

Datenbasierter Vergleich von statistischen Tests mithilfe von Simulationen

Auswahl eines Verfahrens für das A/B-Testing

Felix J. M. Welter¹

Abstract: Die Wahl des richtigen statistischen Verfahrens ist für einen A/B-Test essentiell. Bei Nutzung eines mächtigen Tests kann ein Experiment schneller zu Erkenntnissen führen und es können früher datengetriebene Entscheidungen getroffen werden. In dieser Untersuchung wurden der Permutationstest und der Wilcoxon-Rangsummentest bezüglich Mächtigkeit und Alpha-Robustheit verglichen. Die Vorgehensweise kann zudem abstrahiert und auf andere Tests oder Anwendungen übertragen werden.

Keywords: Statistische Tests, A/B-Test, Permutationstest, Wilcoxon-Rangsummentest, Simulation, Python

1 Einleitung

Das Wachstum der Branche E-Commerce seit der Jahrtausendwende ist enorm. [HD18] Auf die steigenden Anforderungen reagieren Onlinehändler mit unterschiedlichsten Produktentwicklungen und neuen Features für ihre Webseiten. Ein klassisches Beispiel dafür ist das Empfehlungssystem von Amazon. [LSY03] Auch die aggregierten Kundenbewertungen, wie sie auf otto.de eingesetzt werden, gehören dazu. Jedoch sind die Auswirkungen einer neuen Entwicklung nur selten präzise vorherzusehen. Doch diese genaue Quantifizierung ist notwendig, um Ressourcen sinnvoll und mit maximalem Nutzen einzusetzen. Um dies zu ermöglichen, wird im E-Commerce eine Methode namens A/B-Testing verwendet. Die zu prüfende Änderung wird dabei nur einem Teil der Webseitennutzer angezeigt. Wenn bei allen Nutzern relevante Kennzahlen erfasst werden, häufig handelt es sich dabei um Bestellungen oder Umsatz, dann können diese im Nachhinein verglichen werden. Der Vergleich wird dabei zwischen allen Nutzern, welche das neue Feature gesehen haben, und den restlichen Nutzern gezogen. [KHS07]

Ob eine gemessene Differenz der Kennzahlen einem tatsächlichen Unterschied zwischen den Varianten entstammt, oder ob es sich um eine natürliche Schwankung handelt, kann durch einen statistischen Test geprüft werden. Es ist ersichtlich, dass ein

¹ Nordakademie, Köllner Chaussee 11, 25337 Elmshorn, felix.welter.i15c@nordakademie.org

A/B-Test wünschenswerterweise bereits nach kurzer Zeit zu einem Ergebnis führt, da positive Änderungen für alle Nutzer eingeführt und negative Änderungen früher wieder deaktiviert werden können. Daraus ergibt sich, dass der eingesetzte statistische Test mit möglichst wenig Webseitenbesuchern zu einer Aussage kommen muss. Diese Untersuchung vergleicht den Permutationstest mit dem Wilcoxon-Rangsummentest basierend auf dieser Fragestellung. Das Vorgehen kann sehr gut abstrahiert und in anderen Kontexten genutzt werden.

2 Grundlagen

Die in dieser Untersuchung genutzten Daten sind durch Bestellungen von Nutzern auf otto.de zustande gekommen. Betrachtet wird dabei der monetäre Wert der bestellten Artikel. Durch eine interne Vorprozessierung der Daten liegen sowohl für die Testgruppe, welche die Änderung ausgespielt bekommt, wie auch für die Kontrollgruppe, welche die unveränderte Webseite sieht, hundert Messwerte vor. Diese können nun mit einem statistischen Test verglichen werden.

Diese generierten Daten folgen nicht der Normalverteilung, weshalb ein nicht-parametrischer Test genutzt werden sollte. In so einem Fall kommt häufig der Wilcoxon-Rangsummentest zur Anwendung. Dies entspricht auch dem Vorgehen bei der Firma Otto (GmbH & Co KG). Die Funktionsweise ist wie folgt. Die Messwerte beider Stichproben werden übergreifend sortiert und durch ihren Rang ersetzt. Es wird die Rangsumme R der ersten Stichprobe errechnet. Für diese Rangsumme existieren Erwartungswert μ_R und Standardabweichung σ_R .

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$
$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Die errechnete Rangsumme R kann dann in eine z-Statistik überführt und entsprechend der Normalverteilung interpretiert werden. Um den Vergleich der beiden Tests zu vereinfachen, wird daraus noch der p-Wert errechnet. Es handelt sich hierbei um ein approximatives Verfahren, welches erst ab zehn Messwerten je Gruppe zuverlässig funktioniert. Dies ist im vorliegenden Anwendungsfall jedoch gegeben. [Wi45], [Ka86]

Gegenüber dem Wilcoxon-Rangsummentest soll untersucht werden, ob der Permutationstest eine Alternative darstellt. Dabei handelt es sich um ein Resamplingverfahren. Abhängig von Datenmenge und Testszenario existieren unterschiedliche Versionen. Die für diese Untersuchung genutzte wird wie folgt durchgeführt. Im ersten Schritt wird die Differenz der Mittelwerte beider Stichproben errechnet. In Schritt zwei wird die Gruppenzugehörigkeit der Messwerte neu zugewiesen, solange bis beide Gruppen ihre ursprüngliche Größe erreicht haben. Dann wird wieder die Differenz der Mittelwerte kalkuliert. Schritt zwei wird mehrmals

wiederholt, woraufhin die kalkulierten Differenzen die Permutationsverteilung bilden. Der p-Wert des Tests ist dann das Verhältnis zwischen der Anzahl von Permutationsdifferenzen, die größer sind als die real gemessene Differenz und die Anzahl der errechneten Permutationen. [Co13] Wie häufig die Gruppenzugehörigkeit für diesen Test permutiert wird, kann je nach Anwendungsfall variieren. Für Simulationen ist jedoch eine Zahl von 1000 als sinnvoll zu erachten. [Ma07]

Für die Untersuchung ist weiterhin ein statistisches Konzept namens Mächtigkeit relevant, häufig auch als Power, Teststärke oder Trennschärfe bezeichnet. Die Mächtigkeit gibt die Wahrscheinlichkeit dafür an, einen Effekt zu erkennen, wenn ein Effekt vorliegt. Im Kontext eines A/B-Tests ist es die Fähigkeit einen Unterschied zwischen den Varianten einer Webseite auszumachen, wenn diese tatsächlich zu verschiedenen Ergebnissen der Kennzahlen führen. Mächtigkeit ist damit das Komplement zum Fehler zweiter Art. [Ka86] Die Mächtigkeit steigt, wenn für den Test mehr Messwerte zur Verfügung stehen. Dies ist nachvollziehbar, da mit mehr Messwerten die Menge an Informationen steigt und die Aussage des Tests sicherer wird. [Co92] Wie zuvor erläutert, soll ein Test jedoch trotz geringen Stichprobenumfangs möglichst mächtig sein, sodass eine kurze Laufzeit eines A/B-Tests möglich wird.

3 Zielsetzung

Da der Wilcoxon-Rangsummentest derzeit in dem Unternehmen Otto (GmbH Co & KG) eingesetzt wird, ist der Permutationstest als Herausforderer zu betrachten. Es gilt also zu zeigen, dass der Permutationstest besser zur Auswertung von A/B-Tests geeignet ist, sprich, dass er kürzere Laufzeiten ermöglicht. Diese Vermutung liegt darin begründet, dass der Wilcoxon-Rangsummentest bei der Transformation der Messwerte in Ränge Informationen verliert [GS98], während der Permutationstest gemeinhin als sehr mächtiges Verfahren angesehen wird. [Go07], [Co13] Die Frage soll anhand zweier Kriterien untersucht werden.

Das erste Kriterium ist die Mächtigkeit der Tests. Der Permutationstest sollte mächtiger sein als der Wilcoxon-Rangsummentest. Häufig werden Tests über die asymptotische relative Effizienz verglichen. Dieser Wert gibt an, wie viele Messwerte die beiden statistischen Test relativ zueinander benötigen, um eine gewünschte Mächtigkeit bei gesetztem Alphaniveau und bekanntem Effekt zu erreichen. [Ni11] Für diese Untersuchung wird jedoch eine reale Datenbasis genutzt, weshalb die Stichprobengröße gesetzt ist. Der Vergleich wird deshalb über die Mächtigkeit durchgeführt. Dies ist ebenso gut möglich, da jeder Einflussfaktor (Mächtigkeit, Alphaniveau, Effekt und Stichprobengröße) als Funktion der anderen drei dargestellt werden kann. [Co92] Im Zuge der Simulation wird deshalb der Anteil der signifikanten Ergebnisse an allen Simulationsläufen errechnet und die Anteile der beiden Tests dann mit einem Test auf Gleichheit von Proportionen verglichen.

Ein mächtiger Test ist nicht nutzbar, wenn er deutlich häufiger zu Alphafehlern führt, als durch das Alphaniveau vorgegeben ist. Aus diesem Grund wird die Robustheit gegenüber dem Alphafehler als zweites Kriterium angeführt. Diese Robustheit wird für das Signifikanzniveau α und eine tatsächliche Alphafehlerwahrscheinlichkeit α^* wie folgt beschrieben

$$r = \frac{|\alpha - \alpha^*|}{\alpha}$$

und gibt somit die relative Abweichung vom gesetzten Alphaniveau an. [Bü91] Als strenge Grenze für r wird 0,1 angesetzt. [Br78] Das zweite Entscheidungskriterium wird somit erfüllt, wenn die Robustheit des Permutationstests diese Grenze nicht überschreitet.

4 Simulation

Die Datenbasis für die Simulation bilden fünf verschiedene Zeiträume, welche möglichst unterschiedliche Charakteristika für die Kennzahl Bestellwert aufzeigen. Dadurch können Erkenntnisse der Untersuchung verallgemeinert werden. Im weiteren Verlauf wird auf die Betrachtung der einzelnen Zeiträume jedoch verzichtet, da sehr ähnliche Simulationsverläufe beobachtet wurden und die abgeleiteten Aussagen dieselben sind.

Für jeden Zeitraum wurden verschiedene Effekte untersucht. Dieser Effekt wird im E-Commerce meist als Prozentsatz angegeben und als Uplift bezeichnet. Für diese Untersuchung wurden zur Feststellung der Mächtigkeit die Uplifts 1,0%, 1,2%, 1,4% und 1,6% simuliert. Zusätzlich wurde der Uplift 0% berücksichtigt, um ein Urteil über die Alpharobustheit fällen zu können. Dabei wurde der Effekt additiv berechnet. Ob diese additive Berechnung die Realität abbildet, ist zwar infrage zu stellen, jedoch handelt es sich dabei um das gemeinhin anerkannte Vorgehen zur Analyse der Mächtigkeit. [Fa07], [Co92], [DM00]

Die beiden statistischen Tests wurden in Python mithilfe der Pakete *numpy* [OI15], *pandas* [We10], *scipy* [JO01] und *numba* [LPS15] implementiert. Auch die Simulation wurde unter Zuhilfenahme dieser Pakete realisiert. Durchgeführt wurden pro Zeitraum und Uplift 6.000 Einzelsimulationen, in welchen die Messwerte zufällig der Test- oder Kontrollgruppe zugewiesen wurden, der Effekt zu der Kontrollgruppe addiert wurde und zuletzt der Permutationstest und der Wilcoxon-Rangsummentest auf die Daten angewendet wurden. Es wurden jeweils die einseitigen p-Werte aufgezeichnet und in einer Struktur, wie in Tab. 1 dargestellt, zur weiteren Analyse persistiert.

| Uplift | Permutationstest | Wilcoxon-Rangsummentest |
|--------|------------------|-------------------------|
| 0% | 0,15 | 0,18 |
| 0% | 0,59 | 0,61 |
| ... | ... | ... |
| 1,6% | 0,01 | 0,02 |
| 1,6% | 0,02 | 0,02 |

Tab. 1: Struktur der Simulationsergebnisse

Da für jeden einzelnen Uplift 6.000 Einzelsimulationen durchgeführt wurden, besteht eine solche Ergebnistabelle aus 30.000 Zeilen. Pro Test und Uplift wurde nun die Anzahl der Ergebnisse errechnet, die bei einem Alphaniveau von 5% signifikant sind. Als Liniendiagramm dargestellt ergibt sich Abb. 1.

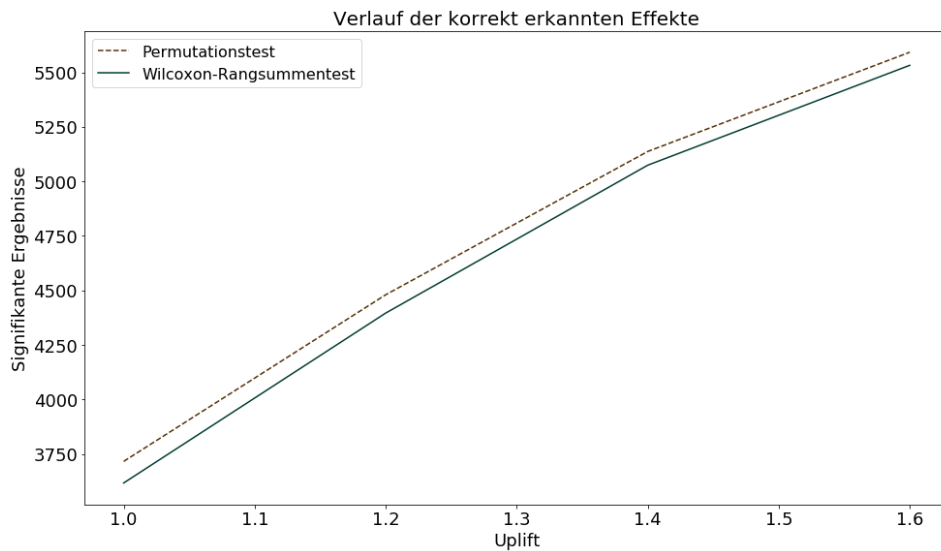


Abb. 1: Verlauf der korrekt erkannten Effekte

Da die Linie des Permutationstests oberhalb der des Wilcoxon-Rangsummentests verläuft, ist erkennbar, dass der erstgenannte scheinbar mächtiger ist. Mit einem Test auf Gleichheit von Proportionen lässt sich diese Beobachtung prüfen. Aus diesem Grund wurde pro Zeitraum und Uplift ein solcher Test durchgeführt. Der Anteil der signifikanten Ergebnisse sowie die p-Werte des Proportionentests sind in Tab. 2 dargestellt.

| Datenbasis | Uplift (%) | Permutations- test | Wilcoxon- Rangsummen- test | Proportionen- test |
|------------|------------|-----------------------|----------------------------------|-----------------------|
| Zeitraum 1 | 1,0 | 0,619 | 0,603 | 0,032 |
| | 1,2 | 0,747 | 0,733 | 0,04 |
| | 1,4 | 0,856 | 0,846 | 0,053 |
| | 1,6 | 0,932 | 0,922 | 0,018 |
| Zeitraum 2 | 1,0 | 0,745 | 0,722 | 0,003 |
| | 1,2 | 0,871 | 0,856 | 0,009 |
| | 1,4 | 0,941 | 0,931 | 0,014 |
| | 1,6 | 0,976 | 0,972 | 0,126 |
| Zeitraum 3 | 1,0 | 0,781 | 0,764 | 0,016 |
| | 1,2 | 0,897 | 0,889 | 0,074 |
| | 1,4 | 0,961 | 0,955 | 0,047 |
| | 1,6 | 0,986 | 0,984 | 0,129 |
| Zeitraum 4 | 1,0 | 0,674 | 0,662 | 0,07 |
| | 1,2 | 0,810 | 0,793 | 0,008 |
| | 1,4 | 0,896 | 0,880 | 0,002 |
| | 1,6 | 0,955 | 0,946 | 0,01 |
| Zeitraum 5 | 1,0 | 0,776 | 0,757 | 0,007 |
| | 1,2 | 0,896 | 0,884 | 0,018 |
| | 1,4 | 0,958 | 0,952 | 0,086 |
| | 1,6 | 0,986 | 0,982 | 0,029 |

Tab. 2: Anteil signifikanter Ergebnisse und p-Werte des Proportionentests

Aus der Tabelle geht hervor, dass der Permutationstest in vielen Fällen statistisch signifikant mächtiger ist als der Wilcoxon-Rangsummentest. Mit dieser Erkenntnis und der visuellen Bestätigung durch Abb. 1 ist das erste Entscheidungskriterium erfüllt.

Ob der Permutationstest auch alfarobust ist, wie vom zweiten Entscheidungskriterium gefordert, lässt sich mittels Tab. 3 erkennen. Dort ist die tatsächlich realisierte Häufigkeit des Fehlers erster Art dargestellt, sowie die daraus abgeleitete Alfarobustheit. Alle Werte liegen unterhalb 0,1 und genügen somit der strengen Grenze für die Alfarobustheit. Das zweite Kriterium ist somit ebenfalls erfüllt.

| Datenbasis | Anteil Fehler erster Art | | Alpharobustheit | |
|------------|--------------------------|--------------------------|-------------------|--------------------------|
| | Permutations-test | Wilcoxon-Rangsummen-test | Permutations-test | Wilcoxon-Rangsummen-test |
| Zeitraum 1 | 0,049 | 0,049 | 0,02 | 0,02 |
| Zeitraum 2 | 0,051 | 0,051 | 0,02 | 0,02 |
| Zeitraum 3 | 0,050 | 0,050 | 0 | 0 |
| Zeitraum 4 | 0,052 | 0,053 | 0,04 | 0,06 |
| Zeitraum 5 | 0,054 | 0,053 | 0,08 | 0,06 |

Tab. 3: Anteil Fehler erster Art und Alpharobustheit

5 Fazit

Die Annahme, dass der Permutationstest mächtiger ist als der Wilcoxon-Rangsummentest, kann basierend auf den gewonnenen Erkenntnissen bestätigt werden. Somit ergibt sich auch die Empfehlung, ihn künftig an Stelle des Wilcoxon-Rangsummentests einzusetzen. Aus A/B-Tests können somit schneller datenbasierte Erkenntnisse abgeleitet und positive Features früher der gesamten Nutzerbasis zur Verfügung gestellt werden. Ebenso ist es möglich, negative Änderungen frühzeitig aus dem Betrieb des Onlineshops zu entfernen. Direkt daraus abgeleitet ergibt sich ein monetärer Nutzen.

Im Zuge dieser Fragestellung wurde ebenfalls der t-Test mit dem beschriebenen Vorgehen untersucht. Bei dem Test handelt es sich um einen etablierten Standard in der Statistik. Er basiert jedoch auf der Annahme, dass die Daten normalverteilt sind. [St08] Obwohl diese Annahme bei den untersuchten Daten von otto.de stark verletzt ist, konnte der t-Test Effekte mindestens so gut wie der Permutationstest erkennen. In einigen Fällen zeigte sich sogar eine kleine Abweichung zugunsten des t-Tests. Ob es sich dabei um ein Zeugnis der hohen Robustheit des t-Tests handelt, welche im häufig nachgesagt wird [Bü91], [Ka86], [HD87], oder ob dies möglicherweise durch die additive Effektrechnung zustande kam, ist eine Frage für künftige Untersuchungen.

Weiterhin bieten sich für künftige Untersuchungen die folgenden Fragestellungen an. Zum einen sollten noch weitere Zeiträume untersucht werden. Während bereits fünf verschiedene zeitliche Abschnitte berücksichtigt wurden, kann die allgemeine Gültigkeit der Erkenntnisse damit gefestigt werden. Zweitens kann die Untersuchung für andere Kennzahlen durchgeführt werden. Während in dieser Untersuchung der monetäre Bestellwert der abgeschickten Warenkörbe berücksichtigt wurde, können auch Bestellmengen oder Retouren als Datenbasis genutzt werden. Ob dies sinnvoll ist, hängt

jedoch von den geschäftlichen Zielen ab. Zuletzt soll an dieser Stelle zudem darauf hingewiesen werden, dass die Untersuchung weiterer statistischer Verfahren ebenfalls eine Option für das weitere Vorgehen darstellt.

6 Weiterentwicklung und Operationalisierung

Um auf den Erkenntnissen dieser Untersuchung aufbauen zu können, werden in diesem Abschnitt einige Ansätze aus der Literatur vorgestellt, welche zu einer besseren Anwendbarkeit oder Verlässlichkeit des Permutationstests beitragen sollen. Während der Durchführung des Permutationstests wird die Gruppenzugehörigkeit wiederholt neu zugewiesen. Der Rechenaufwand dafür ist auf modernen Rechnern zu bewältigen, jedoch kann ein Test besonders bei größeren Stichproben durchaus eine Rechenzeit erreichen, welche einen akzeptablen Rahmen sprengt. Zur Reduzierung der benötigten Zeit schlagen Boos und Zhang vor, nach einer kleineren Anzahl von Permutationen mittels Extrapolation auf die restlichen zu schließen. [BZ00] Eine andere Methode kalkuliert nach einer gewissen Anzahl von Permutationen die Wahrscheinlichkeit für das Eintreten eines signifikanten Ergebnisses, sollte noch keines aufgetreten sein und beendet den Test, wenn diese Wahrscheinlichkeit gering ist. Diese Prüfung wird auch von der anderen Seite vollzogen. Der Test wird also bei einem bereits hochsignifikanten Ergebnis beendet, ohne weitere Permutationen zu berechnen. Auch dieses Verfahren trägt zur Reduzierung der Rechenzeit bei. [Kn09]

Häufig soll bei einem Experiment nicht nur ein Effekt nachgewiesen, sondern auch eine Aussage über die Stärke des Effekts gemacht werden. Dabei kommen Konfidenzintervalle zum Einsatz, welche den Zahlenwert der Teststatistik mit einer bestimmten Wahrscheinlichkeit innerhalb gewisser Grenzen einordnen. Im Gegensatz zu einigen anderen Verfahren kann aus dem Permutationstest nicht direkt ein Konfidenzintervall abgeleitet werden. Heiler und Weichselberger haben ein Verfahren vorgestellt, welches die Grenzen des Konfidenzintervalls im Vorfeld schätzt und diese Schätzung dann iterativ verbessert. [HW69]

Weitere Ansätze versuchen die Mächtigkeit des Permutationstest durch Umgehen von doppelten Gruppenkonstellationen zu erhöhen, [Op03] berechnen Konfidenzintervalle für den p-Wert des Tests, welcher aufgrund der zufälligen Ziehungen schwankt [Er04] oder erweitern die Einsatzmöglichkeiten auf andere Verfahren wie die Regression. [Pi37], [Ma06]

Beim operativen Einsatz des Permutationstests muss bedacht werden, dass dieser, wie bereits erwähnt, nicht deterministisch ist. Um sicherzustellen, dass unabhängige Auswertungen dennoch zu dem gleichen Ergebnis kommen, kann vor der Auswertung ein sogenannter Seed, ein Startwert für einen Zufallsgenerator, festgelegt werden. Die Pseudozufallszahlen des Auswertungsprogramms werden somit reproduzierbar, ebenso wie das Ergebnis des Permutationstests. [PM88]

Bei der Einführung sollte bedacht werden, dass ein Parallelbetrieb beider Testverfahren zu einer praktischen Vergleichbarkeit sinnvoll sein kann, wenn dies nicht zu viele Ressourcen bindet. Unterstützt wird der Permutationstest jedoch bereits von vielen Statistikanwendungen, wie SPSS [Ha98], STATA [Ka07] und R [FS10]. Die Implementierung in beliebigen Programmiersprachen ist ebenfalls umstandslos möglich.

7 Danksagung

Mein Dank gilt dem Unternehmen Otto (GmbH & Co KG), welches mir ein großartiges duales Studium ermöglicht hat, ebenso wie den Testanalysten Antje Krumnack und Rene Gilster, welche mich bei dieser Untersuchung unterstützt haben.

Literaturverzeichnis

- [Br78] Bradley, J. V.: Robustness? In *British Journal of Mathematical and Statistical Psychology*, 1978, 31; S. 144–152.
- [Bü91] Büning, H.: *Robuste und adaptive Tests*, 1991.
- [BZ00] Boos, D. D.; Zhang, J.: Monte Carlo Evaluation of Resampling-Based Hypothesis Tests. In *Journal of the American Statistical Association*, 2000, 95; S. 486.
- [Co13] Collingridge, D. S.: A Primer on Quantitized Data Analysis and Permutation Testing. In *Journal of Mixed Methods Research*, 2013, 7; S. 81–97.
- [Co92] Cohen, J.: A power primer. In *Psychological bulletin*, 1992, 112; S. 155–159.
- [DM00] Davidson, R.; MacKinnon, J.: Bootstrap tests: how many bootstraps? In *Econometric Reviews*, 2000, 19; S. 55–68.
- [Er04] Ernst, M. D.: Permutation Methods: A Basis for Exact Inference. In *Statistical Science*, 2004, 19; S. 676–685.
- [Fa07] Faul, F. et al.: G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. In *Behavior Research Methods*, 2007, 39; S. 175–191.
- [FS10] Fay, M. P.; Shaw, P. A.: Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The intervalR Package. In *Journal of Statistical Software*, 2010, 36.
- [Go07] Govindarajulu, Z.: *Nonparametric inference*, 2007.
- [GS98] Gebhard, J.; Schmitz, N.: Permutation tests — A revival?! In *Statistical Papers*, 1998, 39; S. 75–85.
- [Ha98] Hayes, A. F.: SPSS procedures for approximate randomization tests. In *Behavior Research Methods, Instruments, & Computers*, 1998, 30; S. 536–543.

-
- [HD18] HDE: Umsatz durch E-Commerce (B2C) in Deutschland in den Jahren 1999 bis 2017 sowie eine Prognose für 2018 (in Milliarden Euro). <https://de.statista.com/statistik/daten/studie/3979/umfrage/e-commerce-umsatz-in-deutschland-seit-1999/>, 14.12.2018.
- [HD87] Heeren, T.; D'Agostino, R.: Robustness of the two independent samples t-test when applied to ordinal scaled data. In *Statistics in Medicine*, 1987, 6; S. 79–90.
- [HW69] Heiler, S.; Weichselberger, K.: Über den Permutationstest und ein daraus ableitbares Konfidenzintervall. In *Metrika*, 1969, 14; S. 232–248.
- [JO01] Jones, E.; Oliphant, T.; Peterson, P.; others: SciPy: Open source scientific tools for Python. <http://www.scipy.org/>, 15.01.2019.
- [Ka07] Kaiser, J.: An exact and a Monte Carlo proposal to the Fisher–Pitman permutation tests for paired replicates and for independent samples. In *Stata Journal*, 2007, 7; S. 402–412.
- [Ka86] Kachigan, S. K.: *Statistical analysis. An interdisciplinary introduction to univariate & multivariate methods*. Radius Press, New York, 1986.
- [KHS07] Kohavi, R.; Henne, R. M.; Sommerfield, D.: Practical guide to controlled experiments on the web. In (Berkhin, P.; Caruana, R.; Wu, X. Hrsg.): *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*. ACM Press, New York, New York, USA, 2007; S. 959.
- [Kn09] Knijnenburg, T. A. et al.: Fewer permutations, more accurate P-values. In *Bioinformatics (Oxford, England)*, 2009, 25; i161–8.
- [LPS15] Lam, S. K.; Pitrou, A.; Seibert, S.: Numba. In (Finkel, H. Hrsg.): *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15*. ACM Press, New York, New York, USA, 2015; S. 1–6.
- [LSY03] Linden, G.; Smith, B.; York, J.: Amazon.com recommendations: item-to-item collaborative filtering. In *IEEE Internet Computing*, 2003, 7; S. 76–80.
- [Ma06] Manly, B. F. J.: *Randomization, Bootstrap and Monte Carlo Methods in Biology*. CRC Press, Boca Raton, FL, 2006.
- [Ma07] Marozzi, M.: Some remarks about the number of permutations one should consider to perform a permutation test, 2007.
- [Ni11] Nikitin, Y.: Asymptotic Relative Efficiency in Testing. In (Lovric, M. Hrsg.): *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011; S. 72–75.
- [OI15] Oliphant, T. E.: *Guide to NumPy*. Continuum Press, Austin, Tex., op. 2015.
- [Op03] Opdyke, J. D.: Fast Permutation Tests that Maximize Power Under Conventional Monte Carlo Sampling for Pairwise and Multiple Comparisons. In *Journal of Modern Applied Statistical Methods*, 2003, 2; S. 27–49.

- [Pi37] Pitman, E. J. G.: Significance Tests Which May be Applied to Samples from any Populations. II. The Correlation Coefficient Test. In Supplement to the Journal of the Royal Statistical Society, 1937, 4; S. 225.
- [PM88] Park, S. K.; Miller, K. W.: Random number generators: good ones are hard to find. In Communications of the ACM, 1988, 31; S. 1192–1201.
- [St08] Student: The Probable Error of a Mean. In Biometrika, 1908, 6; S. 1.
- [We10] Wes, M.: Data structures for statistical computing in Python, 2010.
- [Wi45] Wilcoxon, F.: Individual Comparisons by Ranking Methods. In Biometrics Bulletin, 1945, 1; S. 80.