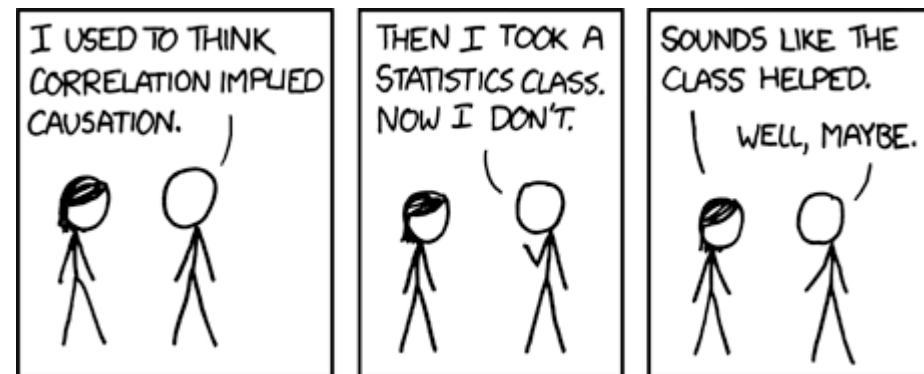


Einführung in die Statistik zur Tier- versuchsplanung und -auswertung

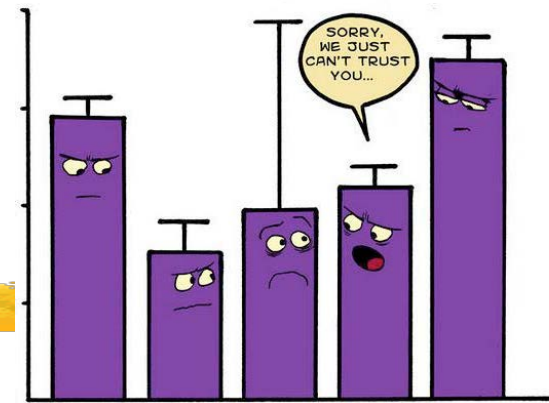
Prof. Peter Pilz
Fachbereich Biologie
Universität Tübingen

So viele Tiere wie nötig,
so wenig Tiere wie möglich



Gliederung

1. Deskriptive Statistik & Power
2. Schließende Statistik Hypothesen & p-Wert
 - Typischer statistischer Test
3. Bsp. Wissen um Statistiktest und Versuchsplanung
4. α - und β -Fehler, Effektgröße, Poweranalyse



GRUNDLAGEN DER
Versuchstier-
kunde

Herausgegeben von
L.F.M. van Zutphen · V. Baumans

1. Deskriptive Statistik

- Kenngrößen

- Zentrale Tendenz:

Mittelwert ↔ **Median**

- **Variation:**

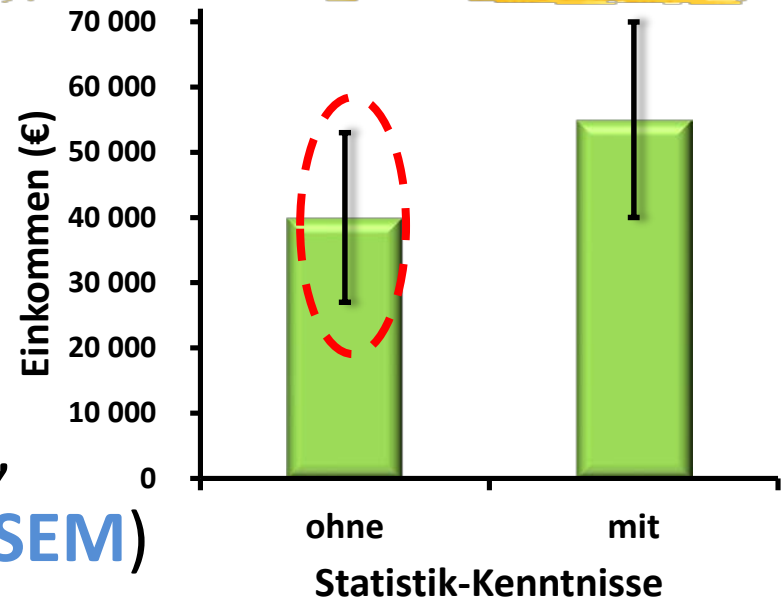
Standardabweichung (**SD**),

Standardfehler (des MW: **SEM**)

(Konfidenzintervalle)

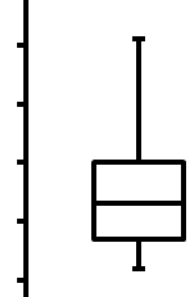
Quartilen

- Stichprobenumfang (**N**)

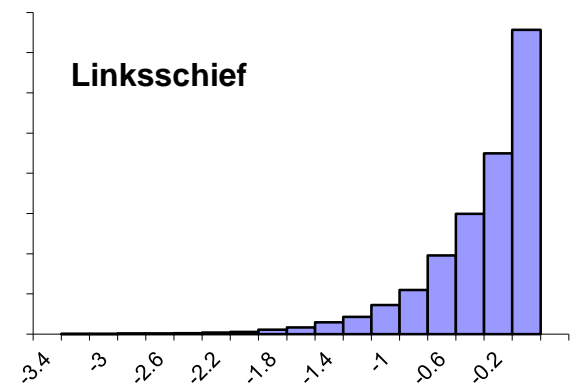
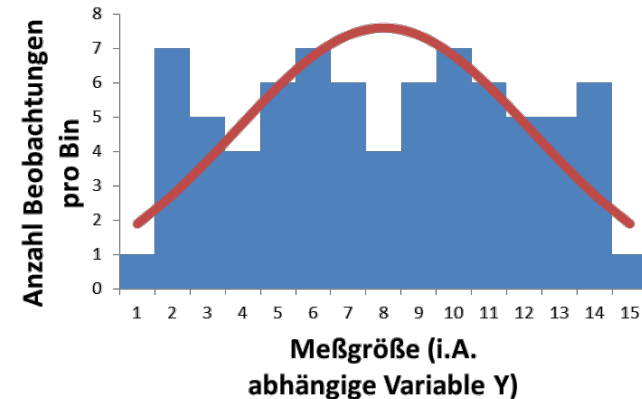




1. Mittelwert ↔ Median



- Mittelwert (parametrisch):
Daten = symmetrisch (Transformation)
sowie Datentyp = kontinuierlich
- Median (nicht-parametrisch):
entw. Daten = nicht symmetrisch
oder Datentyp = ordinal



1. Ordinale Daten → Median?

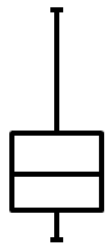
- Schulnoten sind ordinal, z.B. Deutschnote 1 und 3

Rang der Note	Klassenarbeit	Noten Schüler 1	Noten Schüler 2
1	beste	1	3
2	zweitbeste	1.3	3.3
3	Median	3.5	3.5
4	zweitschlechteste	3.7	5.7
5	schlechteste	4	6
	Mittelwert	2.7	4.3

- Median: **weniger Power** (Information) als Mittelwert
=> trotz ordinal: MW statt Median ← viele Stufen (≥8)



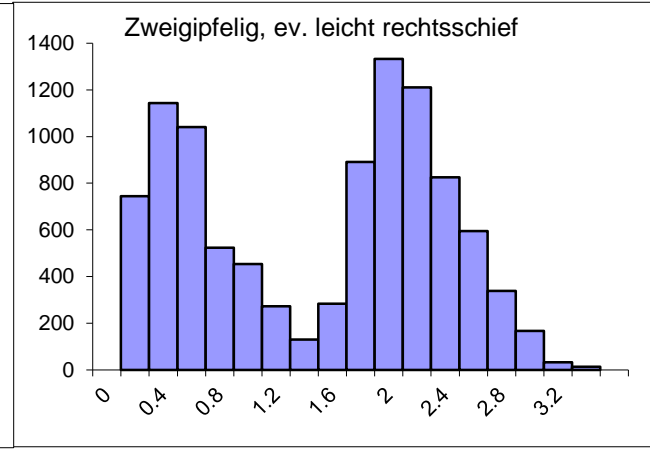
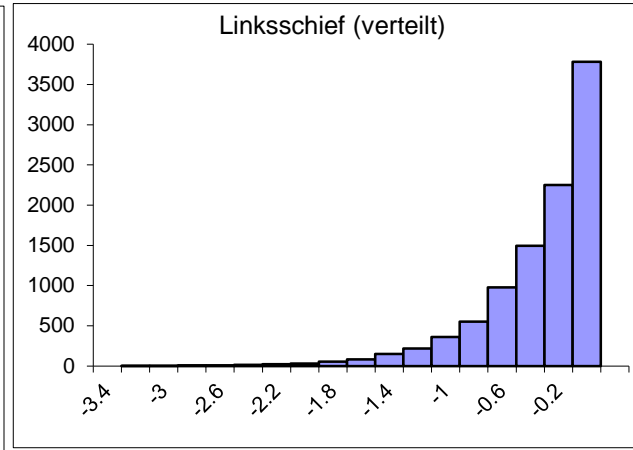
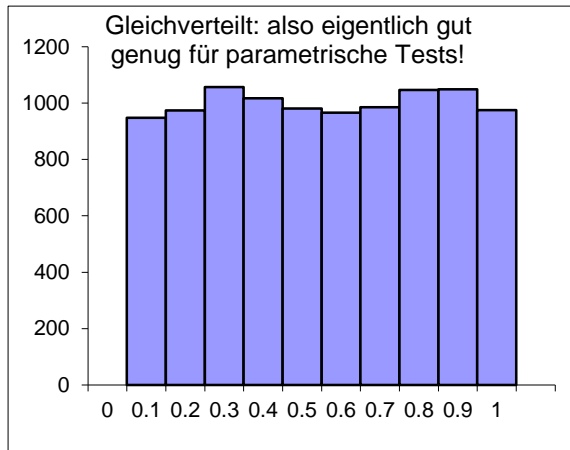
1. Mittelwert ↔ Median



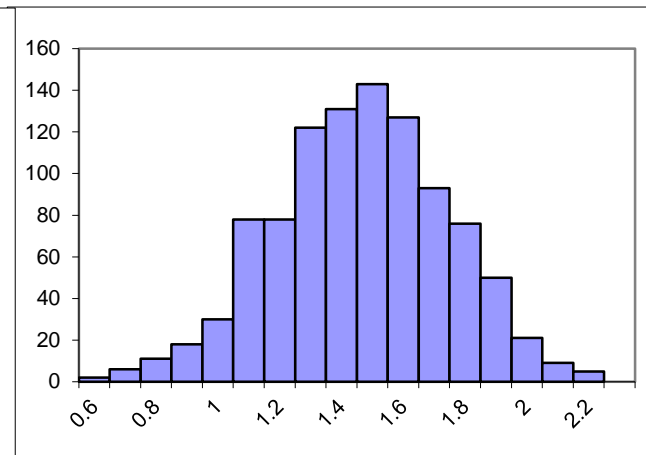
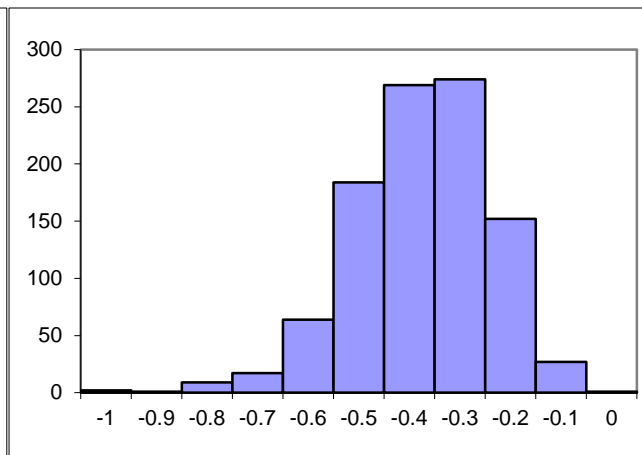
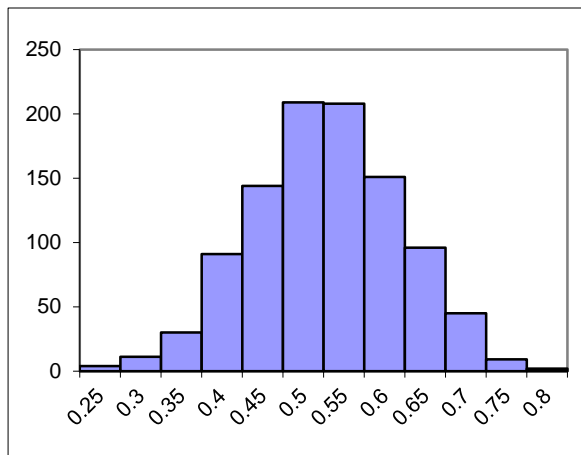
- Mittelwert: **mehr Power (wenn Daten symmetrisch)**
Daten = symmetrisch verteilt (ggf. nach Transformation)
sowie Datentyp = kontinuierlich
oder ordinal: ≥ 8 Gruppierungen [konservativ: ≥ 10]
- Median:
entw. Daten = nicht symmetrisch
oder Datentyp = ordinal

Zentraler Grenzwertsatz = central limit theorem

Oben: jeweils 10 000 Einzelwerte (Wert=X-Achse), div. Verteilungen (Häufigk. = Y-Achse)

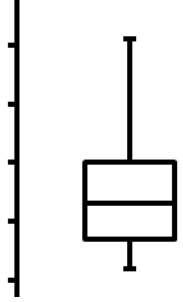


Unten: jeweils 1000 Mittelwerte, jeder davon aus 10 Werten der ursprünglichen 10 000 Einzelwerte oben. Die Verteilungen der Mittelwerte sind annähernd normal verteilt (auf jeden Fall gut genug für jeden parametrischen Test; Bereich X- & Y-Achse wird kleiner: warum?).





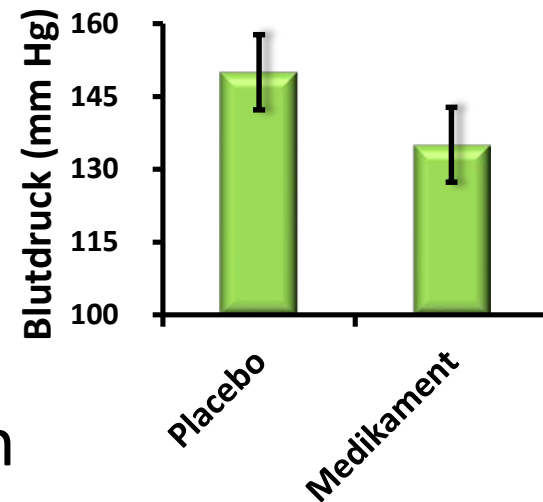
1. Mittelwert ↔ Median



- Mittelwert: **mehr Power (wenn Daten symmetrisch)**
Daten = symmetrisch verteilt (ggf. nach Transformation)
bzw. N hoch (≥ 10 -20/Gruppe), balanciert (ca. gleiche Gruppengrößen)
sowie Datentyp = kontinuierlich
oder ordinal: ≥ 8 Gruppierungen [konservativ: ≥ 10]
- Median:
entw. Daten = nicht symmetrisch *gut bei Ausreißern*
oder Datentyp = ordinal (wenige Gruppen)
- MW \rightarrow parametrische statistische Tests
Median \rightarrow nichtparametrische Tests

2. Statistischer Test: Hypothesen

- Bsp.: 2 Gruppen, Kontrolle & Behandlung
- Frage: Behandlung \rightarrow Unterschied?
- Hypothesen
 - Null-Hypothese H_0 : kein Unterschied
 - Alternativ-Hypothese H_1 : Unterschied
- Arbeitshypothese: darf H_0 oder H_1 sein
- H_0 ist nicht beweisbar (Popper)
 H_1 nur mit Irrtumswahrscheinlichkeit $p > 0$



Nie schreiben: $p=0$
... auch wenn Statistikprogramm das so rechnet,
sondern dann immer: $p < 0.0001$

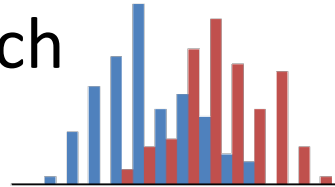
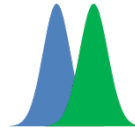
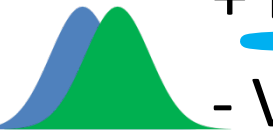
2. Statistischer Test: Prinzip

H1/ Effekt wenn:

+ Effekt hoch

- Varianz niedrig

+ Stichprobe hoch



Bsp.: $n_a = n_b = 6$ (df=10)

t	p
1.8	0.1
2.2	0.05
3.2	0.01
4.6	0.001

Testprinzip:

je höher t (F, r, Chi²...), desto kleiner p

$$t = \frac{\bar{a} - \bar{b}}{Stabw_{pool}} \times \sqrt{N}$$

$$t(10) \geq 2.2 \rightarrow p \leq 0.05 \rightarrow H_1$$

$$t(10) < 2.2 \rightarrow p > 0.05 \rightarrow H_0$$

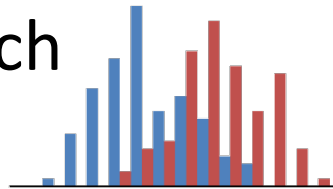
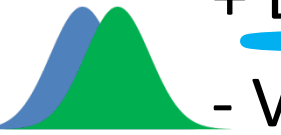
2. Statistischer Test: Prinzip

H1/ Effekt wenn:

+ Effekt hoch

- Varianz niedrig

+ Stichprobe hoch

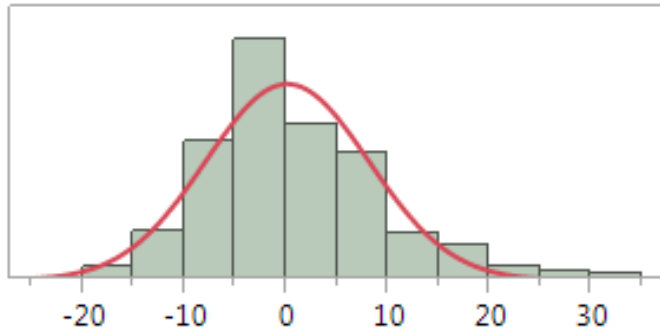


$$t = \frac{\bar{a} - \bar{b}}{Stabw_{pool}} \times \sqrt{N}$$

Effektstärke

Beispiele

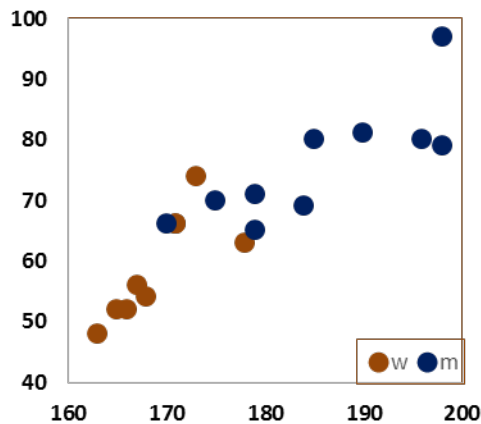
- 277 Studis, Gewicht in kg (Residuen nach Geschlecht)



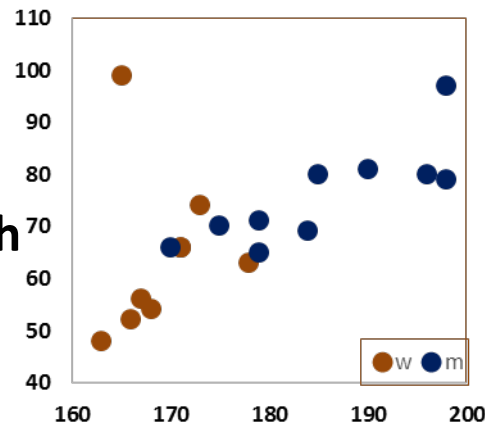
hoch signifikant **nicht** normalverteilt ($p < 0.0001$).
n pro Geschlecht = Gruppe 136 bzw. 141, also ≥ 10 , und balanciert:
Wir nehmen natürlich **parametrische** Statistik!

- n klein ($n \sim 10$ /Gruppe)
ca. symmetrisch

mit Ausreiser



Parametrisch
t-Test,
 $p = 0.0010$
Nicht-parametrisch
Wilcoxon-Test
 $p = 0.0025$



Parametrisch
t-Test,
 $p = 0.063$
Nicht-parametrisch
Wilcoxon-Test
 $p = 0.027$

3. Test? → Versuchsplanung

- „Normaler“ Vgl.

- $t = 0.93$, $p = 0.38$
→ n.s., also H_0

STARTLE	
ohne	mit
290	222
248	151
182	166
92	76
91	51

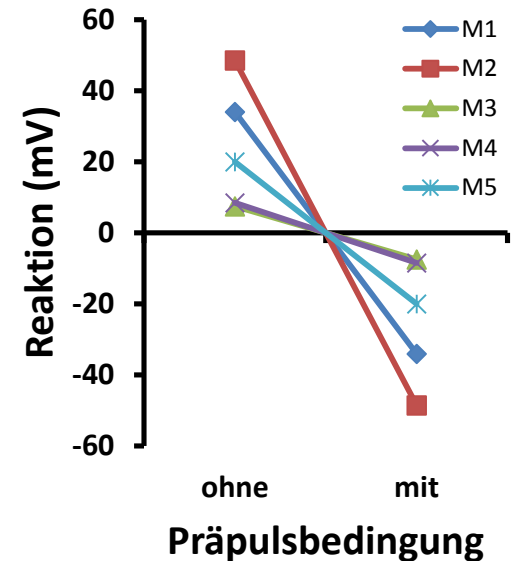
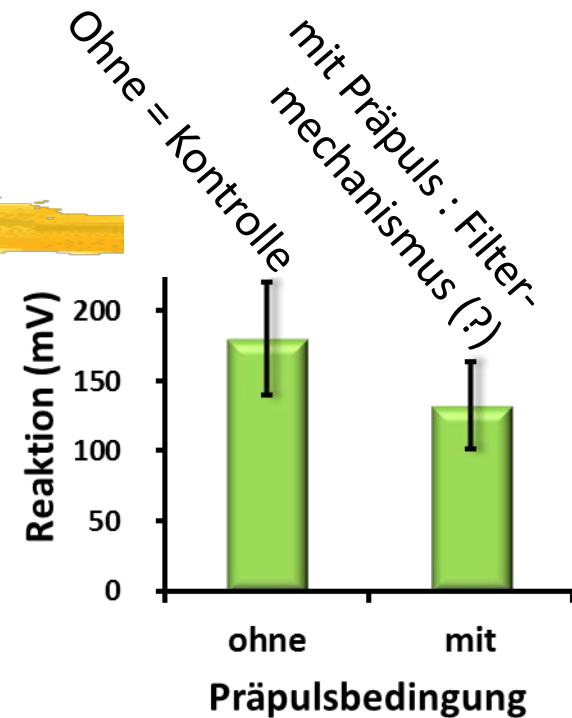
- Korrekter Vgl.

- Messwiederholung **Tier**
- $t = 3.02$, $p = 0.039$ → *, also H_1

- Hälfte Tiere → Signifikanz!

- Vorher wissen = planen!

- Ggf. Tier = seine eigene Kontrolle
- Alter, Gewicht, Geschlecht, ...



3. Versuchsplanung

- Wichtig: gute Versuchsplanung inkl. Kenntnisse des statistischen Testverfahrens
- Bei Planung: Effekte berücksichtigen, die man ev. herausrechnen kann (z.B. Gewicht Tiere)
- Bücher zu Versuchsplanung (z.B. v. Zuthpen et al. 1995) empfehlen u.a. Blockdesigns: mehr Antworten oder weniger Varianz bei gleicher Tierzahl



Übersicht statistische Tests: Vergleich von Mittelwerten bzw. Medianen bzw. Zählungen sowie Verteilung und Korrelation

Faktoren	Stufen	Wdhl.	Parametrisch	Nicht parametrisch	Zählungen
1	1	-	Konfidenzintervall = 1-sample t-Test (Vergleich mit Fixwert)	Wilcoxon rank signed pairs test	Konfidenzintervall: Sokal Rohlf Formel 17.19
1	2	Nein Ja	t-Test (Welch-Test) Dependent t-Test	Man-Whitney-U- Test Wilcoxon rank- signed pairs test (Sign –Test)	(4 Felder, Fisher exakt –Test) χ^2 -Test
1 (2)	≥ 2	Nein Ja	AnoVa (1-faktorielle Varianzanalyse) Tuckey/Dunnet, Bonferroni Repeated measures AnoVa	Kruskal & Wallis – Test U – Test + Bonferroni Friedman –Test	χ^2 -Test
≥ 2	> 2	Egal	Mehrfakt. AnoVa (Wechselwirkungen!)	-	-
1	Kontinu- ierlich	-	Pearson – Korrelation Pearson- Regression	Spearman rank correlation es gibt keine nichtparam. Regress.	Logistische Regression

1 Faktor: z.B. Wirkung ein (1) Medikament mit 2 Stufen (ohne und mit) oder mehreren **Stufen** (0, 1, 3 und 10 mg)

2 Faktoren: z.B. 1 Medikament (1. Faktor) sowie Alkohol (2. Faktor). Wenn wiederholt gemessen wurde, kann man Wechselwirkungen rechnen.

Tukey/Dunnet/Bonferroni: Post-hoc-Tests, macht man, um bei mehreren Stufen den Unterschied zwischen zwei bestimmten Stufen zu testen (z.B. ist 1 mg von 0 mg verschieden?) statt nur die allgemeine Aussage der AnoVa („das Medikament hat eine Wirkung“) zu berichten.

4-Felder-Test: bei **Chi²-Test** mit nur 4 Feldern (2x2) sind verschiedene exaktere Berechnungen von p im Umlauf, u.a. **Fisher's exact test**.

Neben dem Konfidenzintervall einer Zählung gibt es auch Konfidenzintervalle von Verhältnissen von Zählungen (also Prozentzahlen): s. RATIO.XLS © P. Pilz.

Verhältnisse (odds, log-odds etc.): eigenes Statistikkapitel. Oft funktioniert aber einfach Chi².

Parametrische Tests: berücksichtigen die Messwerte
Alle nichtparametrischen Tests: berücksichtigen nur Ränge der sortierten Messwerte, außer:
Vorzeichentest (**Sign-Test**): berücksichtigt sogar nur, ob die Messwerte größer oder kleiner werden.

Generelles lineares Modell (GLM)

Hier sollen mehrere Spezialfälle des GLM besprochen werden. Gerechnet wird GLM in JMP durch Fit Model im Analyze Menü.

1.) repeated measures

Es gibt häufig den Fall, dass von einem Individuum mehrere Messungen genommen werden. Z.B. wird der Einfluss von 3 verschiedenen Pharmaka bei 10 Ratten gemessen. Die Grund-Reaktivität der Ratten ist i.a. sehr verschieden („bias“). I.a. interessiert mich der bias nicht, sondern nur, ob bei jeder individuellen Ratte z.B. Pharmakon A eine relative Reaktivitätserhöhung erzeugt wird. Ich bekomme den Faktor Individuum weg, indem ich ihn als einen (nominalen) Faktor im GLM rechne. Diese Methode reicht, wenn ich 2 Messungen pro Tier durchführe (diese Methode wird für 3 Wiederholungen oft akzeptiert, ist dann aber unsauber). Bei mehreren Wiederholungen eine „Repeated measures“-Anova durchführen.

Es gibt eine nichtparametrische Alternative zur einfaktoriellen repeated measures Anova (also das obige Beispiel): der Friedman-Test.

2.) Covariate

Es gibt öfters den Fall, dass ich einen kontinuierlichen Einflussfaktor habe, der meine Daten beeinflusst, der mich aber nicht interessiert. Beispiel: Das Gewicht von Ratten bei der Reaktivitätsmessung. In diesem Fall kann ich das Gewicht als kontinuierlichen Faktor im GLM mitrechnen. Ich berichte dann diesen Faktor nicht, sondern sage, dass ich ihn als Covariate herausgerechnet habe (im Statistikkapitel von Material und Methode); ich schreibe, dass ich eine Kovarianzanalyse durchgeführt habe (AnCova).

3.) Korrelation mit Messwiederholung

Manchmal interessiert mich nur oder v.a. ein kontinuierlicher Einflussfaktor, z.B. der Einfluss der (kontinuierlichen) Dosierung eines Pharmakons auf die Reaktivität. Bei meiner Messung, die ich mit den verschiedenen Dosierungen immer wieder an denselben 10 Ratten durchgeführt habe, stört aber der individuelle bias der Ratten. In diesem Fall lasse ich das GLM den nominalen Faktor Individuum sowie den kontinuierlichen Faktor Dosis berechnen (als unabhängige Variablen, Einfluss auf abhängige Variable Reaktivität). Ich berichte in diesem Fall nur den Einfluss der Dosierung, und lasse den Faktor Individuum weg. Ich schreibe statt dessen in MatMet, dass ich ein GLM berechnet habe mit dem repeated measures factor Individuum.

Im Prinzip handelt es sich hier um die gleiche Berechnung wie bei der Covariaten, also um eine AnCova. Da Covariate traditionell aber leider für einen kontinuierlichen Faktor benutzt wird, welcher *ignoriert* werden soll, spreche ich von GLM und nicht von AnCova.

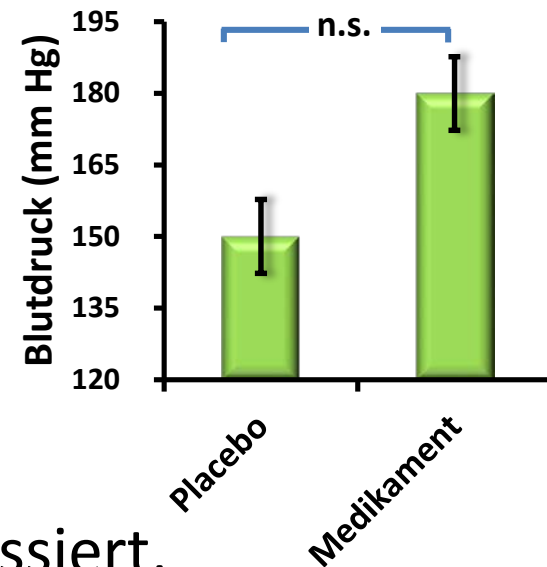
Einige Referees wollen auch für diesen Fall den Mauchly-Test auf Sphericity, und wenn der signifikant ist, die Greenhouse-Geiser-Korrektur. Beides (Mauchly und GG-Korrektur) werden automatisch von den bekannten Statistikprogrammen gerechnet, wenn man ein repeated measures GLM bzw. Anova rechnet.

Kombination von Faktoren

Ich kann die obigen Faktoren kombinieren, d.h. einen repeated measures Faktor Individuum, eine Covariate zum Wegrechnen, sowie eine mich interessierende Korrelation sowie noch einen oder mehrere nominale Faktoren in das GLM gemeinsam mit in's GLM rein nehmen. Ich kann natürlich auch zusätzlich Wechselwirkungen der Korrelation mit einem interessierenden nominalen Faktor rechnen (z.B. ob die Korrelation=Steigung der Reaktivität mit steigender Pharmakonkonzentration vom Geschlecht der Ratte abhängt).

3.) 1-seitiger oder 2-seitiger Test?

- Diese Frage muss vor Datenerhebung gelöst sein!
 - 1-seitig: weniger Tiere benötigt
- Nur Veränderung in eine Richtung interessant → 1-seitiger Test
 - Veränderung in andere Richtung muss als nicht-signifikant publiziert werden!
 - Oft unglaublich wenn $0.05 > p > 0.025$
 - wenn z.B. nur Blutdrucksenkung interessiert, Blutdruckerhöhung aber uninteressant → 1-seitig
- Jede Veränderung interessant → 2-seitiger Test



Teil 4: α - und β -Fehler, Effektstärke, Poweranalyse

Planung Stichprobenumfang
= Anzahl Tiere
(Biometrisches Gutachten)

4. Fehler bei statistischen Tests

- Alpha-Fehler ($\alpha=p$): 5% Irrtumswahrscheinlichkeit = 5% Wahrscheinlichkeit dass man sich irrt (Zufall) ✓
- β -Fehler: zu wenig Power (oder Effekt nicht relevant) ?

Blutdruckmittel (Nullhypothese)		
wirkt nicht (stimmt)	wirkt (stimmt nicht)	Aktzeptiert
o.k.	Typ 2 Fehler β	H0 (wirkt nicht)
Typ 1 Fehler α (p)	o.k.	H1 (wirkt)



Power

Statistische Power

= Wahrscheinlichkeit (vorhandenen) Effekt zu finden

= Wahrscheinlichkeit, falsche H_0 zurückzuweisen

= $1 - \beta$

Beispiel: Blutdrucksenker wirkt (vgl. zu Placebo)

z.B.: $\beta = 0.2 \rightarrow \text{Power} = 1 - \beta = 0.8$

=> In 80% der Fälle statistische Signifikanz

=> In 20% keine Signifikanz (trotz Wirksamkeit)

4. Effektgröße \rightarrow β -Fehler

- β -Fehler/Power kalkulierbar wenn bekannt
 - Effektstärke
 - N (z.B. $N=10+10=20$)
 - α (z.B. $p=0.05$)

Effektstärke

$$d = \frac{\bar{a} - \bar{b}}{Stabw_{pool}}$$

Beispiel:

Blutdruck Placebogruppe = 160 mmHg

Medikamentengruppe = 140 mmHg

SD = 40 mm Hg

Effektstärke = $(160 - 140) / 40 = 0.5$

(eine halbe SD)

Underpowered

- Power von 0.19 → nur in 19% finde ich den Effekt, in 81% finde ich den Effekt nicht!
- Neurobiologie: durchschnittliche Power war¹ 0.21 = viel zu niedrig
- Grund: i.A. Tierzahl zu niedrig
- Beispiel: Geschlecht im Watermaze $d=0.49$
 - Notwendige Tiere, wenn Power = 0.8: 134
 - Im Mittel eingesetzte Tiere: 22



¹: Butten et al. 2013, Nat Neurosc Rev.

4. Effektstärke

- Muss bekannt sein zur Berechnung von Power
- Statistik: Effektgröße in Relation zur Varianz
 - Cohen **d**: Effekt geteilt durch SD
 - Blutdrucksenkung um 1 SD $\rightarrow d=1.0$
 - Zur Berechnung von **d** muss SD bekannt sein
 - z.B. SD des Blutdrucks = 40 mm Hg
mich interessieren Blutdrucksenker ab -20 mm Hg
also ab **d** ≥ 0.5
- Es gibt andere Berechnungsgrößen außer d

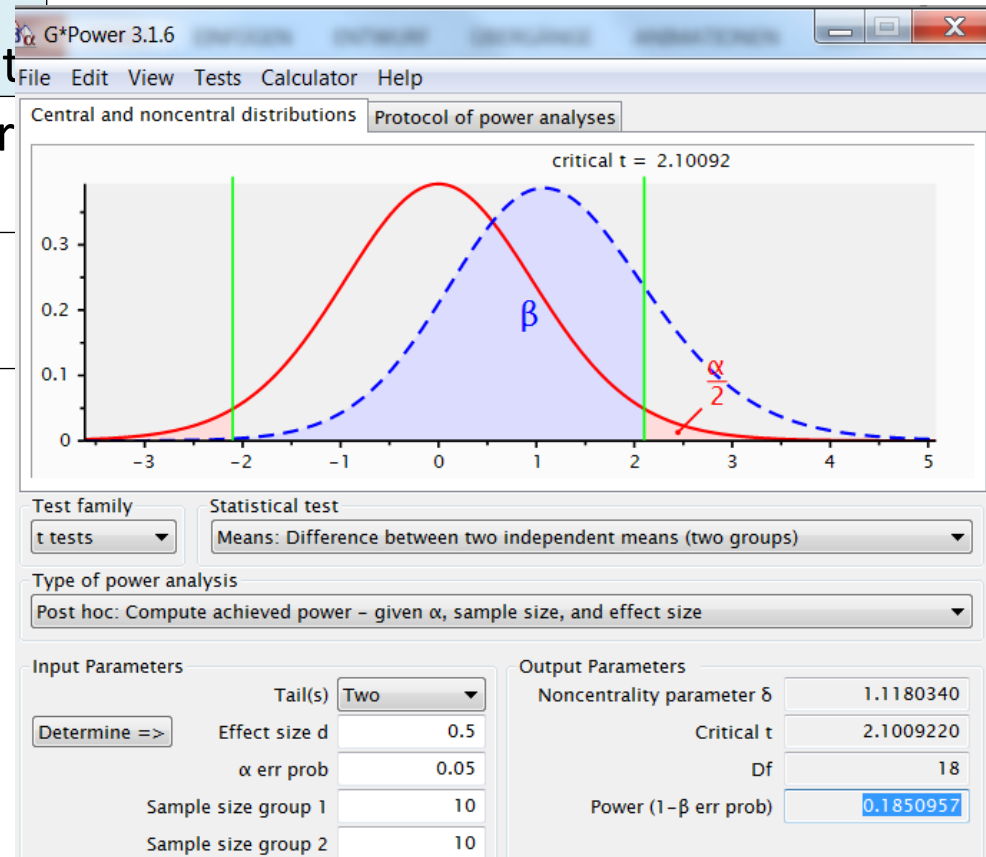
4. Effektgröße festlegen

- Wissen um statist. Effektgröße notwendig für Poweranalyse = Planung der Anzahl Tiere
- Entweder: aus vorherigen Versuchen
 - $SD=40$ mm Hg, Differenz= -20 mm, $d=0.5$
- Oder: festlegen, was sinnvoll/ notwendig/ relevant ist
 - ev. Hilfe: Cohens **d**-Grenzen (Field empfiehlt **r**)

Effekt	d	r	r	d
klein	0.2	0.10	0.1	0.20
mittel	0.5	0.24	0.3	0.63
groß	0.8	0.37	0.5	1.15

4. Effektgröße \rightarrow β -Fehler

	Blutdruckmittel (Nullhypothese)	
Aktzeptiert	wirkt nicht (stimmt)	wirkt (stimmt nicht)
H0 (wirkt nicht)	o.k.	Typ 2 Fehler β
H1 (wirkt)	Typ 1 Fehler α (p)	o.k.



4. Was machen mit β -Fehler?



1. Power des Tests berechnen

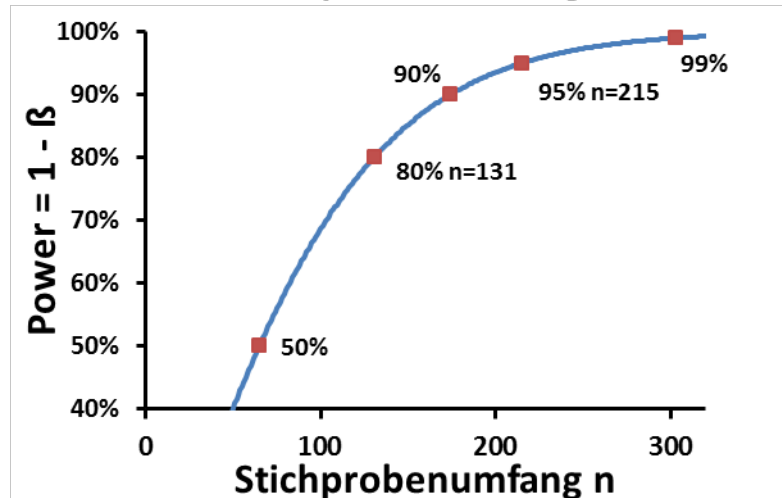
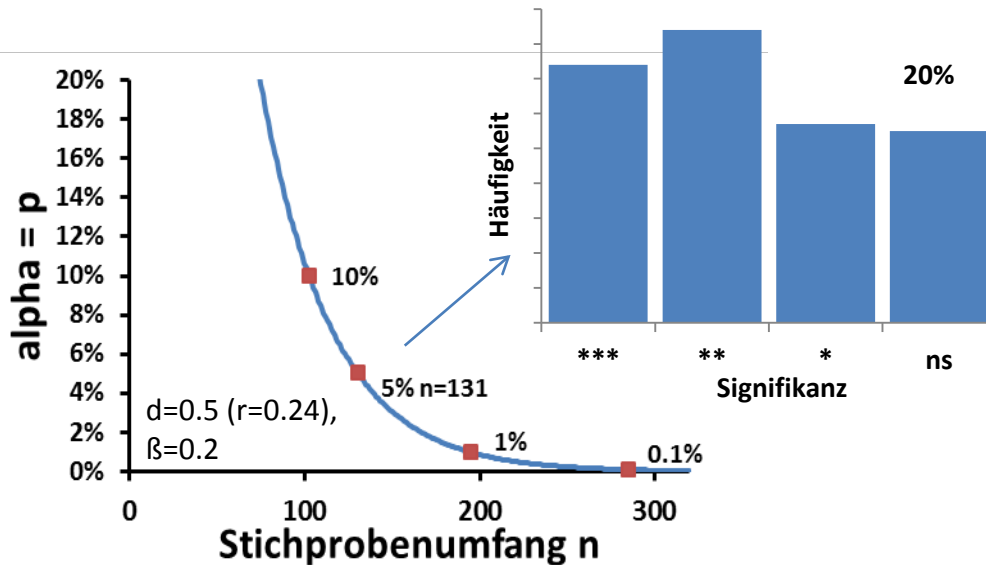
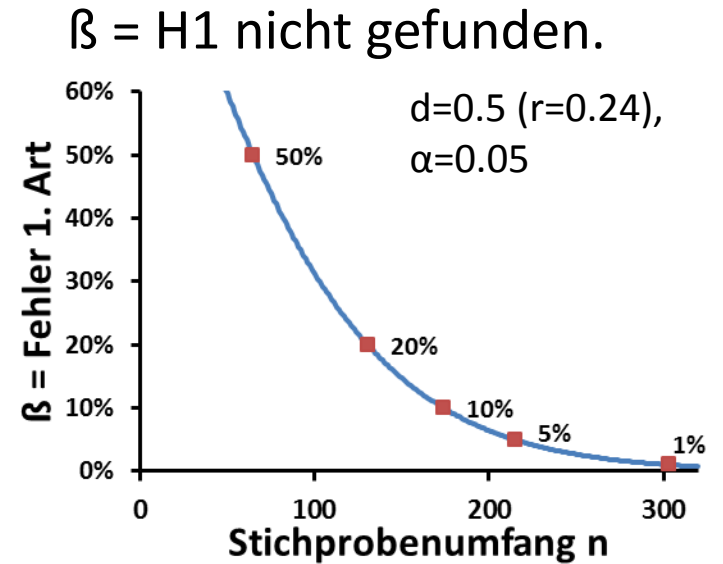
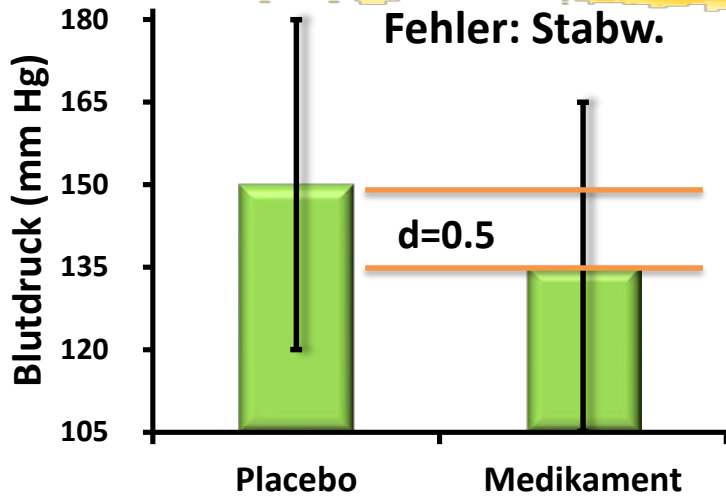
„Power“: Wahrscheinlichkeit, dass der Test einen Effekt der Größe d findet

2. Anzahl Tiere berechnen für Experiment

4. Zwischen-Summary

- 2 Fehlermöglichkeiten, $\alpha + \beta$
- $\alpha = p$ i.A. auf 0.05 festsetzen
- Effektgröße festsetzen
 - aus Literatur bzw. bisherigen Daten oder
 - Vorversuch (wird oft empfohlen: TV-Antrag?)
 - z.B. $d=0.5$ (also Effekt = $\frac{1}{2}$ SD) nach Cohen ($d=0.8$ / $d=1$)
- Jetzt kann man β und die Power bestimmen
 - β = Wahrscheinlichkeit einen Effekt nicht zu finden
 - Power = $1 - \beta$ = Wahrscheinlichkeit, Effekt zu finden

4. β und Power



alpha \rightarrow β -Fehler

- Achtung: wenn α klein (z.B. 0.001 statt 0.05)
 \rightarrow β wird groß (0.84 statt 0.31)
bzw. $N = 48$ statt 20, i.A. inakzeptabel
- Festhalten:
 - $\alpha = p = 0.05$
 - Berechnung β -Fehler (bzw. Power) braucht Effektgröße

4. Signifikanz & Relevanz

Relevanter Effekt: 2 Voraussetzungen

- Signifikanz: wenn $p < 0.05 \rightarrow H1$ bzw. „Effekt“
- Relevanz, d.h. Effektstärke
 - $d \geq 0.2$ „kleiner Effekt“ nach Cohen
 - $d \geq 0.5$ „mittlerer Effekt“
 - $d \geq 0.8$ „großer Effekt“
 - d und p vor Versuch festlegen
 - Blutdruck-Beispiel: $d = 0.5$ =mittlerer Effekt
Blutdrucksenkung um 1 mmHg ist nicht relevant

4. Beispiel

Effekt	Effektgröße	N gesamt	
	r	1-β=0.8	1-β=0.95
klein	0.1	779	1289
mittel	0.3	82	134
groß	0.5	26	42

	d		
klein	0.2 (r=0.1)	779	1289
mittel	0.5 (r=0.24)	131	215
groß	0.8 (r=0.37)	52	84

→ Keine kleinen Effekte suchen (i.A.)

→ Größere Effekte als $r=0.5$ sind natürlich erlaubt und erwünscht!

Bsp. für einseitigen Test: bei $1-\beta=0.8$, $r=\text{groß}$ ist $N=21$

Vorher: Biometrische Planung

- Statistische Hypothesen H_0 und H_1
- Welcher Test? \leftarrow Fachperson
 - ein- oder zweiseitig? $\rightarrow(\leftrightarrow)$ Fachperson
 - parametrisch (MW) oder nicht-parametrisch (Median)
 - wie viele Gruppen (Dosierungen) \rightarrow Fachperson
- Versuchsdesign \leftarrow Fachperson
 - wichtige Neben-Variablen \rightarrow Fachperson
 - Blockdesign, Abfolge, Randomisieren, ...
- N-Zahl (Poweranalyse) \leftarrow Fachperson
 - Avisierte Effektstärke \rightarrow Fachperson
 - $d=0.8/r=0.5$ bei Tierversuchen Standard, höher = besser
 - $1-\beta=0.8$ bei Tierversuchen Standard. N groß wenn $1-\beta=0.95$

Cohen: 0.5 =
Standard in Sozial-
wissenschaften

Power: 0.8?

- Choice of a power level is somewhat **arbitrary** and usually ranges from **80 to 95%**. However, when testing some vaccines for virulence, a power as high as 99% may be specified because of the serious consequences of failure to detect a virulent batch.
- Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals (2002)
- *Michael F. W. Festing and Douglas G. Altman*
- *ILAR Journal* (Institute for Laboratory Animal Research)
- Und weitere dutzende Publikationen



Power = 0.8 !?

- Ab Power 0.8: ab hier Erhöhung n erhöht Power weniger.

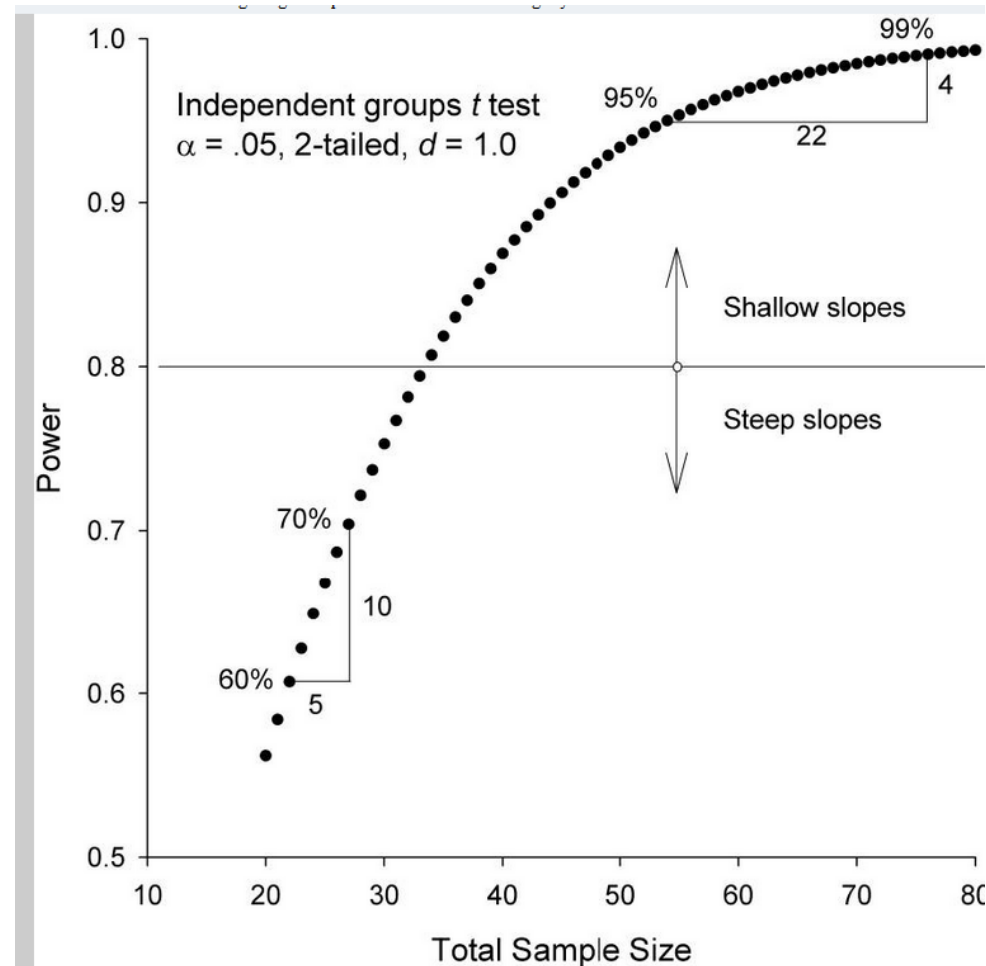
60% → 70% n+5

95% → 99% n+22

$d\text{Power}/n = 2$ bzw. 0.18

- Power > 0.8: Forscher muss rechtfertigen

(Fitts 2011)



Risikoabwägung

80% Power heißt (wenn Gruppenunterschied existiert)

1. Wahrscheinlichkeit für "signifikantes Ergebnis" 80%.
2. in 20% der Fälle KEIN "signifikantes Ergebnis".

→ Risiko-Kosten-Abwägung:

- was kostet Experiment (Zeit, Geld und **Leid** der Tiere)?
- wie wichtig ist es, Effekt zu finden?
- ist es o.k., Effekt mit Risiko =20% zu "verpassen" (also: "nicht signifikant")?

Beispiel: $d=0.8$, 2 Gruppen

- Power = 0.8 → $n = 26$
- Power = 0.9 → $n = 34$, also 30% mehr Tiere
- Dafür halbiert sich Risiko, signifikantes Ergebnis zu verpassen, um 50%

Metanalyse



- Immer berichten:
 - MW, SD oder SEM, n
 - p (i.A. nicht: $p > 0.05$, sondern $p = 0.066$)
 - Statistische Testgröße (t, F, Chi^2 , ...), Freiheitsgrade
- Mit diesen Werten können Metaanalysen der Daten gemacht werden
 - Metaanalyse: umfassen viele Analysen → sehr viel Power

Zusammenfassung

- Mittelwert hat oft mehr Power als Median
- Statistischer Test $H_0 \Leftrightarrow H_1$ ($p \leq 0.05$)
 - Prüfgröße (t) \leftarrow Effektumfang, N, 1/Varianz
 - (t) hoch \rightarrow p niedrig \rightarrow H1 = signifikanter Effekt
- Versuchsplanung: Wissen um Test und ev. Faktoren zum Herausrechnen = wichtig!
- Statistische Effektstärke
Blutdrucksenkung geteilt durch SD. $d=0.5$ bedeutet Effekt = $\frac{1}{2}$ SD
- Mit Effektstärke \rightarrow biometrisches Gutachten \rightarrow N
- Power sollte hoch sein $\geq 80\%$

Anhang

- Poweranalyse: GPower = Opensource-Programm
<http://www.gpower.hhu.de/>
- A. Field: Discovering statistics using SPSS. Meistverkauftes Statistikbuch, nette Beispiele (Aussage über Normalverteilung ist m.E. tlw. Quatsch), für Poweranalyse nur bedingt geeignet
- v. Zupthen et al. Versuchstierkunde, Kapitel 12 „Planung von Tierversuchen“. (Planung Stichprobenumfang = veraltet. Bei mir eine um veraltete Inhalte gekürzte Version erhältlich peter.pilz@uni-tuebingen.de)
- www.ats.ucla.edu/stat/seminars/Intro_power/default.htm erläutert Power, viele Tipps, Power von Experimenten zu erhöhen.
- Glaser, W. Varianzanalyse. UTB. Erläutert, warum man ab $n=10$ fast immer die Normalverteilung annehmen darf.