Statistik

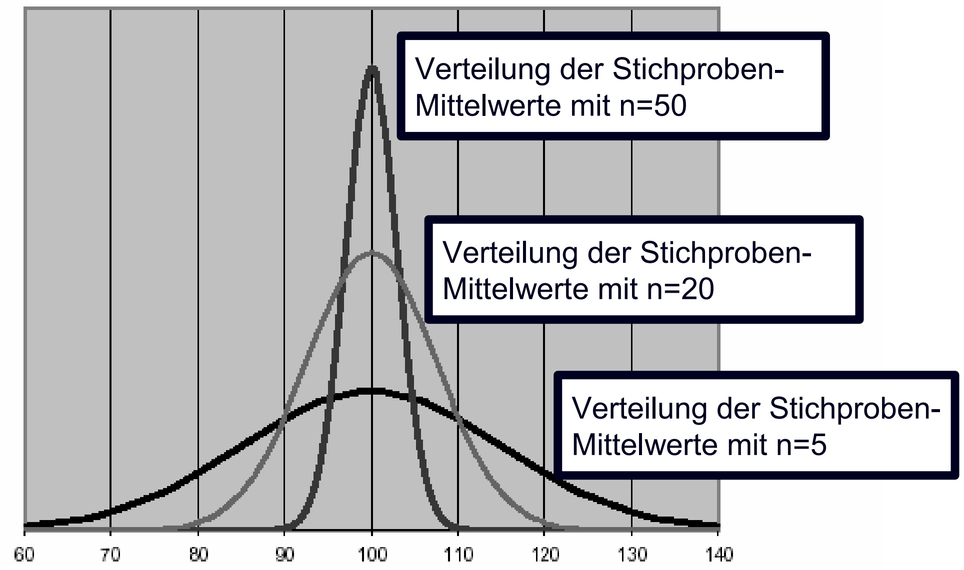
Gregor Dürrenberger
Forschungsstiftung Mobilkommunikation

Hypothesentest

- Vergleich von zwei Datensätzen (z.B. vor Experiment, nach Experiment; Vieltelefonierer, Wenigtelefonierer)
- Signifikanz: können Unterschiede durch Zufall erklärt werden?
- Effektstärke: sind Unterschiede relevant?
- Power: Wie gross ist die Wahrscheinlichkeit, dass ich einen bestimmten Unterschied überhaupt nachweisen kann?

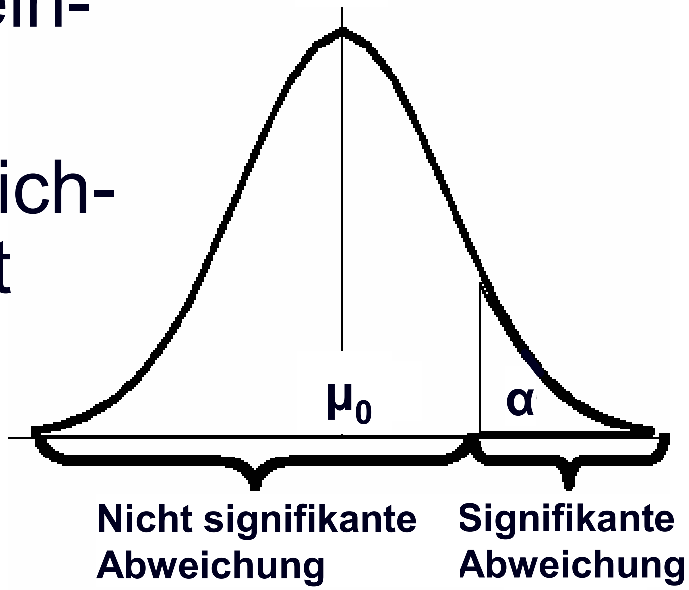
Wahrscheinlichkeitsdichtefunktion

- Verglichen werden die Mittelwerte der Datensätze (Stichproben)
- Annahme: Mittelwerte von Stichproben einer Population streuen um den wahren Mittelwert
  - ➔ Streuung ist Normalverteilt
  - ➔ Je grösser Stichprobe, desto kleiner Streuung



Signifikanz

- Mittelwerte im Zentrum sind viel häufiger als Mittelwerte am Rand
- Konsequenz: Ein Mittelwert am Rand gehört nur mit kleiner Wahrscheinlichkeit zur Population
- Signifikanz: Wahrscheinlichkeit, ab der ein Mittelwert als nicht mehr zur Population gehörend betrachtet wird



α-Fehler

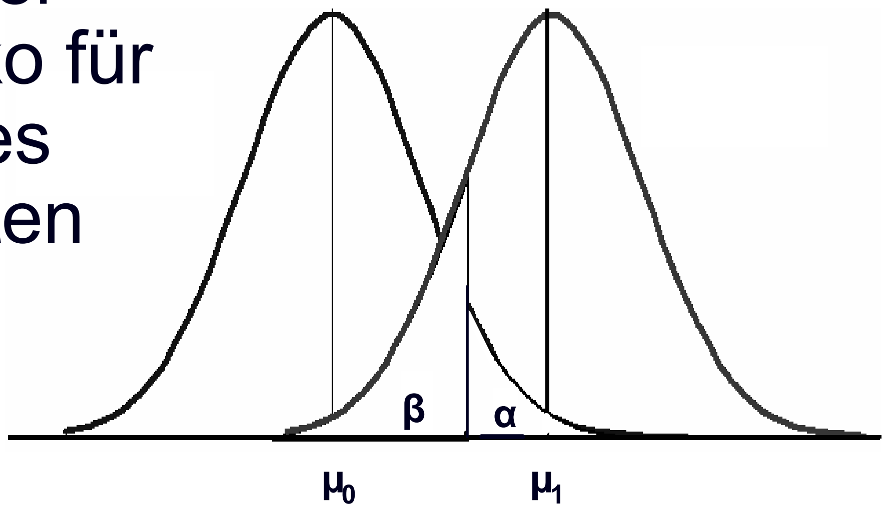
- Signifikanzniveau α legt fest, wann ein Unter-schied als zufällig, wann als gesetzmässig gilt
- Ist α gross, besteht die Gefahr, dass ein Mittel-wert fälschlicherweise als nicht zur Population gehörend klassiert wird (sog. α-Fehler)
- Wahl eines kleinen Signifikanzniveaus ($\alpha \leq 0.05$) gewährleistet, dass diese Fehlklassifikation selten vorkommt (in maximal 5% der Fälle)
  - ➔ Ist ein (Mittel)Wert signifikant, gehört er mit Wahr-scheinlichkeit $p \geq 0.95$ nicht zur Population

Effektgrösse

- Signifikanz sagt nichts aus über die Relevanz eines Unterschieds der Mittelwerte
  - ➔ Bei sehr grossen Stichproben ist die Wahrschein-lichkeitsverteilung der Mittelwerte schmal. Schon kleine Unterschiede können signifikant werden.
- Effektgrösse: In einer Untersuchung muss deshalb zum Voraus bestimmt werden, ab wann ein Unterschied als relevant gelten soll.
- Statistisch: Festlegen der minimalen Differenz der Mittelwerte (μ_0 und μ_1)

β-Fehler

- μ_0 sei bekannt, μ_1 festgelegt. Wenn Signifikanzniveau α hoch, werden die Mittelwerte der zweiten Population die unwahrscheinlicher sind als β als der ersten zugehörig angesehen
- Konsequenz: Je kleiner α, desto grösser Risiko für Fehlklassifikation eines Mittelwertes der zweiten Verteilung
- Dies ist der β-Fehler

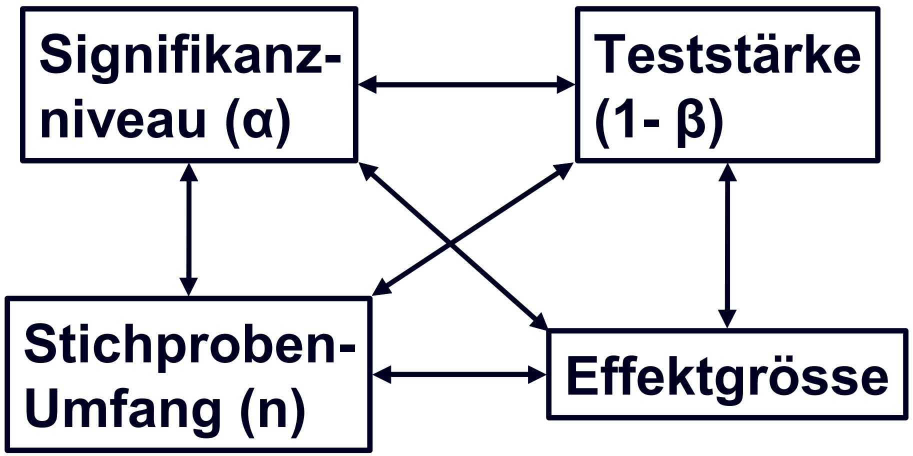


Zusammenhang α-Fehler und β-Fehler

- Wenn Signifikanzniveau reduziert wird, können zuvor erkannte Unterschiede übersehen werden, und umgekehrt.
- Wird die Stichprobengrösse erhöht, sinkt das Risiko von Fehlklassifikationen (β-Fehler)
- Der statistische Nachweis eines Unterschieds kann bei gegebener Effektgrösse und festgelegtem Signifikanzniveau über die Stichprobengrösse optimiert werden (Powerberechnung)

Power

- Power (oder Teststärke 1-β). Wahrscheinlichkeit, mit der ein Unterschied festgelegter Grösse (Effektgrösse) erkannt wird
- Üblich ist eine Power von 80% (bzw. β=4α)
- Power, Effektgrösse, Signifikanz und Stichprobenumfang sind gegenseitig abhängig



Zusammenhänge

- Kleines Signifikanzniveau, kleine Effektstärke
  - ➔ Grosser β-Fehler (Wahrscheinlichkeit von Fehlklassifikation der 2. Verteilung gross)
  - ➔ Kann durch grosse Stichprobe reduziert werden
- Kleines Signifikanzniveau, grosse Effektstärke
  - ➔ Kleine Wahrscheinlichkeit von Fehlklassifikation (β-Fehler)
  - ➔ Stichprobe kann klein gehalten werden
- Kleines Signifikanzniveau, grosse Stichprobe
  - ➔ Bei grosser Effektstärke kleiner β-Fehler
  - ➔ Kleine Effektstärke: bedeutsamer β-Fehler möglich

Hypothesentest

- Vergleich der Wahrscheinlichkeitsverteilungen einer Variable zwischen zwei Gruppen
- α-Fehler: Ablehnen einer richtigen Hypothese (falsch-positiv)
- β-Fehler: Annahme einer falschen Hypothese (falsch-negativ)
- In der Wissenschaft ist die interessierende Hypothese häufig die Nullhypothese: kein Unterschied

Hypothesentest: Allgemeine Darstellung

	Hypothese falsch (Zusammenhang existiert wirklich)	Hypothese richtig (Zusammenhang existiert nicht)
Effekt gefunden (Nullhypothese ab-lehnen)	Richtig-positiv ($1-\alpha$)	Falsch-positiv (α-Fehler)
kein Effekt gefunden (Nullhypothese an-nehmen)	Falsch-negativ (β-Fehler)	Richtig-negativ ($1-\beta$)

Beispiel 1: Diagnosetest

	Krank	Gesund
Test positiv	Richtig-positiv (1-α)	Falsch-positiv (α-Fehler) Fehldiagnose
Test negativ	Falsch-negativ (β-Fehler) Nicht erkannt	Richtig-negativ (1-β)

- Ziel Screening-Tests: möglichst wenige falsch-negative (Krankeit nicht erkannt). Gewünscht ist grosse Power
- Man nimmt dafür in Kauf, dass viele Gesunde positiv getestet werden (→Zweittest für alle positiven Befunde)

Beispiel 2: Gerichtsurteil

	Täter	Unschuldig
Indizien griffig	Richtig-positiv ($1-\alpha$)	Falsch-positiv (α-Fehler)
Indizien nicht griffig	Falsch-negativ (β-Fehler)	Richtig-negativ ($1-\beta$)

- Ziel: wenig Falsch-positive (Unschuldige verurteilt). Hohe Ansprüche an Signifikanz Indizien (täterfreundlich)
- USA: möglichst wenige Falsch-negative (Täter nicht verurteilt). Ansprüche an Signifikanz Indizien tiefer (Tendenz zur Verurteilung von Unschuldigen)

Berechnung der Signifikanz (1)

- Signifikanztest mit Vierfeldertabelle zur Berechnung der Prüfgrösse X^2

	Erfolg	Misserfolg
Probe 1	a	b
Probe 2	g	h

$$X^2=[(a+b+g+h-1)\cdot(a\cdot h-g\cdot b)^2]/[(a+g)\cdot(b+h)\cdot(a+b)\cdot(g+h)]$$

- Falls $X^2 > 3.84$ ist der Unterschied mit weniger als 5% Wahrscheinlichkeit zufällig (also signifikant; $p < 0.05$)
- Notwendige Mindestgrösse der Proben: 6 Fälle

Berechnung der Signifikanz (2)

- Samples klein, Häufigkeitsverteilung unterschiedlich (Erfolg von Probe 1 über 3 mal grösser)

	Erfolg	Misserfolg
Probe 1	9	198
Probe 2	1	80

	Erfolg	Misserfolg
Probe 1	4.3%	95.7%
Probe 2	1.2%	98.8%

- $X^2 = 1.68$
- Unterschied nicht signifikant
- mit 20% Wahrscheinlichkeit zufällig zu Stande gekommen ($p = 0.2$)

Berechnung der Signifikanz (3)

- Samples zehn mal grösser

	Erfolg	Misserfolg
Probe 1	90	1980
Probe 2	10	800

	Erfolg	Misserfolg
Probe 1	4.3%	95.7%
Probe 2	1.2%	98.8%

- $X^2 = 16.8$
- Unterschied wird signifikant
- mit weniger als 0.1% Wahrscheinlichkeit per Zufall zu Stande gekommen

Berechnung der Signifikanz (4)

- Samples zehn mal grösser, Häufigkeitsverteilung ähnlich

	Erfolg	Misserfolg
Probe 1	90	1980
Probe 2	50	760

	Erfolg	Misserfolg
Probe 1	4.3%	95.7%
Probe 2	6.2%	93.8%

- $X^2 = 4.2$
- Unterschied trotzdem noch signifikant
- mit 4% Wahrscheinlichkeit per Zufall zu Stande gekommen ($p = 0.04$)

Powerberechnung (1)

- Das Risiko, einen Unterschied zwischen zwei Messreihen zu übersehen
- Kann nur bestimmt werden, wenn die Effektgrösse bekannt (festgelegt) ist!
- Bei gegebenem Signifikanzniveau und Effekt-grösse lässt sich die minimal notwendige Stich-probengrösse berechnen (Powerberechnung), um den spezifizierten Effekt mit Wahrscheinlichkeit 1-β nachzuweisen

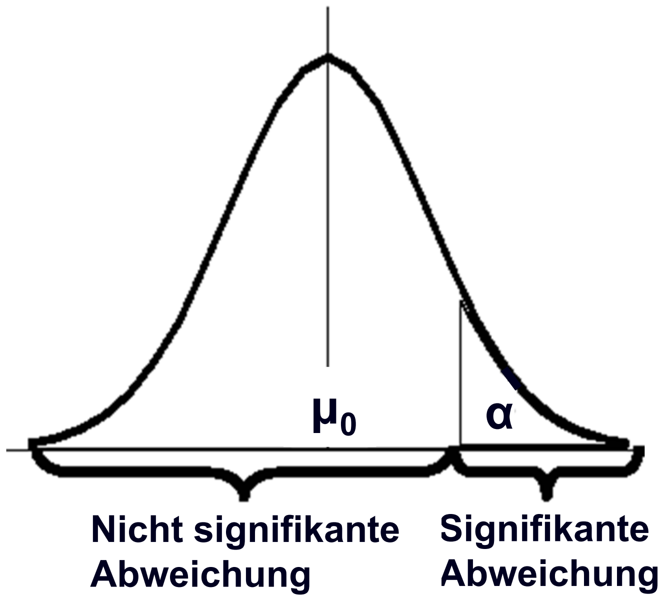
Powerberechnung (2)

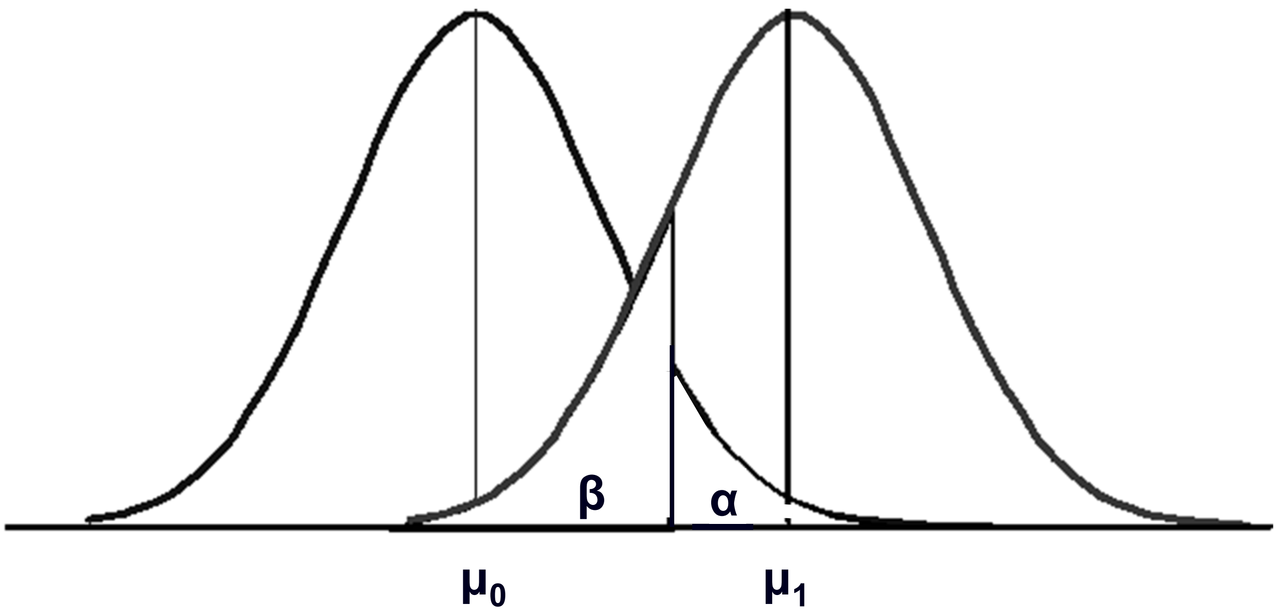
- Samples zehn mal grösser, Häufigkeitsverteilung ähnlich

	Erfolg	Misserfolg
Probe 1	90	1980
Probe 2	50	760

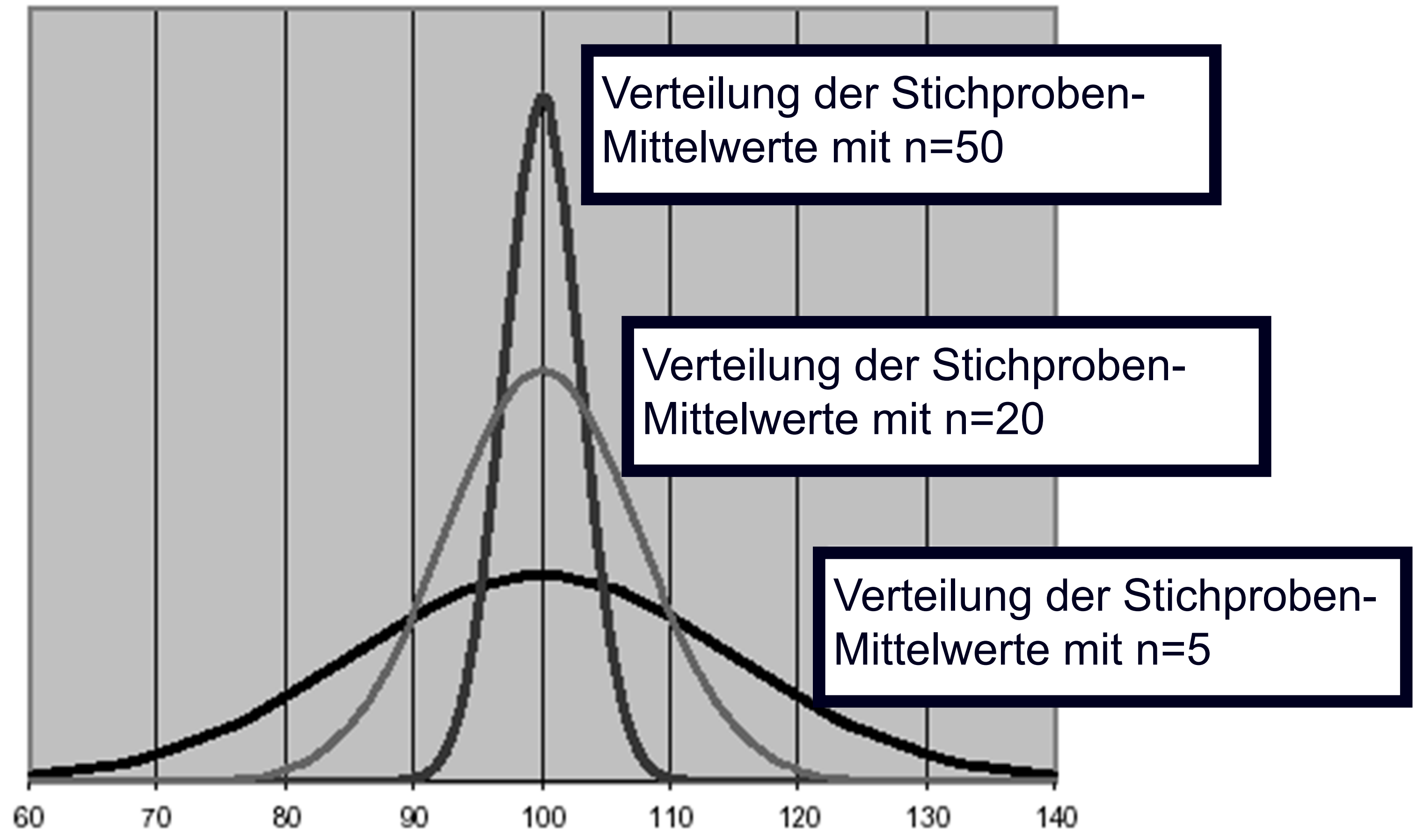
	Erfolg	Misserfolg
Probe 1	4.3%	95.7%
Probe 2	6.2%	93.8%

- $X^2 = 4.2$
- Unterschied trotzdem noch signifikant
- mit 4% Wahrscheinlichkeit per Zufall zu Stande gekommen (p = 0.04)

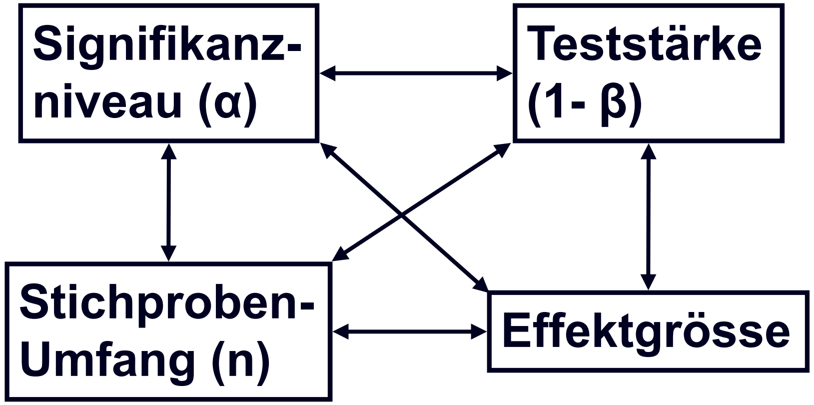

μ_0
α
Nicht signifikante Abweichung
Signifikante Abweichung

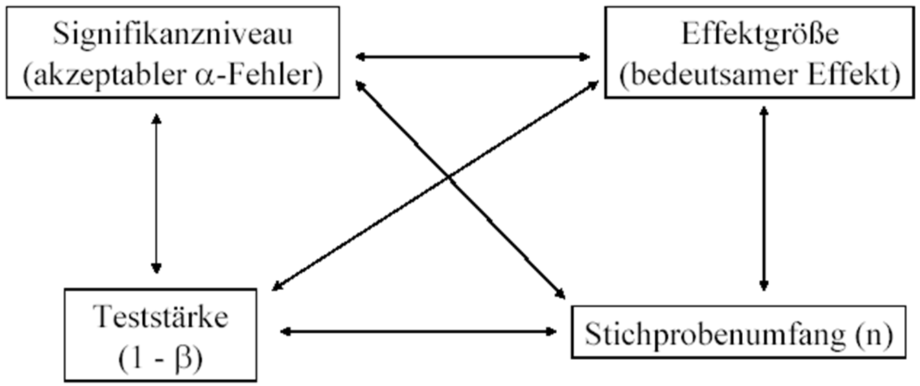


http://www.psych-methoden.uni-koeln.de/statistik_SS2005/vorlesung/vorlesung_2.pdf


Verteilung der Stichproben-
Mittelwerte mit n=50
Verteilung der Stichproben-
Mittelwerte mit n=20
Verteilung der Stichproben-
Mittelwerte mit n=5
60
70
80
90
100
110
120
130
140

				METRISCH				
NOMINAL		ORDINAL		nicht normalverteilt, aber ähnlich		normalverteilt		
unabhängig	abhängig	unabhängig	abhängig	unabhängig	abhängig	unabhängig		abhängig
x2 für: k x l - Felder 2 x 2 Felder	x2 McNemar-Test für: 2 x 2 Felder	Mann-Whitney	Wilcoxon	Mann-Whitney	Wilcoxon	F-Test (Varianzquotiententest) entscheidet über:		t-Test für verbundene Stichproben
						Varianzhomogenität t-Test	Varianzheterogenität Welch-Test	
nichtparametrische Testverfahren						parametrische Testverfahren		


Signifikanz-
niveau (α)
Teststärke
(1- β)
Stichproben-
Umfang (n)
Effektgrösse


Signifikanzniveau
(akzeptabler α-Fehler)
Effektgröße
(bedeutsamer Effekt)
Teststärke
(1 - β)
Stichprobenumfang (n)