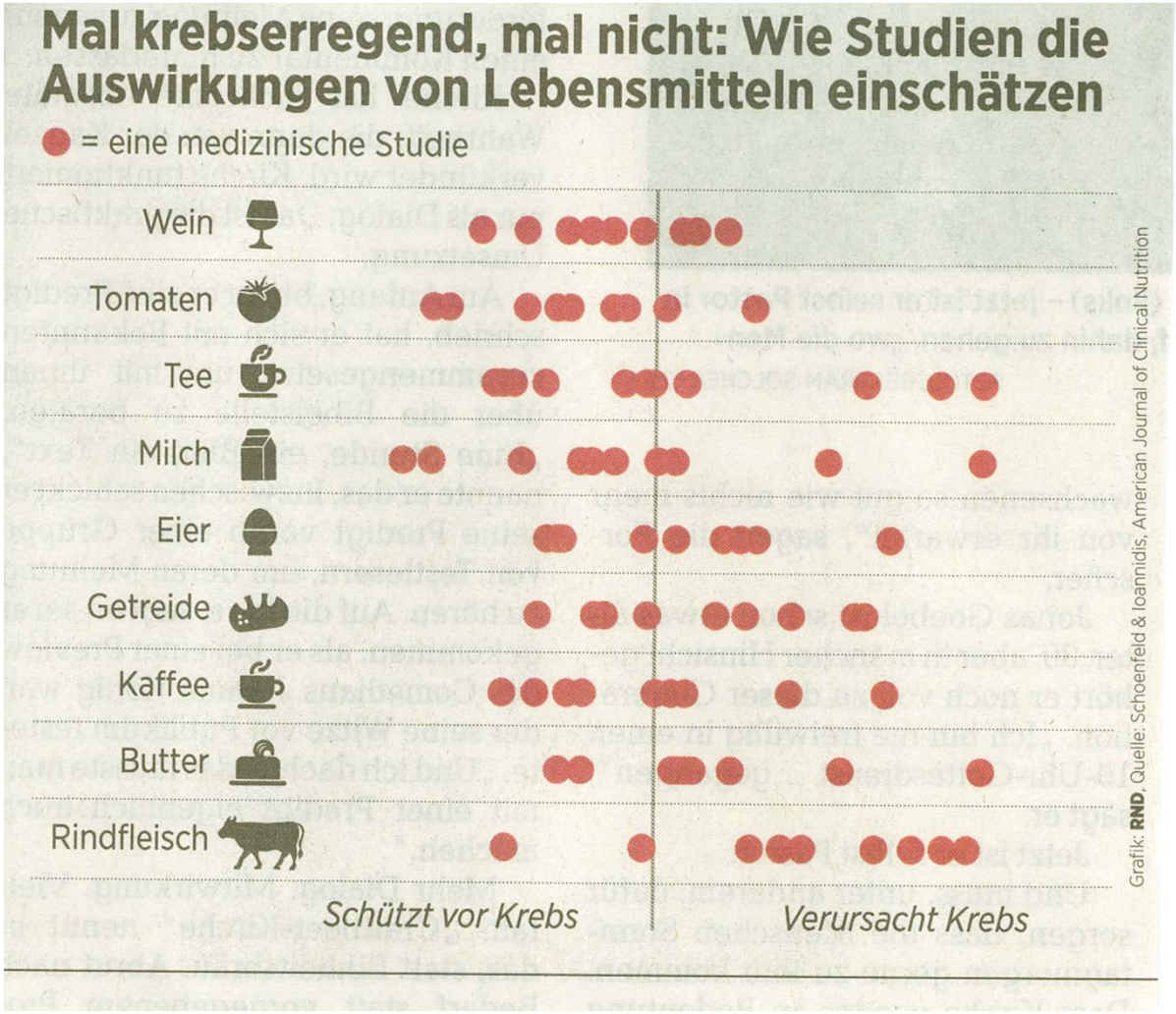
LEIBNIZ-INSTITUT
FÜR NUTZTIERBIOLOGIE

Mitglied der Leibniz Gemeinschaft

- Der P(roblem)-Wert -

Bemerkungen zur Problemdiskussion

OZ-Beilage 20./21. 4. 2019



Jonathan D Schoenfeld and John PA Ioannidis

Is everything we eat associated with cancer? A systematic cookbook review

(Am J Clin Nutr 2013)

Ursachen für derartige Ergebnisse...

1. P-Wertproblematik:
   - Falsch verstandene P-Werte,
   - Fehlinterpretation der Ergebnisse,
   - Jagt nach kleinen P-Werten: P-Hacking...

2. Fehlende Versuchsplanung /
   Ausfallwerte bei vorhandener Versuchsplanung

1. P-Wertproblematik:

   - Falsch verstandene P-Werte,
   - Fehlinterpretation der Ergebnisse,
   - Jagt nach kleinen P-Werten: P-Hacking...

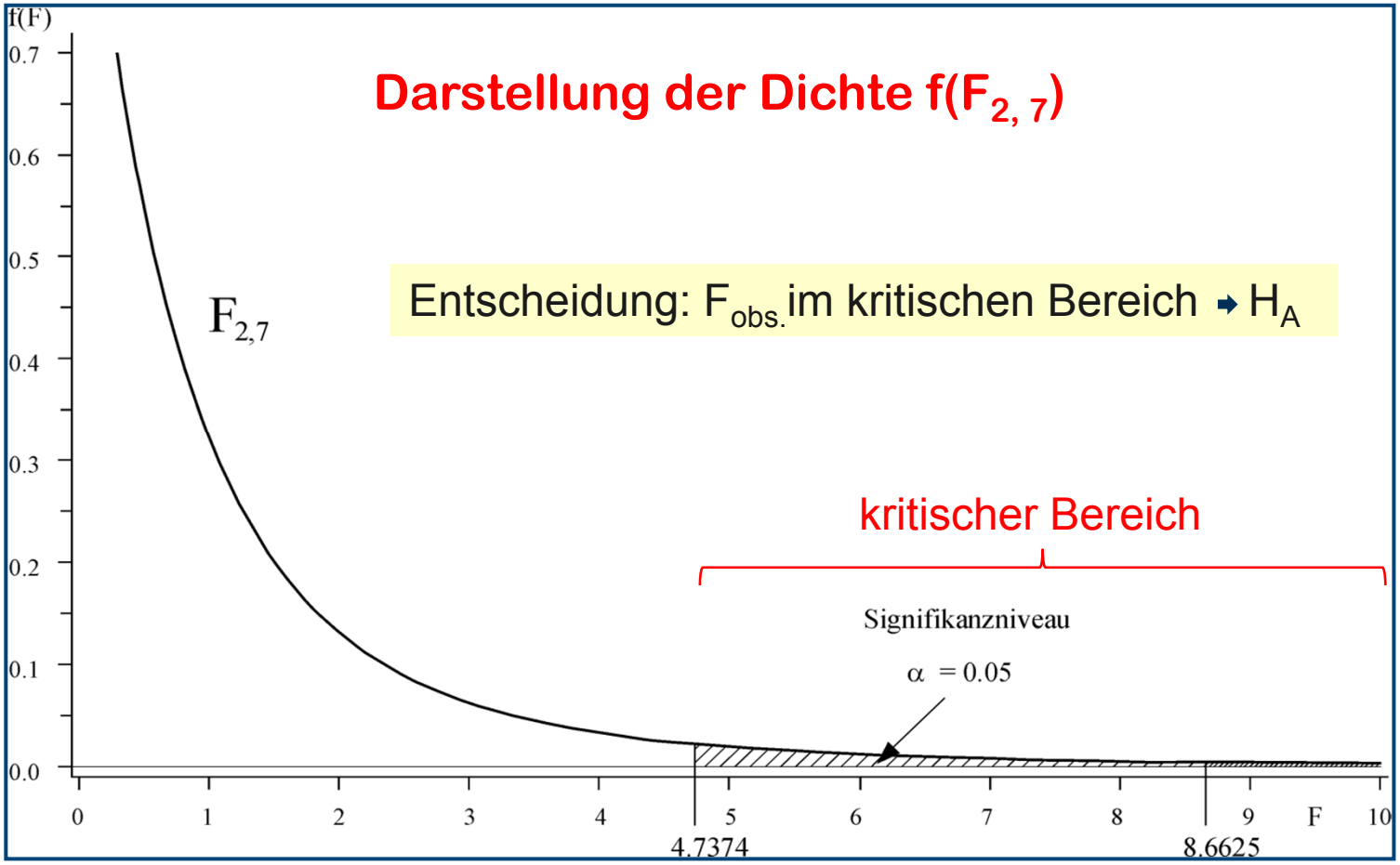
Fehler und ihre Wahrscheinlichkeit beim Test

	Entscheidung für Hypothese	
geltende Hypothese	H_0	H_A
H_0	**$1-\alpha$**	P(Fehler 1. Art)= α **(Risiko 1. Art = Signifikanzlevel)**
H_A	P(Fehler 2. Art)= β **(Risiko 2. Art)**	**$1-\beta$ (Power)**

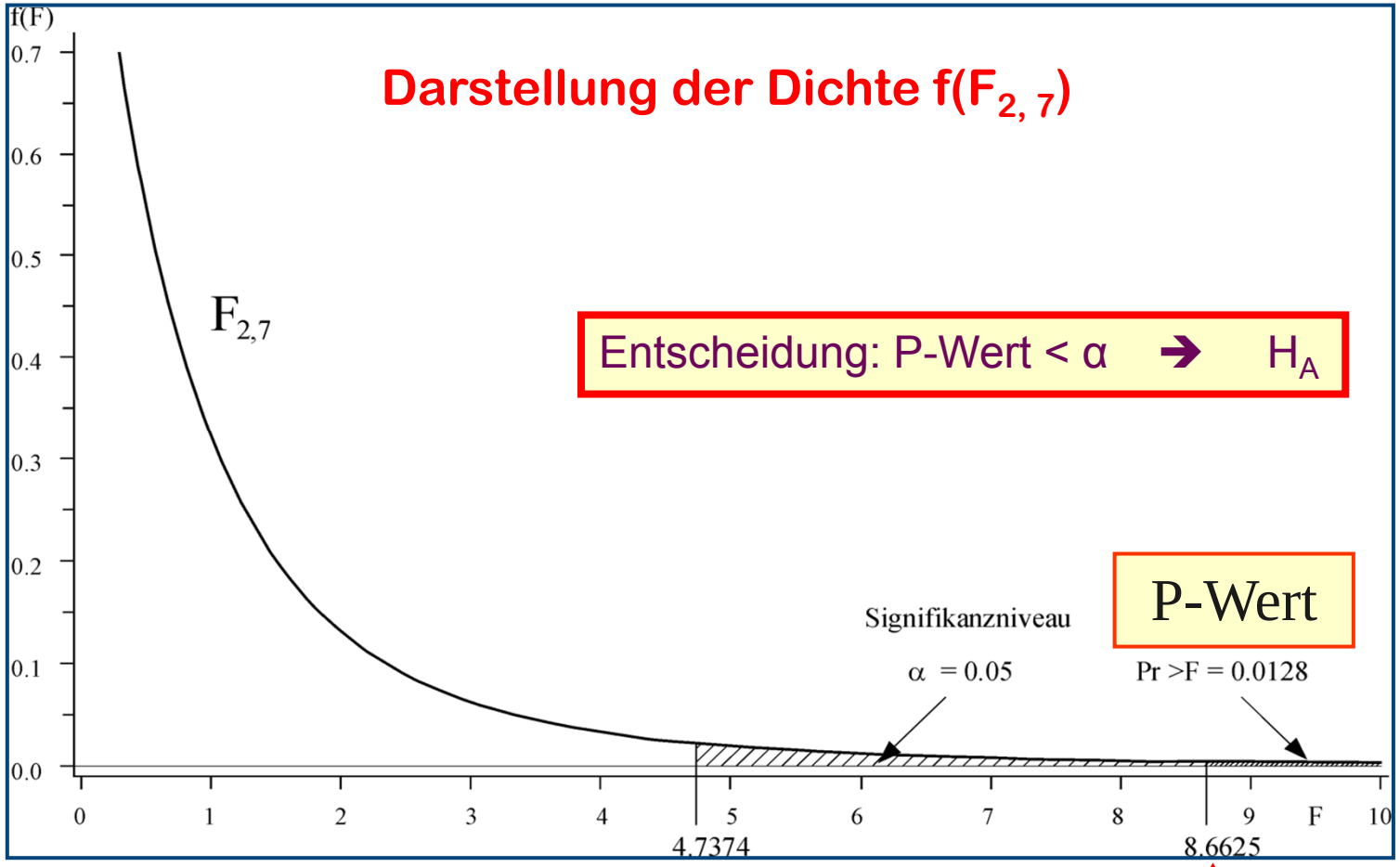
Es handelt sich um bedingte Wahrscheinlichkeiten:

$$\alpha = P(\text{Annahme von } H_A \mid H_0) \qquad \beta = P(\text{Annahme von } H_0 \mid H_A)$$

Die Entscheidung beim Test



Die Entscheidung beim Test
Der P- Wert



Die Entscheidung beim Test
Der P- Wert

Der von R. A. Fisher eingeführte P-Wert enthält keine weiterführende Information. Er ist nur eine "statistische Hilfsgröße".

Nur die Tatsache, ob er kleiner ist als ein vorgegebenes Niveau α, ist von Interesse.

Praxis

Der P-Wert, welcher die statistische Signifikanz induziert, wird sehr häufig fehlinterpretiert und falsch verwendet.

1

Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

Wenn P = 0,05 ist die Chance, dass die Nullhypothese wahr ist, nur 5%.

Falsch!!!

Häufigster und verhängnisvollster Trugschluss über den P-Wert.

(Steven Goodman: A Dirty Dozen: Twelve P-Value Misconceptions. In: Seminars in Hematology. Nr. 45, 2008. S. 135–140)

Der P-Wert wird berechnet unter der Annahme, dass die Nullhypothese wahr ist:

$$\alpha = P\left(\text{Annahme von } H_A \mid H_0\right)$$

2 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

Ein nicht-signifikanter Unterschied bedeutet bei einem Mittelwertvergleich zwischen zwei Gruppen, dass die Mittelwerte gleich sind.

oder

Fehlende Signifikanz heißt kein Effekt.

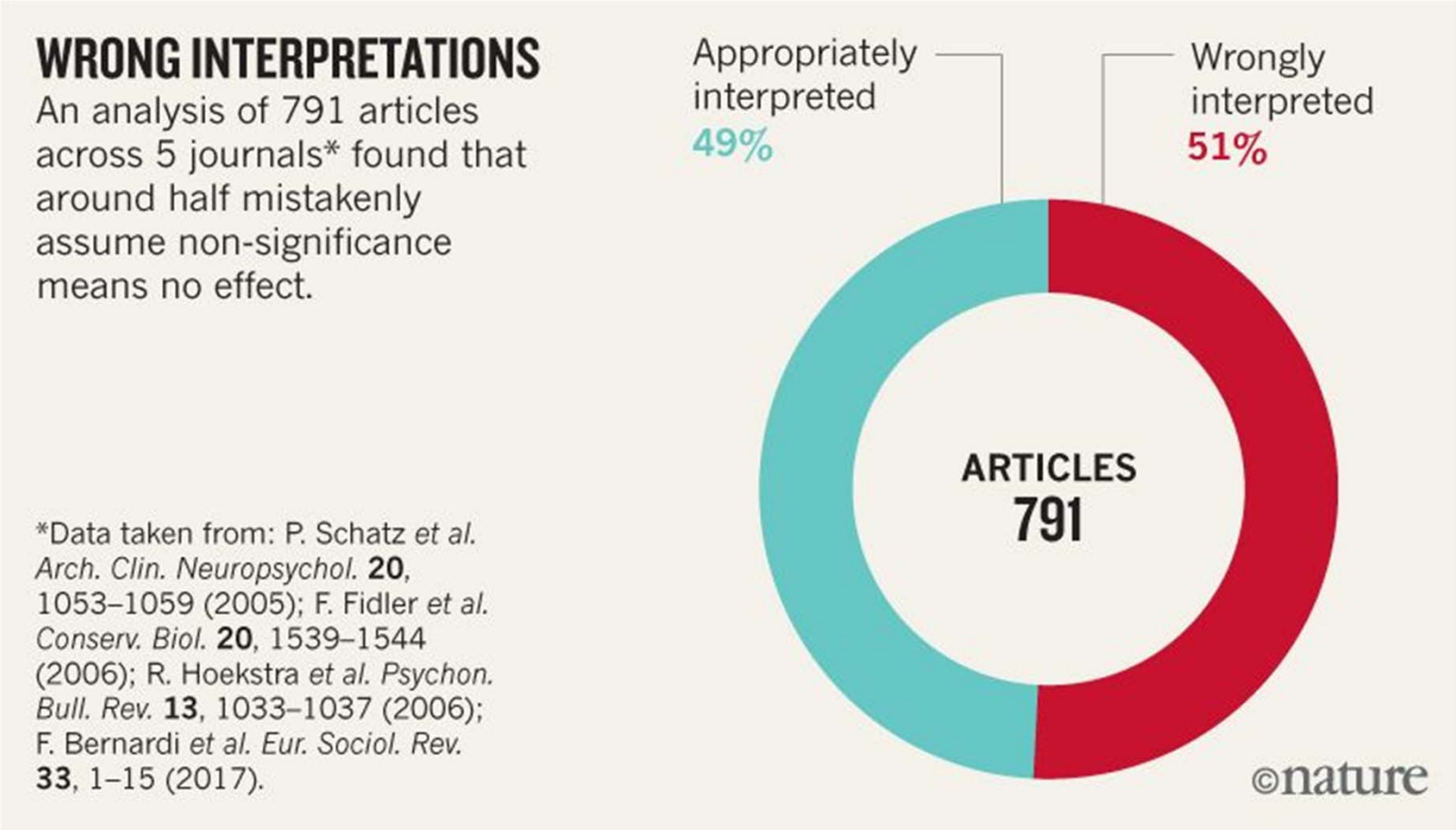
Falsch!!!

Annahme der Nullhypothese aus "Mangel an Beweisen".

Das heißt nicht, dass die Mittelwerte gleich sind bzw. kein Effekt vorhanden ist!

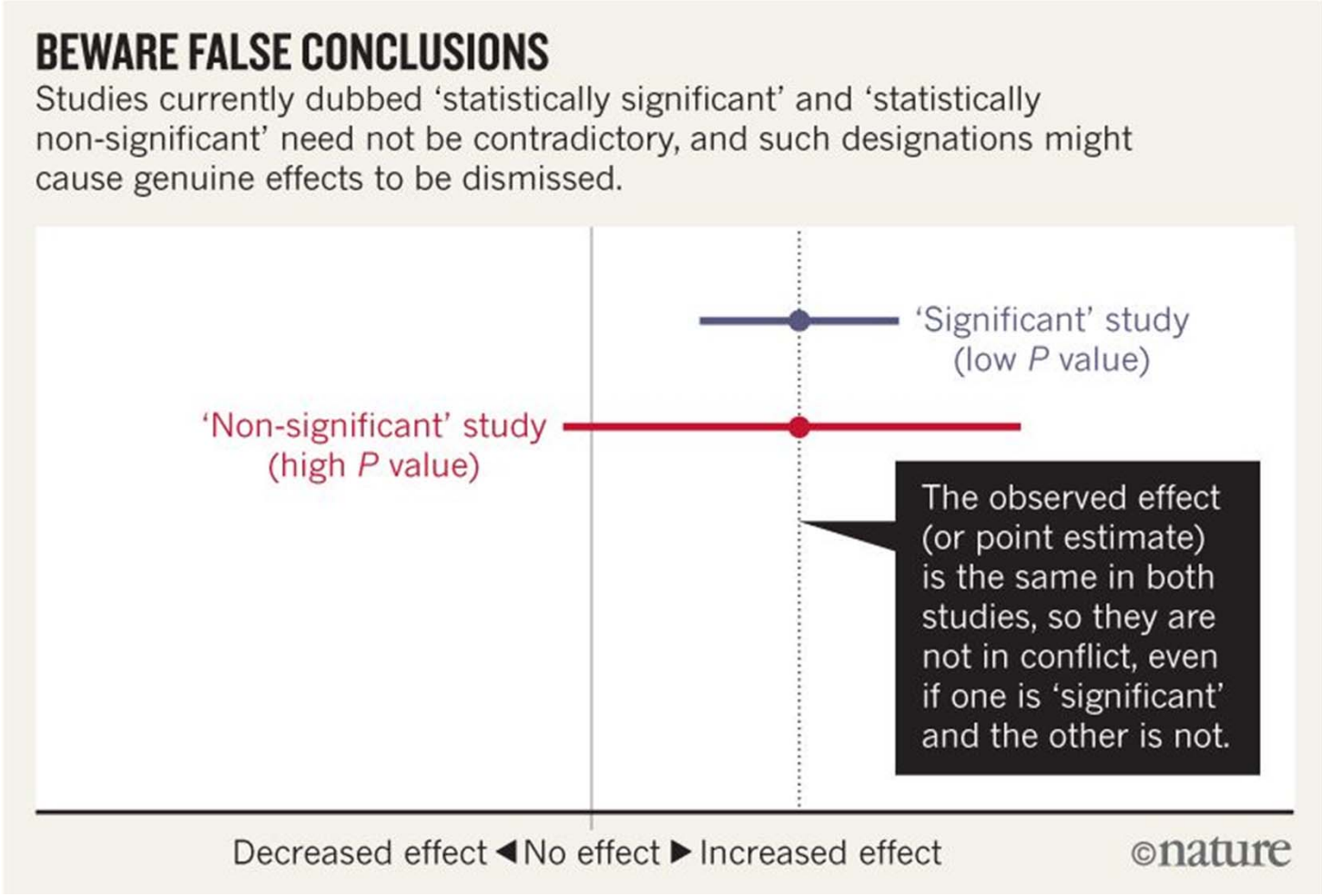
(Äquivalenztests)

2 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:



Source: V. Amrhein *et al.* MARCH 2019

2 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:



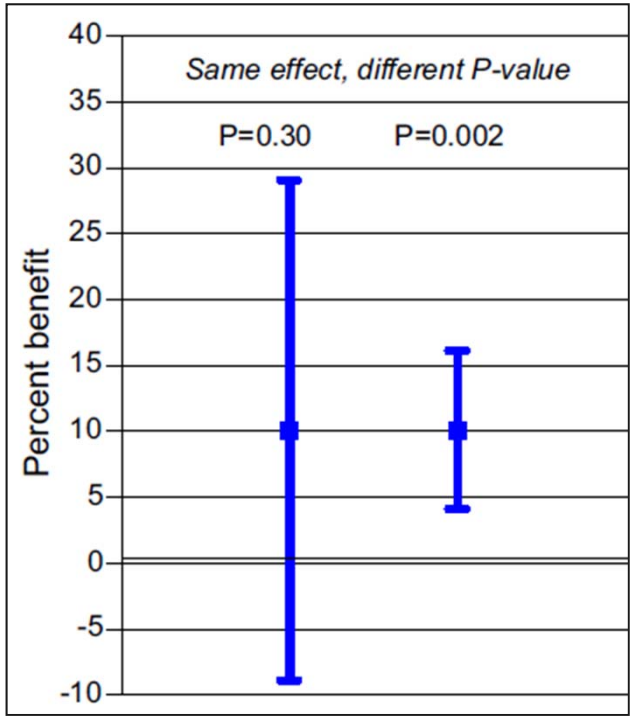
Source: V. Amrhein *et al.* MARCH 2019

3

Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

Nur ein signifikanter Unterschied bedeutet, dass das Ergebnis in der Realität wichtig ist (biologische Relevanz).

Falsch !!!



Der P-Wert gibt keine Information bezüglich der biologischen Relevanz (Größe) eines Effektes oder einer Differenz.

Quelle: S. Goodman (2008)

3 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

Beispiel:
unterschiedliche P-Werte bei gleichen Effekten / Differenzen:

```
proc power;
   TwoSampleMeans test=diff
      Alpha = 0.3 0.002
      Sides = 2
      MeanDiff = 10
      StdDev = 10
      Power = 0.8
      NPerGroup = .
   ;
run;
```

3 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Number of Sides	2
Mean Difference	**10**
Standard Deviation	10
Nominal Power	0.8
Null Difference	0

Computed N Per Group

Index	Alpha	Actual Power	N Per Group
1	**0.300**	0.824	8
2	**0.002**	0.811	34

gleiche Differenz

unterschiedliche P-Werte

Ursache

3 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

Beispiel:
gleiche P-Werte bei unterschiedlichen Differenzen:

```
proc power;
   TwoSampleMeans test=diff
      Alpha = 0.05
      Sides = 2
      MeanDiff = 10 1
      StdDev = 10
      Power = 0.8
      NPerGroup = .
   ;
run;
```

3

Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Number of Sides	2
Alpha	0.05
Standard Deviation	10
Nominal Power	0.8
Null Difference	0

Computed N Per Group

Index	Mean Diff	Actual Power	N Per Group
1	10	0.807	17
2	1	0.800	1571

gleiche P-Werte

unterschiedliche Differenzen

Ursache

4 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

Studien mit P-Werten zu beiden Seiten von 0,05 sind widersprüchlich.

Falsch!!!

Studien können unterschiedliche P-Werte haben bei gleichen Effekten / Differenzen:

```
proc power;
   TwoSampleMeans test=diff
      Alpha = 0.2 0.1 0.05 0.01 0.001
      Sides = 2
      MeanDiff = 0.5
      StdDev = 1
      Power = 0.8
      NPerGroup = .
   ;
run;
```

Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

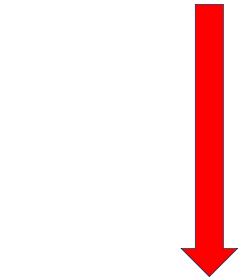
The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Number of Sides	2
Mean Difference	**0.5**
Standard Deviation	1
Nominal Power	0.8
Null Difference	0

Index	Alpha	Actual Power	N Per Group
1	**0.200**	0.805	37
2	**0.100**	0.806	51
3	**0.050**	0.801	64
4	**0.010**	0.805	96
5	**0.001**	0.803	140

gleiche Differenz



unterschiedliche P-Werte

Ursache

5 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

Studien mit dem gleichen P-Wert erbringen den gleichen Nachweis gegenüber der Nullhypothese.

Falsch !!!

Sehr unterschiedliche Effekte / Differenzen können den gleichen P-Wert haben:

```
proc power;
   TwoSampleMeans test=diff
      Alpha = 0.05
      Sides = 2
      MeanDiff = 0.5 1.0 2.0
      StdDev = 1
      Power = 0.8
      NPerGroup = .
   ;
run;
```

5

Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Number of Sides	2
Alpha	**0.05**
Standard Deviation	1
Nominal Power	0.8
Null Difference	0

Index	Mean Diff	Actual Power	N Per Group
1	**0.5**	0.801	64
2	**1.0**	0.807	17
3	**2.0**	0.876	6

gleicher P-Wert

unterschiedliche Diffs

Ursache

6 Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

P-Werte werden zweckmäßigerweise als Ungleichung geschrieben (bei einer Vielzahl von Tests)
z.B. P<0.01 wenn P=0.009

Falsch !!!

Der Schwellenwert für das Risiko erster Art α wird vor dem Versuch festgelegt.

Nach dem Versuch wird der P-Wert berechnet und es wird geprüft, ob P<α ist, unabhängig davon, wie groß P ist.

Da der Schwellenwert für das Risiko erster Art α nicht von den Daten abhängt, sollte man P<0.05 schreiben, falls α=0.05 gewählt wurde.

Berechnete P-Werte direkt angeben.

Aussagen über P-Werte, die ausgesprochen weit verbreitet und dennoch falsch sind:

Bei P-Werten 0,10 > P > 0,05 (α = 0,05) spricht man von
- "borderline significance"
- "tendency"

Falsch !!!

Dieser Unsinn ist häufig sogar im Statistikteil zu finden.

Wenn P>α ist, wird die Nullhypothese (aus Mangel an Beweisen) angenommen.

Derartige Begriffe machen keinen Sinn.

Berechnete P-Werte direkt angeben.

Prinzipien für die Verwendung und Interpretation von P-Werten

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

6 ASA-Prinzipien für die Verwendung und Interpretation von P-Werten

① P-values can indicate how incompatible the data are with a specified statistical model.

② P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

③ Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

④ Proper inference requires full reporting and transparency.

⑤ A P-value, or statistical significance, does not measure the size of an effect or the importance of a result.

⑥ By itself, a P-value does not provide a good measure of evidence regarding a model or hypothesis.

6 ASA-Prinzipien für die Verwendung und Interpretation von P-Werten

③ Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

Ableitung von wissenschaftlichen Schlussfolgerungen nur unter Einbeziehung von:

- Versuchsplanung (Umfang und Design der Studie, Power)
- Qualität der Messungen / Datenerhebung
- "Datenbearbeitung"
- Validität der Voraussetzungen der statistischen Verfahren
- externem Wissen zum untersuchten Phänomen (Literatur, biologischer Sachverstand)

6 ASA-Prinzipien für die Verwendung und Interpretation von P-Werten

④ Proper inference requires full reporting and transparency.

Vollständige Darstellung der Datengewinnung, "Datenbearbeitung", aller statistischen Analysen und aller P-Werte!

Keine selektive Darstellung von P-werten, denn Rosinenpickerei führt zu:

- data dredging
- significance chasing
- significance questing
- selective inference
- P-hacking

2. Fehlende Versuchsplanung /

Ausfallwerte bei vorhandener Versuchsplanung

Fehlende Versuchsplanung

Beispiel: Vergleich von 3 Gruppen (Gruppenumfänge willkürlich festgelegt)

$n_1 = 10$, $n_2 = 10$, $n_3 = 10$
$\alpha = 0{,}05$

$s = 10$
Gruppenmittelwerte: 0, 7, 10

Signifikantes Ergebnis beim ANOVA-F-Test.

Frage nach der	Güte des Test Power Teststärke Trennschärfe	1-β →	Powerberechnung z.B. mit SAS: proc power proc glmpower

Fehlende Versuchsplanung

Powerberechnung:

```
proc power;
   OneWayANOVA
      Alpha = 0.05
      GroupMeans = (0 7 10)
                   (0 0 10)   /* günstigster Fall */
                   (0 5 10)   /* ungünstigster Fall */
      StdDev = 10
      Power = .
      GroupNs = (10 10 10)     /* N = 30 */
   ;
run;
```

Fehlende Versuchsplanung

The POWER Procedure
Overall F Test for One-Way ANOVA

Fixed Scenario Elements

Method	Exact
Alpha	0.05
Standard Deviation	10
Group Sample Sizes	**10 10 10**

Computed Power

Index	-----Means-----			**Power**
1	**0**	**7**	**10**	**0.479**
2	0	0	10	0.581
3	0	5	10	0.458

günstigster Fall
ungünstigster Fall

Fehlende Versuchsplanung

Powerberechnung (Effekt der Unbalanciertheit der Gruppenumfänge):

```
proc power;
   OneWayANOVA
      Alpha = 0.05
      GroupMeans = (0 5 10)  /* ungünstigster Fall */
      StdDev = 10
      Power = .
      GroupNs = (21 21 21)   /* N = 63 optimaler Fall */
                (15 21 27)   /* N = 63 unbalancierter Fall */
                (15 33 15)   /* N = 63 unbalancierter Fall */
   ;
run;
```

Fehlende Versuchsplanung

The POWER Procedure
Overall F Test for One-Way ANOVA

Fixed Scenario Elements

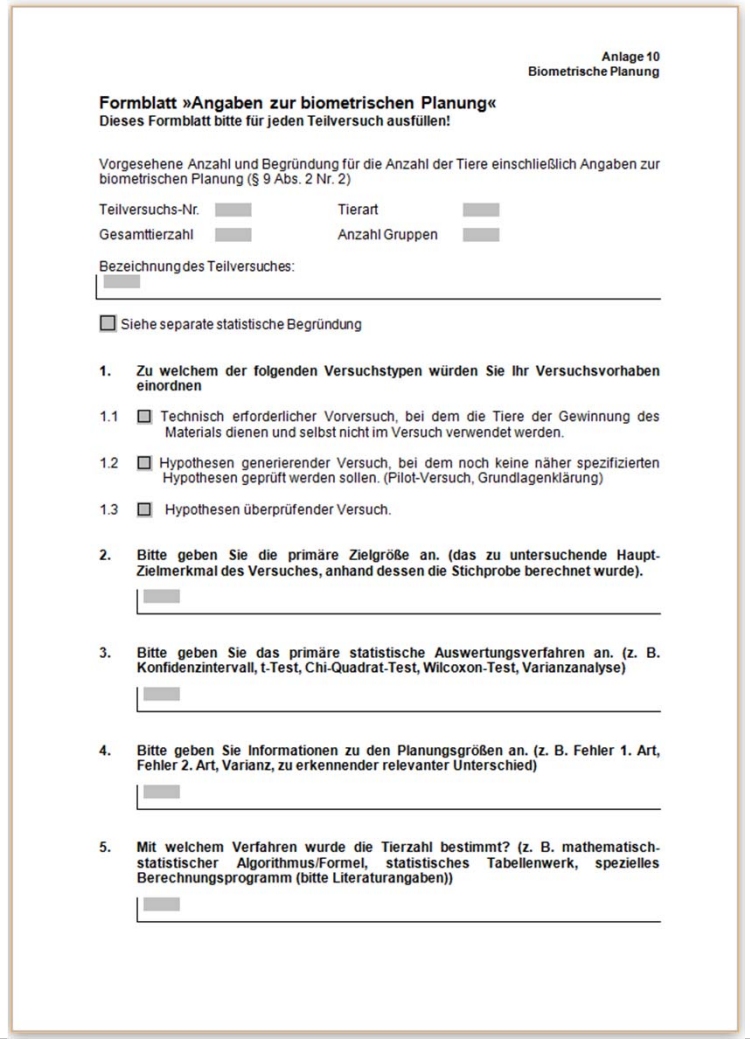
Method	Exact
Alpha	0.05
Group Means	0 5 10
Standard Deviation	10

Computed Power

Index	Group N's			Power
1	21	21	21	**0.815**
2	15	21	27	**0.791**
3	15	33	15	**0.664**

Unbalanciertheit

Versuchsplanung



Anlage 10
Biometrische Planung

Formblatt »Angaben zur biometrischen Planung«
Dieses Formblatt bitte für jeden Teilversuch ausfüllen!

Vorgesehene Anzahl und Begründung für die Anzahl der Tiere einschließlich Angaben zur biometrischen Planung (§ 9 Abs. 2 Nr. 2)

Teilversuchs-Nr. Tierart
Gesamttierzahl Anzahl Gruppen

Bezeichnung des Teilversuches:

☐ Siehe separate statistische Begründung

1. **Zu welchem der folgenden Versuchstypen würden Sie Ihr Versuchsvorhaben einordnen**

1.1 ☐ Technisch erforderlicher Vorversuch, bei dem die Tiere der Gewinnung des Materials dienen und selbst nicht im Versuch verwendet werden.

1.2 ☐ Hypothesen generierender Versuch, bei dem noch keine näher spezifizierten Hypothesen geprüft werden sollen. (Pilot-Versuch, Grundlagenklärung)

1.3 ☐ Hypothesen überprüfender Versuch.

2. **Bitte geben Sie die primäre Zielgröße an. (das zu untersuchende Haupt-Zielmerkmal des Versuches, anhand dessen die Stichprobe berechnet wurde).**

3. **Bitte geben Sie das primäre statistische Auswertungsverfahren an. (z. B. Konfidenzintervall, t-Test, Chi-Quadrat-Test, Wilcoxon-Test, Varianzanalyse)**

4. **Bitte geben Sie Informationen zu den Planungsgrößen an. (z. B. Fehler 1. Art, Fehler 2. Art, Varianz, zu erkennender relevanter Unterschied)**

5. **Mit welchem Verfahren wurde die Tierzahl bestimmt? (z. B. mathematisch-statistischer Algorithmus/Formel, statistisches Tabellenwerk, spezielles Berechnungsprogramm (bitte Literaturangaben))**

Statistische Versuchsplanung

ist Bestandteil

eines Tierversuchsantrags /

einer Tierversuchsanzeige!

Risiko erster Art
Risiko zweiter Art
Varianz
biologisch relevante Mindestdifferenz

Stichprobenumfang

(Ausfallwerte?)

Versuchsplanung - Ausfallwerte

Mögliche Ausfallquote bei der Versuchsplanung berücksichtigen.

Wurde die Ausfallquote bei der Versuchsplanung nicht berücksichtigt,
➡ Powerberechnung empfehlenswert :

```
proc power;
   OneWayANOVA
      Alpha = 0.05
      GroupMeans = (0 5 10)     /* ungünstigster Fall */
      StdDev = 10
      Power = .
      GroupNs = (21 21 21)      /* N = 63 optimaler Plan */
                (18 18 18)      /* N = 54  */
                (15 21 15)      /* N = 51  */
   ;
run;
```

Versuchsplanung - Ausfallwerte

The POWER Procedure
Overall F Test for One-Way ANOVA

Fixed Scenario Elements

Method	Exact
Alpha	0.05
Group Means	0 5 10
Standard Deviation	10

Computed Power

Index	Group N's			Power
1	21	21	21	0.815
2	18	18	18	**0.744**
3	15	21	15	**0.658**

N=63 nach Plan

N=54 Ausfallwerte

N=51 Ausfallwerte

Neue Zürcher Zeitung

Servan Grüninger, 25.3.2016

Der P(roblem)-Wert

Der P-Wert: Alle kennen ihn,
die meisten missbrauchen ihn,
kaum einer versteht ihn.

Statistiker fordern nun ein Umdenken.

Denn:

Forschende ohne Statistikkenntnisse schaden der Wissenschaft.

LEIBNIZ-INSTITUT
FÜR NUTZTIERBIOLOGIE

Leibniz-Institut für Nutztierbiologie FBN

Wilhelm-Stahl-Allee 2
18196 Dummerstorf

Kontakt

Dr. Armin Tuchscherer

Telefon: +49 38208 68 904
Telefax: +49 38208 68 902
E-Mail: atuchsch@fbn-dummerstorf.de
Internet: www.fbn-dummerstorf.de