Nullhypothesis Significance Testing (NHST), Effektgrößen, Deskriptive Verteilungsfunktionen, Bayes-Faktoren

Ingo Klein
Würzburg, 03.07.2012
Lehrstuhl für Statistik und Ökonometrie

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

RECHTS- UND WIRTSCHAFTS-
WISSENSCHAFTLICHE FAKULTÄT

Agenda

Beispielhafte Fragestellung

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

RECHTS- UND WIRTSCHAFTS-
WISSENSCHAFTLICHE FAKULTÄT

Erhebung: Selbsteinschätzung von Mathematik- (U) und Computerkenntnissen (V) I

- Befragung aller Studierenden des Kurses Statistik in der ersten Vorlesungsstunde des Wintersemesters 2002/03.
- Gefragt wurde unter anderem nach:
  - Abiturnote
  - Selbtseinschätzung Kenntnisse Mathematik
  - Selbsteinschätzung Kenntnisse IT
  - Selbsteinschätzung Kenntnisse Programmierung
  - Alter
  - Berufserfahrung
  - Studienfach
  - Geschlecht

Erhebung: Selbsteinschätzung von Mathematik- (U) und Computerkenntnissen (V) II

- Totalerhebung ohne nennenswerten Datenausfall.
- Deskriptive Auswertung des Zusammenhangs zwischen der Selbsteinschätzung der Kenntnisse in Mathematik (stat[,4]) und IT (stat[,5])
  - Bravais-Pearson-Korrelationskoeffizient: 0.108.
  - Goodman-Kruskals γ: 0.161.
- Frage: Wie stark ist der Zusammenhang?
- Lösung 1: Liegt relativ dicht bei 0, deshalb nur schwacher (linearer bzw. monotoner) Zusammenhang.

Klassifikation nach Schlittgen

Quelle: Schlittgen, R. (2000). Einführung in die Statistik, München, S. 179:

$\lvert r \rvert$			Interpretation
	0		keine Korrelation
0	-	0.5	schwache Korrelation
0.5	-	0.8	mittlere Korrelation
0.8	-	1	starke Korrelation
	1		perfekte Korrelation

Problem: Woher kommt Tabelle?

Erhebung: Selbsteinschätzung von Mathematik- (U) und Computerkenntnissen (V) IV

- Lösung 2: Man ignoriert, dass Daten nicht Ergebnis einer Zufallsstichprobe sind und führt Signifikanztest durch.
- Test mittels Bravais-Pearson Korrelationskoeffizient:

```
> cor.test(stat[,4],stat[,5],alternative="greater")

        Pearson's product-moment correlation

data:  stat[, 4] and stat[, 5] t = 3.6294, df = 1114, p-value =
0.0001485 alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.05915486 1.00000000
sample estimates:
      cor
0.1081036
```

Erhebung: Selbsteinschätzung von Mathematik- (U) und Computerkenntnissen (V) V

- Test mittels γ-Maß von Goodman & Kruskal:

```
>library(vcdExtra)
>GKgamma(fbeob)
gamma        : 0.122

std. error   : 0.032

CI           : 0.059 0.185

>1-pnorm(0.122/0.032)
[1] 6.878411e-05
```

- Fazit: p-Werte zeigen einen hochsignifikanten (linearen bzw. monotonen) Zusammenhang an.
- Frage 1: Ist ein Signifikanztest für Beobachtungsdaten aus einer Totalerhebung wirklich adäquat?
- Frage 2: Besagt der kleine p-Wert, dass Zusammenhang in der Grundgesamtheit stark ist?

Null Hypothesis Significance Testing (NHST)

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

RECHTS- UND WIRTSCHAFTS-
WISSENSCHAFTLICHE FAKULTÄT

Null Hypothesis Significance Testing-Ritual

Quellen:

- Chow, S.L. (1998). The null-hypothesis significance-test procedure is still warranted. *Behavioral and Brain Sciences* **21**, 228-235.
- Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences* **21**, 199-200.
- Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics* **33**, 587-606.

Nullritual nach Gigerenzer I

Ausgangspunkt:

- Lehrbücher und Curricula zur Statistik informieren fast nie über die vielfältige Toolbox statistischer Verfahren (wie deskriptive Statistik, explorative Datenanalyse, Bayes-Verfahren, Neyman-Pearson Entscheidungstheorie, Walds Sequentialanalyse).
- Kenntnisse bezüglich des Inhalts der Toolbox verlangt statistisches Denken, d.h. die Kunst das oder die richtigen Tools für ein gegebenes Problem auszuwählen.
- Stattdessen: Einziges und universelles Tool = Nullritual (rituelle Handwaschung).
- Dieses Nullritual ersetzt einen kritischen Blick auf die Daten.

Nullritual nach Gigerenzer II

Schritte des Nullrituals:

1. Formulierung einer Nullhypothese als ”no mean difference”, ”zero correlation” oder ”zero regression coefficient”.
2. Verwendung der Irrtumswahrscheinlichkeit von 5% als Konvention. Falls signifikantes Ergebnis, Annahme der Forschungshypothese. Kommunikation der Ergebnisse als $p < 0.05$, $p < 0.01$ oder $p < 0.001$, je nachdem, was dem tatsächlichen p-Wert am nächsten kommt.
3. Verfahre **immer** so.

Gerechfertigt wird dieses Vorgehen unter Verweis auf die Arbeiten von R.A. Fisher und von Neyman & Pearson.

Fishers Nullhypothesentesten (nach Gigerenzer)

1. Postuliere eine Nullhypothese, die nicht(!) notwendig eine Nilhypothese sein muss.
2. Kommuniziere den exakten p-Wert (d.h. $p = 0.049$ statt $p < 0.05$). Verwende nicht 5% als Konvention und sprich nicht davon, Hypothesen abzulehnen oder anzunehmen.
3. Verwende dieses Verfahren nur, wenn sehr wenig Informationen über das vorliegende Problem existieren.

Zitat aus R.A. Fisher (1956). *Statistical methods and scientific induction*, Oliver & Boyd, Edinburgh.

> „. . . so scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."

Neyman-Pearson Entscheidungstheorie (nach Gigerenzer)

1. Formuliere zwei konkurrierende Hypothesen H_0 und H_1.
   Lege vor dem Experiment α, β und n aufgrund subjektiver Kostenüberlegungen fest.
   (Damit sind Ablehnungsbereiche für beide Hypothesen festgelegt.)
2. Fallen die Daten in den jeweiligen Ablehnunsgbereich von H_0, wird H_1 angenommen.
   Beachte: Annahme heißt nicht, dass man an die Hypothese glaubt; es wird lediglich so gehandelt als sei sie wahr.
3. Dieses Verfahren ist beschränkt auf Situationen konkurrierender Hypothesen und sinnvoller Kosten-Nutzen-Abwägungen, um α und β festzulegen.

Beispiel nach Gigerenzer: Qualitätskontrolle.

Nullritual als Hybrid aus Fisher und Neyman & Pearson

1. Festlegung der Nullhypothese nach Fisher (bis auf Nilhypothese); Inkonsistenz zu Neyman & Pearson.
2. Binäre Testentscheidung entspricht Vorgehen nach Neyman & Pearson (bis auf Festlegung von α aufgrund von Kosten-Nutzen-Überlegungen). Fisher lehnt binäre Testentscheidungen (bis auf Spezialfälle der Qualitätskontrolle) ab. Stattdessen Kommunikation des exakten *p*-Wertes.
3. Fisher und Neyman & Pearson: Ablehnung des mechanistischen Gebrauchs statistischer Verfahren.

Gigerenzer zitiert renommierte Psychologen, die das Nullritual stets abgelehnt haben: Bartlett, Köhler, Pavlov, S.S. Stevens, Boring, Skinner, Luce, H.A. Simon.

Mythen des *p*-Wertes I

Quellen:

1. Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* Chichester, Wiley.
2. Haller, H. & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research* **7**, 1-20.

- Studien von Oakes (1986) und Haller & Krauss (2002) zeigen an einem MC-Fragebogen, dass ein Großteil von Studierenden und von Lehrern der Statistik den *p*-Wert nicht korrekt interpretieren können.
- Mythos 1: *p*-Wert misst bedingte Wahrscheinlichkeit, dass H_0 richtig ist, wenn die Daten extreme Werte des Prüfmaßes zeigen (d.h. $P(H_0|D)$).

  Richtig: $P(D|H_0)$, d.h. bedingte Wahrscheinlichkeit, dass das Prüfmaß extremere Werte als den beobacheteten Wert annimmt, wenn H_0 richtig ist.

Mythen des p-Wertes II

- Bayes-Formel: Sei $p = P(D|H_0)$, $\pi = P(D|\overline{H}_0)$.

$$\begin{aligned} P(H_0|D) &= \frac{P(D|H_0)P(H_0)}{P(D)} = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0) + P(D|\overline{H}_0)P(\overline{H}_0)} \\ &= \frac{pP(H_0)}{pP(H_0) + \pi P(\overline{H}_0)} = \frac{1}{1 + \frac{\pi}{p}\frac{1-P(H_0)}{P(H_0)}}. \end{aligned}$$

- Fazit:
  1. Entscheidend ist das Verhältnis π/p. Je größer, desto kleiner Wahrscheinlichkeit, dass Nullhypothese nicht abgelehnt wird, wenn Prüfgröße extreme Werte annimmt.
  2. Verteilung der Prüfgröße wird nicht nur unter H_0, sondern für alle Parameterkonstellationen benötigt.
  3. Wenn unwahrscheinliche Nullhypothese ($P(H_0)$ klein) hineingesteckt wird, sinkt ebenfalls die Wahrscheinlichkeit, dass Nullhypothese nicht abgelehnt wird, wenn Prüfgröße extreme Werte annimmt.

Mythen des p-Wertes III

- Mythos 2: p-Wert misst die Stärke eines Effektes.

  Misst eine gegebene Mittelwertdifferenz (von z.B. 1) für $p = 0.01$ einen größeren Einfluss der qualitativen Variablen als für $p = 0.05$?

  Da $p = P(D|H_0)$ ist, hat der p-Wert nichts mit der Stärke eines Effektes oder der Richtigkeit einer Hypothese zu tun.

  Für NHSTP geht es nur um eine binäre Entschiedung „H_0 ablehnen oder nicht" und nicht um die Stärke des Einflusses unabhängig von der Höhe des p-Wertes.

Mythen des *p*-Wertes IV

- Mythos 3: Signifikant heißt wichtig.

  Bestenfalls heißt signifikant nicht-zufällig, aber darf keineswegs mit inhaltlich bedeutsam verwechselt werden.
- Einfluss des Stichprobenumfanges auf den *p*-Wert: *p*-Wert sinkt mit wachsendem Stichprobenumfang.

  In einer Totalerhebung ist jeder noch so kleine Effekt signifikant (=nicht-zufällig).
- Quelle: J. G. Combs (2010). From the editors. Big samples and small effects: Let's not trade relevance and rigor for power. Academy of Management Journal **53**, 9-13.

  > "I see more and more studies in which correlations and standardized regression coefficients of .05 or less receive the prized label highly significant."

Wirkung des Stichprobenumfangs auf p-Wert I

- Ausgangspunkt: Indifferenztabelle ($\gamma = 0$)

	1	2	3
1	20	75	5
2	16	60	4
3	4	15	1

- Modifikation des Zelleneintrags ($\gamma = 0.031$, p-Wert=0.414)

	1	2	3
1	20	75	5
2	16	60	4
3	4	14	2

Wirkung des Stichprobenumfangs II

Vervielfachung der absoluten Häufigkeiten (d.h. relative Häufigkeiten bleiben konstant)

- Da γ nur von relativen Häufigkeiten abhängt, tangiert die vervielfachung den Wert von γ nicht.
- Varianz und *p*-Wert sinken mit wachsendem *n*:

n	*p*-Wert
200	0.414
2000	0.246
10000	0.062
20000	0.0148
30000	0.00385

Effektgrößen vs. *p*-Wert

Quelle: Tversky, A. & Kahnemann, D. (1971). Belief in the law of small numbers. *Psychological Bulletin* **76**, 105-110.

Zitat (S. 109):

> ”The emphasis on significance levels tends to obscure a fundamental distinction between the size of an effect and its statistical significance. Regardless of sample size, the size of an effect in one study is a reasonable estimate of the size of the effect in replication. In contrast, the estimated significance level in a replication depends on the sample size.”

Effektgrößen

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

RECHTS- UND WIRTSCHAFTS-
WISSENSCHAFTLICHE FAKULTÄT

Lösungvorschlag: Effektgrößen nach Cohen I

- Effektgröße ist ein Maß für die Stärke der Beziehung zwischen zwei oder mehr Variablen in einer Grundgesamtheit.
  Beispiel: Zwei-Stichproben-Problem für unverbundene Stichproben

$$\mu_1 - \mu_2 \quad \text{bzw.} \quad \frac{\mu_1 - \mu_2}{\sigma}.$$

- Schätzer der Effektgröße für die Grundgesamtheit wird ebenfalls Effektgröße genannt.
  Beispiel:

$$\overline{X}_1 - \overline{X}_2 \quad \text{bzw.} \quad \frac{\overline{X}_1 - \overline{Y}_2}{S}$$

  mit S geeigneter Schätzung der Standardabweichungen von X_i und/oder Y_j.
- Kennzeichen: Schätzer von Effektgrößen hängen nicht funktional von Stichprobenumfängen ab.

Lösungvorschlag: Effektgrößen nach Cohen II

- Unterschied zum Hypothesentest: Schätzung der Stärke der Beziehung statt Beurteilung, ob die Beziehung lediglich durch Zufall erklärt werden kann.
- Sowohl in Experiment- als auch Beobachtungsstudien ist häufig neben der Signifikanz auch die Beurteilung der Effektgröße wichtig.
- Mögliche Situation: Signifikante, aber sehr kleine Effektgröße.
- Anwendungsbereich Metastudien: Aggregation von Ergebnissen aus mehreren Einzelstudien.
- Anwendungsbereich Power Analysis: Vorgabe eines Signifikanzniveaus und eines Effektes zur Berechnung der Power bei gegebenem Stichprobenumfang bzw. zur Berechnung des Stichprobenumfangs bei gegebener Power.

Effektgrößen auf der Basis der Differenz von Mittelwerten

- Standardisierte Mittelwertdifferenz:

$$\theta = \frac{\mu_1 - \mu_2}{\sigma},$$

  wobei σ geeignete Varianz.
- Cohens d als Schätzer für θ

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S},$$

  wobei S z.B. die gepoolte Stichprobenstandardabweichung

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2}}$$

Bewertung von Mittelwertdifferenzen I

Quelle: Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

- Cohens Grenzen für $\delta = (\mu_1 - \mu_2)/\sigma$:

δ	0.2	0.5	0.8
Effekt	klein	mittel	stark

- Interpretation: Nichtüberlappungswahrscheinlichkeit

$$U_1 = \frac{2\Phi(\delta/2) - 1}{\Phi(\delta/2)}.$$

δ	0.2	0.5	0.8
U_1	0.147	0.33	0.474
Effekt	klein	mittel	stark

Bewertung von Mittelwertdifferenzen II

Inhaltliche Interpretation von „klein”:

- Nichtüberlappungsbereich im Sinne von U_1: 14.7%.
- Standardisierte mittlere IQ-Differenz zwischen Nichtzwillingen und Zwillingen.
- Mittlere Differenz der Körpergröße zwischen 15- und 16-jährigen Mädchen.
- Differenz für Teil des Wechsler-IQ-Testes für Männer und Frauen.

Bewertung von Mittelwertdifferenzen V

Inhaltliche Interpretation von „mittel”:

- Nichtüberlappungsbereich im Sinne von U_1: 33%.
- Sollte mit bloßem Auge sichtbar sein.
- Standardisierte mittlere IQ-Differenz zwischen Büroangestellten und angelernten Arbeitern.
- Mittlere Differenz der Körpergröße zwischen 14- und 18-jährigen Mädchen.

Bewertung von Mittelwertdifferenzen VI

Inhaltliche Interpretation von „groß”:

- Nichtüberlappungsbereich im Sinne von U_1: 50%.
- Aber: Standardisierte mittlere IQ-Differenz zwischen Ph.D-Studenten und Studienanfängern.
- Aber: Mittlere Differenz der Körpergröße zwischen 13- und 18-jährigen Mädchen.

Generell:

- Beurteilung, ob klein, mittel oder stark, ist kontextabhängig.
- Trotzdem: Vorteil eines generell anzuwendenden, pragmatischen Maßstabes.
- Wichtig: Standardisierung gleicht Unsicherheiten aus, d.h. z.B. Mittelwertdifferenz in Einheiten Standardabweichung.

Übertragung auf Korrelationskoeffizienten und Anteil erklärter Varianz

Quellen:

1. Rosenthal, R. & Rubin, D.R. (1983). A Simple, General Purpose Display of Magnitude of Experimental Effect. *Journal of Educational Psychology* **74**, 166-169.
2. Rosnow, R.L., Rosenthal, R. & Rubin, D.R. (2000). Contrasts and Correlations in Effect-Size Estimation. *Psychological Science* **11**, 446-453.
3. Rosenthal, R. & Rubin, D.R. (2003). $r_{\text{equivalent}}$: A Simple Effect Size Indicator. *Psychological Methods* **8**, 493-496.

- Zusammenhang von Mittelwertdifferenzen und punkt-biserialem Korrelationskoeffizient, Korrelationskoeffizient und Cramérs V.

Übertragung auf Korrelationskoeffizienten

- Beispiel: Punkt-biserialer Korrelationskoeffizient:

$$\rho_{bs} = Corr(X, Y) = \frac{\mu_1 - \mu_2}{\sqrt{(\mu_1 - \mu_2)^2 + 1/(p_1(1 - p_1))}}$$

  D.h. ρ_{bs}^2 gibt den Anteil der Varianz von X an, die durch die dichotome Zufallsvariable Y erklärt werden kann.

- Für $\sigma = 1$ und $p_1 = 1/2$ ist

δ	0.2	0.5	0.8
r_{bs}	0.1	0.243	0.371
r_{bs}^2	0.01	0.059	0.138
Effekt	klein	mittel	stark

Übertragung auf Anteil erklärter Varianz? I

- BESD = Binomial effect size display.
- BESD = Effekt auf die Erfolgsrate einer Behandlung.
- Beispiel:

	Behandlungsergebnis		
Zustand	Lebendig	tot	Summe
Behandlung	66	34	100
Kontrolle	34	66	100
Summe	100	100	200

Übertragung auf Anteil erklärter Varianz? II

- Quadrat des Korrelationskoefizient zweier binärer Variabler (Lebendig/tot vs. Behandlung/Kontrolle):

$$r^2 = \frac{\chi^2(1)}{n^2} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{1.}n_{.1}n_{2.}n_{.2}} = 0.1.$$

  D.h. obwohl nur 10% der Varianz des Behandlungsergebnisses durch die Behandlung erklärt werden kann, steigt die Überlebenswahrscheinlichkeit von 34% auf 66%.
- Zusammenhang: r^2 und Überlebenswahrscheinlichkeit bei identischen Gruppengrößen:

$$n_{11}/n = 0.5 + r/2, \quad \text{und} \quad n_{21}/n = 0.5 - r/2$$

  bzw.

$$n_{11}/n - n_{21}/n = r.$$

Übertragung auf Anteil erklärter Varianz? III

- BESD in Abhängigkeit von r bei identischer Gruppengrößen

d	r^2	r	n_{21}/n	n_{11}/n
schwach	0.0100	0.1000	0.4500	0.5500
mittel	0.0600	0.2450	0.3775	0.6225
stark	0.1400	0.3700	0.3150	0.6850

Übersicht der Vorschläge von Cohen

Quelle: Sawyer, A.G. & A.D. Ball (1981). Statistical power and effect sizes in marketing research. *Journal of Marketing Research* **18**, 275-290.

Test	Effektgröße	schwach	mittel	stark	schwach	mittel	stark
t-MWD.	δ	0.2	0.5	0.8			
t-*r*	ρ_{bs}	0.1	0.3	0.5			
	ρ_{bs}^2				0.01	0.059	0.138
t-*r*-Differenzen	$1/2 \ln \frac{1+\rho}{1-\rho}$	0.1	0.3	0.5			
	$\lvert Z_1 - Z_2 \rvert$						
	$\rho_2^2 - \rho_1^2$				0.05-0.08	0.15-0.23	0.28-0.38
Vorzeichen	$\lvert p - 0.5 \rvert$	0.05	0.15	0.25			
Vorzeichen-differenzen	$\varphi_i = 2 \arcsin(\sqrt{p_i})$	0.3	0.5	0.8			
	$\lvert \varphi_1 - \varphi_2 \rvert$						
	ϕ^2				0.01	0.059-0.061	0.137-0.152
χ^2	$C = \sqrt{\chi^2/(\chi^2 + n)}$	0.1	0.3	0.5			
	$w = \sqrt{C^2/(1 - C^2)}$						
	ϕ^2				0.01	0.059	0.138
F-MWD.	$f = \sigma_i/\sigma$	0.10	0.25	0.40			
	$\eta^2 = \sigma_i^2/(\sigma^2 + \sigma_i^2)$				0.01	0.059	0.138
F (zusätzl. Regressor)	$f^2 = \frac{\mathcal{R}^2_{y.B}}{1-\mathcal{R}^2_{y.B}}$	0.02	0.15	0.35			
	$\mathcal{R}^2_{y.B} = f^2/(1 + f^2)$				0.02	0.13	0.26

Möglichkeiten und Grenzen von Effektgrößen

Quellen:

1. Olejnik, S. & Algina, J. (2000). Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations. *Contemporary Educational Psychology* **25**, 241-286.
2. Breaugh, J.A. (2003). Effect Size Estimation: Factors to Consider and Mistakes to Avoid. *Journal of Management* **29**, 79-97.

- Kein neues Ritual aufbauen.
- Auch kleine Effektgrößen können inhaltlich wichtig sein.

  Beispiel: Studie, die Wirkung von Aspirin auf Herzbeschwerden nachweist, basiert auf einer Korrelation von 0.03 zwischen zwei Variablen (siehe Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist* **45**, 775-777.)

Verteilung von Effektgrößen

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

RECHTS- UND WIRTSCHAFTS-
WISSENSCHAFTLICHE FAKULTÄT

Konfidenzintervalle für standardisierte Mittelwerte I

Quelle:

1. Kelley, K. (2007). Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. *Journal of Statistical Software* **20**, 1-24.
2. Kelley, K. (2007). **MBESS**. *Methods for Behavioral, Educational and Social Sciences.* R package version 0.0.8, URL http://CRAN.R-project.org/.

Grundidee für normalverteilte Grundgesamtheiten mit Mittelwert μ und Varianz σ^2.

- Betrachte den Variationskoeffizienten $\theta = \mu/\sigma$.
- Dann ist

$$\sqrt{n}\frac{\overline{X}_n}{S} \sim t(n-1;\lambda)$$

mit Nichtzentralitätsparameter

$$\lambda = \frac{\mu}{\sigma/\sqrt{n}} = \sqrt{n}\theta.$$

Konfidenzintervalle für standardisierte Mittelwerte II

- Sei $t_{p;\nu,\lambda}$ das p-Quantil der t-Verteilung mit ν Freiheitsgraden und Nichtzentralitätsparameter λ, dann ist das $1-\alpha$-Schwankungsintervall

$$P\left(t_{\alpha/2;n-1,\sqrt{n}\theta} \leq \frac{\sqrt{n}\overline{X}_n}{S} \leq t_{1-\alpha/2;n-1,\sqrt{n}\theta}\right) = 1-\alpha.$$

  Die Grenzen (Quantile) sind streng monoton zunehmende Funktionen f_1 und f_2 in θ, so dass

$$P\left(f_2^{-1}\left(\sqrt{n}\frac{\overline{X}_n}{S}\right) \leq \theta \leq f_1^{-1}\left(\sqrt{n}\frac{\overline{X}_n}{S}\right)\right) = 1-\alpha.$$

Konfidenzintervalle für standardisierte Mittelwerte III

- Beispiel: $\alpha = 0.025$, $n = 25$, $\overline{x}_{25} = 50$, $s = 10$.

$$f_1^{-1}\left(\sqrt{25}\frac{50}{10}\right) = f_1^{-1}(25) = 6.453777$$

wegen

```
qt(0.025,24,6.453777*5)
[1] 25
```

und

$$f_2^{-1}\left(\sqrt{25}\frac{50}{10}\right) = f_2^{-1}(25) = 3.536517$$

wegen

```
qt(0.975,24,3.536517*5)
[1] 25
```

Realisiertes KI für $\theta = \mu/\sigma$:

$$(3.536517, 6453777)\,.$$

Konfidenzintervalle für standardisierte Mittelwerte IV

Fazit: Statt Verteilung unter einer unrealistischen Nullhypothese wird Verteilung des Schätzers der Effektgröße unter allen Konstellationen für die Effektgröße in der Grundgesamtheit betrachtet.

Empirische Verteilung von Effektgrößen

Quelle: Hemphill, J.F. (2003). Interpreting the Magnitudes of Correlation Coefficients. *American Psychologist* **58**, 78-79.

- Reanalyse zweier großer Zusammenfassungen der Forschungsliteratur zu psychologischer Bewertung (Meyer et al. (2001), 78 meta-analytische Studien) und psychologischer Behandlung (Lipsey & Wilson (1993), 302 meta-analytische Studien).
- Konvertierung von Cohens *d* in Korrelationskoeffizienten (siehe Rosnow et al. (2000)).
- Berechnung der 33.3% und 66.7% Quantile der empirischen Verteilung der 380 Werte für den Korrelationskoeffizienten

	Meyer et al.	Lipsey & Wilson	kombiniert	Vorschlag
Unteres Drittel	0.02 bis 0.21	-0.08 bis 0.17	-0.08 bis 0.17	< 0.20
Mittleres Drittel	0.21 bis 0.33	0.17 bis 0.28	0.18 bis 0.29	0.20 bis 0.30
Oberes Drittel	0.35 bis 0.78	0.29 bis 0.60	0.30 bis 0.78	> 0.30

Theoretische Verteilung von Effektgrößen für qualitative Variablen I

Quellen:

1. Vogel, F. & Wiede, T. (1994). Ein neues Zusammenhangsmaß für ordinalskalierte Merkmale. *Jahrbücher für Nationalökonomie und Statistik* **213**, 1-30.
2. Pavlides, M.G. & Perlman, M.D. (2009). How Likely is Simpson's Paradox? *American Statistician* **63**, 226-233.

- Ausgangspunkt: k Zellen einer oder mehrerer qualitativer Variablen
- $p_1, \ldots, p_{k-1}$: relative Häufigkeiten der Zellen (=Wahrscheinlichkeiten)
  $n_1, \ldots, n_{k-1}$: absolute Häufigkeiten der Zellen mit $n_1 + \ldots + n_{k-1} < n$.

Theoretische Verteilung von Effektgrößen für qualitative Variablen II

- Idee: „Vergleich ist die Seele Statistik” (Sigmund Schott nach Zizek, F. (1922). *Fünf Hauptprobleme der statistischen Methodenlehre* Duncker-Humblot-Verlag, Berlin.)

  Betrachtung aller möglichen alternativen Grundgesamtheiten als Realisationen von Zufallszügen aus einer möglichst flexiblen Verteilungsfamilie für die relativen (absoluten) Häufigkeiten.

Theoretische Verteilung von Effektgrößen für qualitative Variablen III

- Sei $T = t(p_1, \ldots, p_{k-1})$:

$$P(T \leq t) = \int_A f(p_1, \ldots, p_{k-1}) \prod_{i=1}^{k-1} dp_i$$

$$A = \{(p_1, \ldots, p_{k-1}) \in [0,1]^{k-1} | t(p_1, \ldots, p_{k-1}) \leq t, \sum_{i=1}^{k-1} p_i \leq 1\}$$

- Seien n fixiert und $T = t(n_1, \ldots, n_{k-1})$:

$$P(T \leq t) = \sum_B f(n_1, \ldots, n_{k-1})$$

$$B = \{(n_1, \ldots, n_{k-1}) \in \{0, 1, \ldots, n\}^{k-1} | t(n_1, \ldots, n_{k-1}) \leq t, \sum_{i=1}^{k-1} n_i \leq n\}$$

Quantile der theoretischen Verteilung von Effektgrößen für qualitative Variablen

- Seien T ein bivariates Konkordanzmaß und t_α das α-Quantil der theoretischen Verteilung.
- Klassifizierung nach der Gleichwahrscheinlicheitsmethode:

Stärke des Zusammenhangs	Wertebereich des Konkordanzmaßes
stark negativ	$-1 = \min T < T \leq t_{1/6}$
mittel negativ	$t_{1/6} < T \leq t_{2/6}$
schwach negativ	$t_{2/6} < T < t_{1/2} = 0$
schwach positiv	$0 = t_{1/2} < T \leq t_{4/6}$
mittel positiv	$t_{4/6} < T \leq t_{5/6}$
stark positiv	$t_{5/6} < T \leq \max T = 1$

Dirichlet-V. als Modell für die relativen Häufigkeiten von qualitativen Variablen

- Konkret für (relative) Häufigkeitsdaten ohne Ganzzahligkeitsrestriktion: Dirichlet-Verteilung

$$f_{p_1,\ldots,p_{k-1}}(p_1,\ldots,p_{k-1}) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}p_i^{\alpha_i-1}$$

mit $\sum_{i=1}^{k-1} p_i \leq 1$ und $p_k = 1 - \sum_{i=1}^{k-1} p_i$, $0 \leq p_i \leq 1$, $\alpha_i > 0$ für $i = 1, 2, \ldots, k$.

Was spricht für die Dirichlet-V.?

Quelle: Walley, P. (1996). Inferences form Multinomial Data: Learning about a Bag of Marbles. *Journal of the Royal Statistical Society, Series B* **58**, 3-57.

1. Konjugiertheit führt zu leichter mathematischer Handhabbarkeit.
2. Zusammenfassung von Kategorien führt wieder zur Dirichlet-Verteilung (sichert das ”representative invariance principle” = Eigenschaft der Unabhängigkeit vom Stichprobenraum).
3. Menge der Dirichlet-Verteilungen ist sehr umfassend und flexibel; Mischungen sind wieder Dirichlet; jede a priori kann durch eine Mischung von Dirichlet-Verteilungen approximiert werden.
4. Die meisten bayesianischen Modelle bezüglich ”prior ignorance” bezüglich p arbeiten mit Dirichlet.

Multivariate Polya-V. als Modell für die absoluten Häufigkeiten von qualitativen Variablen I

- Konkret für Häufigkeitsdaten mit Ganzzahligkeitsrestriktion und fester Größe n der Grundgesamtheit: multivariate Polya-Verteilung (oder Dirichlet compound multinomial V.)
- Wahrscheinlichkeitsfunktion:

$$f(n_1, \ldots, n_{k-1}) = \frac{n!}{\prod_{i=1}^{k} n_i!} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Gamma(n + \sum_{i=1}^{k} \alpha_i)} \prod_{i=1}^{k} \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)}$$

für $\sum_{i=1}^{k-1} n_i \leq n$ und $n_k = n - \sum_{i=1}^{k-1} n_i$.

Multivariate Polya-V. als Modell für die absoluten Häufigkeiten von qualitativen Variablen II

- Darstellung als Mischverteilung:

$$f(n_1, \ldots, n_{k-1}) = \int f(n_1, \ldots, n_{k-1} | p_1, \ldots, p_{k-1}) f(p_1, \ldots, p_{k-1}) \, dp_1 \ldots dp_{k-1}$$

  mit $f(n_1, \ldots, n_{k-1} | p_1, \ldots, p_{k-1})$ als multinomischer (= Likelihood) und $f(p_1, \ldots, p_{k-1})$ als Dirichlet-Verteilung (= a priori) mit Hyperparameter $\alpha_1, \ldots, \alpha_k$.
- Multivariate Polya-V. als „natürliche" Diskretisierung der Dirichlet-V. mit der Eigenschaft

$$f(n_1, \ldots, n_{k-1}) \to f(p_1, \ldots, p_{k-1})$$

  für $n \to \infty$.

Numerische Bestimmung der Quantile der theoretischen Verteilung

- Totalenumeration für n fixiert:
  Konstruktion aller Zerlegungen von n in k nicht-negative ganzzahlige Summanden und Berechung der Werte von T samt zugehöriger Wahrscheinlichkeiten.
- Simulation durch hinreichend viele Züge aus einer Dirichlet-V. bzw. der multivariaten Polya-V..
- Algorithmen in R:
  - Dirichlet-V.: `rdirichlet` im MCMCpack
  - Polya-V.: Sukzessives Ziehen aus einer Dirchlet- und anschließend einer multinomischen Verteilung.
  - Anzahl der Wiederholungen: 1000000.
  - Wichtig: Programmierung der Maßzahlen ohne Schleifen.

Problem: Hyperparameter α_i

Quellen:

1. Jaeger, M. (2005). A Representation Theorem and Applications to Measure Selection and Noninformative Priors. *International Journal in Approximating Reasoning* **38**, 217-243.
2. Yang, R. & Berger, J.O. (1998). A Catalog of Noninformative Priors. Working Paper.

- Ansatz 1: Sensitivitätsanalyse für alternative Setzungen (siehe Pavilides & Perlman (2009)).
- Ansatz 2: Verwendung nicht-informativer a priori Verteilungen für die multinomische Verteilung (Yang & Berger (1998), Jaeger (2005)).
- Ansatz 3: Verwendung unscharfer Wahrscheinklichkeiten (Walley (1996)).

Wirkung der Hyperparameter auf die Dirichlet-V. I

- Ohne weitere Information $\alpha_i = \alpha$, $i = 1, 2, \ldots, k$
- $\alpha = 1$: Gleichverteilung über dem Wahrscheinlichkeitssimplex (Unabhängigkeitsmodell).
- $\alpha = 1/2$: Jeffreys prior (1961) legt mehr Wahrscheinlichkeitsmasse an die Ränder des Wahrscheinlichkeitssimplex.
- $\alpha = 1/k$: Perks prior (1947).
- $\alpha = 0$: Haldanes prior (1932) führt zu unechter Verteilung.

Wirkung der Hyperparameter auf die Dirichlet-V. II

- Mit wachsendem α konzentriert sich die Wahrscheinlichkeitsmasse um den Mittelwertvektor $1/k, \ldots, 1/k$ wegen

$$Var_\alpha(p_r) = \frac{k-1}{k^2(k\alpha+1)} = O\left(\frac{1}{\alpha}\right).$$

  D.h. extreme Werte der Konkordanzmaße werden mit wachsendem α unwahrscheinlicher.

Erhebung: Selbsteinschätzung von Mathematik- (U) und Computerkenntnissen (V)

- Zusammenfassung der Daten in einer 6×6-Kontingenztabelle

$U \setminus V$	1	2	3	4	5	6
1	7	9	10	0	0	1
2	30	90	90	45	15	8
3	21	152	142	93	37	12
4	10	73	86	37	28	9
5	5	28	37	19	10	1
6	1	0	6	3	0	1

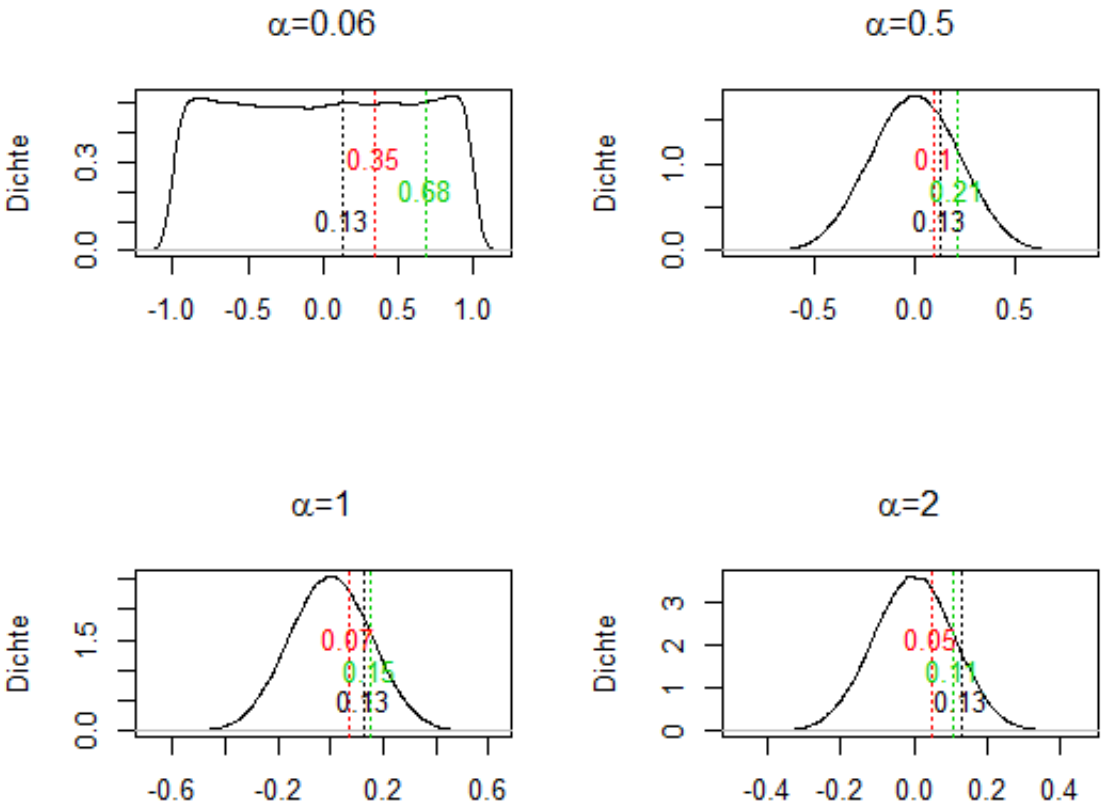
- Goodman-Kruskal γ: 0.128.
- „p"-Wert: Wahrscheinlichkeit, dass empirischer Wert (= 0.128) (nicht nur unter H_0) überschritten wird.

Quantile für alternative Hyperparameter I

α	66.7%-Quantil	88.3%-Quantil	„p"-Wert	Effekt
		Dirichlet-V.		
$1/36$	0.538	0.893	0.463	schwach
$2/36$	0.338	0.672	0.437	schwach
$1/2$	0.0952	0.213	0.281	mittel
1	0.0677	0.151	0.205	mittel
2	0.0474	0.107	0.122	stark
		Polya-V.		
$1/36$	0.551	0.905	0.460	schwach
$2/36$	0.339	0.676	0.437	schwach
$1/2$	0.0964	0.215	0.215	mittel
1	0.0686	0.153	0.209	mittel
2	0.0490	0.110	0.128	stark

Quantile für alternative Hyperparameter II

- Dichtefunktion von Goodman & Kruskals γ:



Cramérs V und Korrelationskoeffizient

- Cramérs V:

$$0 \leq V = \sqrt{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n)}}\sqrt{\frac{1}{\min\{k-1, l-1\}}} \leq 1$$

- Für $k = l = 2$ (Vierfeldertafel) stimmen V und der Absolutbetrag des Korrelationskoeffizienten r überein.
- Quantile für Cramérs V:

α	66.7%-Quantil	83.3%-Quantil
$1/4$	0.257	0.573
$1/2$	0.200	0.500
1	0.184	0.412
2	0.145	0.318
Cohen	0.240	0.370

Welche Maßzahlen können so behandelt werden?

1. Qualitative Merkmale
   - Entropie
   - Cramérs V, Transinformation, PRE-Maße
2. Komparative Merkmale
   - Summenhäufigkeitsentropie (Vogel (1981), Klein (1999))
   - Schiefemaße (Klein (2001), (2012))
   - Goodman & Kruskals γ, Kendalls τ, Spearmans ρ
3. Klassierte Daten bei fixierten Klassenmitten
   - Mittelwert, Varianz, Schiefe, Wölbung
   - Korrelationsverhältnis

Bayes-Inferenz für Effektgrößen

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

RECHTS- UND WIRTSCHAFTS-
WISSENSCHAFTLICHE FAKULTÄT

A posteriori Dirichlet-Verteilung

- Problem: Wie wahrscheinlich ist es, dass in der Grundgesamtheit ein mittlerer positiver Zusammenhang herrscht, wenn in der Stichprobe für $n = 50$ und eine 3×4-Tabelle $\gamma = 0.612$ beobachtet wurde?
- Konjugiertheit von multinomischer und Dirichlet-Verteilung führt zur a posteriori-Verteilung

$$p_1, \ldots, p_{k-1} | n_1, \ldots, n_{k-1} \sim \text{DIR}\left(\alpha_1 + n_1, \ldots, \alpha_k + n - \sum_{i=1}^{k-1} n_i\right)$$

- Vorgehen:
  - Fixierung von α.
  - Festlegung der Klassifizierung in „schwach, mittel und stark” (d.h. der 66.7%- und 88.3%-Quantile $t_{2/3}$ und $t_{5/6}$
  - Berechnung der a posteriori Verteilung von $T = t(p_1, \ldots, p_{k-1})$ gegeben γ.
  - Berechnung der a posteriori Wahrscheinlichkeit

$$P(t_{2/3} \leq T < t_{5/6} | \gamma).$$

A posteriori Dirichlet-Verteilung: Beispiel

- Originalbeispiel von Goodman & Kruskal (1964)

	1	2	3	4
1	8	5	3	3
2	0	8	1	0
3	0	4	14	4

- $\gamma = 0.612$, p-Wert$= 0.0000252$
- A posteriori Wahrscheinlichkeiten:

	$P(\text{schwach}\|\gamma)$	$P(\text{mittel}\|\gamma)$	$P(\text{stark}\|\gamma)$
$\alpha = 1$	0.0186	0.146	0.834
$\alpha = 1/2$	0.0269	0.295	0.677

Bayes-Faktor I

Quellen:

1. Jeffreys, H. (1961). *Theory of Probability.* Oxford, Oxford University Press.
2. Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

- Bayes-Faktor als Maß für die Stärke einer Hypothese gegenüber ihrem Komplement (z.B.):

$$\text{BF}_{\text{stark pos.}} = \frac{P(\text{stark pos.}|\gamma)/(1 - P(\text{stark pos.}|\gamma)}{P(\text{stark pos.})/(1 - P(\text{stark pos.})}$$

- Mit $P(\text{stark}) = 1/6$ ist

$$\text{BF}_{\text{stark pos.}} = 5 \cdot \frac{P(\text{stark pos.}|\gamma)}{1 - P(\text{stark pos.}|\gamma)}$$

Bayes-Faktor II

- Klassifikation nach Kass & Raftery (1995) mittels

$2 \log_e \mathrm{BF}$	BF	Klassifizierung der Evidenzstärke
0 bis 2	1 bis 3	Not worth more than a bare mention
2 bis 6	3 bis 20	positive
6 bis 10	20 bis 150	strong
> 10	> 150	very strong

- Beachte: Jeffreys (1961) verwendet $log_{10}\mathrm{BF}$.

A posteriori Dirichlet-V. und Bayes-Faktor

- Originalbeispiel von Goodman & Kruskal ($n = 50$, $\gamma = 0.612$)

	$BF_{schwach}$	BF_{mittel}	BF_{stark}
$\alpha = 1$	0.0948	0.855	25.121
$\alpha = 1/2$	0.138	2.093	10.483

- Positive bis starke Evidenz für einen stark positiven Zusammenhang in der Grundgesamtheit, wenn in der Stichprobe $\gamma = 0.612$ beobachtet werden.

Weiteres Beispiel: Selbsteinschätzung von Mathematik- (U) und Computerkenntnissen (V)

- Daten: $n = 1116$, $\gamma = 0.128$
- A posteriori Wahrscheinlichkeiten:

	$P(\text{schwach}\|\gamma)$	$P(\text{mittel}\|\gamma)$	$P(\text{stark}\|\gamma)$
$\alpha = 1$	0.0491	0.752	0.199
$\alpha = 1/2$	0.191	0.805	0.00362

- Bayes-Faktoren:

	$\text{BF}_{\text{schwach}}$	$\text{BF}_{\text{mittel}}$	BF_{stark}
$\alpha = 1$	0.258	15.125	1.244
$\alpha = 1/2$	1.181	20.665	0.0182

Wirkung des Stichprobenumfangs auf den Bayes-Faktor I

- Ausgangspunkt: Indifferenztabelle ($\gamma = 0$)

	1	2	3
1	20	75	5
2	16	60	4
3	4	15	1

- Modifikation des Zelleneintrags ($\gamma = 0.031$, p-Wert=0.414)

	1	2	3
1	20	75	5
2	16	60	4
3	4	14	2

Wirkung des Stichprobenumfangs auf den Bayes-Faktor II

- Fall: $\alpha = 1$
- Quantile der a priori V.: $t_{4/6} = 0.198$ und $t_{5/6} = 0.423$.

	schwach	mittel	stark
	$n = 50$		
a posteriori Ws.	0.495	0.107	0.0012
Bayes-Faktor	4.908	0.599	0.006
	$n = 500$		
a posteriori Ws.	0.753	0	0
Bayes-Faktor	15.27	0	0
	$n = 5000$		
a posteriori Ws.	0.983	0	0
Bayes-Faktor	27.685	0	0

- D.h.: Unabhängig vom Stichprobenumfang Entscheidung für schwachen positiven Zusammenhang ($n \geq 500$).

Fazit

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

RECHTS- UND WIRTSCHAFTS-
WISSENSCHAFTLICHE FAKULTÄT

- p-Werte statistischer Hypothesentests erlauben nur Aussage über Nichtzufälligkeit eines Effekts und keine Aussage über die Effektstärke.
- Erhöhung des Stichprobenumfang lässt jeden Effekt nicht-zufällig werden.
- Wenn Effektstärke a priori fixiert wird, können Güte oder Stichprobenumfang gewählt werden, wenn Gütefunktion verfügbar ist.
- Inferenzstatistik für Effektgrößen via Konfidenzintervalle (d.h. nicht nur H_0).
- Effektgrößenklassifikation nach Cohen ist zwar intuitiv einleuchtend, aber doch rein pragmatisch mit starken Annahmen (z.B. Normalverteilung).
- Objektive Festlegung der Effektgrößenklassifikation (nur) für qualitative Variablen via Quantile theoretischer Verteilungsfunktion.
- Grenzen der Objektivität: Hyperparameter α der Dirichlet-Verteilung (Wahl: $\alpha = 1/2$ oder $\alpha = 1$).
- Inferenzstatistik für Effektstärke via Bayes-Faktoren.
- Baustelle: Analoges Vorgehen für Korrelationskoeffizient ρ via geeigneter Verteilung ($\rho \sim \beta(a, a)$ mit a entsprechend Jeffreys prior?).