

Aus der Forschung

Was ist CUPED?

CUPED bereinigt eine A/B-Test-Kennzahl mit ihrem Wert vor dem Test. Das senkt Rauschen, nicht Verzerrung, und verlangt eine vom Test unberührte Kovariate.

Autor	Veröffentlicht	Lesezeit	Wörter
Sinan Isoglu, MBA (Quantic) Führungskraft für Umsatzwachstum, Dozent und Doktorand	18. Oktober 2026	10 Min.	2.175

Management Summary CUPED (Controlled-experiment Using Pre-Experiment Data) ist eine Auswertung randomisierter Experimente, die das Ergebnis jedes Nutzers mit Daten aus der Zeit vor dem Experiment bereinigt, sodass sich derselbe Effekt mit weniger Nutzern oder in kürzerer Zeit nachweisen lässt. Dieser Konzeptleitfaden definiert CUPED (Deng et al.), erklärt, warum die Varianz bestimmt, wie lange ein Test laufen muss (Kohavi et al.), zeigt, wie viel Varianz die Bereinigung abhängig von einer einzigen Korrelation entfernen kann, berichtet, was sie bei Bing erreichte, nennt die Bedingungen, unter denen sie ein Ergebnis verzerrt, und bietet ein Blatt für eine CUPED-Auswertung. Die Quellen sind eine Methodenarbeit mit Ergebnissen aus einer Suchmaschine, ein Überblick mit Praxisleitfaden und ein Erfahrungsbericht; die beiden letzten stammen aus derselben Experimentiergruppe bei Microsoft.

Inhalt	§ 1	Was ist CUPED?	2
	§ 2	Warum bestimmt die Varianz, wie lange ein Test laufen muss?	3
	§ 3	Wie senkt CUPED die Varianz?	4
	Abbildung 1	Nötige Nutzer für dieselbe Teststärke, nach Korrelation mit der Kennzahl der Vorperiode	5
	§ 4	Was erreichte CUPED bei Bing?	6
	§ 5	Wann verzerrt CUPED ein Ergebnis?	7
	§ 6	Was gehört in das Blatt einer CUPED-Auswertung?	7
	Abbildung 2	Ein Blatt für eine CUPED-Auswertung	8
	§ 7	Wo wurde CUPED geprüft?	9
		Referenzen	10

Was ist CUPED?

§ 1 von 7

CUPED ist eine Bereinigung bei der Auswertung eines A/B-Tests, nicht bei seiner Planung. Deng et al.¹ (2013) führten sie ein: Sie schlagen ein Verfahren namens CUPED vor, das Kennzahlen mit Daten aus der Zeit vor dem Experiment für die Nutzer in Kontroll- und Behandlungsgruppe bereinigt, um die Streuung der Kennzahl zu senken („we propose a technique, called CUPED (Controlled-experiment Using Pre-Experiment Data), which adjusts metrics using pre-experiment data for users in the control and treatment to reduce metric variability.“). Das Experiment läuft wie gewohnt; die Auswertung korrigiert dann den Durchschnitt jeder Gruppe mit dem, was dieselben Nutzer vor Beginn des Experiments taten.

¹ Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 123–132.

Die Bereinigung lässt den Erwartungswert der Schätzung unverändert und macht sie weniger verrauscht. Nach den Autoren ist die bereinigte Schätzung weiterhin erwartungstreu für die Verschiebung der Mittelwerte („is still an unbiased estimator for the shift in the means“) und hat eine kleinere Varianz als die unbereinigte Differenz, sodass die t-Statistik bei gleicher erwarteter Effektgröße größer ausfällt („the corresponding t-statistic would be larger for the same expected effect size“).

In diesem Leitfaden heißt CUPED: den Effekt eines randomisierten Experiments schätzen, nachdem das Ergebnis jedes Nutzers um eine vor der Zuteilung gemessene Kovariate bereinigt wurde, meist dieselbe Kennzahl aus einer Vorperiode, mit einem einzigen Bereinigungskoeffizienten für beide Gruppen. CUPED verändert, wie genau die Schätzung ist, nicht, was geschätzt wird, und es repariert keine fehlerhafte Randomisierung.

Warum bestimmt die Varianz, wie lange ein Test laufen muss? § 2 von 7

Weil das Rauschen einer Kennzahl festlegt, wie viele Nutzer ein Test braucht: Der Standardfehler eines Durchschnitts sinkt nur mit der Quadratwurzel der Nutzerzahl. Kohavi et al. (2009)³ nennen die Faustregel für ein Konfidenzniveau von 95 Prozent und eine Teststärke von 80 Prozent: Die nötige Nutzerzahl je Variante wächst mit der Varianz der Kennzahl und sinkt mit dem Quadrat der Änderung, die man nachweisen will. Ihr Beispiel ist ein Shop, in dem 5 Prozent der Besucher kaufen: Eine Änderung des Umsatzes je Nutzer um 5 Prozent nachzuweisen, braucht mehr als 409.000 Nutzer je Variante („over 409,000 users“), dieselbe Änderung der Conversion, einer weniger streuenden Kennzahl, weniger als 122.000 („less than 122,000 users“).

³ Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140–181.

Sie nennen drei Wege, die Sensitivität eines Experiments zu erhöhen: mehr Nutzer, eine Kennzahl mit von Natur aus geringerer Streuung („inherently lower variability“) oder das Herausfiltern von Nutzern, die die Änderung gar nicht sehen konnten. Deng et al.¹ ergänzen, warum das praktisch zählt: Kleine Unterschiede in zentralen Kennzahlen, im Bereich von Bruchteilen eines Prozents, können erhebliche geschäftliche Folgen haben („Small differences in key metrics, on the order of fractions of a percent, may have very significant business implications“), und eine Differenz von 0,5 Prozent nachzuweisen, erfordert hundertmal so viele Nutzer („detecting a 0.5% delta requires 100 times (10 squared) more users“) wie eine von 5 Prozent. Der [Leitfaden zum A/B-Test](#) erklärt, wie ein Test bemessen wird; CUPED ist ein vierter Weg zum selben Ziel.

Wie senkt CUPED die Varianz?

§ 3 von 7

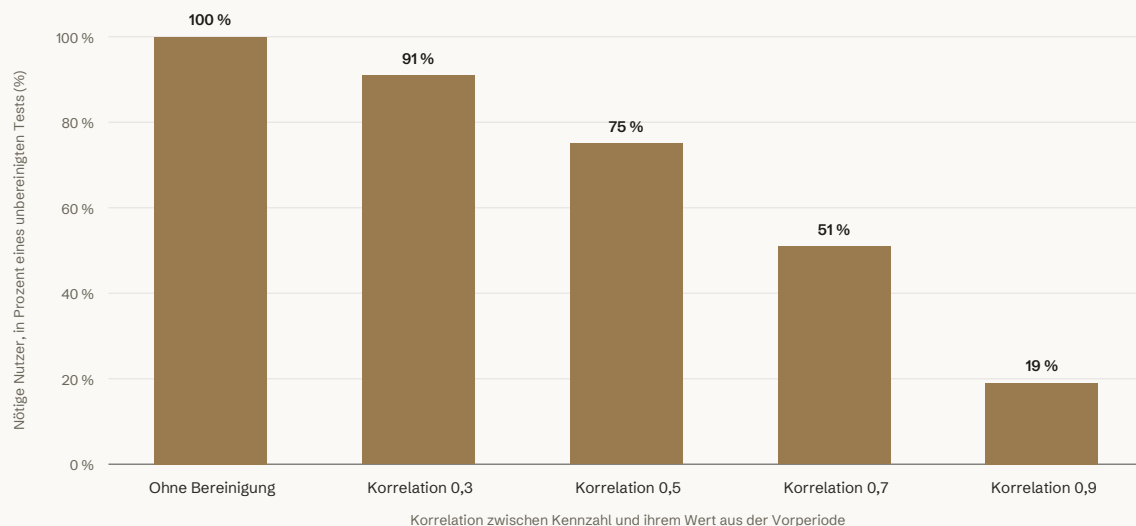
Mit einem Verfahren aus der Simulation, den Kontrollvariaten. Deng et al.¹ zeigen: Zieht man vom Ergebnis eine vor dem Experiment gemessene Kovariate ab, skaliert mit dem richtigen Koeffizienten, wird die Varianz der geschätzten Differenz mit eins minus dem Quadrat der Korrelation zwischen beiden multipliziert. Je größer die Korrelation, desto stärker die Varianzreduktion („The larger rho, the better the variance reduction“), und der beste Koeffizient ist die Kleinste-Quadrate-Lösung einer Regression der zentrierten Kennzahl auf die zentrierte Kovariate („turns out to be the ordinary least square (OLS) solution of regressing (centered) Y on (centered) X“). Damit steht CUPED der Regressionsbereinigung nahe, mit einem Unterschied, den die Autoren betonen: Ihre Herleitung braucht kein Modell des Ergebnisses, während ein lineares Modell starke Annahmen trifft, die in der Praxis meist nicht erfüllt sind („makes strong assumptions that are usually not satisfied in practice“).

Weil die Kovariate vor dem Experiment gemessen wird, verzerrt sie die Schätzung nicht. Wird die Kontrollvariate nur aus Informationen vor dem Start des Experiments gebildet, sorgt die Randomisierung dafür, dass beide Gruppen denselben Erwartungswert der Kovariate haben („By using only information from before the launch of the experiment to construct the control variate, the randomization between treatment and control ensures“), sodass sich ihre Korrektur im Erwartungswert aufhebt. Die Wahl der Kovariate folgt aus der Korrelation, nicht aus dem Zeitpunkt: Damit die Korrelation groß wird, nehmen die Autoren als Kovariate dieselbe Kennzahl aus dem Zeitraum vor dem Experiment („an obvious approach is to choose X to be the same as Y, which naturally leads to using the same variable during pre-experiment observation window as the control variate“). Für die Erwartungstreue gilt eine weitere Bedingung: Derselbe Koeffizient muss für beide Gruppen gelten („the same theta has to be used for both control and treatment“). Sie zeigen auch, dass Stratifizierung, also das Gruppieren der Nutzer nach einer kategorialen Variable aus der Vorperiode, ein Sonderfall derselben Idee ist.

Abbildung 1 übersetzt das Varianzergebnis in Nutzer. Das ist eine Rechnung dieses Leitfadens, keine Daten aus den Quellen: Wird die Varianz der Schätzung mit eins minus dem Quadrat der Korrelation multipliziert und steigt die für eine gegebene Teststärke nötige Nutzerzahl proportional zur Varianz, halbiert eine Korrelation von 0,7 die nötige Nutzerzahl ungefähr (51 Prozent).

Abbildung 1 Nötige Nutzer für dieselbe Teststärke, nach Korrelation mit der Kennzahl der Vorperiode

Bei einer Korrelation von 0,7 zwischen einer Kennzahl und ihrem Wert aus der Vorperiode braucht ein Test etwa die Hälfte der Nutzer. Bei 0,3 spart er weniger als ein Zehntel.



Rechnung nach der Formel der Autoren, unter der Annahme, dass jeder Nutzer einen Wert aus der Vorperiode hat. Fehlende Werte schmälern den Gewinn.

Quelle Rechnung des Autors aus dem Varianzergebnis bei Deng et al. (2013), S. 126 und 127, und der Faustregel zur Stichprobengröße bei Kohavi et al. (2009), S. 152: Die nötigen Nutzer entsprechen eins minus dem Quadrat der Korrelation, in Prozent eines unbereinigten Tests. Veranschaulichende Rechnung, keine beobachteten Daten.

Was erreichte CUPED bei Bing?

§ 4 von 7

Etwa die Hälfte der Varianz, bei Kennzahlen, in denen sich Nutzer stark unterscheiden und von Woche zu Woche ähnlich bleiben. Deng et al.¹ berichten, dass sie die Varianz um etwa 50 Prozent senken und so dieselbe Teststärke mit der Hälfte der Nutzer oder der halben Laufzeit erreichen („we can reduce variance by about 50%, effectively achieving the same statistical power with only half of the users, or half the duration“). Im laufenden System zeigten drei wichtige Experimente Varianzreduktionen von 45, 52 und 49 Prozent bei einer Woche Experiment und einer Woche Vorperiode („Three important recent experiments showed variance reductions of 45%, 52% and 49% with one week of experiment and one week of pre-experiment data.“). In einem Experiment, das die Seite absichtlich verlangsamte, war das bereinigte Ergebnis vom ersten Tag an signifikant („the delta was statistically significant from day 1“), während der unbereinigte Test die Schwelle erst nach zwei Wochen knapp erreichte.

Der Gewinn hängt von der Kovariate und der Kennzahl ab. In einem dreiwöchigen A/A-Test zu Suchanfragen je Nutzer senkte der Wert derselben Kennzahl aus der Woche vor dem Test die Varianz um mehr als 45 Prozent („variance reduction rate can reach more than 45%“), der Tag, an dem ein Nutzer zum ersten Mal im Experiment auftrat, dagegen um etwa 9 bis 10 Prozent; dieser Tag stammt nicht aus der Vorperiode, kann aber von der Behandlung nicht verändert werden. Der Umsatz war die Ausnahme: Beim Umsatz je Nutzer senkte CUPED die Varianz um weniger als 5 Prozent, weil der Umsatz je Nutzer zwischen Vorperiode und Experiment nur schwach korrelierte („One notable exception is revenue-per-user, where CUPED reduced the variance by less than 5% due to the low correlation of revenue-per-user between the pre-experiment and the experiment periods.“). Die Autoren empfehlen eine Vorperiode von ein bis zwei Wochen: Bei einer zu kurzen Periode lassen sich zu wenige Nutzer zuordnen, eine zu lange schwächt die Korrelation mit der Kennzahl im Experiment („Using a pre-experiment period of 1-2 weeks works well for variance reduction. Too short a period will lead to poor matching, whereas too long a period will reduce correlation with the outcome metric during the experiment period.“). Davon getrennt zählt die Laufzeit des Tests: Je länger er läuft, desto mehr seiner Nutzer sind neu und haben keinen Wert aus der Vorperiode, und der Gewinn sinkt.

Wann verzerrt CUPED ein Ergebnis?

§ 5 von 7

Wenn die Behandlung die Kovariate verändern kann. Die Autoren formulieren die Regel als Warnung: Nie Kovariaten verwenden, die von der Behandlung beeinflusst werden könnten, denn das könnte die Ergebnisse verzerren („Never use covariates that could be affected by the treatment, as this could bias the results.“). Deng et al. zeigen, was sonst geschieht. In einem Bing-Experiment, das die Suchanfragen je Nutzer erhöhte, bereinigten sie mit einer Kennzahl aus dem Experimentzeitraum, die eng mit den Suchanfragen zusammenhing, und die Schätzung wurde negativ, ein Ergebnis, das dem bekannten Effekt entgegengesetzt ist („a result that is directionally opposite of the known effect“).

Zwei weitere Grenzen ergeben sich aus den Daten. Nutzer ohne Werte aus der Vorperiode, etwa neue Nutzer, haben keine Kovariate; die Autoren fügen einen Indikator dafür hinzu, ob ein Nutzer in der Vorperiode vorkam, und teilen die Nutzer so in zwei Gruppen. Schon diese Teilung senkt die Varianz, aber in der Gruppe ohne Vorwert gibt es nichts zu bereinigen: Je geringer die Abdeckung, also der Anteil der Nutzer mit Vorwert, desto kleiner der Gewinn. Und Kennzahlen je Seitenaufruf statt je Nutzer brauchen für ihre Varianz die Delta-Methode, kombiniert mit der Bereinigung. Kohavi et al. (2013) halten am Ende ihres Berichts über Experimente bei Bing, der die Sensitivität ausklammert, fest, dass Bing begonnen habe, Daten aus der Vorperiode zur Varianzsenkung zu nutzen („In Bing, we started using pre-experiment data to reduce the variance and improve sensitivity“), und sehen hier noch viel Raum für Verbesserungen („we believe there is room for significant improvements in this area“). Der Leitfaden zur statistischen Schlussfolgerungvalidität zeigt, was ein Test mit höherer Teststärke belegt und was nicht.

Was gehört in das Blatt einer CUPED-Auswertung?

§ 6 von 7

Sechs Felder, ein Eintrag je Experiment: die Kennzahl, die Kovariate und warum die Behandlung sie nicht verändern kann, das Fenster der Vorperiode, der Anteil der Nutzer mit einem Wert aus der Vorperiode, die Analyseeinheit und die Prüfungen, die mit dem bereinigten Ergebnis berichtet werden. Die Felder folgen den Bedingungen und Empfehlungen von Deng et al.¹ und den Regeln von Kohavi et al.³² für die Bemessung eines Tests.

2 Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1168–1176.

Abbildung 2 Ein Blatt für eine CUPED-Auswertung

Kennzahl, Kovariate, Fenster, Abdeckung, Einheit und Prüfungen festhalten, bevor ein bereinigtes Ergebnis gelesen wird.

Kennzahl	Kovariate	Fenster	Abdeckung	Einheit	Prüfungen
Welche Kennzahl, und streut sie zwischen Nutzern?	Vor der Zuteilung gemessen? Kann der Test sie verändern?	Wie viele Wochen vor dem Teststart?	Welcher Anteil der Nutzer hat einen Vorwert?	Je Nutzer, Account, Seite oder Bestellung?	Auch unbereinigt berichtet? A/A-Test bestanden?

Kann die Behandlung die Kovariate verändern, nur das unbereinigte Ergebnis berichten.

Eine Zeile je Experiment. Ein bereinigtes Ergebnis ohne das unbereinigte daneben lässt sich nicht prüfen.

Quelle Entscheidungsblatt des Autors. Die Felder für Kovariate und Fenster folgen Deng et al. (2013), S. 127 bis 130; das Feld für die Kennzahl folgt Deng et al. (2013), S. 130; die Felder für Abdeckung und Einheit folgen Deng et al. (2013), S. 127 bis 129; das Feld für die Prüfungen folgt Kohavi et al. (2009), S. 151 und 152, und Deng et al. (2013), S. 128. Die Felder sind Eingaben der Lesenden; Unternehmensdaten werden nicht geliefert.

Ein Anwendungsbeispiel ohne Unternehmensdaten: Eine B2B-Softwarefirma testet einen neu gestalteten Berichtsbaukasten mit den Nutzern ihrer bestehenden Kundenaccounts und teilt ganze Accounts zufällig zu, damit alle, die gemeinsam Berichte nutzen, dieselbe Version sehen. Der Eintrag nennt erstellte Berichte je Nutzer als Kennzahl, weil sich Viel- und Wenignutzer stark unterscheiden (nach Deng et al.¹ der Fall, in dem die Varianzreduktion am besten wirkt), und dieselbe Kennzahl aus den zwei Wochen vor dem Test als Kovariate, die die Neugestaltung nicht beeinflusst haben kann. Er hält fest, dass Nutzer, die während des Tests zu Accounts hinzukommen, keinen Wert aus der Vorperiode haben und stattdessen einen Indikator erhalten, wertet je Nutzer aus, mit der Delta-Methode für die Varianz,

wie Deng et al. es tun, wenn die ausgewertete Einheit kleiner ist als die zugeteilte, und berichtet die unbereinigte und die bereinigte Differenz nebeneinander, nach einem A/A-Test mit derselben Aufteilung. Bevor die Firma CUPED auf den Umsatz je Account anwendet, prüft sie, wie eng der Umsatz jedes Accounts vor dem Test mit seinem Umsatz im Test zusammenhängt, denn der kleine Gewinn beim Umsatz kam bei Bing von einer schwachen Korrelation zwischen beiden Zeiträumen, nicht vom Umsatz an sich. Das sind Beispiele des Autors, und sie sind nicht geprüft.

Wo wurde CUPED geprüft?

§ 7 von 7

Bei einer Suchmaschine, von seinen Urhebern. Deng et al. (2013)¹ liefern die Herleitung und berichten Ergebnisse aus dem Experimentssystem von Bing; Kohavi et al. (2009)³ ist ein Überblick mit Praxisleitfaden, Kohavi et al. (2013)² ein Erfahrungsbericht, beide aus derselben Experimentiergruppe bei Microsoft. Auch der t-Test, den die Arbeit bereinigt, setzt große Stichproben voraus: Bei Onlineexperimenten, schreiben die Autoren, umfassen beide Gruppen mindestens Tausende Nutzer, sodass die Normalverteilungsannahme wegen des zentralen Grenzwertsatzes meist entbehrlich ist („are at least in thousands, hence the normality assumption on Y is usually unnecessary because of the Central Limit Theorem.“). So liest es dieser Leitfaden: Ein B2B-Test mit einigen Hundert Accounts ist eine andere Situation, in der die Bereinigung die Varianz zwar senken kann, der übliche Test aber womöglich nicht trägt; die Quellen untersuchen sie nicht.

 Ende des Beitrags

Referenzen

Zitate stehen in der Sprache des Originals; die deutschen Wiedergaben im Fließtext sind sinngemäß und stammen vom Autor.

- 1 Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 123-132. <https://doi.org/10.1145/2433396.2433413>
- 2 Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1168-1176. <https://doi.org/10.1145/2487575.2488217>
- 3 Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140-181. <https://doi.org/10.1007/s10618-008-0114-1>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist CUPED? Sinan Isoglu.
<https://isoglu.com/de/insights/journal/was-ist-cuped/>
Online-Fassung: <https://isoglu.com/de/insights/journal/was-ist-cuped/>

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand
© 2026 Sinan Isoglu · isoglu.com · Alle Rechte vorbehalten