Hausarbeit im Seminar:
Neuste Trends in Big Data Analytics

Understanding p-value

Tuan Anh Nguyen

nguyentuananh1121@gmail.com
Studiengang Wirtschaftsinformatik
Matr.-Nr. 6403062

Betreuer: Dr.Julian Kunkel

Inhaltsverzeichnis

1 Motivation

Heutzutage in statistichem Test wird der p-Wert so oft einfach verstanden, dass, wenn dieser Wert kleiner als ein vorher festgelegte Signifikanzniveau α (häufig wird α mit dem Wert 0,05 festgelegt) ist, dann die Ergebnisse als „signifikant“ bewertet werden. Aber ein p-Wert selbst liefert keine konstante Aussage, weil p-Werte sich verändern können, wenn Experimente wiederholt werden oder Stichprobengrößen verändert werden. Das bedeutet, dass p-Wert falsch positive Ergebnissen ermöglichen kann. "P-Werte leisten nicht, was sie sollen. Ganz einfach, weil sie das nicht können, (…)." (Nuzzo , 2014).

1.1 Grundlagen der Testtheorie

Statistiker und Wissenschaftler arbeiten experimentell und stellen aufgrund von Beobachtung Hyppthesen und Theorien auf. Mit diesen Hypothesen wird eine endliche Anzahl von Experimente und Beobachtungen durchgefürht, um auf eine allgemeine Regel oder Aussage zu schließen. Daher sind Hypothesen vereinfachte Modelle der Realität. Häufig werden die aufgestellten Hypthesen als **statistische Hypothesen** genannt.

Hypothesen formulieren:
„Eine Hypothese muss so formuliert sein, dass sie prinzipiell durch empirische Beobachtung widerlegt (falsifiziert) werden kann.“ (Uni Köln, 2001)
In einem statistischen Test werden zwei gegensätzliche Hypothesen gegenüber gestellt. Das Ziel der statistische Hypothesentests ist es in der Regel, die Nullhypothese H_0 zu wiederlegen, und damit implizit die Alternativhypothese H_1 zu bestätigen.
Nul- und Alternativhypothese sind stets disjunktiv. Die Ablehnung der einen bedeutet die Annahme der anderen und umgekehrt. Um Missverständnisse zu vermeiden, wird in dieser Arbeit nur über die Ablehnung oder die Annahme von H_0 geredet.

Beispie 1: Formulierung einer Hypothese:
In einem Supermarkt wird eine Untersuchung des Kaufverhaltens von 150 (fiktive) Personen durchgeführt und dabei unter anderem die Zufallsvariablen X: „Geschlecht“ und Y: „Kaufverhalten“ erhoben. Als nächstes wird es getestet, ob diese Variablen unabhängig voneinander sind, ob also Männer und Frauen das gleiche Kaufverhalten aufweisen.
Diese Informationen stellen wir in einer Kontingenztabelle (Tabelle 1) zusammen.

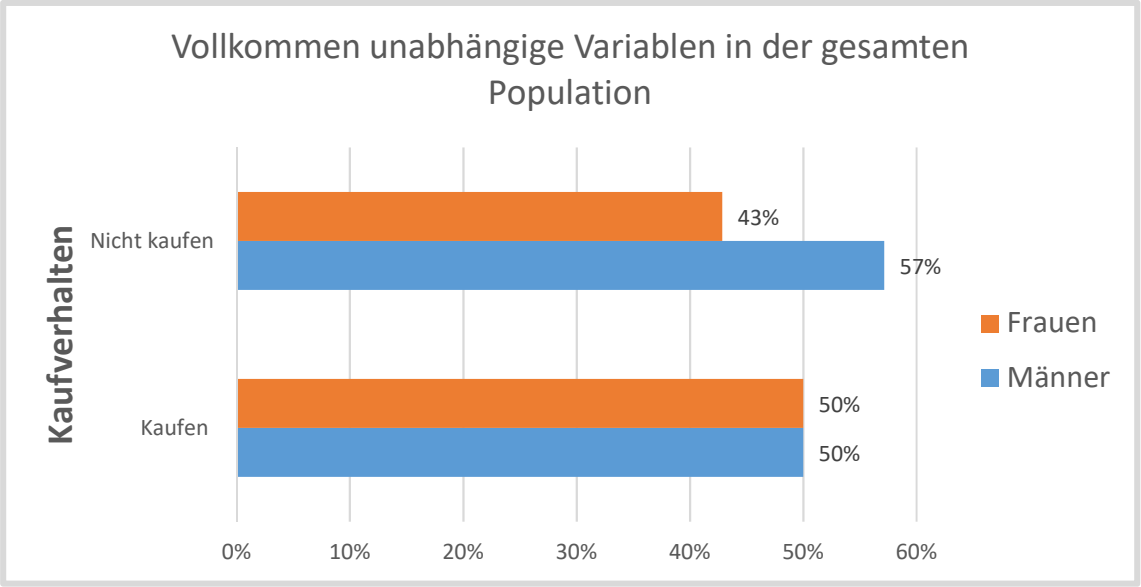

Abbildung 1: Reine unabhänngige Beobachtung des Kaufverhaltens von allen Kunden

Tabelle 1: Kontingenztabelle

	Kaufen	Nicht kaufen
Männer	40	40
Frauen	40	30
Gesamt	80	70

Die Hypothesen lauten:

- H_0: die Zufallsvariablen „Geschlecht“ und „Kaufverhalten“ sind stochastisch unabhängig voneinander.
- H_1: die Zufallsvariablen „Geschlecht“ und „Kaufverhalten“ weisen Abhängigkeiten auf.

Fehlertyp bei einem statistischen Hypothesentest

Fehler 1 und 2. Art bezeichnen eine statistische Fehlentscheidung. Hier sind zwei Zustände möglich: Treffen oder Irrtum. Beim Test einer Hypothese liegt ein Fehler 1. Art vor, wenn die Nullhypothese abgelehnt wird, obwohl sie in Wirklichkeit zutrifft (beruhend auf falsch positiven Ergebnissen). Dagegen bedeutet ein Fehler 2. Art, dass der Test die Nullhypothese angenommen wird (nicht abgelehnt), obwohl die Alternativhypothese korrekt ist.

In der Tabelle 1.2 werden die vier möglichen Zustände bei einem statistischen Test zusammengefasst dargestellt.

Tabelle 2: Vier möglichen Zustände bei einem statistischen Test

Entscheidung	Realität	
	H_0 ist richtig	H_1 ist richtig
Annahme von H_0	Richtige Entscheidung Wahrscheinlichkeit: 1 - α	**Fehler 2. Art** Wahrscheinlichkeit: β
Ablehnung von H_0	**Fehler 1. Art** Wahrscheinlichkeit: α	richtige Entscheidung Wahrscheinlichkeit: 1-β

Fehlermessung bei einer stochastischen Entscheidung

Die Größen der Fehler eines Tests werden mit Hilfe ihrer Wahrscheinlichkeit gemessen. Es gibt zwei Fälle:

Wahrscheinlichkeit des Fehlers 1.Art = P(H_0 wird abgelehnt | H_0 trifft zu)

Wahrscheinlichkeit des Fehlers 2.Art = P(H_0 wird nicht abgelehnt | H_0 trifft nicht zu)

In der empirischen Forschung legt man großen Wert daruaf, dass der Fehler bei der Annahme einer nicht zutreffenden Alternativhypothese H_1 (Ablehnung von H_0, wenn H_0 zutrifft) so klein wie möglich bleibt. Dazu setzt man eine obere Grenze $\boldsymbol{\alpha}$ für die Wahrscheinlichkeit dieses Fehlers. **Der Wert $\boldsymbol{\alpha}$,** der nicht überschritten werden soll wird Signifikanzniveau des Tests genannt. Es gilt dann **P(Fehler 1.Art) ≤ $\boldsymbol{\alpha}$** . Die obere Grenze für die Größe des Fehlers 2.Art wird mit **β** bezeichnet. Es gilt **P(Fehler 2.Art) ≤ β.** (Dr. Delgado, 2008)**.**

Prüfgröße, kritischer Bereich

Um eine statistiische Entscheidung über die Richtigkeit einer Hypothese auf Grund einer zufällig gezogenen Stichprobe **(X_1, X_2,…,X_n)** zu treffen, wird eine geeignete Stichprobendunktion $\hat{\gamma}_n$ und teilt den Wertebereich dieser Funktion in zwei ausschließende Teile: einen Teilbereich K und seiner Komplement $\overline{K}$, so dass, wenn der Wert der Funktion in den Teilbereich K hinfällt, H_0 abgelehnt wird. Fällt der Wert der Stichprobenfunktion in den anderen Teilbereich, dann wird H_0 nicht abgelehnt (H_0 wird angenommen). Die Stichprobenfunktion und die Teilbereiche werden in diesem Zusammenhang **Prüfgröße, Ablehnungsbereich (Ablehnung von H_0)** und **Annahmebereich (Annahme von H_0)** genannt. Der Ablehnungsbereich von H_0 wird auch **kritischer Bereich** genannt.

$\hat{\gamma}_n \in K \Rightarrow H_0$ wird abgelehnt (H_1 wird angenommen)

$\hat{\gamma}_n \ni K \Leftrightarrow \hat{\gamma}_n \in \overline{K} \Rightarrow H_0$ wird angenommen (H_0 wird nicht abgelehnt)

P($\hat{\gamma}_n \in K$ | H_0 trifft in der Realität zu) = P(H_0 wird abgelehnt | H_0 trifft in der Realität zu) = **P(Fehler 1.Art) ≤ $\boldsymbol{\alpha}$**

(Dr. Delgado, 2008).

Beispiel 2: Berechnung für Prüfgröße und kritischen Bereich

Mit den Daten vom Beispiel 1 wird ein X^2 – Unabhängigkeitstest durchgeführt. Um die Prüfgröße und den kritischen Bereich zu bestimmen, wird zuerst eine Kreuztabelle benötigt. Die Kreuztabelle stellt die beobachte Häufigkeit der jeweiligen Kombination zwischen Geschlecht und Kaufverhalten der Kunden dar.

Tabelle 3: Kreuztabelle

	Kaufen		**Nicht kaufen**		**Randhäufigkeiten**	
Männer	40	h_{11}	40	h_{12}	80	$h_{1.}$
Frauen	40	h_{21}	30	h_{22}	70	$h_{2.}$
	80	$h_{.1}$	70	$h_{.2}$	150	

Von den Daten der Kreuztabelle wird die erwartete Kreuztabelle berechnet.

Tabelle 4: Erwartete Kreuztabelle

	Kaufen	**Nicht kaufen**
Männer	42.6	37.3
Frauen	37.3	32.6

Wobei **die erwartete Zellenhäufigkeit** wird mit der Formel: $\tilde{h}_{ij} \coloneqq \frac{h_{i.} \cdot h_{.j}}{n}$ berechnet

Die erwarteten Werte sind die Häufigkeiten, die man unter Annahme der Gültigkeit der H_0 erwarten würde. (Uni Jena, 2015)

Dann ist die **Prüfgröße** (auch als „Teststatistik“) bei k Ausprägungen von X und l Ausprägungen von Y:

$$V = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

$$= \sum\sum\frac{(\text{beobachtete Zellhäufigkeit - erwartete Zellhäufigkeit})^2}{\text{erwartete Zellhäufigkeit}}$$

$$= \frac{(40 - 42.6)^2}{42.6} + \frac{(40 - 37.3)^2}{37.3} + \frac{(40 - 37.3)^2}{37.3} + \frac{(30 - 32.6)^2}{32.6} = \mathbf{0.7568}$$

Zu dem gegebenen/gewünschten Signifikanzniveau α (in diesem Fall α=0.05) lautet der kritische Bereich dann:

$$K = (X^2_{(k-1)(l-1),1-\alpha}, +\infty)$$

Wobei:

$X^2_{(k-1)(l-1),1-\alpha}$ ist $(\mathbf{1-\alpha})$**–Quantil** der X^2 –Verteilung mit $(k-1)(l-1)$ **Freiheitsgraden**

Die Nullhypothese H_0 wird abgelehnt, wenn $\mathbf{V > X^2_{(k-1)(l-1),1-\alpha}}$ (also V ∈ K)

Mit den Daten vom obigen Beispiel ($k = 2$, $l = 2$, $\alpha = 0.05$) ist $X^2_{1,0.95} = 3.841$ und damit ist $\mathbf{V < X^2_{1,0.95}}$. Also liegt die Prüfgröße nicht in dem Intervall des kritischen Bereichs, kann die Nullhypothese zum Signifikanzniveau nicht verworfen werden. Dann X und Y sind stochastisch unabhängig voneinander.

Anmerkung: das Ergebnis $X^2_{1,0.95} = 3.841$ wird von Tabelle der Quantile der Chiquadrat-Verteilung abgelesen

df	(rote/dunkle) Fläche (1-α)					
	0,7	0,75	0,8	0,85	0,9	0,95
1	1,07	1,32	1,64	2,07	2,71	3,84
2	2,41	2,77	3,22	3,79	4,61	5,99
3	3,66	4,11	4,64	5,32	6,25	7,81
4	4,88	5,39	5,99	6,74	7,78	9,49
5	6,06	6,63	7,29	8,12	9,24	11,07
6	7,23	7,84	8,56	9,45	10,64	12,59
7	8,38	9,04	9,80	10,75	12,02	14,07
8	9,52	10,22	11,03	12,03	13,36	15,51

Abbildung 2: Tabelle der Quantile der Chiquadrat-Verteilung (Quelle: eswf.uni-koeln.de)

1.2 Was ist p-Wert?

Selbst wenn die Stichprobe zu einem Prüfgrößenwert von $V = X^2_{1,0.95} = 3.841$ geührt hätte, wäre H_0 (gerade noch) zum Signifikanzniveau 0.05 nicht abgelehnt worden. Es wäre also informativer, ein feineres Maß für die Verträglichkeit von Daten und Nullhypothese anzugeben.

Statt von einem festen Signifikanzniveau auszugehen, z.B. α=0.05 oder α=0.01, und daraufhin einen kritischen Wert fur die Prüfgröße zu bestimmen, geht die **p-Wert-Methode** vom konkret beobachteten Wert einer Prüfgröße aus (in diesem Beispiel V= 0.7568). Die wahrscheinlichkeitstheoretische Beurteilung, ob der Prufgrößenwert 0.7568 im Sinne der Nullhypothese extrem oder selten ist, erfolgt nicht über den Umweg kritischer Werte

sondern direkt. Die p-Wert-Methode fragt nach der Wahrscheinlichkeit, einen Prüfgrößenwert V^* zu beobachten, der im Sinne der Nullhypothese noch extremer, noch seltener als 0.7568 ist.

Diese Wahrscheinlichkeit, unter H_0 einen Prüfgrößenwert V^* mit

$$V^* > 0.7568$$

zu beobachten ist der p-Wert. Dieser wird in Abhängigkeit vom konkreten Prüfgrößenwert 0.7568 mit $P(X_1^2 > 0.7568)$ bezeichnet.

```
> with(Statistics) :
> X := Matrix([[40, 40], [40, 30]]) :
> ChiSquareIndependenceTest(X, level = 0.05, summarize = embed) :
```

Chi-Square Test for Independence

Null Hypothesis:	Two attributes within a population are independent of one another
Alternative Hypothesis:	Two attributes within a population are not independent of one another

Dimensions	Total Elements	Distribution	Computed Statistic	Computed p-value	Critical Value
2.	150.	*ChiSquare*(1)	0.765306	0.381673	3.84146

Result:	**Accepted**: This statistical test does not provide enough evidence to conclude that the null hypothesis is false.

Abbildung 3: Berechnung von p-Wert mittels Maple-Software

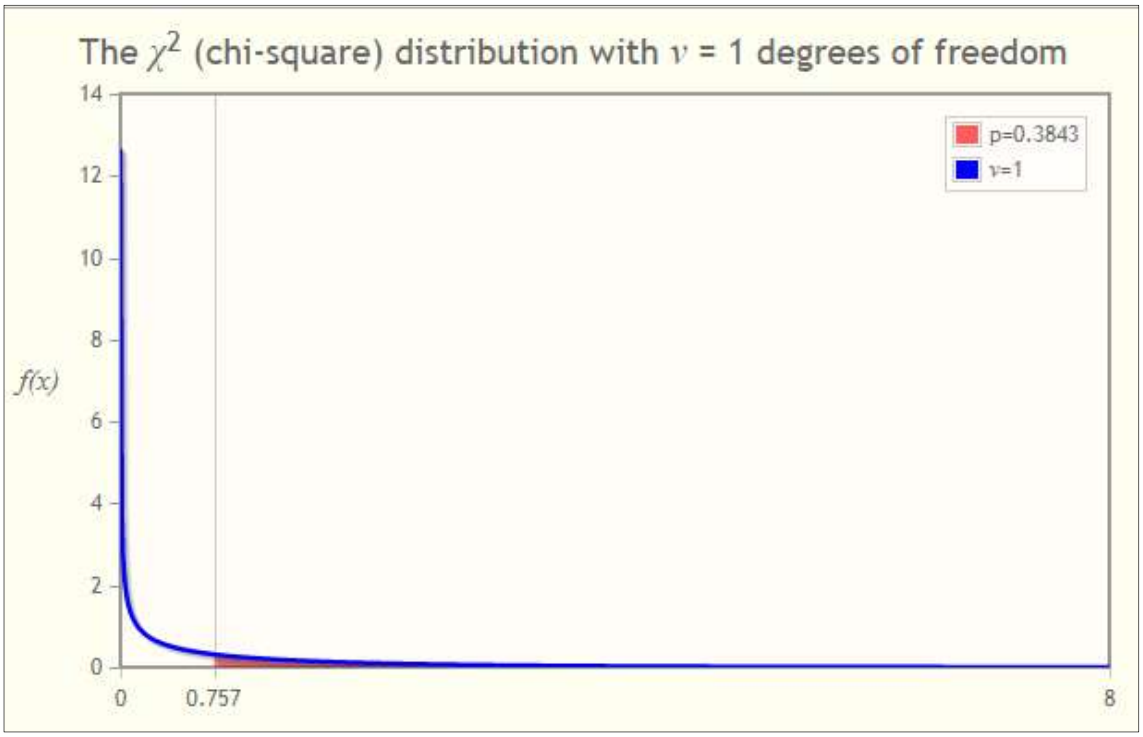

Abbildung 4: X^2-Verteilung mit p-Wert = 0.38 und Prüfgröße V = 0.7568

Mit anderen Worten: Der p–Wert 0.38 ist das **größte Signifikanzniveau**, welches bei einem Prüfgrößenwert von V = 0.7568 noch zu einer Annahme von H_0 führt. (Die Wahl eines größeren Signifikanzniveaus als 0.38 fuhrt zu einem Annahmesbereich, der den Prüfgrößenwert von 0.7568 nicht mehr enthält).

Formale Definition des p-Wertes :
„Der p-Wert ist definiert als die Wahrscheinlichkeit, unter H_0 den beobachteten Prüfgrößenwert oder einen in Richtung der Alternative extremeren Wert zu erhalten." (Böker, 2006).

Testentscheidung aufgrund des p-Wertes :
Der Mathematiker Ron Fisher hatte den p-Wert in den 1920er-Jahren erfunden, um als eine objektive Methode die Daten mit der **Nullhypothese** zu vergleichen. (Frost, 2014)

Abbildung 5: Ronald Aylmer Fisher (Quelle: Wikipedia)

Also eine Nullhypothese H_0 wird als unannehmbar betrachtet, falls für den p– Wert eines statistischen Tests gilt $p \leq \alpha$ mit $0.01 \leq \alpha \leq 0.05$. Dieses Ergebnis wird auch als **Statistische Signifikanz** genannt.
Wird H_0 zu einem p–Wert abgelehnt, so bedeutet dies, dass eine Fehlerwahrscheinlichkeit 1. Art in Höhe des p–Wertes akzeptiert wird. Der p-Wert wird auch exaktes oder tatsächliches Signifikanzniveau genannt.

1.2.1 p-Wert und kritischer Wert

Um den Unterschied zwischen dem klassischen Verfahren mit kritischem Bereich und der p-Wert Methode zu verdeutlichen, wird ein Beispiel durchgeführt:

Beispiel 3:
Es soll getestet werden, ob die **durchschnittliche Laufzeit μ** von Handy-Akkus möglicherweise von den vom Hersteller angegebenen 29,9 Stunden **abweicht**. Dazu werden bei **30 Akkus** dieser Marke unter kontrollierten gleichen Bedingungen die Laufzeiten gemessen.
Laufzeit der Handy-Akkus in der Grundgesamtheit ist **normalverteilt**.

$$H_0: \mu = \text{ 29,9 Stunden}$$
$$H_1: \mu \neq \text{ 29,9 Stunden}$$

Dies wird auch als zweiseitiger Test genannt.

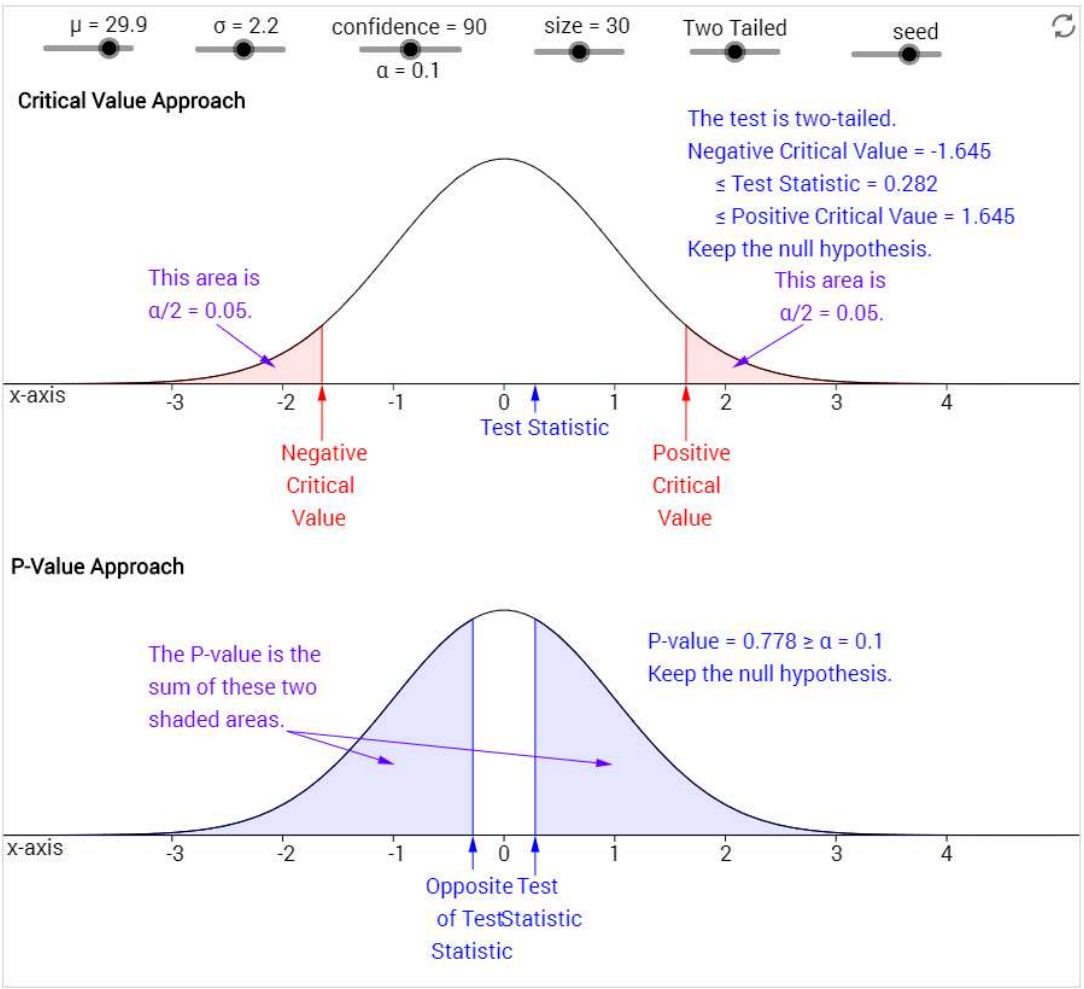

Abbildung 6: Die Ergebnisse von klassichem Verfahren (oben) und von p-Wert Methode

Tatsächlich nimmt p-Wert Methode einen größeren Bereich als den kritischen Bereich des klassichem Verfahrens. Man hat also ein größeres Vertrauen in die Entscheidung, H_0 anzunehmen.

1.2.2 Der Begriff „signifikant“

Die Bedeutung des Wortes „signifikant“ hängt von der Breite der Verteilung der Messwerte ab. Also ein Ergebnis ist signifikant, wenn die Chance, dass dieses zufällig entstanden ist, gering ist.

Beispiel 4:
In beiden Fällen unten betragen die Mittelwerte jeweils 570.7 und 582.1 - im ersten Fall wird man die Mittelwerte als unterschiedlich ansehen, im zweiten Fall ist dies sehr wohl nicht unmittelbar klar. Betrachtet man die zugehörigen Verteilungen, so wird die Bedeutung des Begriffes signifikant klar: Der Unterschied zwischen den Mittelwerten ist zwar numerisch in

jedem Fall gegeben, im ersten Fall ist dieser aber deutlich (=signifikant), im zweiten ist das nicht so deutlich, da die beiden Verteilungen sich beträchtlich überlagern.

	Fall 1:		Fall 2:	
	Norden	Süden	Norden	Süden
Einzelwerte:	571.5	581.7	566.3	571.5
	570.8	582.8	549.5	608.1
	570.3	580.4	538.9	544.7
	570.6	583.3	588.9	571.7
	570.4	582.9	592.5	589.7
	571.0	579.5	560.1	588.5
	571.6	582.8	572.9	577.7
	569.4	583.3	575.1	583.7
	570.5	583.0	602.7	561.7
	570.9	581.3	560.1	623.7
Mittelwerte:	570.7	582.1	570.7	582.1

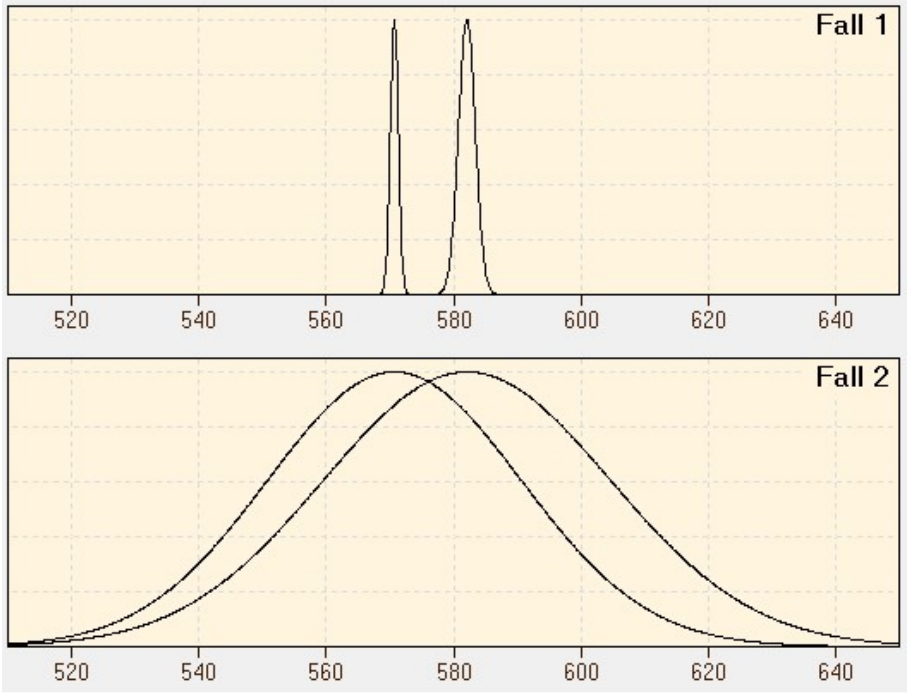

Abbildung 7: Verteilungen von Fall 1 und Fall 2

(Quelle: http://www.statistics4u.info/fundstat_germ/ee_significance.html)

In diesem Beispiel ist der Unterschied der Mittelwerte im Fall eins klar, weil sich die zugehörigen Verteilungen nicht überlappen und dadurch **die Chance, dass die Werte für den Norden und den Süden aus der selben Verteilung stammen praktisch gleich Null ist**. **Im zweiten Fall ist diese Chance ganz beträchtlich (nämlich ca. 12%).**

1.2.3 Bedeutung von p-Wert

Ein niedriger P-Wert weist darauf hin, dass die Stichprobe genügend Beweise liefert, dass die Nullhypothese für die gesamte Population abgelehnt werden können.
P-Werte adressieren nur eine Frage: Wie wahrscheinlich sind die Daten unter der Annahme einer echten Nullhypothese? Es misst nicht die Unterstützung für die alternative Hypothese.

2 Typische Fehlinterpretationen von p-Wert

2.1 Typische Fehlinterpretationen

1. Fehlinterpretation: *„Wenn p= 0.05, die Nullhypothese hat nur eine 5% Chance, wahr zu sein."*
Dies ist die am weitesten verbreite Missverständnisse über den p-Wert. Es setzt die falsche Vorstellung fort, dass die Daten allein uns sagen können, wie wahrscheinlich wir richtig oder falsch in unseren Schlussfolgerungen sind.
Der einfachste Weg, um zu sehen, dass dies falsch ist, besteht darin, dass der p-Wert unter der Annahme berechnet wird, dass die Nullhypothese wahr ist. Es kann daher nicht gleichzeitig eine Wahrscheinlichkeit sein, dass die Nullhypothese falsch ist.
Es gibt zum Beispiel, dass eine Münze viermal geworfen wird und das Ergebnis der Beobachtung sind vier Köpfe. Wenn eine Hypothese erstellt wird, mit der Frage, ob die Münze fair mit p-Wert 0.125 ist. Dies bedeutet nicht, dass die Wahrscheinlichkeit, ob die Münze fair ist, beträgt nur 12,5%. Die einzige Möglichkeit, diese Wahrscheinlichkeit zu berechnen, ist der Satz von Byes.

2. Fehlinterpretation: „*p=0.05 bedeutet, dass wenn die Nullhypothese abgelehnt wird, ist die Wahrscheinlichkeit eines Fehlers Typ I nur 5%."*
Diese Aussage ist äquivalent zu der vorherigen Aussage. Ein Fehler Typ I ist ein „falsch-positives" oder genauer zu beschreiben, dass die Nullhypothese abgelehnt wird, obwohl sie in Wirklichkeit zutrifft. Wie in der 1.Fehlinterpretation existiert p-Wert unter der Annahme, dass die Nullhypothese wahr ist. Deswegen ist es mit p=0.05 also 5% der Chance auf eine falsche Ablehnung von Nullhypothese zu sagen, ist nicht korrekt. Eine andere Möglichkeit zu sehen, dass diese Aussage falsch ist, kann mit einem Beispiel gezeigt werden. Eine Serie von Experimenten über die Wirksamkeit von Insulin gegen Diabetes. In der Realität wurde Insulin schon bewiesen, effektiv gegen Diabetes zu sein. Wenn eine Nullhypothese mit der Aussage, dass Insulin unwirksam gegen Diabetes ist, abgelehnt wird, ist die Wahrscheinlichkeit, dass diese Ablehnung falsch ist, immer Null. Da alle Ablehnungen dieser Nullhypothese immer wahr sind, egal was der p-Wert ist.

3. Fehlinterpretation: „*eine wissenschaftliche Schlussfolgerung sollte darauf basieren, ob der p-Wert signifikant ist oder nicht."*
Dieses Missverständnis umfasst alle anderen. Es ist äquivalent zu der Aussage, dass die Größe des Effektes nicht relevant ist, und allein das Ergebnis des Experimentes im Labor ist relevant für eine wissenschaftliche Schlussfolgerung. Die Beweise aus einer gegebenen Studie müssen mit denen aus früheren Arbeiten kombiniert werden, um eine Schlussfolgerung zu generieren. In einigen Fällen kann eine wissenschaftlich vertretbare Schlussfolgerung darin bestehen, dass die Nullhypothese wahrscheinlich auch nach einem signifikanten Ergebnis noch wahr ist und umgekehrt.

2.2 p-Werte und die Replikation von Experimenten

Jim Frost und sein Artikel „P Values and the Replication of Experiments" zusammen mit einer Studie im Jahr 2015 „Estimating the reproducibility of psychological science" erklären die Beziehung zwischen p-Werten und der Replikation von experimentellen Ergebnissen. Diese beiden Studien highlighten, dass es entscheidend ist, p-Werte richtig zu interpretieren, und signifikante Ergebnisse müssen repliziert werden, um als vertrauenswürdig zu zeigen.
In der Studie musste die Gruppe von 300 Forschern, die mit dem Projekt verbunden waren, zuerst eigene Replikationsstudien durchführen. Diese Forscher führten Replikationen von 100 Psychologie-Studien durch, die bereits statistisch signifikante Ergebnisse erzielt hatten und von drei renommierten psychologischen Fachzeitschriften zur Veröffentlichung angenommen worden waren.
Insgesamt ergab die Studie, dass nur 36% der Replikationsstudien selbst statistisch signifikant waren. Die Grafik zeigt auch, wie **niedrigere p-Werte** in den Originalstudien mit einer höheren Rate statistisch signifikanter Ergebnisse in den Follow-up-Studien assoziiert sind. Diese niedrige Rate bestätigt die Wichtigkeit, die Ergebnisse zu replizieren, bevor eine experimentelle Bestätigung anerkannt wird.

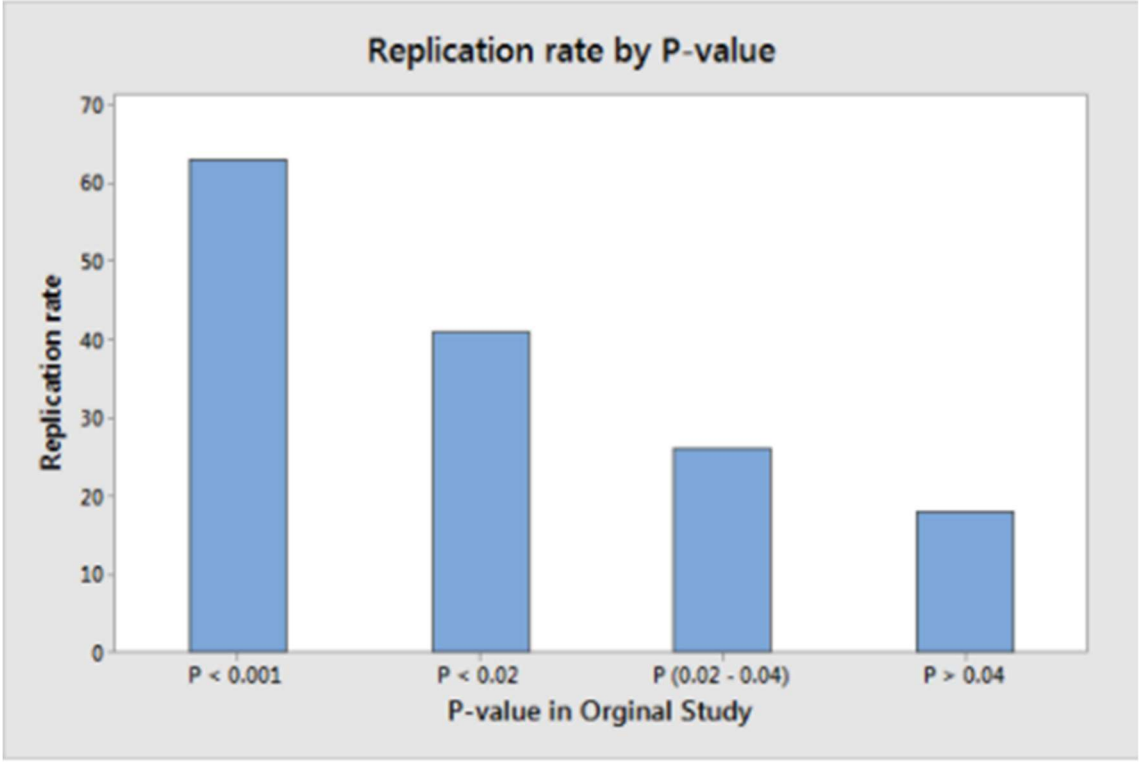

Abbildung 8: P Values and the Replication of Experiments (Quelle: Jim Frost, 2015)

3 p-Wert Simulation

3.1 Fehler 1.Art bei zusätzlicher beobachtung

Forscher entscheiden häufig, wann sie die Datenerhebung auf der Grundlage einer vorläufigen Datenanalyse einstellen. In einer Studie im Jahr 2011 wurde eine Umfrage unter den Forschern durchgefürht und das Ergebnis ergab, dass ca. 70% der Befragten dieses Verhalten zugegeben haben. Viele Forscher glauben, dass diese Praxis nur einen trivialen Einfluss auf falsch-positive Raten ausübt. (Simmons, et al., 2011)

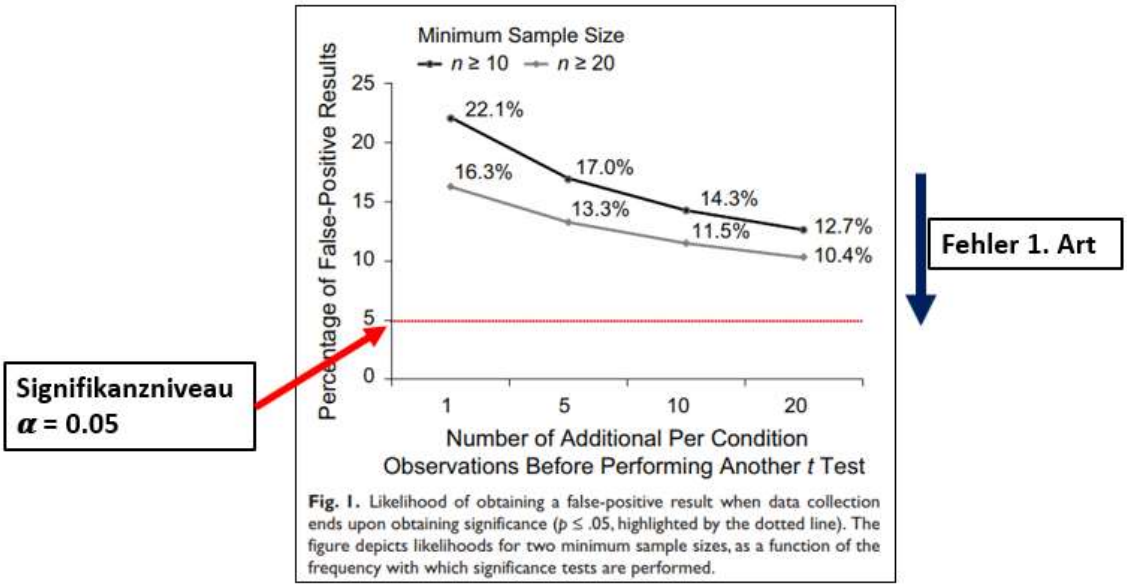

Abbildung 9: Die Reduzierung des Fehler 1.Arts bei zusätzlicher Beobachtung

(Quelle: Joseph P. Simmons, False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant)

Im Widerspruch zu dieser Intuition zeigt Abbildung 1 die falsch-positiven Raten von zusätzlichen Simulationen für einen Forscher, der entweder 10 oder 20 Beobachtungen innerhalb jeder der beiden Bedingungen gesammelt hat und dann nach jeder 1, 5, 10 oder 20 Beobachtungen pro Bedingung auf Signifikanz getestet hat. Der Forscher hört auf, Daten zu sammeln, sobald statistische Signifikanz erreicht ist oder wenn die Anzahl der Beobachtungen in jeder Bedingung 50 erreicht.

3.2 Simulation unter der Annahme, dass die Null wahr ist

In seinem Artikel über die Simulation von p-Werten hat der Autor Jim Frost das Verhalten des Werts untersucht unter der Bedingung, dass die Nullhypothese immer wahr ist.

100.000 Experimente mit jeweils 100 Teilnehmern werden simuliert. Der Mittelwert jeder Gruppe wurde auf 100 festgelegt, und die Standardabweichung wurde auf 20 festgelegt. Für jeden Teilnehmer setzte der Computer als Punktzahl eine zufällig ausgewählte Zahl aus einer Normalverteilung mit einem Mittelwert von 100 und Standardabweichung von 20. Am Ende von jedem Experiment führt der Computer einen t-Test durch und zeichnet den beobachteten p-Wert auf. Am Ende wird ein Histogramm aufgezeichnet, das die Häufigkeit jedes p-Wertes über die gesamte Simulation zeigt.

Entgegen vieler Erwartungen folgt der p-Wert einer stetigen Gleichverteilung (Uniformverteilung). Wenn der Nullhypothese wahr ist, ist der p-Wert eine Zufallsvariable zwischen Null und Eins. Das bedeutet, wenn die Nullhypothese wahr ist, nimmt der p-Wert gleichermaßen jeden Wert an. Somit ist ein p-Wert von 0.009 genauso wahrscheinlich wie ein p-Wert von 0.9.

Namenlos - R Editor

```
nSims <- 100000 #number of simulated experiments
p <-numeric(nSims) #set up empty container for all simulated p-values

for(i in 1:nSims){ #for each simulated experiment
  x<-rnorm(n = 100, mean = 100, sd = 20) #produce 100 simulated participants
                                         #with mean=100 and SD=20
  y<-rnorm(n = 100, mean = 100, sd = 20) #produce 100 simulated participants
                                         #with mean=100 and SD=20
 z<-t.test(x,y) #perform the t-test
 p[i]<-z$p.value #get the p-value and store it
}

#now plot the histogram
hist(p, main="Histogram of p-values under the null", xlab=("Observed p-value"))
```

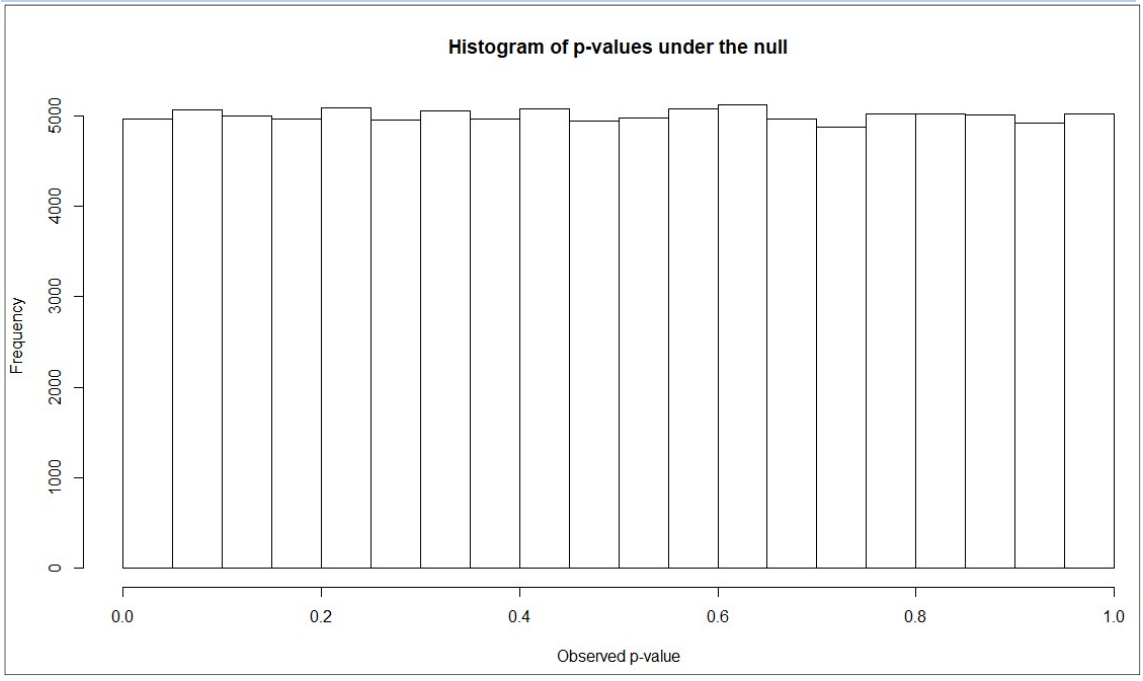

Abbildung 10: Histogramm der p-Werte der Annahme, dass die Null wahr ist

(Quelle: Jim Frost, Understanding p-values via simulations)

3.3 Simulation unter der Annahme, dass die Null nicht wahr ist

Das Verhalten des Wertes, wenn die Nullhypothese als nicht wahr angenommen wird, wird in folgenden Grafiken gezeigt. Der Autor hat dazu den Mittelwert der Verteilung für eine

Gruppe von 100 auf 103 erhöht. Das resultierende Histogramm zeigt, dass es ein größeres Auftreten von niedrigen p-Werten gibt.

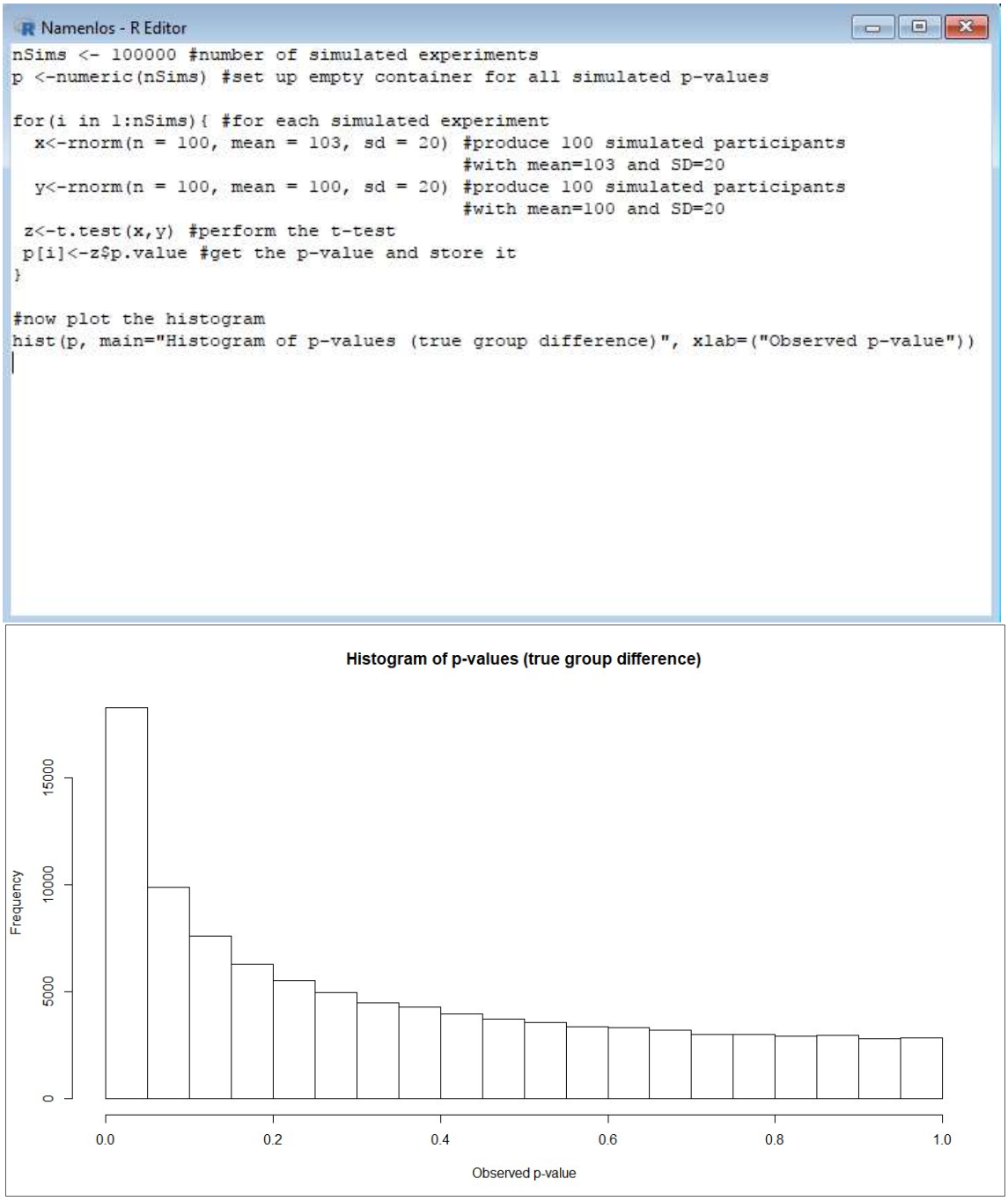

Abbildung 11: Histogramm der p-Werte der Annahme, dass die Null nicht wahr ist und N=100

(Quelle: Jim Frost, Understanding p-values via simulations)

In einer anderen Simulation mit gleicher Annahme nimmt die beobachtete Häufigkeit von p-Werten zu ihrem niedrigeren Wert zu, wenn die Stichprobengröße jedes simulierten

Experiments auf 500 erhöht wird. Dies liegt daran, dass jedes Experiment aufgrund seiner größeren Stichprobengröße eine größere Wahrscheinlichkeit hat, den wahren Unterschied in den Daten zu finden.

Hier ist das neue Histogramm mit N = 500 in jedem Experiment:

Namenlos - R Editor

```
nSims <- 100000 #number of simulated experiments
p <-numeric(nSims) #set up empty container for all simulated p-values

for(i in 1:nSims){ #for each simulated experiment
  x<-rnorm(n = 500, mean = 103, sd = 20) #produce 100 simulated participants
                                         #with mean=103 and SD=20
  y<-rnorm(n = 500, mean = 100, sd = 20) #produce 100 simulated participants
                                         #with mean=100 and SD=20
 z<-t.test(x,y) #perform the t-test
 p[i]<-z$p.value #get the p-value and store it
}

#now plot the histogram
hist(p, main="Histogram of p-values (true group difference)", xlab=("Observed p-value"))
```

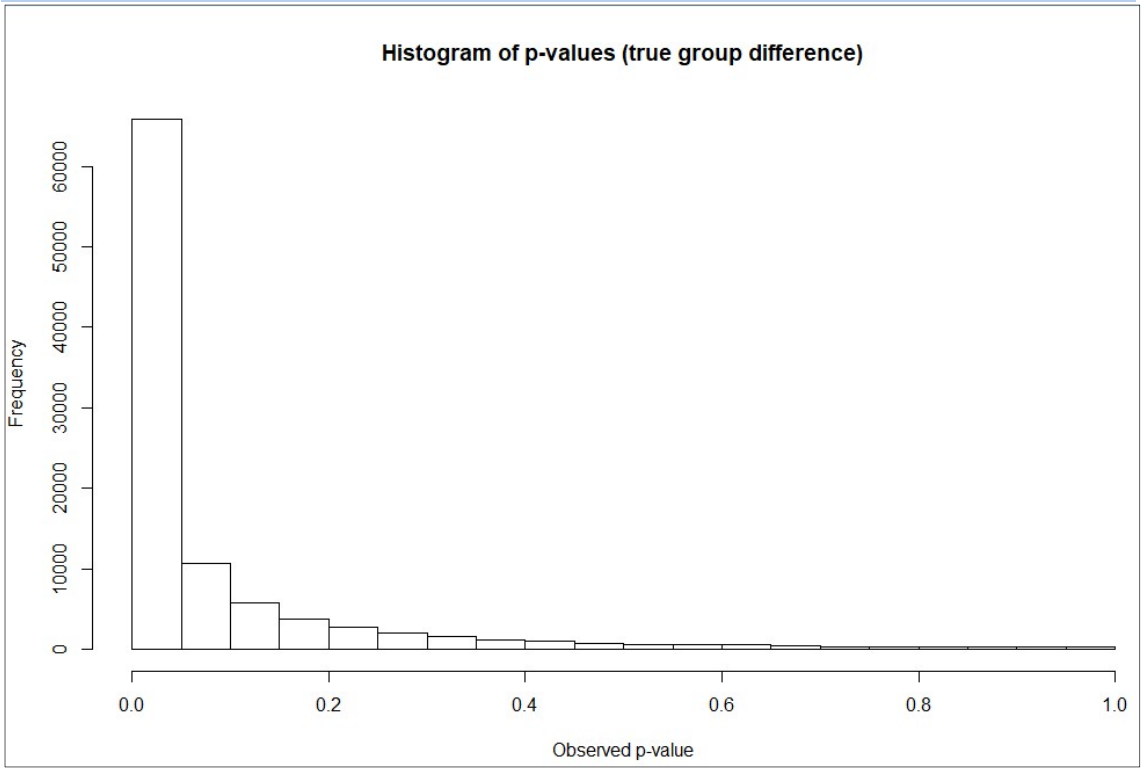

Abbildung 12: Histogramm der p-Werte der Annahme, dass die Null nicht wahr ist und N=500

(Quelle: Jim Frost, Understanding p-values via simulations)

3.4 "p-Hacking" hilft bei der Datenmassage

Im Allgemein bedeutet der Ausdruck „p-Hacking“, dass "die Autoren ließen eine Bedingung weg, damit der Gesamt-p-Wert auf unter 0,05 fiel". Oder: "Sie ist ein P-Hacker, sie schaut sich die Daten immer schon an, wenn das Experiment noch läuft." (Nuzzo , 2014)

Mit Hilfe von Simulationen hat Joseph P. Simmons gezeigt, dass es bereits genügt, die Stichprobegröße zu ändern. Allein das kann dazu führen, dass p-Wert gestiegen oder gefallen ist. Dann basierend auf der Veränderung des p-Werts wird die Nullhypothese auch bestätigt oder abgelehnt. Eine Person kann mit dieser Methode die Daten anschauen und dann die Stichprobegröße so ändern, damit er das gewünschte Ergebnis erhalten kann.

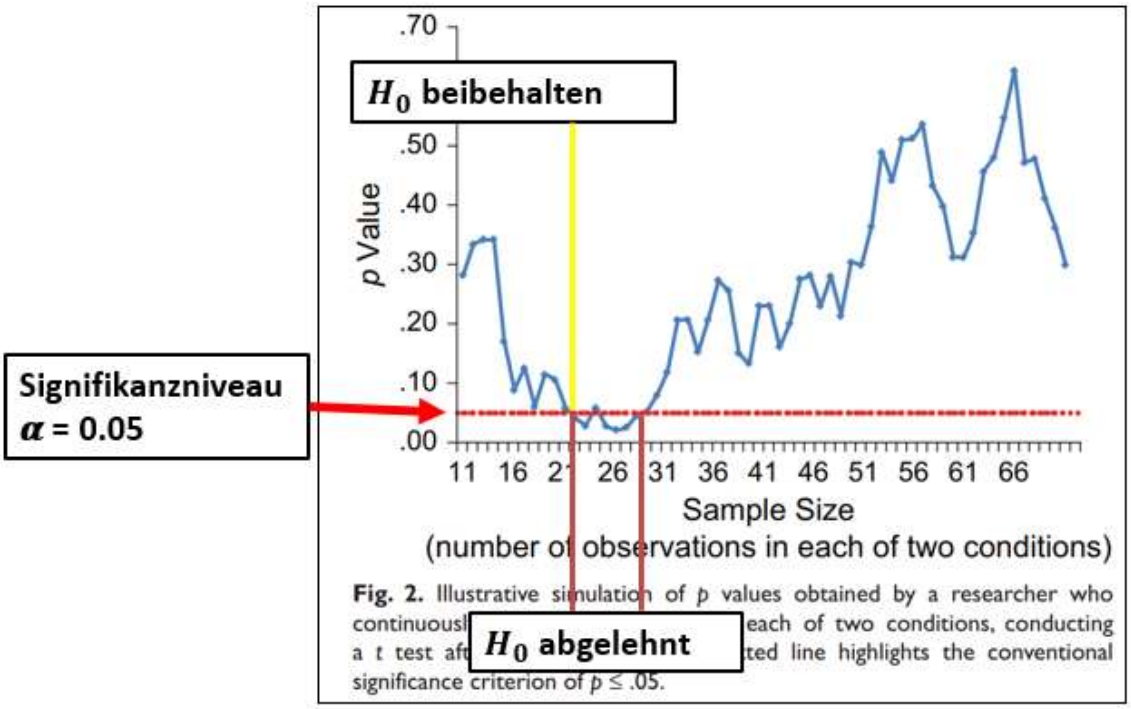

Abbildung 13: Illustrative Simulation von p-Werten, die von einem Forscher erhalten werden, der kontinuierlich eine Beobachtung hinzufügt.

(Quelle: Joseph P. Simmons, False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant)

4 Zusammenfassung

In Forschungen soll der p-Wert nur eine von vielen Kennzahlen betrachtet werden, weil erstens p-Wert keinerlei Informationen darüber liefert, ob ein Ergebnis wahr bzw. wichtig ist. Zweitens ist es vergleichbar einfach, die vorhandenen Daten so zu bearbeiten, dass der p-Wert möglichst klein wird (p-Hacking). Drittens war der p-Wert gar nie darauf ausgelegt, eine derart beherrschende Rolle bei der Auswertung von Daten einzunehmen. Sein Erfinder Ronald A. Fisher, einer der Gründerväter der modernen Statistik, sah ihn lediglich als eine von vielen verschiedenen statistischen Kennzahlen. „Der p-Wert sollte zeigen, ob es sich lohnt, ein bestimmtes Resultat genauer zu betrachten – und vor allem: das Experiment zu wiederholen!“ (Grüninger, 2016)

Entscheidend ist es, Fachkenntnisse zu allen Aspekten des Hypothesentests anzuwenden. Die Forscher müssen ihr wissenschaftliches Urteil über die Plausibilität der Hypothesen, die Ergebnisse ähnlicher Studien, die vorgeschlagenen Mechanismen, das richtige experimentelle Design und so weiter treffen. Expertenwissen verwandelt Statistiken aus Zahlen in aussagekräftige, vertrauenswürdige Ergebnisse.

5 Literaturverzeichnis

Böker, F., 2006. *Formelsammlung für Wirtschaftswissenschaftler.* s.l.:Pearson Studium.

Dr. Delgado, R., 2008. *wiwi.uni-rostock.de.* [Online]
Available at: https://www.wiwi.uni-rostock.de/fileadmin/Institute/VWL/LS_Statistik/vorl_gs/Testverfahren_I.pdf
[Zugriff am 18 03 2018].

Frost, J., 2014. *The Minitab Blog.* [Online]
Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/five-guidelines-for-using-p-values
[Zugriff am 20 03 2018].

Grüninger, S., 2016. *Neue Zürcher Zeitung.* [Online]
Available at: https://www.nzz.ch/karriere/studentenleben/der-problem-wert-ld.133818
[Zugriff am 01 04 2018].

Nuzzo , R., 2014. *spektrum.de.* [Online]
Available at: https://www.spektrum.de/news/statistik-wenn-forscher-durch-den-signifikanztest-fallen/1224727
[Zugriff am 05 04 2018].

Simmons, J. P., Nelson, L. D. & Simonsohn, U., 2011. *False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.* [Online]
Available at: http://pss.sagepub.com/content/22/11/1359
[Zugriff am 11 03 2018].

U. J., 2015. *metheval.uni-jena.de.* [Online]
Available at: www.metheval.uni-jena.de/get.php?f=1015
[Zugriff am 28 03 2018].

U. K., 2001. *eswf.uni-koeln.de.* [Online]
Available at: http://eswf.uni-koeln.de/glossar/node85.html
[Zugriff am 03 15 2018].

6 Bildverzeichnis