Understanding p-value

Tuan Anh Nguyen - 20.11.2017

Seminar „Neuste Trends in Big Data Analytics“

Betreuer: Dr.Julian Kunkel

Agenda

- Motivation
- Grundlagen
- Typische Fehlinterpretationen
- p-Wert Simulation
- Zusammenfassung

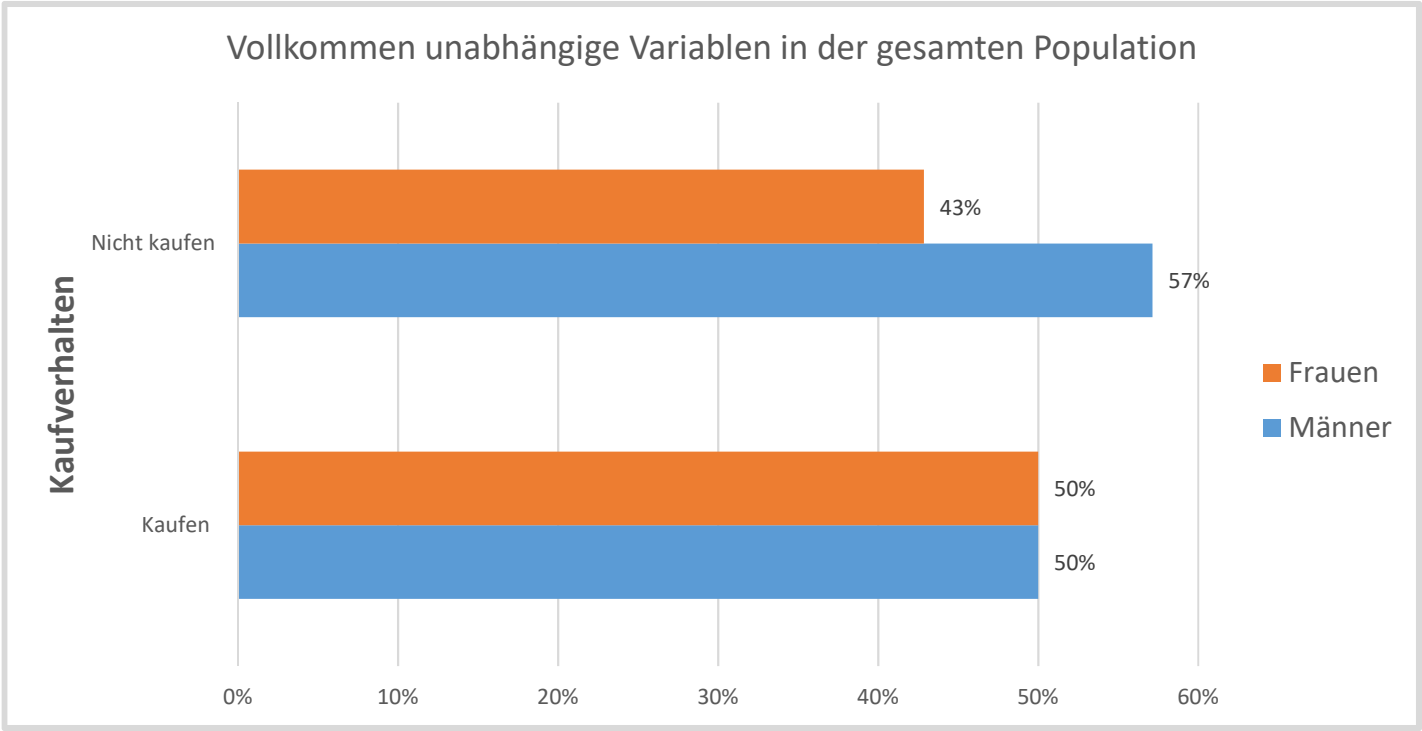
Motivation

P-Werte wurden von Statistiker **Ronald Aylmer Fisher** eingeführt, um als eine objektive Methode die Daten mit der **Nullhypothese** zu vergleichen.

(Quelle: five-guidelines-for-using-p-values – Jim Frost, [1])

Motivation

Beispielaufgabe



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

Motivation

Beispielaufgabe

Wir führen ein X^2 - Unabhängigkeitstest:

n =150 Kunden

X = Geschlecht

Y = Kaufverhalten

H_0: X, Y sind unabhängig.

Motivation

Beispielaufgabe

Kreuztabelle: die beobachte Häufigkeit der jeweiligen Kombination zwischen Geschlecht und Kaufverhalten der Kunden

	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	

Motivation

Beispielaufgabe

Erwartete Kreuztabelle:

	Kaufen	Nicht kaufen
Männer	42.6	37.3
Frauen	37.3	32.6

Wobei **die erwartete Zellhäufigkeit**: $\tilde{h}_{ij} := \frac{h_{i.} \cdot h_{.j}}{n}$

Motivation

Beispielaufgabe

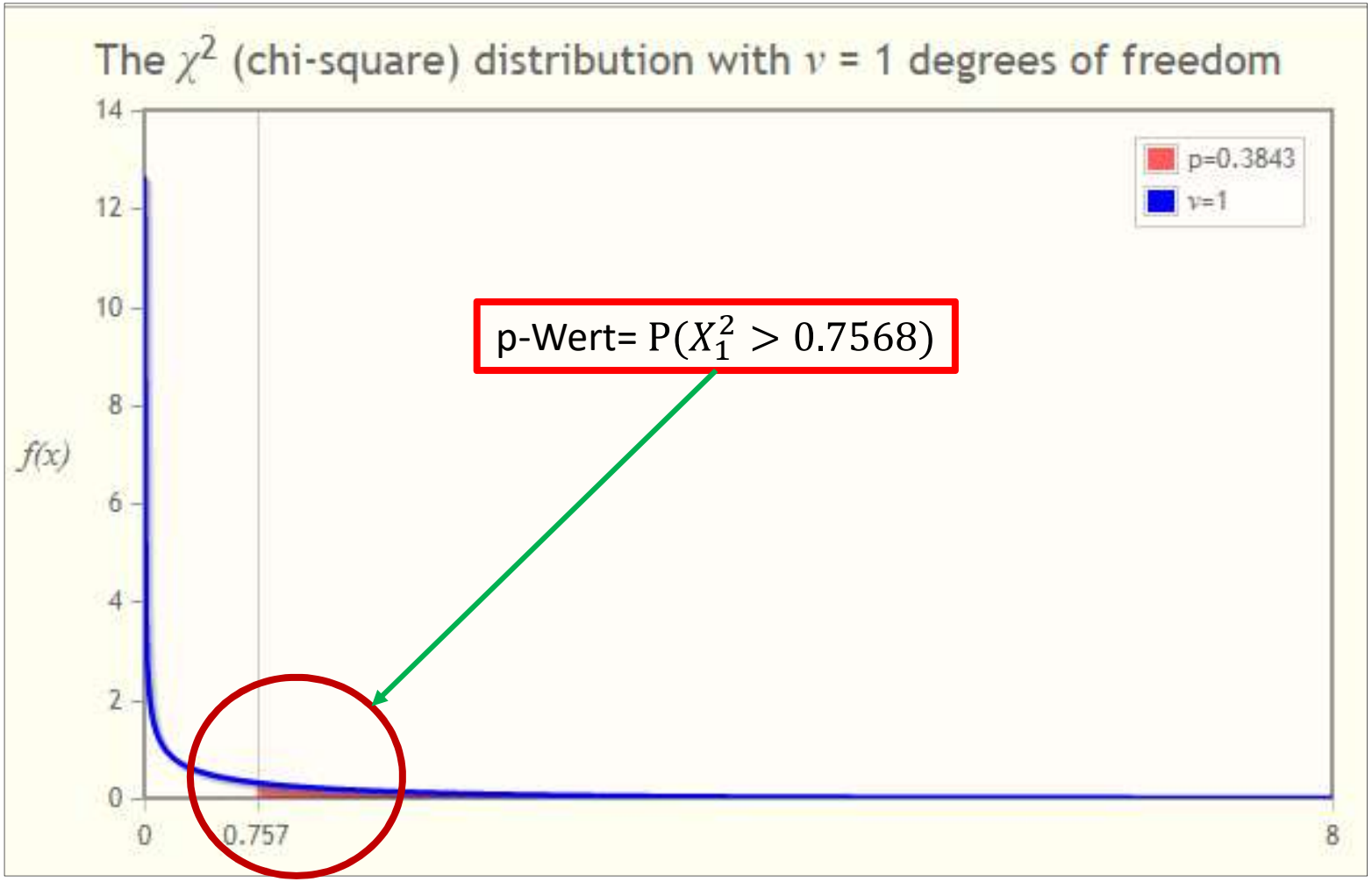
Prüfgröße: $V = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} =$

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

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

Motivation

Beispielaufgabe



Motivation

Beispielaufgabe

Mit Hilfe von Maple: p-Wert= $\mathrm{P}(X_1^2 > 0.7568) < \alpha$

```
> 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.

Motivation

Beispielaufgabe

Andere Methode:

- Bei einem **Signifikanzniveau $\boldsymbol{\alpha}$** wird H_0 abgelehnt, wenn $\mathbf{V} > \boldsymbol{X^2_{(k-1)(l-1),1-\alpha}}$

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

- **Im Beispiel:** $k = 2, l = 2, \alpha = 5\%$

$\rightarrow X^2_{1,0.95} = 3.841$ (Tabelle!) $> V = 0.7568$

$\rightarrow \boldsymbol{H_0}$ wird nicht abgelehnt

Motivation

Beispielaufgabe

Tabelle der Quantile der Chiquadrat-Verteilung

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

(Quelle: eswf.uni-koeln.de, [2])

Grundlagen

Hyothesentest

In einem statistischen Test werden zwei **gegensätzliche Hypothesen** gegenüber gestellt.

H_0: Nullhypothese

H_1: Alternativhypothese oder Forschungshypothese

Grundlagen

Hyothesentest

Fehlertyp bei einem statistischen Hypothesentest:

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

Grundlagen

Hyothesentest

Fehler 1.Art (Ablehnung von H_0 , wenn H_0 zutrifft) **so klein wie möglich bleibt**.

Obere Grenze α **(Signifikanzniveau)**

Es gilt dann: **P(Fehler 1. Art) ≤ α**

Grundlagen

Was ist p-Wert?

Definition: Der p-Wert ist definiert als die Wahrscheinlichkeit, unter H_0 den beobachteten Prüfgrößenwert oder einen in Richtung der Alternative extremeren Wert zu erhalten.

(Quelle: Formelsammlung für Wirtschaftswissenschaftler: Mathematik und Statistik, Fred Böker)

- $p \leq \alpha \Rightarrow H_0$ ablehnen (Alternativhypothese wird angenommen -> **Statistische Signifikanz**)
- $p > \alpha \Rightarrow H_0$ beibehalten

Grundlagen

P-Wert und kritischer Wert

Beispiel: Zweiseitiger Test

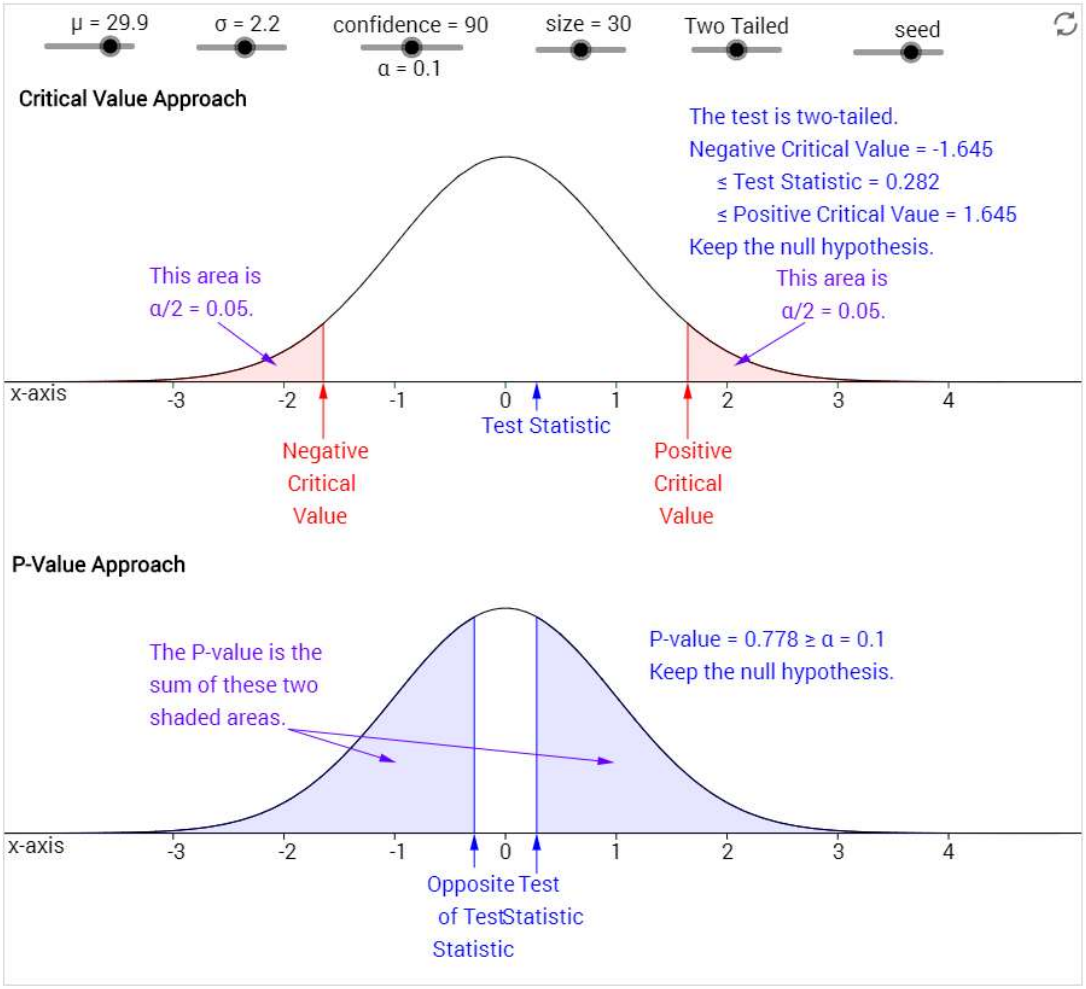
Es soll getestet werden, ob die **durchschnittliche Laufzeit μ** von Handy-Akkus möglicherweise von den vom Hersteller angegebenen 29,9 Stunden **abweicht**. Dazu werden bei **30 Akkus** dieser Marke unter kontrollierten gleichen Bedingungen die Laufzeiten gemessen.

Laufzeit der Handy-Akkus in der Grundgesamtheit **normalverteilt** ist.

$\mathrm{H}_0: \mu = 29{,}9 \text{ Stunden}$
$\mathrm{H}_1: \mu \neq 29{,}9 \text{ Stunden}$

Grundlagen

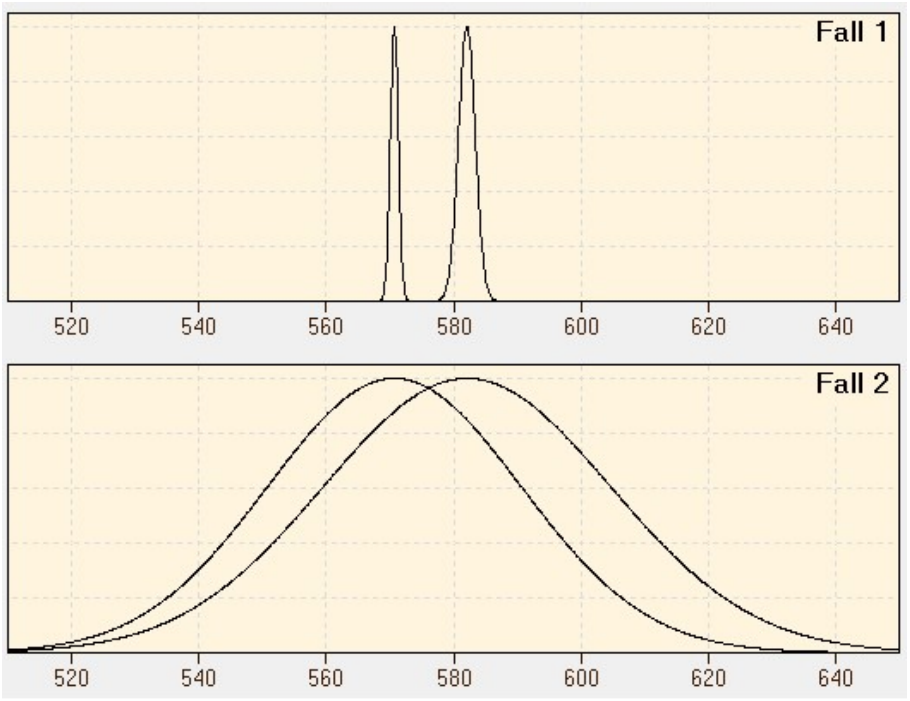
P-Wert und kritischer Wert



Grundlagen

Der Begriff 'signifikant'

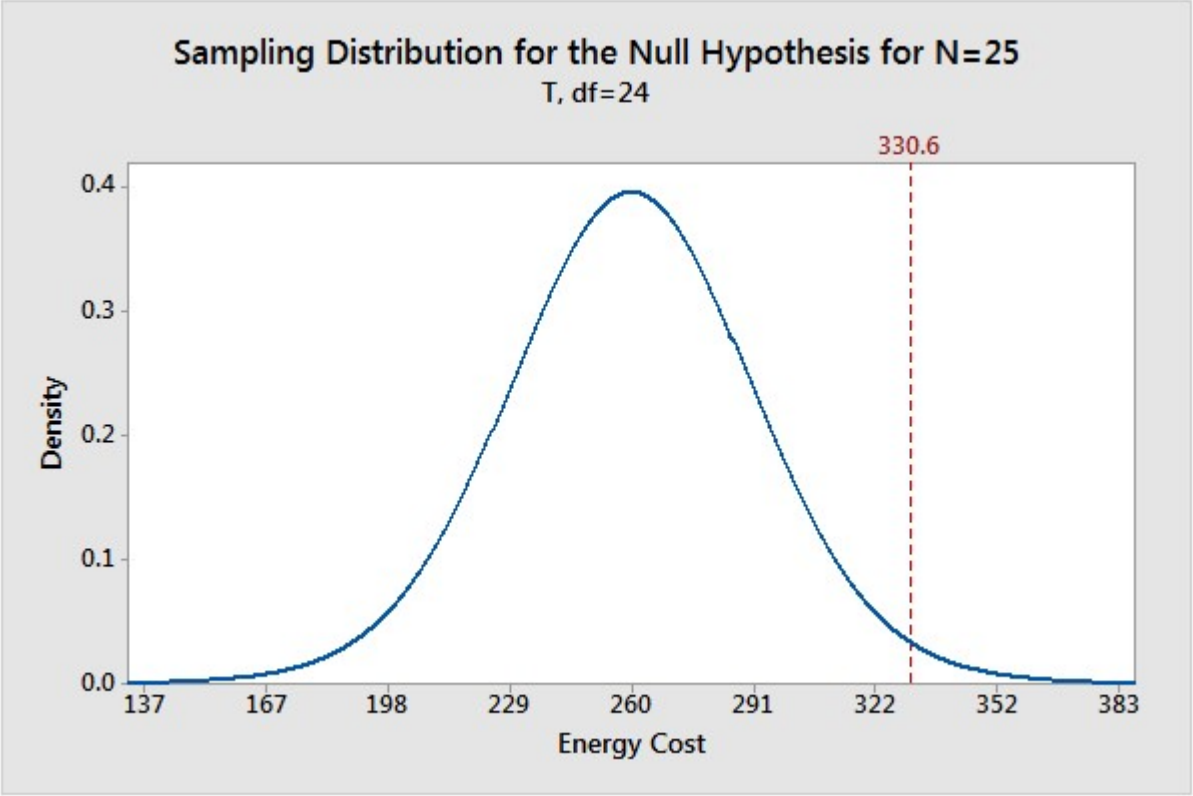
	Fall 1:		Fall 2:	
	Norden	Süden	Norden	Süden
Einzelwerte:	571.5	581.7	566.3	571.5
	570.8	582.8	549.5	608.1
	570.3	580.4	538.9	544.7
	570.6	583.3	588.9	571.7
	570.4	582.9	592.5	589.7
	571.0	579.5	560.1	588.5
	571.6	582.8	572.9	577.7
	569.4	583.3	575.1	583.7
	570.5	583.0	602.7	561.7
	570.9	581.3	560.1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



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

Grundlagen

Was ist p-Wert?



Hypothesentest – zweiseitig

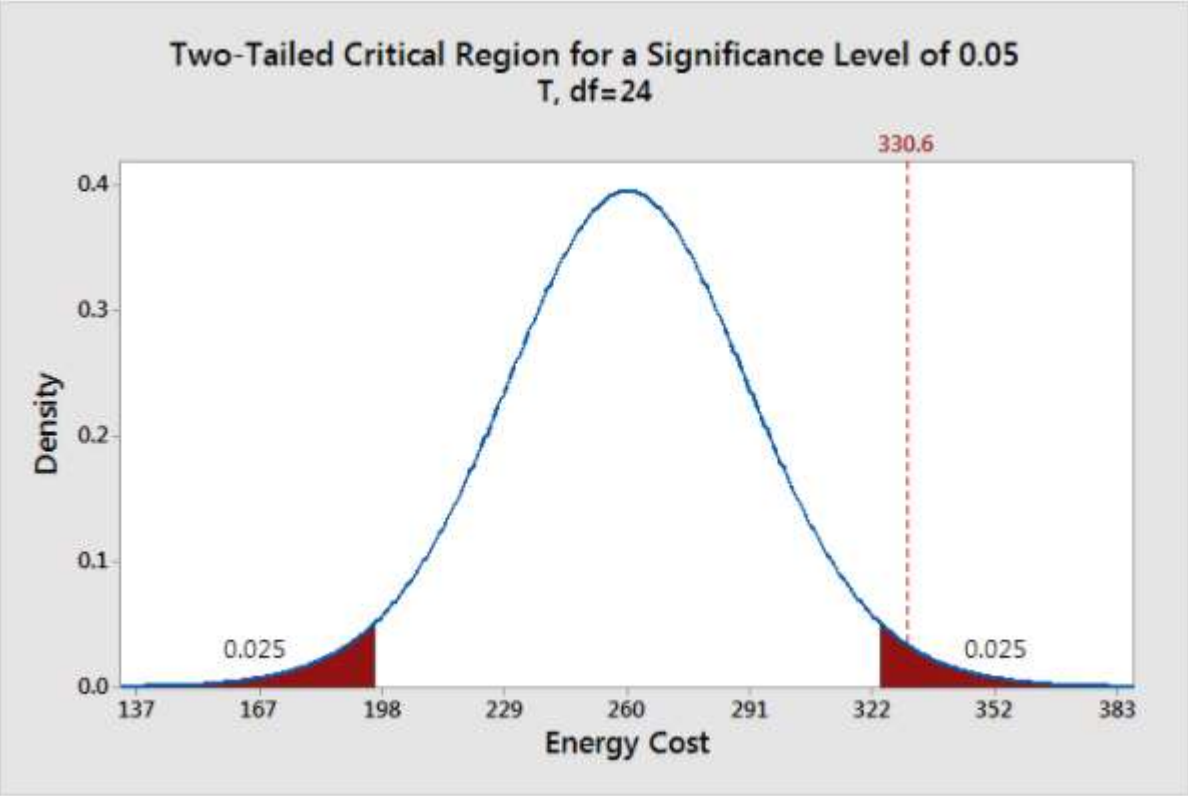
$H_0: \mu = 260$

$H_1: \mu \neq 260$

(Quelle: Understanding Hypothesis Tests: Significance Levels (Alpha) and P values in Statistics – Jim Frost, [3])

Grundlagen

Was ist p-Wert?



Hypothesentest – zweiseitig

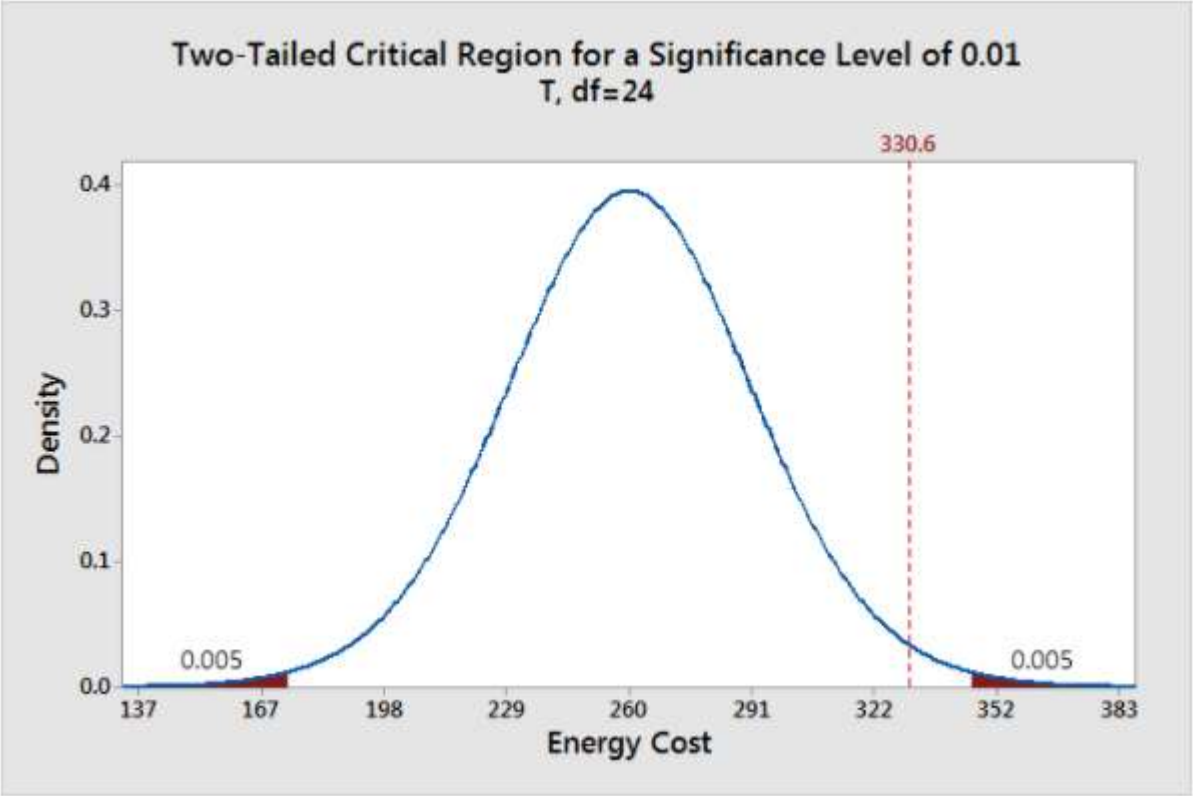
$H_0: \mu = 260$

$H_1: \mu \neq 260$

(Quelle: Understanding Hypothesis Tests: Significance Levels (Alpha) and P values in Statistics – Jim Frost, [3])

Grundlagen

Was ist p-Wert?



Hypothesentest – zweiseitig

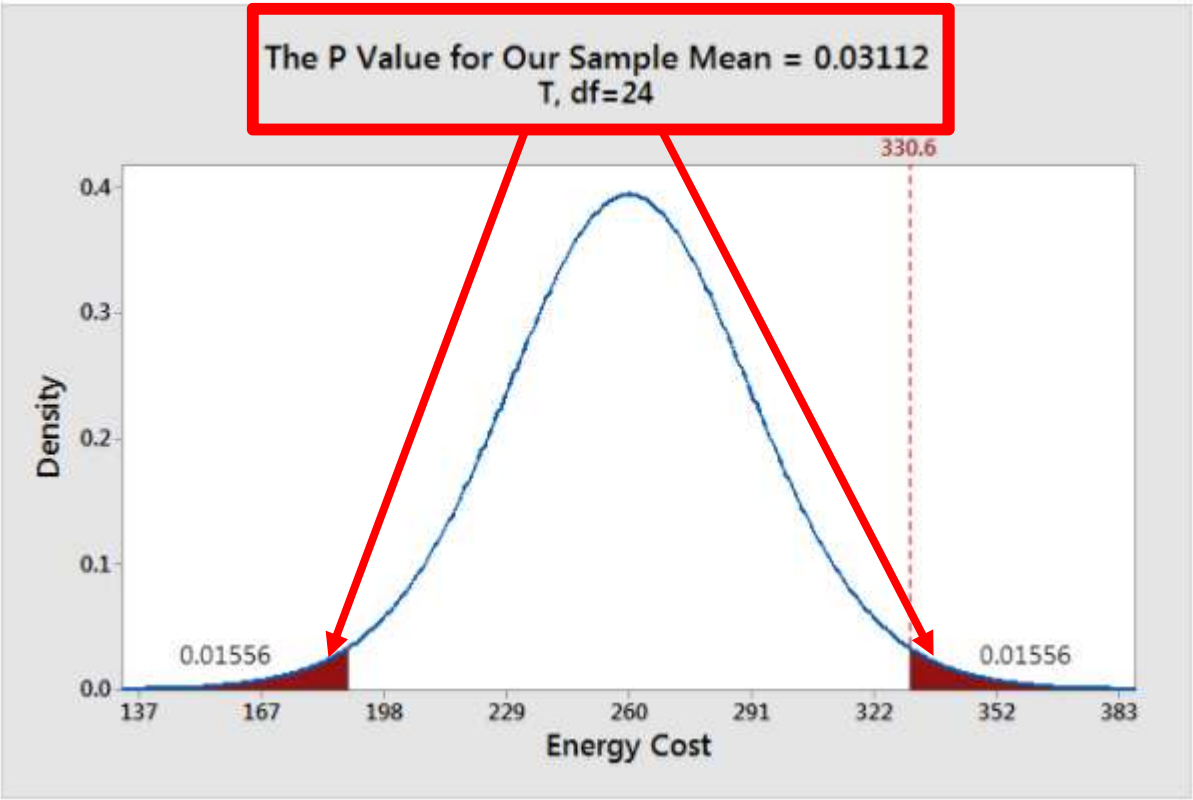
$H_0: \mu = 260$

$H_1: \mu \neq 260$

(Quelle: Understanding Hypothesis Tests: Significance Levels (Alpha) and P values in Statistics – Jim Frost, [3])

Grundlagen

Was ist p-Wert?



Hypothesentest – zweiseitig

$H_0: \mu = 260$

$H_1: \mu \neq 260$

$\alpha = 0.05$ -> H_0 beibehalten

$\alpha = 0.01$ -> H_0 wird abgelehnt

(Quelle: Understanding Hypothesis Tests: Significance Levels (Alpha) and P values in Statistics – Jim Frost, [3])

Grundlagen

Bedeutung von p-Wert

- Ein niedriger P-Wert: Stichprobe liefert **genügend Beweise**, dass die Nullhypothese **abgelehnt werden können**.

- P-Werte adressieren nur eine Frage: **Wie wahrscheinlich** sind unsere Daten unter der Annahme einer Nullhypothese?

- P-Wert **misst nicht** die Unterstützung für die **alternative Hypothese**.

→ Häufige **Fehlinterpretation** von P-Werten

Typische Fehlinterpretationen

1. Fehlinterpretation: Wenn p= 0.05, die Nullhypothese hat **nur** eine 5% Chance, wahr zu sein.

Falsche Vorstellung: wie wahrscheinlich wir **richtig oder falsch** in unseren **Schlussfolgerungen** sein können.

Der P-Wert unter der Annahme: die Nullhypothese ist wahr. Es kann daher nicht gleichzeitig eine Wahrscheinlichkeit sein , dass die Nullhypothese falsch ist.

Beispiel: Nehmen wir an, Wir drehen viermal eine Münze und beobachten vier Köpfe:

→ p= 0.125: Dies bedeutet **nicht**, dass die Wahrscheinlichkeit, ob die Münze fair ist, beträgt nur 12,5%.

(Quelle: A Dirty Dozen: Twelve P-Value Misconceptions – Steven Goodman)

Typische Fehlinterpretationen

2. Fehlinterpretation: p=0.05 bedeutet, dass wenn die Nullhypothese abgelehnt wird, ist die Wahrscheinlichkeit eines Fehlers Typ I nur 5%.

Fehler 1. Art: H_0 ist richtig aber wird abgelehnt.

Diese Aussage ist äquivalent zu der Aussage in Folie 25

(Quelle: A Dirty Dozen: Twelve P-Value Misconceptions – Steven Goodman)

Typische Fehlinterpretationen

3. Fehlinterpretation: Eine wissenschaftliche Schlussfolgerung sollte darauf basieren, ob der p-Wert signifikant ist oder nicht.

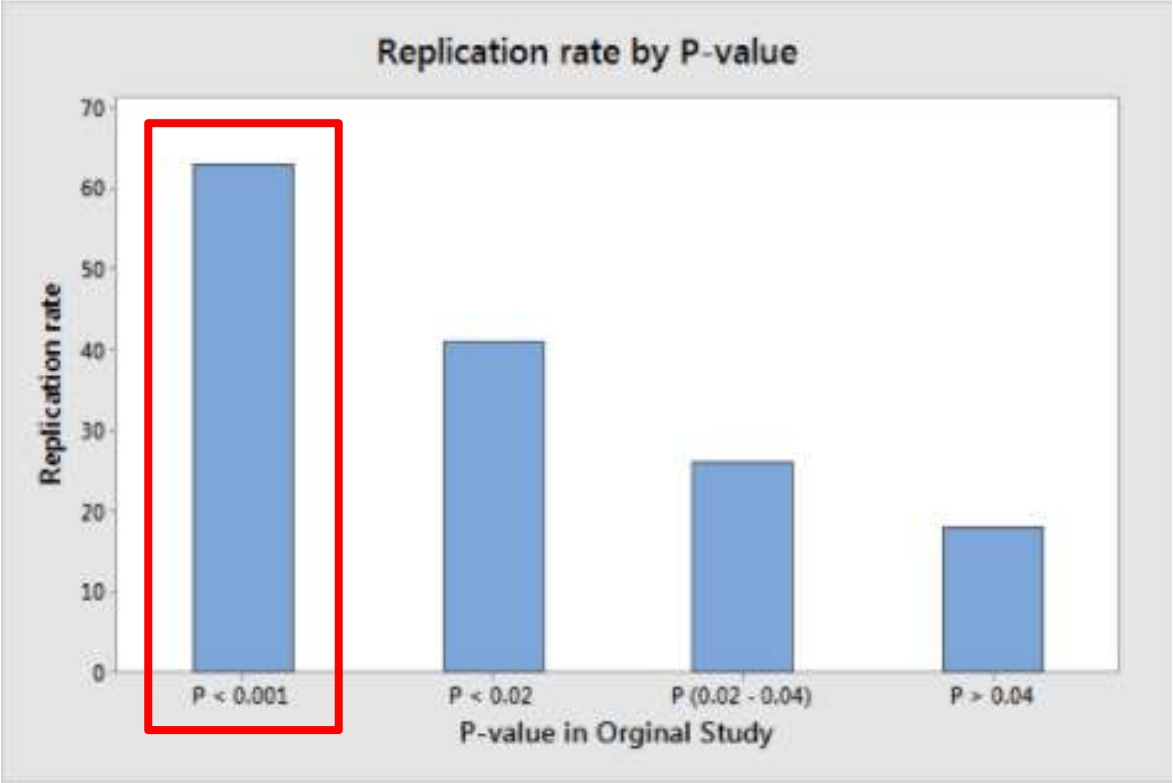
Die Nachweise aus einer gegebenen Studie muss mit derjenigen aus früheren Forschungen **kombiniert** werden, um eine Schlussfolgerung zu erhalten.

In einigen Fällen könnte eine wissenschaftlich vertretbare Schlussfolgerung sein, dass die Nullhypothese auch **nach einem signifikanten Ergebnis** wahrscheinlich **noch wahr** ist und umgekehrt.

(Quelle: A Dirty Dozen: Twelve P-Value Misconceptions – Steven Goodman)

Typische Fehlinterpretationen

P-Werte und die Replikation von Experimenten



Niedrigere p-Werte in den Originalstudien mit einer höheren Rate statistisch signifikanter Ergebnisse in den Follow-up-Studien assoziiert sind.

(Quelle: P Values and the Replication of Experiments – Jim Frost, [4])

p-Wert Simulation

Fehler 1.Art bei zusätzlicher beobachtung

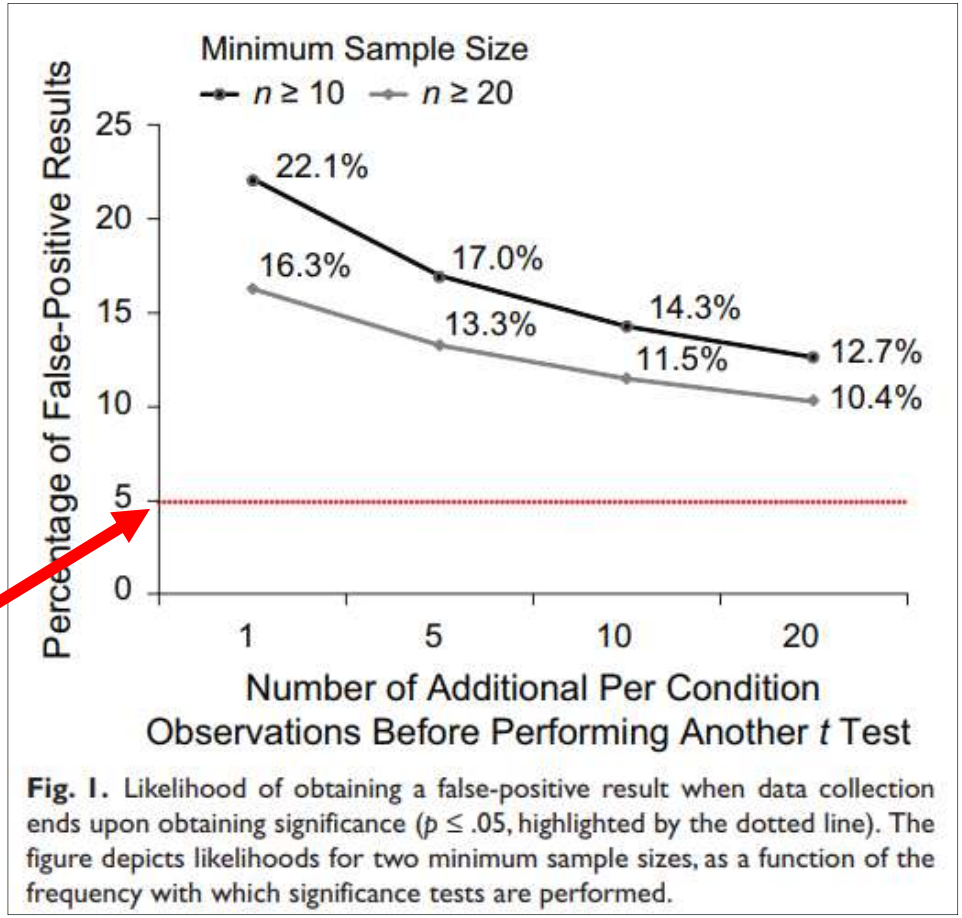


Fig. 1. Likelihood of obtaining a false-positive result when data collection ends upon obtaining significance ($p \leq .05$, highlighted by the dotted line). The figure depicts likelihoods for two minimum sample sizes, as a function of the frequency with which significance tests are performed.

Signifikanzniveau
α = 0.05

Fehler 1. Art

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

p-Wert Simulation

Simulation unter der Annahme, dass die Null wahr ist.

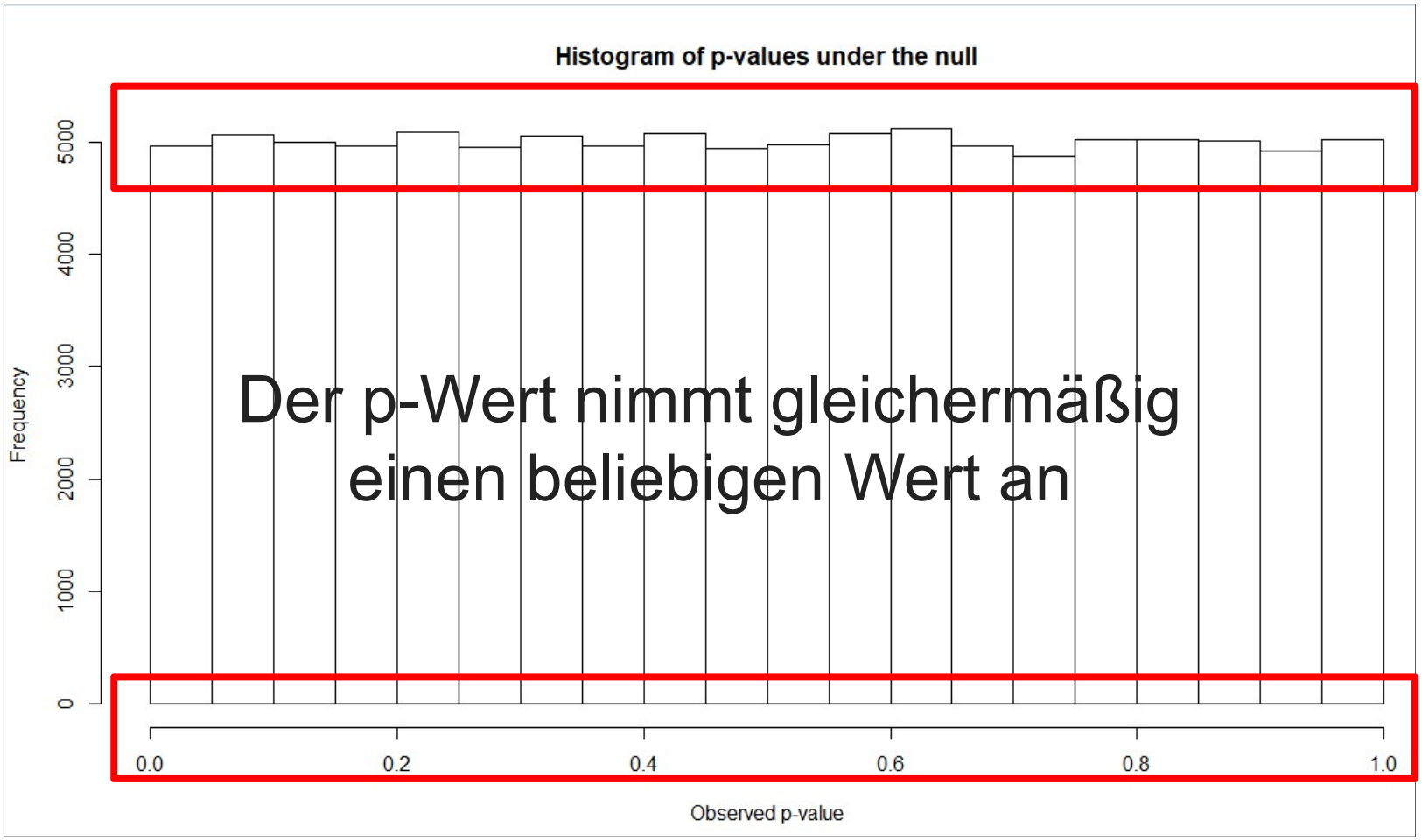
Namenlos - R Editor

```
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"))
```

(Quelle: researchutopia – Jim Frost, [5])

p-Wert Simulation

Simulation unter der Annahme, dass die Null wahr ist.



(Quelle: researchutopia – Jim Frost, [5])

p-Wert Simulation

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

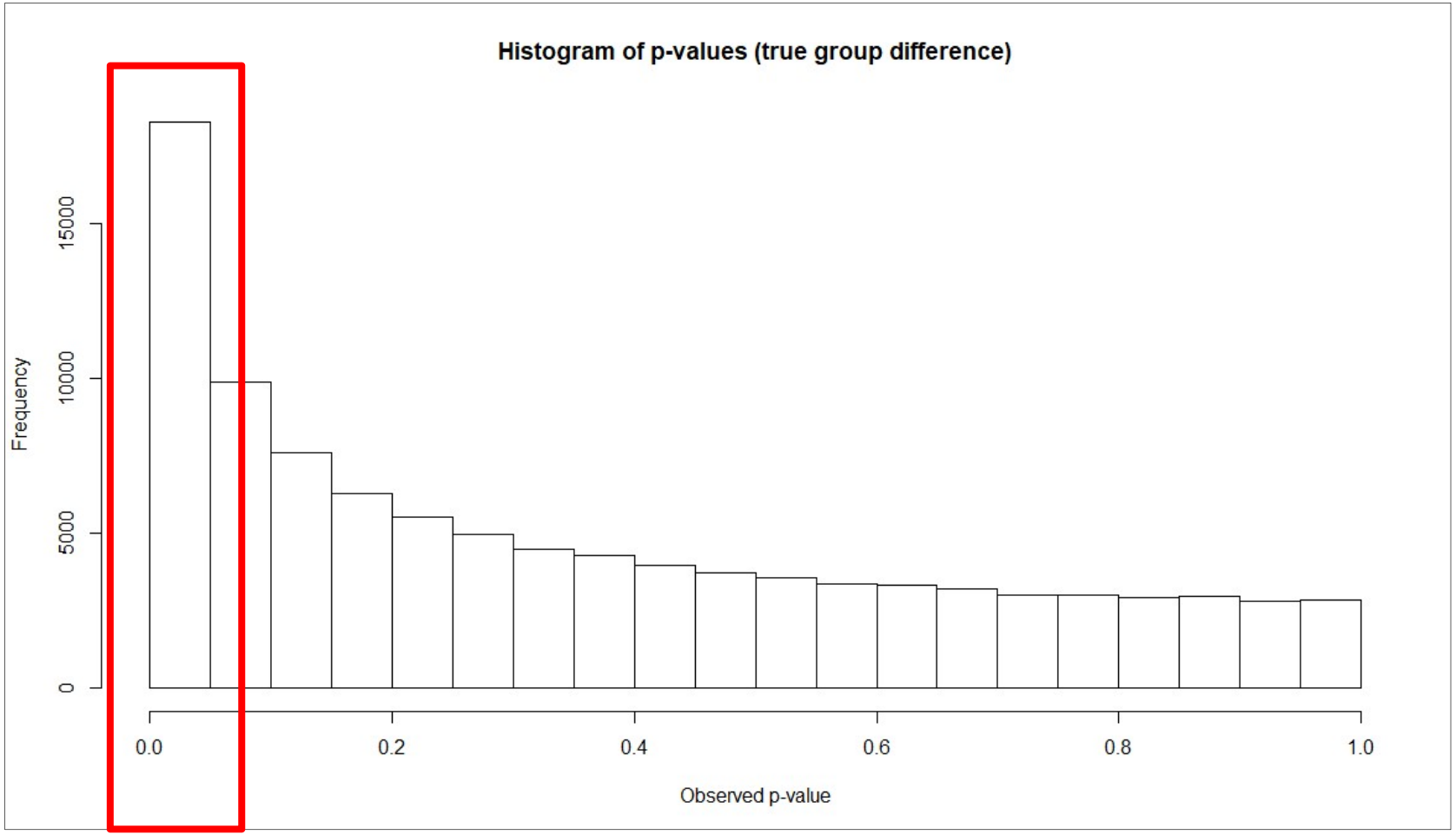
Namenlos - R Editor

```
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 = 103, sd = 20) #produce 100 simulated participants
                                         #with mean=103 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 (true group difference)", xlab=("Observed p-value"))
```

(Quelle: researchutopia – Jim Frost, [5])

p-Wert Simulation

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



(Quelle: researchutopia – Jim Frost, [5])

p-Wert Simulation

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

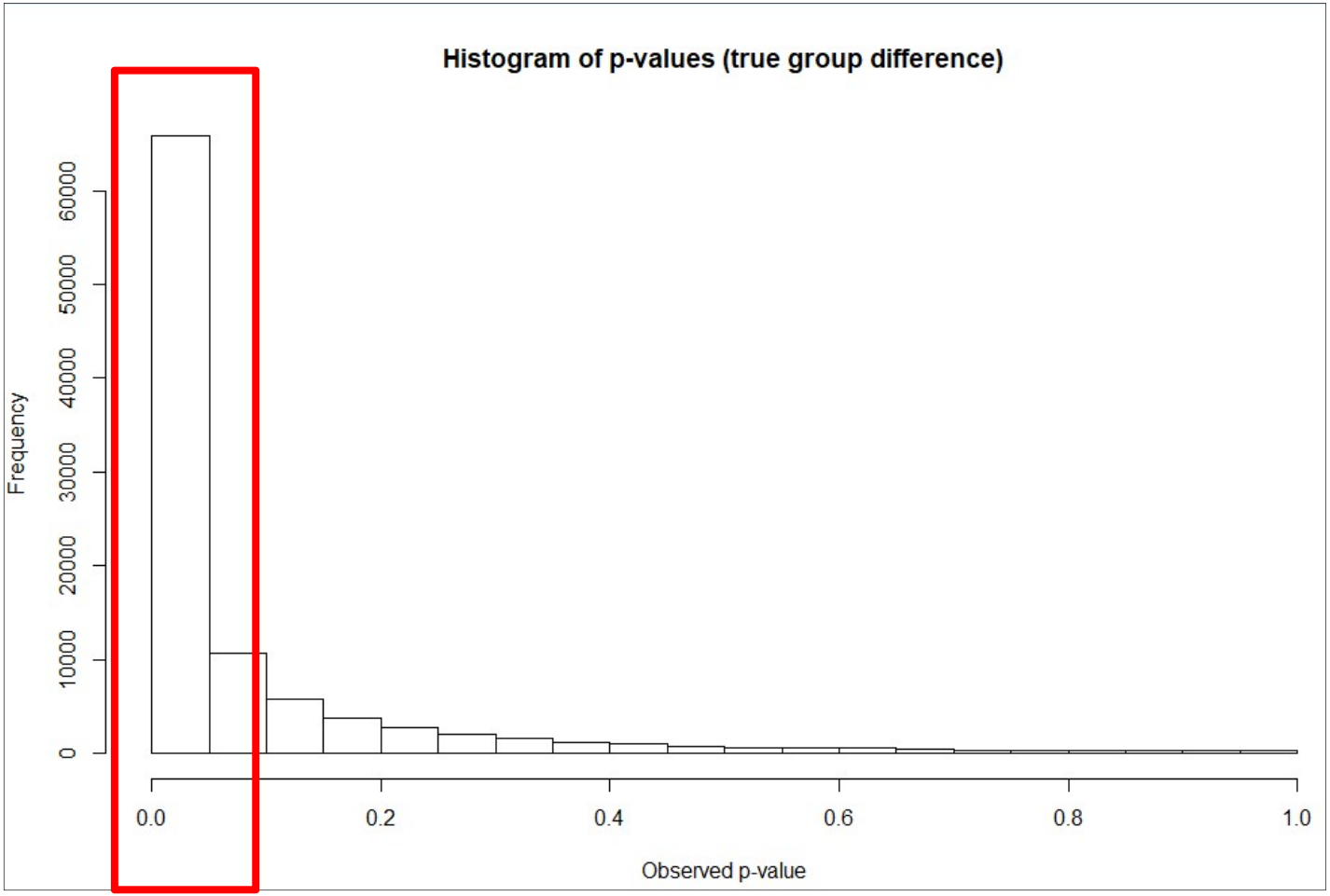
Namenlos - R Editor

```
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"))
```

(Quelle: researchutopia – Jim Frost, [5])

p-Wert Simulation

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



(Quelle: researchutopia – Jim Frost, [5])

p-Wert Simulation

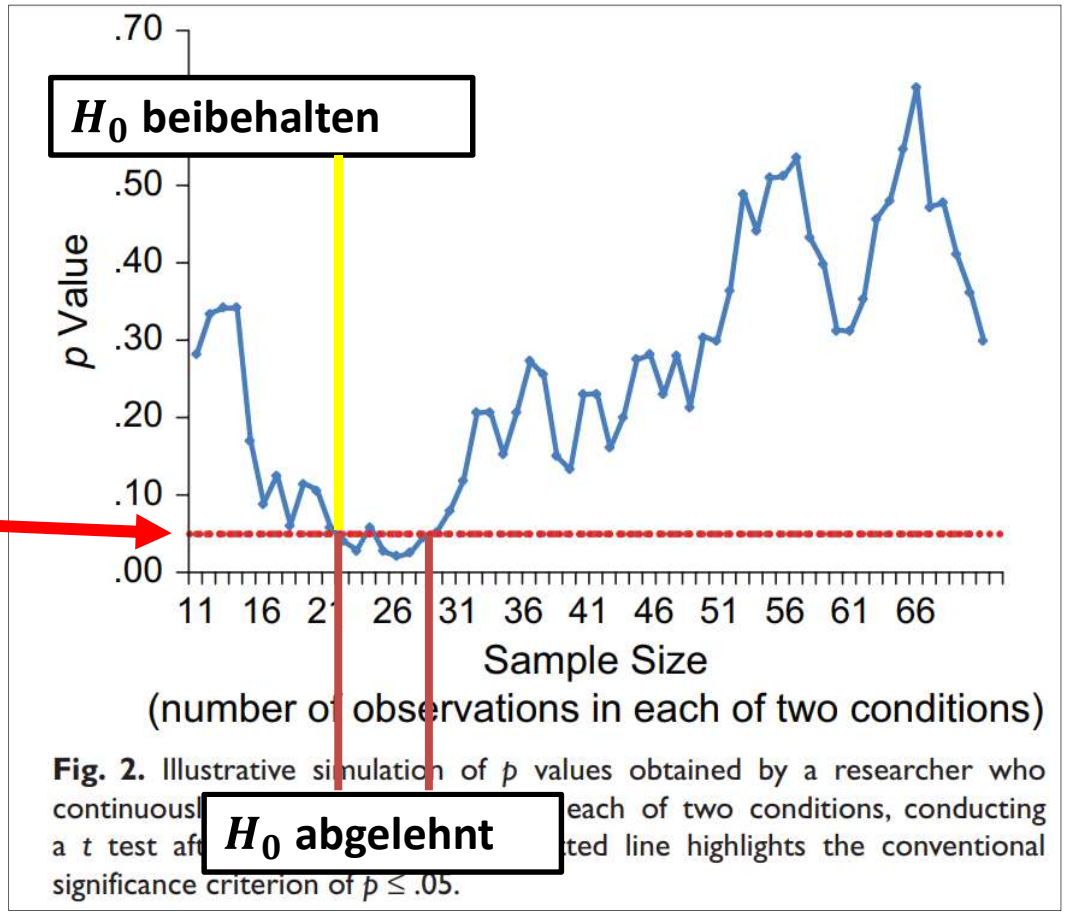
"P-Hacking" hilft bei der Datenmassage

- „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.“

(Quelle: Wenn Forscher durch den Signifikanztest fallen – Regina Nuzzo, [6])

p-Wert Simulation

"P-Hacking" hilft bei der Datenmassage



Signifikanzniveau
α = 0.05

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

Zusammenfassung

P-Wert

- Der exakter P-Wert ist wichtig
- Replikation ist entscheidend
- Fachgebietskenntnisse ist wichtig
- Die Effektgröße spielt eine große Rolle

(Quelle: five-guidelines-for-using-p-values – Jim Frost, [4])

Quellen

- [1] http://blog.minitab.com/blog/adventures-in-statistics-2/five-guidelines-for-using-p-values
- [2] http://eswf.uni-koeln.de/glossar/chivert.htm
- [3] http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests%3A-significance-levels-alpha-and-p-values-in-statistics
- [4] http://blog.minitab.com/blog/adventures-in-statistics-2/p-values-and-the-replication-of-experiments
- [5] https://researchutopia.wordpress.com/2013/11/10/understanding-p-values-via-simulations/
- [6] http://www.spektrum.de/news/statistik-wenn-forscher-durch-den-signifikanztest-fallen/1224727
- False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant, Joseph P. Simmons, Leif D. Nelson and Uri Simonsohn
- Formelsammlung für Wirtschaftswissenschaftler: Mathematik und Statistik, Fred Böker
- A Dirty Dozen: Twelve P-Value Misconceptions, Steven Goodman
- https://www.vox.com/science-and-health/2017/7/31/16021654/p-values-statistical-significance-redefine-0005
- http://blog.minitab.com/blog/adventures-in-statistics-2/not-all-p-values-are-created-equal
- http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124