Promotionsstudiengang
Gesundheits- und Pflegewissenschaften
Partizipation als Ziel von Pflege und Therapie

Methodenkolloquium

Power und Fallzahlberechnung

Ist Fallzahlberechnung Zauberei?

Aus: Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005 Apr 9-15;365(9467):1348-53.

Motivation

Warum braucht man (v.a. in Interventionsstudien) Fallzahlplanungen?

- Weil zu große Studien **???** sind!
- Weil zu kleine Studien **???** sind!

Motivation

Warum braucht man (v.a. in Interventionsstudien) Fallzahlplanungen?

- Weil zu große Studien **unethisch** sind!
- Weil zu kleine Studien **unethisch** sind!

Der diagnostische Test

		Krankheit	
		K+	K-
Diagnostischer Test	T+	**RP**	FP
	T-	FN	**RN**

Maßzahlen für die Güte eines diagnostischen Tests:

Sensitivität: $P(T+ \mid K+)$

W‘t für positives Testergebnis bei einem Kranken

Spezifität: $P(T- \mid K-)$

W‘t für negatives Testergebnis bei einem Gesunden

Der diagnostische Test

		Krankheit	
		K+	K-
Diagnostischer Test	T+	RP	**FP**
	T-	**FN**	RN

Maßzahlen für die Güte eines diagnostischen Tests:

$P(T+ \mid K-)$

W‘t für falsch positives Testergebnis bei einem Gesunden

$P(T- \mid K+)$

W‘t für falsch negatives Testergebnis bei einem Kranken

Der statistische Test

		Wahrheit (Nullhypothese)	
		+ (H_0 richtig)	- (H_0 falsch)
Statistischer Test	ST+ (H0 akzeptieren)	☺	Fehler 2. Art
	ST- (H0 verwerfen)	Fehler 1. Art	☺

Fehler 1. Art: $\alpha = P(ST- \mid H_0+)$

W‘t für Verwerfen der Nullhypothese, wenn Sie vorliegt.

Fehler 2. Art: $\beta = P(ST+ \mid H_0-)$

W‘t für Annahme der Nullhypothese, wenn Sie nicht vorliegt.

Der statistische Test

Durchführung eines statistischen Tests:

- Lege α vor Beginn des Experimentes fest (z.B: $\alpha = 0.05$)
- Nach der Durchführung des Experiments (Daten: x) berechnen wir eine Teststatistik T(x) und entscheiden:

$$\text{Verwerfe } H_0, \text{ falls } P(T \geq T(x) \mid H_0+) < \alpha$$

$$P(T \geq T(x) \mid H_0+) \text{ ist der } \textbf{p-Wert}.$$

Der p-Wert ist die Wahrscheinlichkeit, dass, unter der Annahme dass die Nullhypothese richtig ist, die Teststatistik einen Wert annimmt, der größer oder gleich dem tatsächlich beobachteten Wert ist.

Der statistische Test: p-Wert

Der p-Wert ist proportional zur Wahrscheinlichkeit für das Eintreten der Nullhypothese.

Der p-Wert ist aber **nicht gleich** der Wahrscheinlichkeit für das Eintreten der Nullhypothese.

Wesentlicher Unterschied:

p-Wert wird aus den Daten (**nach der Studie**) berechnet,
α wird **vor der Studie** festgelegt.

Der statistische Test: Fehler 1. und 2. Art

α-Fehler, Fehler 1. Art:
Die Nullhypothese verwerfen, obwohl sie richtig ist.

β-Fehler, Fehler 2. Art:
Die Nullhypothese beibehalten, obwohl sie falsch ist.

Durch das Festlegen des α-Fehler vor Beginn des Experiments haben wir diesen eingeschränkt, d.h. wir machen diesen nur mit 5% Wahrscheinlichkeit.
Wenn der Test also die Nullhypothese ablehnt, können wir mit (mindestens) 95%-iger Wahrscheinlichkeit sicher sein, dass wir keinen Fehler 1. Art machen.

Der statistische Test: Fehler 1. und 2. Art

Aber: Wir können **nur** eine Aussage mit einer festen Sicherheit machen, wenn der Test zur **Verwerfung der Nullhypothese** geführt hat.

Falls der Test die Nullhypothese **nicht** verworfen hat, können wir **keine** Sicherheitsaussage über die Gültigkeit der Nullhypothese machen.

→ Statistische Tests sind **unsymmetrisch**.

Wir können also insbesondere **keine Aussage** über den **Fehler 2. Art** ($\beta = P(ST+ \mid H_0-)$) oder über die **Power** ($1-\beta = P(ST- \mid H_0-)$) des Tests machen.

Der statistische Test: Power

Die **Power** eines stat. Tests ist die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, wenn Sie tatsächlich falsch ist.

$$\text{Power} = 1-\beta = P(ST- \mid H_0-)$$

Weniger technisch: Die **Power** eines stat. Tests ist die Wahrscheinlichkeit, einen tatsächlich vorhandenen Effekt auch zu entdecken.

Der statistische Test: Power

Die Power eines stat. Tests hängt ab von:

- der Fallzahl
- dem Fehler erster Art (α)
- der wahren Größe des Effekts

Ferner:

- bei metrischen Zielgrößen von der Streuung der Werte in der Stichprobe
- bei binären Zielgrößen von der Ereigniswahrscheinlichkeit in der Kontrollgruppe

Statistische Tests: Power

Es gelten folgende Zusammenhänge:

- Fallzahl und α konstant:
  Power **steigt/sinkt** mit der Größe des wahren Effekts?

- Fallzahl und Größe des Effekts konstant:
  Power **steigt/sinkt** mit α?

- α und Größe des Effekts konstant:
  Power **steigt/sinkt** mit der Fallzahl?

Statistische Tests: Power

Es gelten folgende Zusammenhänge:

- Fallzahl und α konstant:
  Power **steigt** mit der Größe des wahren Effekts!

- Fallzahl und Größe des Effekts konstant:
  Power **steigt/sinkt** mit α?

- α und Größe des Effekts konstant:
  Power **steigt/sinkt** mit der Fallzahl?

Statistische Tests: Power

Es gelten folgende Zusammenhänge:

- Fallzahl und α konstant:
  Power **steigt** mit der Größe des wahren Effekts!

- Fallzahl und Größe des Effekts konstant:
  Power **sinkt** mit α!

- α und Größe des Effekts konstant:
  Power **steigt/sinkt** mit der Fallzahl?

Statistische Tests: Power

Es gelten folgende Zusammenhänge:

- Fallzahl und α konstant:
  Power **steigt** mit der Größe des wahren Effekts!

- Fallzahl und Größe des Effekts konstant:
  Power **sinkt** mit α!

- α und Größe des Effekts konstant:
  Power **steigt** mit der Fallzahl?

Fallzahlplanung

Idee: Wenn Power (1-β), Fehler 1. Art (α), Fallzahl und wahren Größe des Effekts dergestalt zusammenhängen, dass jeweils drei von ihnen den vierten determinieren, dann kann man die **benötigte Fallzahl** für eine Studie aus gewünschter Power, gewünschtem Fehler 1. Art (α), und der wahren Größe des Effekts bestimmen.

Fallzahlplanung: Zwei-Gruppen-Vergleich mit binärer Zielgröße

Zugrundeliegender Test: χ^2-Test in Vierfeldertafel
Benötigte (Gesamt!) Fallzahl (s. z.B. Eng, 2003):

$$N = \frac{2 * \left[z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_{1-\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\right]^2}{D^2}$$

mit: p_1, p_2 erwartete Häufigkeiten in den beiden Gruppen, $D = |p_1 - p_2|$ als erwarteter minimaler Differenz, $\bar{p} = (p_1 + p_2)/2$ und z_{1-x} dem (1-x)-Quantil der Standardnormalverteilung (Tabelle s.u.)

1-x	0,8	0,9	0,95	0,975	0,995
z_{1-x}	0,842	1,282	1,645	1,960	2,576

Eng J. Sample size estimation: how many individuals should be studied? Radiology. 2003 May;227(2):309-13.

Fallzahlplanung: Zwei-Gruppen-Vergleich mit binärer Zielgröße

Beispiel: SCAN-Studie (Supportive Cancer Care Networkers - A randomized controlled multicenter trial)

- **Population:** Patienten mit kolorektalem Karzinom nach Klinikentlassung
- **Intervention:** Pflegerisch durchgeführte telefonische Nachbetreuung
- **Kontrolle:** Keine Nachbetreuung
- **Zielgröße:** Verbesserung der Lebensqualität um mindestens 10 EORTC-Punkte zwischen Entlassung und 8 Wochen Nachbeobachtung

Fallzahlplanung: Zwei-Gruppen-Vergleich mit binärer Zielgröße

$p_{\text{Int}}=0{,}65$, $p_{\text{Kontr}}=0{,}5$, $D=0{,}15$, $\bar{p}=0{,}575$

Power$(=1-\beta)=80\%$, $\alpha=5\%$

$$N = \frac{2 * \left[1{,}960\sqrt{2 * 0{,}575 * (1-0{,}575)} + 0{,}842\sqrt{0{,}5 * (1-0{,}5) + 0{,}65 * (1-0{,}65}\right]^2}{0{,}15^2}$$

$$=338{,}6$$

(Kontrolle mit SAS PROC POWER, Nquery 7.0: 340)

Ergebnis: Schließe insgesamt 340 Patienten ein, 170 pro Gruppe.

Fallzahlplanung: Zwei-Gruppen-Vergleich mit binärer Zielgröße

Interpretation:

Wenn die wahre Erfolgsw‘t in der Kontrollgruppe 50% ist, die in der Interventionsgruppe 65%, dann finde ich diesen Unterschied von 15%-Punkten mit 80%-iger Wahrscheinlichkeit (mit einem χ^2-Test zum α=5%-Niveau), wenn ich 340 Patienten in die Studie einschließe.

Fallzahlplanung: Woher kommt der Wert für den minimalen klinisch relevanten Unterschied?

Nicht aus der Literatur und nicht vom Biometriker!

Stephen Senn (Dicing with Death, 2003, S.96):
“If the trial is negative, it is quite likely that this project will be cancelled. This means that the intervention will **never** be studied again. It will be lost to mankind **forever**. Bearing in mind that there are other interventions (waiting to be) studied what is the maximum level of a effect at which we would be able to tolerate the loss of such an intervention.”

Fallzahlplanung: Woher kommt der Wert für den minimalen klinisch relevanten Unterschied?

Nicht aus der Literatur und nicht vom Biometriker!

“Wenn sich Ihre Intervention als überlegen herausgestellt, brauchen Sie XXX.000 €, um diese klinik-/landes-/bundesweit einzuführen.

Welche Größe des Effektes (in der Sprache der Zielgröße) überzeugt die Pflegedienstleistung/das Gesundheitsministrium/den Rentenversicherungsträger …, Ihnen diese Summe zur Verfügung zu stellen.”

Fallzahlplanung: Zwei-Gruppen-Vergleich mit stetiger Zielgröße

Zugrundeliegender Test: (Unabhängiger) t-Test
Benötigte (Gesamt!) Fallzahl (s. z.B. Eng, 2003):

$$N = \frac{4\sigma^2\left[(z_{1-\alpha/2} + z_{1-\beta}\right]^2}{D^2}$$

mit: σ als (in beiden Gruppen als gleich angenommene) Standardabweichung der Zielgröße, $D = |\mu_1 - \mu_2|$ als erwarteter minimaler Differenz der Mittelwerte und z_{1-x} dem (1-x)-Quantil der Standardnormalverteilung (Tabelle s.u.)

1-x	0,8	0,9	0,95	0,975	0,995
z_{1-x}	0,842	1,282	1,645	1,960	2,576

Eng J. Sample size estimation: how many individuals should be studied? Radiology. 2003 May;227(2):309-13.

Fallzahlplanung: Zwei-Gruppen-Vergleich mit stetiger Zielgröße

Beispiel: KMT-Studie

- **Population:** Patienten mit hämatopoetischer Stammzelltransplantation (HSCT)
- **Intervention:** Somato-psychosoziale Pflegeintervention mit den Modulen Aktivitätsförderung, orale Mukositisprophylaxe und Appetitförderung & Ernährung
- **Kontrolle:** Klinikübliche Pflege
- **Zielgröße:** Globale gesundheitsbezogene Lebensqualität (EORTC) bei Entlassung

Fallzahlplanung: Zwei-Gruppen-Vergleich mit stetiger Zielgröße

σ=16, *D=10,* Power(=1-β)= 80%, α=5%

$$N = \frac{4\sigma^2\left[(z_{1-\alpha/2}+z_{1-\beta}\right]^2}{D^2} = \frac{4*16^2*[(1{,}960+0{,}842]^2}{10^2} = 70{,}6$$

(Kontrolle mit SAS PROC POWER, Nquery 7.0: 74)

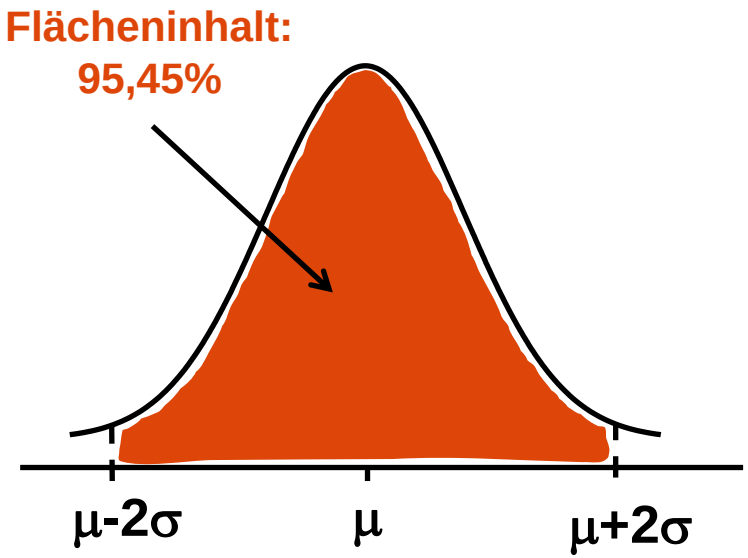
Ergebnis: Schließe 72 Patienten ein (36 pro Gruppe)

Fallzahlplanung: Zwei-Gruppen-Vergleich mit stetiger Zielgröße

Interpretation: Wenn die wahre Differenz der EORTC-Werte bei Entlassung 10 Punkte beträgt (mit einer Standardabweichung von 16 Punkten in beiden Gruppen), dann finde ich diesen Unterschied mit 80%-iger Wahrscheinlichkeit (mit einem t-Test zum α=5%-Niveau), wenn ich 72 Patienten in die Studie einschließe.

Fallzahlplanung: Woher kommt der Wert für die angenommene Standardabweichung?

Aus der Literatur oder aus der 2σ-Regel, aber nicht vom Biometriker!



2σ-Regel: Bei normalverteilten Daten liegen 95% der Beobachtungen zwischen μ-2σ und μ+2σ

Idee:

- Schätze die Länge des Bereichs ab, in dem 95% der **beobachteten** (Kein KI!) Werte liegen, dieser hat die Länge 4σ.
- Teile diese Länge durch 4 und erhalte eine Abschätzung für σ

Anmerkungen

- **Statistische Power ist nicht dasselbe wie Repräsentativität!**
  Die Frage „Wie groß muss die Fallzahl sein, damit ich eine repräsentative Stichprobe habe?“ ist nicht beantwortbar.
  Ob die Stichprobe repräsentativ ist, weiß man erst, wenn man sie hat!
- Fallzahlplanungen für Ein-Stichproben-Studien (z.B. wie groß muss die Fallzahl sein, wenn ich die Sensitivität eines diagnostischen Tests schätzen will) sind möglich und weniger kompliziert (nutze maximal tolerierbare Länge des 95%-KIs)
- Fallzahlplanungen für kompliziertere Modelle (z.B. Regression) sind möglich, es ist allerdings i.d.R. unmöglich, die notwendige Vorinformation beizubringen.

Anmerkungen

- Retrospektive Power-Betrachtungen sind sinnlos!

Smith AH, Bates MN. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiology. 1992 Sep;3(5):449-52.

Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994 Aug 1;121(3):200-6.

Levine M, Ensom MH. Post hoc power analysis: an idea whose time has passed? Pharmacotherapy. 2001 Apr;21(4):405-9.

Hoenig JM, Heisey DM, 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 55:19-24.

Senn SJ. Power is indeed irrelevant in interpreting completed studies. BMJ. 2002 Nov 30;325(7375):1304.