzip für den t-Test zum Vergleich zweier Mittelwerte beschrieben und anhand eines Beispiels simulationsbasiert illustriert. Ausgehend von zwei Verteilungen, die die PISA-Leistungen von Mädchen und Jungen zeigten, wurde ein Nullmodell erstellt und erneut die Verteilung der Teststatistikwerte betrachtet.

In diese Verteilung der Teststatistikwerte aus dem Nullmodell wurde wiederum der eigentlich in der Datenerhebung erhaltene Teststatistikwert eingetragen, um den p-Wert zu motivieren. Denn auf diese Weise stellt sich simulationsbasiert die Frage: Wie wahrscheinlich ist es, diesen oder einen noch extremeren Teststatistikwert zu erhalten, wenn in der Grundgesamtheit gar kein Unterschied in den Mittelwerten vorliegt. Zur Abgrenzung vom p-Wert wurde als Maß für die Effektstärke der durch den Interquartilsabstand geteilte Mittelwertunterschied verwendet (der Interquartilsabstand wurde hier als schnell einzuführendes Streumaß genutzt, weil noch nicht alle Schüler:innen Vorkenntnisse zur Standardabweichung hatten).

In der vierten Vorlesungseinheit wurde der Korrelationskoeffizient als Maß für den linearen Zusammenhang zweier metrischer Merkmale anhand von Scatterplots präsentiert und das Vorgehen für dessen Berechnung erläutert. Es wurde thematisiert, dass der Korrelationskoeffizient bereits als Effektstärke interpretiert werden kann. Überdies wurden in der Vorlesung auch Signifikanztests zur Überprüfung von Zusammenhangshypothesen vorgestellt („Unterscheidet sich der Korrelationskoeffizient signifikant von Null?"). Anschließend wurde über die Methode der kleinsten Quadrate qualitativ ein Verfahren zur Berechnung einer linearen Einfachregressionsgeraden sowie die Interpretation der auftretenden Werte vorgestellt.

Die fünfte Vorlesung widmete sich abschließend typischen Fehlvorstellungen der beschreibenden sowie Inferenzstatistik, wie die Verwechslung von Korrelation und Kausalität, das Simpson-Paradoxon, das Will-Rogers-Phänomen, aber auch HARKing (nachträgliches Entwickeln passender Theorien zur Erklärung signifikanter Befunde), p-hacking/data-mining (Berichten von signifikanten Ergebnissen ohne vorheriges Aufstellen konkreter Hypothesen) und den

publication bias (signifikante Befunde werden häufiger kommuniziert als nicht signifikante).

Die Übungsaufgaben umfassten sowohl Verständnisfragen zu den in der Vorlesung behandelten Inhalten als auch das Aufstellen eigener Hypothesen und deren Überprüfung mit geeigneten Testverfahren mithilfe der Statistiksoftware R. Speziell zur Förderung des Verständnisses von p-Werten wurden auch typische Fehlvorstellungen thematisiert.

Prävention der Fehlvorstellungen

Während der Einführung der statistischen Grundlagen und der verschiedenen Testverfahren wurde besonders auf eine präzise, probabilistische Beschreibung und Interpretation des p-Wertes geachtet, um die in Tabelle 1 beschriebenen, typischen Fehlvorstellungen nach Möglichkeit gar nicht erst aufkommen zu lassen.

Im Sinne des Aufbaus negativen Wissens wurden 6 der 31 Items (zu verschiedenen Fehlvorstellungen) in ähnlicher Form in den Übungen thematisiert. Die Fehlvorstellung, dass ein (signifikantes) Testergebnis Auskunft darüber geben kann, wie wahrscheinlich das Vorliegen der Nullhypothese ist, wurde darüber hinaus explizit in der Vorlesung eingeführt und begründet als falsch benannt (vgl. Merkmale von refutational texts, Guzzetti, 2000). Zudem wurden im Rahmen einer Diskussion typischer Kritikpunkte an Signifikanztests auch die Begriffe „signifikant“ und „Effektstärke“ kontrastiert: während die statistische Signifikanz in hohem Maße von der Stichprobengröße abhängt, sind Effektstärken davon unabhängig und beschreiben das Ausmaß eines empirischen Effekts. Beispielsweise kann ein sehr kleiner Unterschied (kleine Effektstärke) zwischen zwei Gruppen in einem Merkmal dennoch, bei ausreichend großer Stichprobe, in einem t-Test ein signifikantes Ergebnis hervorbringen. Ein signifikantes Testergebnis bedeutet also, dass es mit den vorliegenden Daten zwar hinreichend unplausibel erscheint, eine solche Stichprobe zu erhalten, wenn die Nullhypothese wahr ist. Jedoch kann daraus keine Aussage darüber abgeleitet werden, wie groß der vorliegende Effekt (z. B. Unterschied zwischen den Gruppen) ist.

Neben der expliziten Thematisierung typischer Fehlvorstellungen, wurde den Lernenden bei der eigenständigen Bearbeitung der Übungsaufgaben die Möglichkeit gegeben, selbst Fehler zu begehen und dafür individuelles Feedback zu erhalten.

3. Fragestellung der Erprobung

In der vorliegenden Studie soll der Frage nachgegangen werden, inwiefern man jungen Schüler:innen bereits ein grundlegendes Verständnis für Signifikanztests mithilfe eines einwöchigen Kurses vermitteln kann und welche Aspekte einer weiteren Vertiefung der Inhalte bedürfen, da diese in der kurzen Zeit nur rudimentär vermittelt werden können.

4. Methode

4.1 Eingesetzte Items

Zur Messung des Wissens über Signifikanztests bearbeiteten die Teilnehmenden des Kurses zu Beginn der letzten Vorlesungseinheit einen digitalen Fragebogen mit wahr/falsch-Aussagen zu Signifikanztests und p-Werten. Die 31 Items (siehe Anhang) zielten dabei auf die oben beschriebenen sieben verschiedenen Klassen typischer Fehlvorstellungen zum p-Wert ab (siehe Tabelle 2) und stellten damit – bis auf drei Ausnahmen – stets falsche Aussagen dar. Dabei wurden die Items zu den Kategorien *Inverse-Wahrscheinlichkeits-Fehlschluss*, *Replikationsfehlschluss* und *Effektstärken-Fehlschluss* noch in jeweils zwei Subgruppen unterteilt (siehe auch Abb. 4). Für den Inverse-Wahrscheinlichkeits-Fehlschluss wurde unterschieden, ob sich das Item auf die Wahrscheinlichkeit des Zutreffens der Nullhypothese (H_0) oder der Alternativhypothese (H_1) bezog. Für den Replikations-Fehlschluss wurden Items gestellt, bei denen entweder p oder 1-p die Wahrscheinlichkeit für die Replizierbarkeit angibt und beim Effektstärken-Fehlschluss wurde differenziert, ob das Item den Eindruck erweckt, der p-Wert selbst sei ein Maß für die Effektstärke (direkt) oder ob nur ein unmittelbarer Zusammenhang zwischen den beiden Größen unterstellt wird (indirekt). Insgesamt wurden die Items für drei verschiedene Situationen gestellt und bezogen sich entweder auf allgemeine Aussagen zum p-Wert, ohne dass ein solcher direkt vorgegeben wurde, oder beinhalteten die Interpretation eines konkreten p-Werts unterhalb bzw. oberhalb des gegebenen Signifikanzniveaus (siehe auch Anhang). Eine exemplarische Auswahl der Testitems ist mit einer Erläuterung bereits in Tabelle 1 dargestellt.

Zu Beginn des Kurses gaben die Teilnehmenden ihre letzte Zeugnisnote im Fach Mathematik an und beantworteten Selbsteinschätzungsskalen zum mathematischen Selbstkonzept und Interesse (Rach et al., 2021; Ufer et al., 2017).

4.2 Stichprobe

Am Kursprogramm nahmen 66 Schüler:innen im Rahmen eines außerschulischen Förderprojekts teil, was eine positiv verzerrte Stichprobe hinsichtlich motivationaler und leistungsbezogener Aspekte plausibel erscheinen lässt. Insgesamt füllten 41 Schüler:innen (27 Schüler, 13 Schülerinnen, 1 divers) aus den Jahrgangsstufen 8-12, die im Schnitt 16,1 Jahre alt waren (SD = 1,5), die Fragebögen aus.

5. Ergebnisse

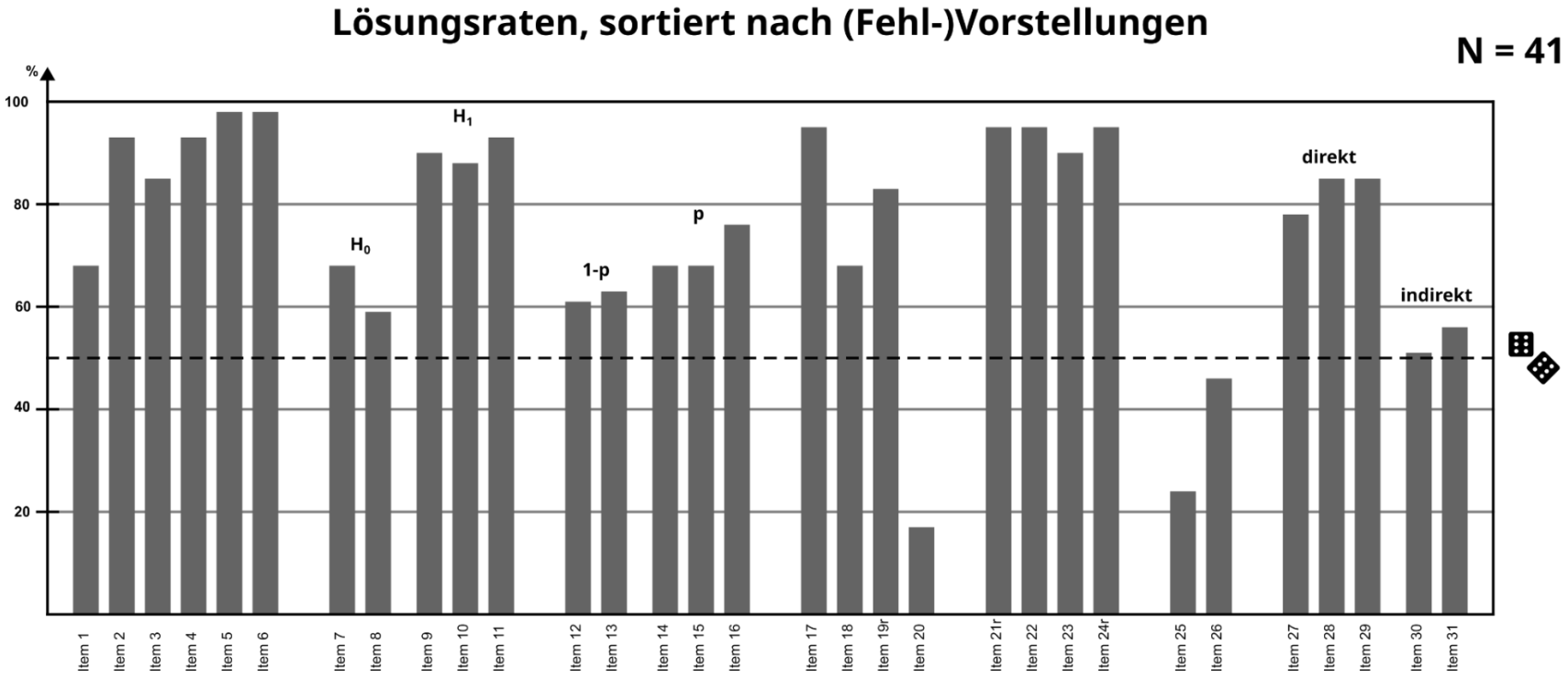
Die Lösungsraten der Fragen zum p-Wert bzw. zur Signifikanz sind in Abbildung 4 gruppiert nach den dort angesprochenen Fehlvorstellungen dargestellt. Zunächst fällt auf, dass die Lösungsraten zu Fragen innerhalb der einzelnen Fehlkonzepte – bis auf wenige Ausnahmen – sehr homogen ausfallen.

Fragen, die auf die Interpretation eines signifikanten Ergebnisses als *eindeutiger Beweis* für die Hypothese abzielten, wurden im Mittel von 89% der Schüler:innen richtig (siehe Abb. 4) und damit sehr gut gelöst. Die einzige Ausnahme hiervon stellt die Aussage des Items „Ein signifikanter p-Wert lässt den Schluss zu, dass die Hypothese der Autor:innen sicher richtig ist“ dar, welches nur von 68% der Schüler:innen korrekt als falsch markiert wurde. Insgesamt beantwortete gut die Hälfte der Teilnehmenden alle sechs Items dieser Kategorie korrekt. Bezüglich dem *Inverse-Wahrscheinlichkeits-Fehlschluss* erkannten zwar im Mittel 90% der Teilnehmenden, dass der p-Wert nicht die Wahrscheinlichkeit angibt, mit der die Hypothese in Wahrheit richtig ist. Dennoch wurde im Mittel nur in 63% der Fälle korrekt angegeben, dass der p-Wert auch *nicht* die Wahrscheinlichkeit für das Zutreffen der Nullhypothese beschreibt.

Aussagen, die auf den *Replikations-Fehlschluss* abzielten, also den p-Wert bzw. 1–p als die Wahrscheinlichkeit interpretieren, mit der die in der Studie gefundenen Ergebnisse bei einer wiederholten Messung repliziert werden können, wurden im Mittel von 71% bzw. 62% der Schüler:innen korrekt beantwortet und als falsch bewertet. Die größte Inhomogenität hinsichtlich der Lösungsrate lag bei den *Fehlschlüssen beim Überschreiten des Signifikanzniveaus*. Die Aussage „Bei einem p-Wert von 0,30 und einem Signifikanzniveau von 0,05 ist die Hypothese der Autor:innen sehr unplausibel“ (falsch) war mit einer Lösungsrate von nur 17% insgesamt das am schlechtesten bearbeite Item. Die anderen Items, insbesondere auch die ähnliche Aussage „Ein p-Wert größer oder gleich dem Signifikanzniveau bedeutet, dass die Studienergebnisse gegen die Hypothese der Autor:innen sprechen“, wurden deutlich besser bearbeitet. Insgesamt lag die mittlere Lösungsrate damit bei 66% und nur drei von 41 Personen lösten hier alle Aussagen korrekt. Die Inhomogenität bei der Beantwortung der Items zu Fehlschlüssen beim Überschreiten des Signifikanzniveaus kann auch daran liegen, dass die Items 18-20 nicht in eindeutiger Weise auf Fehlvorstellungen hindeuten müssen (im Vergleich zu den anderen eingesetzten Items). Zum Beispiel kann die Aussage „Ein p-Wert größer oder gleich dem Signifikanzniveau bedeutet, dass die Studienergebnisse gegen die Hypothese der Autor:innen sprechen.“ (vgl. Item 18) in Kontexten durchaus als korrekt gesehen werden, in denen eine große statistische Power vorausgesetzt werden kann – beispielsweise, weil große Stichproben generiert werden. Zustimmung zu dieser Aussage im Allgemeinen legt jedoch nahe, dass ein hoher p-Wert als starkes Indiz dafür gewertet wird, dass die Nullhypothese wahr ist und somit die asymmetrische Anlage des Tests ignoriert wird. Die verbreiteten schulischen Entscheidungsregeln, die in der Regel ohne Bezug zur Testpower unterrichtet werden, spiegeln diese Asymmetrie hingegen nicht wider (z. B. bei Ausdrücken wie „Entscheidung für die Nullhypothese“ oder „Beibehaltung der Nullhypothese“ ohne Betrachtung der Power eines Tests). Folgt man also diesen Entscheidungsregeln, kann ein hoher p-Wert leicht so missinterpretiert werden, dass die Ergebnisse nicht *für* – und damit *automatisch* gegen – die Hypothese der Autor:innen sprechen. Eine korrekte Interpretation lässt jedoch nur den Schluss zu, dass die Ergebnisse zwar in der Tat nicht für – aber ohne weitere Annahmen (Power) eben auch *nicht zwingend* gegen – die Hypothese der Autor:innen sprechen (vgl. auch Naumann & Bühner, 2020).

Fehlschlüsse beim Unterschreiten des Signifikanzniveaus waren allgemein relativ selten; die mittlere Lösungsrate liegt hier bei 94%. Deutlich weiter verbreitet scheint hingegen der *Relevanz-Fehlschluss* zu sein. Beide Fragen wurden hier schlechter als die Ratewahrscheinlichkeit gelöst. Insgesamt bearbeiteten 33 der 41 Personen mindestens eine Frage, 20 sogar beide Fragen falsch.

Der *direkte Effektstärken-Fehlschluss*, also die Auffassung, dass der p-Wert selbst ein Maß für die Effektstärke sei, wurde im Mittel von 83% der Teilnehmenden korrekt verneint. Jedoch wurden die Items, die einen implizit beschriebenen, aber doch unmittelbaren Zusammenhang zwischen p-Wert und Effektstärke herstellten, nicht besser als die Ratewahrscheinlichkeit gelöst.



M	0,89	0,63	0,90	0,62	0,71	0,66	0,94	0,35	0,83	0,54
SD	0,14	0,40	0,21	0,46	0,36	0,20	0,16	0,39	0,23	0,39

Abb. 4: Lösungsraten der Aussagen zum p-Wert bzw. zur Signifikanz. Die gestrichelte Linie markiert die Ratewahrscheinlichkeit. Items mit dem Zusatz „r“ waren invertiert gegeben, das heißt die korrekte Antwort war hier „richtig“, ansonsten immer „falsch“. Unten: M: Mittlere Lösungsrate und SD: Standardabweichung, bezogen auf die zehn untersuchten (Unter-)konstrukte.

6. Resümee

Im Sinne von Data Literacy wird das Interpretieren von Daten als wichtiger Bestandteil angesehen. In Forschung und Wissenschaft werden Signifikanztests in der „klassischen Statistik" weiterhin als zentrale Methoden zur verlässlichen Generierung von Erkenntnissen angesehen, auch wenn die vielen Fehlvorstellungen (Oakes, 1986; Haller und Krauss, 2002; Ufer, 2022), die Replikationskrise in der Psychologie (siehe z. B. Open Science Collaboration, 2015; Simmons, Nelson, Simonsohn, 2011; Maxwell, Lau & Howard, 2015) und alternative statistische Methoden (wie z. B. Konfidenzintervall, Cumming, 2012; Bayessche Statistik, Masson, 2011; oder Machine Learning Verfahren, Berrar & Dubistzky, 2019) die Grenzen dieser inferenzstatistischen Methoden verdeutlichen. Dennoch stellen Signifikanztests ein probates Mittel für inferenzstatistische Analysen dar, wenn sie richtig eingesetzt und interpretiert werden. Da „Statistik für …" eine der häufigsten Lehrveranstaltungen an deutschen Universitäten ist und Signifikanztests einen wesentlichen Aspekt in diesen Lehrveranstaltungen darstellen, ist ein fundiertes konzeptuelles Verständnis für Signifikanztests und p-Werte im Sinne einer Data Literacy unerlässlich.

Die vorgestellten Ergebnisse verdeutlichen, dass selbst jüngere (aber hier auch mathematikaffine) Schüler:innen im Rahmen eines einwöchigen Intensivkurses die Korrektheit typischer falscher und richtiger Aussagen bezüglich Signifikanztests und p-Werten richtig einschätzen können, wenn ein simulationsbasierter Ansatz gewählt wird (vgl. Chance et al., 2022; Chandrakantha, 2020; Podworny, 2019), in dem Signifikanztests anhand von p-Werten eingeführt werden (vgl. auch Griese et al., 2020; Krauss & Wassner, 2001; Weber, 2020). Für Anschlussstudien wäre eine zusätzliche Untersuchung mithilfe qualitativer Methoden interessant, um überprüfen zu können, ob Schüler:innen nach einem simulationsbasierten Unterricht zu Signifikanztests auch *begründen* können, warum die jeweiligen Aussagen richtig oder falsch sind und ob diese Begründungen auch wirklich auf der Vorstellung von simulierten Stichproben basieren. Zudem könnte damit geklärt werden, ob die niedrigen Lösungsraten einiger Items wirklich auf konzeptionelle Fehlvorstellungen zurückzuführen sind oder eher von einer abweichenden Interpretation nicht genau definierter Begriffe (z. B. „unplausibel", vgl. Item 20) herrühren.

Besonders gut gelang im Workshop der Aufbau der korrekten Interpretation, dass ein signifikanter p-Wert zwar ein gutes Indiz, jedoch niemals ein Beweis für das Zutreffen einer Hypothese ist. Der auch bei Expert:innen durchaus häufig angetroffenen Fehlvorstellung des eindeutigen Beweises (Haller & Krauss, 2002; Oakes, 1986) lässt sich also schon direkt bei der Einführung in die Thematik der Signifikanztests effektiv vorbeugen. Deutlich schwieriger gestaltet sich hingegen der Aufbau eines korrekten, probabilistischen Verständnisses des p-Werts. Dieser beschreibt formal die Wahrscheinlichkeit, den vorliegenden (oder noch extremeren) Teststatistikwert zu finden, wenn in Wahrheit die Nullhypothese gilt. Analog zu vielen früheren Forschungsbefunden (Passon & van der Twer, 2020; Haller & Krauss, 2002; Kirk, 1996; Greenwald, 1996; Shaver, 1993; Oakes, 1986; Carver, 1978) zeigten auch unsere Teilnehmenden die Tendenz, diese Wahrscheinlichkeit mit der Wahrscheinlichkeit des Zutreffens der Nullhypothese oder der Replizierbarkeit der gefundenen Daten zu verwechseln, obwohl diese Fehlvorstellung im Rahmen der Items von Oakes (1986) explizit in einer Übungsstunde besprochen wurden. Dies verdeutlicht einmal mehr die Tücken des intuitiven Umgangs mit (bedingten) Wahrscheinlichkeiten und dass diese nicht durch eine einmalige, kurze Intervention behoben werden können.

Während p-Werte unterhalb des Signifikanzniveaus also durchaus sicher interpretiert wurden, stifteten p-Werte oberhalb des Signifikanzniveaus ein gewisses Maß an Verwirrung, das sich auch durch stark unterschiedliche Lösungsraten der entsprechenden Items äußerte. Ein großes Problem scheint hier zu sein, dass ein nicht-signifikantes Testergebnis nicht korrekt im Sinne der asymmetrischen Testkonzeption als „Die Nullhypothese kann nicht auf dem gewünschten Signifikanzniveau verworfen werden" interpretiert wird, sondern als „Die Alternativhypothese konnte nicht bestätigt werden, damit gilt automatisch die Nullhypothese" (vgl. Hirschauer et al., 2016). Auf diese Fehlvorstellung wurde im Kurs jedoch auch nicht im Detail eingegangen. Eine explizite Thematisierung auch nicht-signifikanter p-Werte im Unterricht erscheint daher zwingend notwendig.

Auch eine weitere Fehlerquelle, die Verwechslung von Signifikanz mit der praktischen Relevanz des zugrundeliegenden Effekts (Kirk, 1996), wurde in diesem Kurs nur implizit durch die begleitende Behandlung von Effektstärken thematisiert und konnte allein dadurch nicht behoben werden, im Gegensatz zur (direkten) Verwechslung von p-Wert und Effektstärke. Jedoch haftete auch hier vielen

Teilnehmenden die Fehlvorstellung an, aus dem p-Wert ließe sich zumindest indirekt auf die vorliegende Effektstärke schließen, nach dem Schema „je kleiner der p-Wert im Vergleich zum Signifikanzniveau ist, desto größer muss der in der Population vorliegende Effekt sein".

Insgesamt lässt sich also sagen, dass der Aufbau eines grundlegenden Verständnisses von Signifikanztests auch mit Schüler:innen der Ober- und Mittelstufe mit einem simulationsbasierten Zugang durchaus machbar erscheint. Typische Fehlvorstellungen sollten jedoch mehrfach explizit besprochen werden, um langfristig eine akkurate Interpretation von p-Werten und vor allem eine adäquate Vorstellung von Signifikanztests zu erreichen. Da ein simulationsbasierter Zugang ebenso bei Binomialtests möglich ist, ließe sich ein solches Vorgehen prinzipiell auch für den Binomialtest anwenden, der in der gymnasialen Oberstufe typischerweise unterrichtet wird.

Es konnten keine signifikanten Zusammenhänge zwischen der Lösungsrate der Fragen zum p-Wert und dem mathematischen Selbstkonzept bzw. Interesse sowie der letzten Zeugnisnote in Mathematik gefunden werden. Dies kann zum einen der kleinen Stichprobengröße (N=41) geschuldet sein, durch die mithilfe von Signifikanztests nur starke Zusammenhänge entdeckt werden können. Zum anderen handelte es sich bei den Teilnehmenden um am Fach Mathematik interessierte Schüler:innen, weswegen die mathematikbezogenen Kenngrößen durch Deckeneffekte wenig Variabilität aufwiesen.

Die Schüler:innen arbeiteten im Laufe der Woche mit R als Werkzeug zur statistischen Datenanalyse und Darstellung der Daten. Obwohl die Arbeit mit der Software durch vorgegebene Code-Sequenzen in den R-Notebooks vorentlastet wurde, um den Fokus auf statistische Entdeckungen lenken zu können (Podworny, 2019; Reinhold et al., 2024), ist uns dies nur eingeschränkt gelungen. Gerade die Übungsphasen an den Nachmittagen, in denen die Schüler:innen eigenständig an entsprechenden Übungsaufgaben arbeiteten, waren in vielen Phasen dadurch geprägt, Fehler in Befehlen zu finden und zu beseitigen. Der Einsatz statistischer Werkzeuge, die auch im Studium oder in Forschung und Wissenschaft eingesetzt werden, hat zwar den Vorteil einer authentischen Anwendung echter Statistik-Software. Allerdings entsteht so eine ungünstige unterrichtliche Fokusverschiebung: Die Aufmerksamkeit der Schüler:innen lag vielfach doch wieder auf Prozeduren – nämlich auf den genauen Befehlen in R – statt auf den statistischen Konzepten. Einen Wechsel auf (teils kommerzielle) statistische WYSIWYG-Software (wie SPSS, PSPP, STATA) würden wir hier allerdings nicht empfehlen, weil hier durch die gestaltete Oberfläche ein zu reiches Angebot statistischer Möglichkeiten überfordert wirken dürfte und dazu führt, dass die Schüler:innen in der Vielfalt der Möglichkeiten schließlich nur wenige Datei-Pfade „auswendig lernen" müssten, um zu den gesuchten Befehlen zu gelangen. Hier sind schlankere Programme im Vorteil. Eine weitere Alternative bieten Statistikprogramme, die eigens für den Mathematikunterricht entwickelt wurden, wie z. B. Fathom oder CODAP. Fathom ist in der deutschen Version seit 2018 kostenlos verfügbar (Biehler et al., 2006) und eignet sich gut für die Durchführung der Tests und auch den hier vorgestellten simulationsbasierten Ansatz. CODAP (ebenfalls kostenlos) ist überdies webbasiert und bietet über ein entsprechendes PlugIn seit Kurzem ebenfalls die Möglichkeit des hier vorgeschlagenen simulationsbasierten Ansatzes zur Einführung von p-Werten (Binder & Erickson, eingereicht). Der Einsatz spezieller didaktischer Software hat den Vorteil, dass alle technischen und rechnerischen Aspekte noch weiter vorentlastet werden können und die von uns beobachtete ungünstige Fokusverschiebung auf Fehler in den Befehlen in den Übungsphasen zugunsten einer stärkeren Aufmerksamkeit auf die eigentlichen statistischen Fragestellungen weichen könnte. Grundsätzlich sehen wir die Einführung der Signifikanztests mittels p-Werten und softwaregestützten Simulationen aber positiv.

Danksagung

Wir danken den Schüler:innen, die mit großem Engagement am Workshop teilgenommen haben und den Gutachtenden für wertvolle Hinweise zur Verbesserung des Manuskripts.

Literatur

APA – American Psychological Association (2010). *Publication manual of the American Psychological Association*, 6. Aufl., Washington2010.

Arbeitskreis Stochastik der GDM (2003). *Empfehlungen zu Zielen und zur Gestaltung des Stochastikunterrichts.* Stochastik in der Schule. http://stochastik-in-der-schule.de/ Dokumente/Leitidee_Daten_und_Zufall_SekII.pdf

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A. & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology*, *7*, 1247.

Berrar, D. & Dubitzky, W. (2019). Should significance testing be abandoned in machine learning? *International Journal of Data Science and Analytics*, *7*, 247-257.

Biehler, R., Maxara, C., Hofmann, T. & Prömmel, A. (2006). *Fathom 2.* Springer Science & Business Media.

Biehler, R., Engel, J. & Frischemeier, D. (2023). Stochastik: Leitidee Daten und Zufall. In *Handbuch der Mathematikdidaktik* (pp. 243-278). Berlin, Heidelberg: Springer Berlin Heidelberg.

Binder, K. & Erickson, T. (eingereicht). A sample-based exploration of p-values with shake boxes and CODAP.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378-399.

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., & Leader, S. (2022). Student performance in curricula centered on simulation-based inference. *Statistics Education Research Journal, 21*(3), 4-4

Chandrakantha, L. (2020). Visualizing the p-value and understanding hypothesis testing concepts using simulation in R. *Electronic Journal of Mathematics & Technology*, *14*(3).

Cumming, G. (2012). *Understanding The New Statistics. Effect Sizes, Confidence Intervals, and Meta-Analysis.* Routledge, New York.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society* (Vol. 22, No. 5, pp. 700-725). Cambridge University Press.

Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures.* Reidel.

Gagnier, J. J. & Morgenstern, H. (2017). Misconceptions, misuses, and misinterpretations of p values and significance testing. *JBJS*, *99*(18), 1598-1603.

Gal, I. (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review*, *70*, 1-25.

Gigerenzer, G., Krauss, S. & Vitouch, O. (2004). The null ritual – what you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Hrsg.), *The Sage handbook of quantitative methodology in the social sciences* (pp. 391–408). Thousand Oaks: SAGE.

Gigerenzer, G. (2004). Mindless statistics*. The journal of socio-economics*, *33*(5), 587-606.

Goodman, S. N. (2008). A dirty dozen: twelve P-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25.

Greenwald, A. G., Gonzalez, R., Harris, R. J. & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, *33*(2), 175–183.

Griese, B., Nieszporek, R. & Biehler, R. (2020). Frei verfügbare Materialien für Unterricht und Fortbildung: Stochastik verständnisorientiert unterrichten. *Stochastik in der Schule*, *40(1)*, 10–17.

Guzzetti, B. (2000). Learning counter-intuitive science concepts: what have we learned from over a dacade of research? *Reading & Writing Quarterly*, 16:2, 89-98.

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers*. Methods of psychological research*, *7*(1), 1-20.

Herrera-Bennett, A. C., Heene, M., Lakens, D. & Ufer, S. (Preprint). Improving statistical inferences: Can a MOOC reduce statistical misconceptions about p-values, confidence intervals, and Bayes factors? https://doi.org/10.31234/osf.io/zt3g9

Hirschauer, N., Mußhoff, O., Grüner, S., Frey, U., Theesfeld, I. & Wagner, P. (2016). Grundsätzliche Missverständnisse bei der Interpretation des p-Werts. *WiSt-Wirtschaftswissenschaftliches Studium*, *45*(8), 407-412.

Hoekstra, R., Morey, R. D., Rouder, J. N. & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review*, *21*, 1157-1164.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

ISB – Staatsinstitut für Schulqualität und Bildungsforschung München (2024). Gymnasium, Mathematik 12 – Vertiefungskurs Modul 5 „Statistik". https://www.lehrplanplus.bayern.de/fachlehrplan/gymnasium/12/mathematik/vertieft#310469 (abgerufen am 5. April 2024).

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746-759.

Kline, R. B. (2013). *Beyond significance testing – statistics reform in the behavioral sciences.* Baltimore: United Book Press.

Knuth, J. (2023). "Der Anteil der Knieverletzungen steigt mit diesen Schuhen signifikant". Süddeutsche Zeitung, https://www.sueddeutsche.de/sport/laufschuhe-super-schuhe-marathon-adidas-nike-asics-kiptum-chicago-2-00-35-1.6262660?reduced=true

Krauss, S., Weber, P., Binder, K., Bruckmaier, G. & Hilbert, S. (eingereicht). Zur Propädeutik des Hypothesentestens in der gymnasialen Oberstufe – Die Diskrepanz zwischen schulischem Stochastikunterricht und tatsächlicher Anwendung.

Krauss, S. & Wassner, C. (2001). Wie man das Testen von Hypothesen einführen sollte. *Stochastik in der Schule*, *21*(1), 29-34.

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior research methods*, *43*, 679-690.

Maxwell, S. E., Lau, M. Y. & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.

Meyfarth, T. (2006). *Ein computergestütztes Kurskonzept für den Stochastik-Leistungskurs mit kontinuierlicher Verwendung der Software Fathom – Didaktisch kommentierte Unterrichtsmaterialien.* Kasseler Online-Schriften zur Didaktik der Stochastik (KaDiSto) (Bd. 2). Kassel: Universität Kassel.

Naumann, F. & Bühner, M. (2020). Inferenzstatistik. In F. Naumann & M. Bühner (Hrsg.), *Lehrbuch. Statistik: Eine kurze Einführung für Studierende der Psychologie und Sozialwissenschaften* (pp. 11–52). Springer. https://doi.org/10.1007/978-3-662-62070-0_3

Neumann, I., Rohenroth, D. & Heinze, A. (2021). Mathe braucht man überall? Welche mathematischen Lernvoraussetzungen erwarten Hochschullehrende für Studiengänge außerhalb des MINT-Bereichs?. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, (111), 45-49.

Neyman, J. & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society Series A*, 231, 289–337.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* Chichester: Wiley.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Oser, F., Hascher, T. & Spychiger, M. (1999). Lernen aus Fehlern Zur Psychologie des „negativen" Wissens. In Fehlerwelten:

Vom Fehlermachen und Lernen aus Fehlern. *Beiträge und Nachträge zu einem interdisziplinären Symposium aus Anlaß des 60. Geburtstags von Fritz Oser* (pp. 11-41). VS Verlag für Sozialwissenschaften.

Passon, O. & von der Twer, T. (2020). Evidenz, Signifikanz und das kleine p. *Zeitschrift für Bildungsforschung*, *10*(3), 377-395.

Podworny, S. (2019). *Simulationen und Randomisierungstests mit der Software TinkerPlots: Theoretische Werkzeuganalyse und explorative Fallstudie.* Springer-Verlag.

Rach, S., Ufer, S., Kosiol, T. (2021). Die Rolle des Selbstkonzepts im Mathematikstudium – Wie fit fühlen sich Studierende in Mathematik? *Zeitschrift für Erziehungswissenschaft*, *24*(6), 1549-1571. doi: 10.1007/s11618-021-01058-9.

Reinhold, F., Leuders, T., Loibl, K., Nückles, M., Beege, M. & Boelmann, J. M. (2024). Learning mechanisms explaining learning with digital tools in educational settings: a cognitive process rramework. *Educational Psychology Review*, *36*(1), 14.

Ridgway, J. (2016). Implications of the Data Revolution for Statistics Education. *International Statistical Review*, 84: 528–549. doi: 10.1111/insr.12110

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*(1), 411-432.

Schüller, K., Busch, P. & Hindinger, C. (2019). Future Skills: Ein Framework für Data Literacy – Kompetenzrahmen und Forschungsbericht (Arbeitspapier Nr. 47). *Hochschulforum Digitalisierung*. 10.5281/zenodo.3349865

Sedgwick, P. (2012). Multiple significance tests: the Bonferroni correction. *BMJ*, *344*.

Sedlmeier, P. & G. Gigerenzer (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin*, *105*(2), 309–316.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education*, *61*(4), 293-316.

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Staatsinstitut für Schulqualität und Bildungsforschung (ISB). (2024). LehrplanPlus Gymnasium Bayern Mathematik 12. Klasse. https://www.lehrplanplus.bayern.de/fachlehrplan/gymnasium/12/mathematik/vertieft

Sterner, P., Friemelt, B., Goretzko, D., Kraus, E., Bühner, M. & Pargent, F. (2024). Das Konfidenz- / Signifikanzniveau impliziert ein bestimmtes Kostenverhältnis zwischen Fehler 1. Art und Fehler 2. Art: Für ein stärkeres Einbeziehen der Entscheidungstheorie in die psychologische Einzelfalldiagnostik. *Diagnostica, 70*(3), 126–138. https://doi.org/10.1026/0012-1924/a000329

Trafimow, D. & Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, *37*, 1–2. [374,375,376]

Ufer, S. (2022). Studierfähigkeit als eine Zieldimension von Mathematikunterricht in der gymnasialen Oberstufe – Konzepte, Modelle und Beitrag des Mathematikunterrichts. In T. Rolfes, S. Rach, S. Ufer, A. Heinze (Hrsg). *Das Fach Mathematik in der gymnasialen Oberstufe* (S. 75–101). Waxmann.

Ufer, S., Rach, S., Kosiol, T. (2017). Interest in mathematics = Interest in mathematics? What general measures of interest reflect when the object of interest changes. *ZDM – Mathematics Education, 49*(3), 397-409.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

Weber, P. (2020). Wie gut bereitet der Stochastikunterricht auf Alltag, Studium und Berufsleben vor? Die Diskrepanz zwischen Schule und Realität an den Beispielen „natürliche Häufigkeiten" und „Signifikanztests". *Dissertation*. Regensburg.

Michael Rößner
Ludwig-Maximilians-Universität München
Fakultät für Mathematik, Informatik und Statistik
Theresienstraße 39
80333 München
Roessner@math.lmu.de

Seit 01.04.2025 neue Adresse:
Universität Paderborn
Warburger Str. 100
33098 Paderborn
roessner@math.uni-paderborn.de

Karin Binder
Ludwig-Maximilians-Universität München
Fakultät für Mathematik, Informatik und Statistik
Theresienstraße 39
80333 München

Seit 01.04.2025 neue Adresse:
Universität Paderborn
Institut für Mathematik
Warburger Straße 100
33098 Paderborn
Karin.Binder@uni-paderborn.de

Stefan Ufer
Ludwig-Maximilians-Universität München
Fakultät für Mathematik, Informatik und Statistik
Theresienstraße 39
80333 München
ufer@math.lmu.de

Anhang

Kategorie*	Einleitender Text
A (ohne konkreten p-Wert)	Stellen Sie sich vor, Sie lesen einen Forschungsartikel zu einer empirischen Studie. Die Autor:innen berechnen einen statistischen Test, mit dem sie die Mittelwerte zweier Populationen in Bezug auf eine Variable vergleichen wollen. Ihre Hypothese ist, dass sich die Mittelwerte der beiden Populationen unterscheiden. Sie berichten als Ergebnis einen p-Wert. Was sagt dieser p-Wert aus?
B (p-Wert unter dem Signifikanzniveau)	Stellen Sie sich vor, Sie lesen einen Forschungsartikel zu einer empirischen Studie. Die Autor:innen berechnen einen statistischen Test, mit dem sie die Mittelwerte zweier Populationen in Bezug auf eine Variable vergleichen wollen. Ihre Hypothese ist, dass sich die Mittelwerte der beiden Populationen unterscheiden. Sie berichten als Ergebnis einen p-Wert von 0,006. Das Signifikanzniveau wurde auf 0,05 festgelegt. Was haben die Autor:innen damit herausgefunden?
C (p-Wert über dem Signifikanzniveau)	Stellen Sie sich vor, Sie lesen einen Forschungsartikel zu einer empirischen Studie. Die Autor:innen berechnen einen statistischen Test, mit dem sie die Mittelwerte zweier Populationen in Bezug auf eine Variable vergleichen wollen. Ihre Hypothese ist, dass sich die Mittelwerte der beiden Populationen unterscheiden. Sie berichten als Ergebnis einen p-Wert von 0,30. Das Signifikanzniveau wurde auf 0,05 festgelegt. Was haben die Autor:innen damit herausgefunden?

(Fehl-) Vorstellung	Nr.	Kat.	Item	Richtig/falsch
Eindeutiger Beweis	1	A	Ein p-Wert kleiner als das Signifikanzniveau lässt den Schluss zu, dass die Hypothese der Autor:innen (also, dass sich die Mittelwerte wirklich unterscheiden) sicher richtig ist.	Falsch
	2	A	Ein p-Wert größer als das Signifikanzniveau lässt den Schluss zu, dass die Hypothese der Autor:innen (also, dass sich die Mittelwerte wirklich unterscheiden) sicher falsch ist.	Falsch
	3	B	Es ist damit absolut bewiesen, dass die Nullhypothese (also, dass es keinen Unterschied zwischen den Mittelwerten der beiden Populationen gibt) falsch ist.	Falsch
	4	B	Die Hypothese der Autor:innen (also, dass sich die Mittelwerte wirklich unterscheiden) ist sicher richtig.	Falsch
	5	C	Es ist damit absolut bewiesen, dass die Nullhypothese gilt (also, dass es keinen Unterschied zwischen den Mittelwerten der beiden Populationen gibt).	Falsch

	6	C	Die Hypothese der Autor:innen (also, dass sich die Mittelwerte wirklich unterscheiden) ist sicher falsch.	Falsch
Inverse-Wahrscheinlichkeits-Fehlschluss	7	B	Die Autor:innen haben die Wahrscheinlichkeit (p = 0,006) dafür bestimmt, dass die Nullhypothese gilt (also, dass es keinen Unterschied zwischen den Mittelwerten der beiden Populationen gibt).	Falsch
	8	C	Die Autor:innen haben die Wahrscheinlichkeit (p=0,30) dafür bestimmt, dass die Nullhypothese gilt (also, dass es keinen Unterschied zwischen den Mittelwerten der beiden Populationen gibt).	Falsch
	9	A	Der p-Wert beschreibt die Wahrscheinlichkeit, dass die Hypothese der Autor:innen (also, dass sich die Mittelwerte wirklich unterscheiden) wirklich korrekt ist.	Falsch
	10	B	Die Autor:innen haben die Wahrscheinlichkeit (p=0,006) dafür bestimmt, dass ihre Hypothese gilt (also, dass sich die Mittelwerte der beiden Populationen unterscheiden).	Falsch
	11	C	Die Autor:innen haben die Wahrscheinlichkeit (p=0,30) dafür bestimmt, dass ihre Hypothese gilt (also, dass sich die Mittelwerte der beiden Populationen unterscheiden).	Falsch
Replikationsfehlschluss	12	B	Die Wahrscheinlichkeit, dass man dieselben Ergebnisse erhält, wenn man die Studie mit einer neuen Stichprobe erneut durchführen würde, ist 1 – p = 0,994.	Falsch
	13	C	Die Wahrscheinlichkeit, dass man dieselben Ergebnisse erhält, wenn man die Studie mit einer neuen Stichprobe erneut durchführen würde, ist 1 – p = 0,70.	Falsch
	14	A	Der p-Wert beschreibt die Wahrscheinlichkeit, dass man einen signifikanten Unterschied erhält, wenn man die Studie erneut mit einer vergleichbaren Stichprobe durchführt.	Falsch
	15	B	Die Autor:innen haben die Wahrscheinlichkeit (p = 0,006) bestimmt, dass man einen statistisch signifikanten	Falsch

			Unterschied beobachtet, wenn man die Studie mit einer neuen Stichprobe erneut durchführen würde.	
	16	C	Die Autor:innen haben die Wahrscheinlichkeit (p=0,30) bestimmt, dass man einen statistisch signifikanten Unterschied beobachtet, wenn man die Studie mit einer neuen Stichprobe erneut durchführen würde.	Falsch
Fehlschlüsse beim Überschreiten des Signifikanzniveaus	17	A	Ein p-Wert größer oder gleich dem Signifikanzniveau bedeutet, dass die Studienergebnisse für die Hypothese der Autor:innen sprechen.	Falsch
	18	A	Ein p-Wert größer oder gleich dem Signifikanzniveau bedeutet, dass die Studienergebnisse gegen die Hypothese der Autor:innen sprechen.	Falsch
	19	A	Ein p-Wert größer oder gleich dem Signifikanzniveau bedeutet, dass man anhand der Studienergebnisse keine Aussage über die Gültigkeit der Hypothese der Autor:innen machen kann.	Richtig
	20	C	Die Hypothese der Autor:innen (also, dass sich die Mittelwerte wirklich unterscheiden) ist sehr unplausibel.	Falsch
Fehlschlüsse beim Unterschreiten des Signifikanzniveaus	21	A	Ein p-Wert kleiner als das Signifikanzniveau bedeutet, dass die Studienergebnisse für die Hypothese der Autor:innen sprechen.	Richtig
	22	A	Ein p-Wert kleiner als das Signifikanzniveau bedeutet, dass die Studienergebnisse gegen die Hypothese der Autor:innen sprechen.	Falsch
	23	A	Ein p-Wert kleiner als das Signifikanzniveau bedeutet, dass man anhand der Studienergebnisse keine Aussage über die Gültigkeit der Hypothese der Autor:innen machen kann.	Falsch
	24	B	Die Hypothese der Autor:innen (also, dass sich die Mittelwerte wirklich unterscheiden) ist sehr plausibel.	Richtig
Relevanz-Fehlschluss	25	B	Da das Ergebnis statistisch signifikant ist, haben die Autor:innen herausgefunden, dass ein relevanter Unterschied zwischen den Populationen besteht.	Falsch

	26	C	Da das Ergebnis nicht statistisch signifikant ist, haben die Autor:innen herausgefunden, dass kein relevanter Unterschied zwischen den Populationen besteht.	Falsch
Effektstärken-Fehlschluss	27	A	Der p-Wert beschreibt, ob es sich bei den Unterschieden um einen großen (in der Praxis relevanten) oder einen kleinen (in der Praxis wenig relevanten) Effekt handelt.	Falsch
	28	B	Der Wert p = 0,006 sagt direkt aus, dass der Unterschied einem großen Effekt entspricht.	Falsch
	29	C	Der Wert p = 0,30 sagt direkt aus, dass der Unterschied nur einem kleinen Effekt entspricht.	Falsch
	30	B	Da der Wert 0,006 sehr viel kleiner ist als das Signifikanzniveau 0,05, liegt ein großer Effekt vor.	Falsch
	31	C	Da der Wert 0,30 sehr viel größer ist als das Signifikanzniveau 0,05, kann der Unterschied zwischen den Populationen nur sehr klein sein.	Falsch

Tab. A1: Verwendete Testitems. *) Die Items wurden nicht in der hier dargestellten Reihenfolge präsentiert, sondern nach drei Kategorien gruppiert. Items einer Kategorie hatten jeweils denselben einleitenden Text. Die Nr. der Items bezieht sich auf die Nummerierung in Abb. 4.