Irrtümer durch Signifikanzstatistik

- Eltern von Töchtern haben ein höheres Scheidungsrisiko als Eltern von Söhnen (Morgan et al. 1988)
- Do pretty women inspire men to discount the future? (Wilson and Daley 2003)
- „Beautiful parents have more daughters than ugly parents…“ (Kanazawa 2006).
- Im Dezember geborene Kinder haben eine höhere Wahrscheinlichkeit als im Juni geborene Kinder, den 105. Geburtstag zu erreichen (Scholz, Doblhammer und Maier 2005).
- Männer untertreiben ihr Körpergewicht bei Anwesenheit eines Interviewers im Mittel um ein Kilo (Kroh 2005).
- Curry steigert die kognitive Leistung des Gehirns bei älteren Menschen (Tze-Pin Ng et al. 2006).
- Linkshändigkeit hat bei Männern mit höherer Bildung einen signifikanten Effekt auf das Lohnniveau (Ruebeck et al . 2006, NBER).
- Die durchschnittliche Körpergrösse von Mikroökonomen ist geringer als die von Makroökonomen (forthcoming Diekmann et al. 2007, Journal of Irreproducable Results)

t-Test für zwei Stichproben ($n_1 = n_2 = 20$) mit Nullhypothese $\mu_1 - \mu_2 = 0$. Die Teststatistik ist für $\alpha = 0.01$ signifikant.

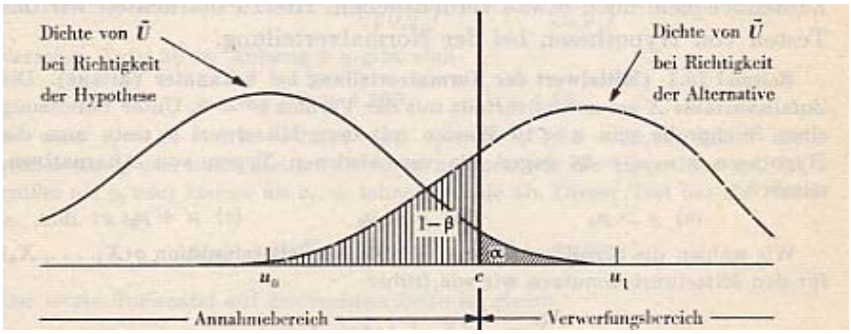
1. **Die Hypothese, es gäbe keine Unterschiede, ist falsch.**
2. **Die Wahrscheinlichkeit, dass die Nullhypothese zutrifft, liegt unter 1 Prozent.**
3. **Die Vermutung, es könnte Unterschiede geben, ist richtig.**
4. **Man kann immerhin die Wahrscheinlichkeit dafür angeben, dass diese Vermutung richtig ist.**
5. **Die Wahrscheinlichkeit, bei Ablehnung der Nullhypothese einen Fehler zu machen, ist kleiner als 1 Prozent.**
6. **Wenn das Experiment sehr oft wiederholt würde, käme in 99 Prozent der Fälle eine signifikante Prüfgröße zustande.**

Haller und Kraus 2002 nach: Krämer, W., 2006, Statistik. Vorm Geburtshelfer zum Bremser in den Sozialwissenschaften? In A. Diekmann, Hrsg., Methoden der Sozialforschung, Wiesbaden.

1. **Die Hypothese, es gäbe keine Unterschiede, ist falsch.**
2. **Die Wahrscheinlichkeit, dass die Nullhypothese zutrifft, liegt unter 1 Prozent.**
3. **Die Vermutung, es könnte Unterschiede geben, ist richtig.**
4. **Man kann immerhin die Wahrscheinlichkeit dafür angeben, dass diese Vermutung richtig ist.**
5. **Die Wahrscheinlichkeit, bei Ablehnung der Nullhypothese einen Fehler zu machen, ist kleiner als 1 Prozent.**
6. **Wenn das Experiment sehr oft wiederholt würde, käme in 99 Prozent der Fälle eine signifikante Prüfgröße zustande.**

Alle sechs Behauptungen sind falsch!

Keiner der befragten Psychologie-Studierenden, 20% ihrer Statistik-Dozenten (incl. studentische Hilfskräfte) und 10 Prozent anderer Wissenschaftler haben das erkannt (n = 44; 30; 39).



β-Fehler α-Fehler

Kreyszig 1965: 209

- α = Fehler I. Art bei statistischen Signifikanztests. Ist die Nullhypothese zutreffend, gibt α die Wahrscheinlichkeit dafür an, dass die Nullhypothese irrtümlich abgelehnt und die Alternativhypothese angenommen wird. Häufig wird α = 0, 05 festgelegt.
- β = Fehler II. Art. Die Wahrscheinlichkeit, dass die Nullhypothese irrtümlich angenommen und die Alternativhypothese irrtümlich abgelehnt wird.
- β hängt ab: 1. von α, 2. von der Fallzahl n, 3. von der Stärke eines Effekts, 4. von der Varianz, 5. von der Macht (power) eines statistischen Tests.

	H_A = Alternativhypothese trifft zu	H_0 = Nullhypothese trifft zu
Entscheidung für H_A	1 - β	α = Fehler I. Art (falsch positiv)
Entscheidung für H_0	β = Fehler II. Art (falsch negativ)	1 - α

	Angeklagter schuldig	Angeklagter unschuldig
Verurteilung	1 - β	α = Fehler I. Art (falsch positiv)
Freispruch	β = Fehler II. Art (falsch negativ)	1 - α

Der β-Fehler hängt – ceteris paribus – ab von:

1. α
2. der Stärke eines Effekts
3. der Fallzahl
4. der Macht (power) eines statistischen Tests.
5. dem Standardfehler der Prüfgröße

John P. A. Ioannidis, 2005. Why most published research findings are false, PLoS Medicine 2. 696-701.

- a = Anzahl wahre Hypothesen
- b = Anzahl falscher Hypothesen
- R = a/b (R ist eine Eigenschaft des Untersuchungsbereichs und dessen Kenntnis)
- c = Anzahl der geprüften Hypothesen

A-Priori-Wahrscheinlichkeit, dass eine Hypothese wahr ist.

R/(R + 1) = a/(a + b) = Anzahl wahrer Hypothesen/Anzahl aller Hypothesen = Apriori-Wahrscheinlichkeit, dass eine Hypothese wahr ist.

1/(R + 1) = A-Priori-Wahrscheinlichkeit, dass eine Hypothese falsch ist.

- **A-Posteriori-Wahrscheinlichkeit, dass die empirisch geprüfte Hypothese wahr ist.**
- Wenn eine Hypothese nach empirischer Prüfung als wahr angenommen wurde (ein Ergebnis signifikant ist): Wie groß ist die Wahrscheinlichkeit, dass sie tatsächlich wahr ist?
- PPV = Positive predictive value = A-Posteriori-Wahrscheinlichkeit, dass die Hypothese wahr ist, gegeben die empirische Evidenz.

Research finding	True Relation: Yes	True Relation: No
Yes	(i) $c(1 - \beta)R/(R+1)$	(ii) $c\alpha/(R+1)$
No	(iii) $c\beta R/(R+1)$	(iv) $c(1- \alpha)/(R+1)$

(1)PPV = (i)/((i) + (ii)) = $(1 - \beta)\ R/(R - \beta R + \alpha)$

(2) PPV = $R/[R + (\alpha/(1 - \beta))]$

PPV = $R/[R + (\alpha/(1 - \beta))]$

Wovon hängt der PPV ab?

- wächst mit R,
- wächst mit $(1 - \beta)$, d.h. mit der Effektstärke, der Fallzahl und der Macht des Tests,
- sinkt mit α (und β).

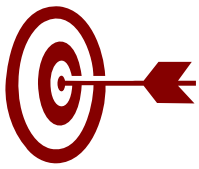
Wann sind mehr Resultate falsch als wahr?

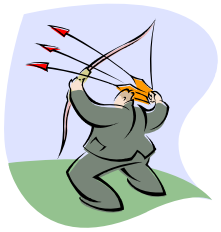
- Signifikante Ergebnisse werden in der Literatur berichtet. Es sind mehr Resultate falsch als wahr, wenn PPV < 0,5.
- Aus (1) $PPV = (1 - \beta) R/(R - \beta R + \alpha) < 0,5$ folgt:

  $(1 - \beta) R < \alpha$

  Mit $\alpha = 0,05$:

  $(1 - \beta) R < 0,05$
- Auf Gebieten mit einer geringen A-Priori-Wahrscheinlichkeit der Wahrheit von Hypothesen werden selbst bei kleinem β mehr falsche als wahre Hypothesen berichtet.





1. Pfeil abschießen. 2. Zielscheibe aufhängen

Besser umgekehrt: Hypothesen Design, Analysemethoden und Erfolgsindikatoren vorher registrieren!

Hinzu kommt aber:

- In vielen Studien ist β relativ hoch (kleine Fallzahlen, geringe Effekte).
- Bias durch selektive Berichterstattung (z.B. nachträgliche Auswahl signifikanter Ergebnisse).
- Tests durch mehrere unabhängige Teams. Nur signifikante Ergebnisse werden berichtet.

  „The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true" (Ioannidis 2005).

- Deshalb neuerdings Register zur Anmeldung von Studien: International Standard Randomised Controlled Trial Number (www.isrctn.com) in U.K. und National Institutes of Health in den USA (www.clinicaltrials.gov). Nur die Ergebnisse angemeldeter Studien werden in Fachzeitschriften publiziert.

Wissenschaftstheoretische Konsequenzen

„Wissenschaftstheoretische“ Konsequenzen

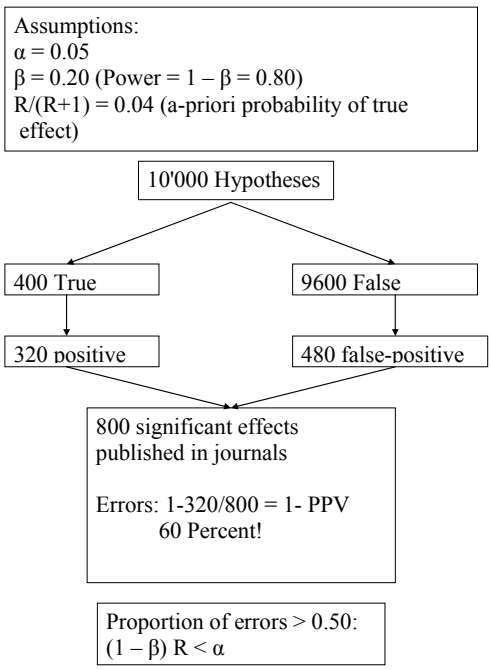
$$PPV = R/[R + (\alpha/(1 - \beta))]$$

- Wenn der Fehler für falsch-positive Resultate α gegen null geht, geht PPV gegen eins (deterministische Zusammenhänge, keine Messfehler).
- Werden die zu prüfenden Hypothesen aus einer empirisch gut geprüften Theorie abgeleitet, dann hat R hohe Werte im Vergleich zu Gebieten, die wenig erforscht sind. Die Theorie steuert die Auswahl der Hypothesen, denen dann eine hohe A-Priori-Wahrscheinlichkeit zukommt.

Was tun?

- **Verringerung von Irrtümern durch Replikationen**
- P(„no“| Hypothese falsch) = (1 – α)/(R + 1).
- Besonders bei kleinem R ist die Wahrscheinlichkeit hoch, dass der Irrtum bei einer Replikation entdeckt wird.

Ein Rechenbeispiel



Assumptions

$\alpha = 0.05$
$\beta = 0.20$ (Power = $1 - \beta = 0.80$)
R/(R+1) = 0.04 (a-priori probability of true effect)

800 significant = 320 true + 480 false positive, 60 percent errors

Replication with new data:
320 x 0.80 = 256 significant and true
480 x 0.05 = 24 significant, false positive
24/280; 8.6 percent errors

Replication reduces errors from 60 to 8.6 percent !

Nur ein Beispiel

Taijfel-Hypothese über Gruppenidentifiktion

- “Geburtstagsexperiment” mit Vertrauensspiel
- Und mit Gefangenendilemma

Sieht nicht gut aus für unser Antidepressivum.