

# Einführung in die Biostatistik mit R

Ambrus Kaposi

Institut für Entwicklung und Weiterbildung  
auf dem Gebiet der Medizinischen Informatik,  
Semmelweis Universität,  
Budapest, Ungarn

2011.03.29

# Outline

Allgemeine Informationen

Grundgedanke der Biostatistik

Klinische Studien

- Planung

- Durchführung

- Auswertung

- Phasen einer Arzneimittelstudie

R

# Outline

## Allgemeine Informationen

## Grundgedanke der Biostatistik

## Klinische Studien

Planung

Durchführung

Auswertung

Phasen einer Arzneimittelstudie

R

# Kontakt

- ▶ Ambrus Kaposi MD
- ▶ <http://akaposi.web.elte.hu>
- ▶ [kaposi.ambrus@gmail.com](mailto:kaposi.ambrus@gmail.com)

# Noten

- ▶ letzte Vorlesung:
  - ▶ multiple-choice test
  - ▶ jede Vorlesung im ganzen Semester mit gleichem Gewicht
- ▶ zusätzliche Punkte von Aufgaben
  - ▶ an Ende meiner Vorlesungen
  - ▶ Eingabe durch Email
  - ▶ können Note verbessern

# Outline

Allgemeine Informationen

Grundgedanke der Biostatistik

Klinische Studien

Planung

Durchführung

Auswertung

Phasen einer Arzneimittelstudie

R

# Was ist Statistik?

- ▶ “Es gibt drei Arten von Lügen: Lügen, große Lügen und Statistiken.” (Mark Twain)
- ▶ “Ich traue keiner Statistik, die ich nicht selbst gefälscht habe.” (Fälschlich Winston Churchill zugeschrieben. Aller Wahrscheinlichkeit nach eine Erfindung von Joseph Goebbels oder aus seinem Propagandaministerium im Zusammenhang mit der Weisung an die deutsche Presse, Churchill als Lügner hinzustellen.)



(Source des Bildes: <http://complex.elte.hu/elemistisztika>)

# Was ist Statistik?

- ▶ nicht Mathematik
- ▶ sondern Sammlung von Methoden mit Konditionen, wenn man eine Methode benutzen darf
- ▶ wir werden über solche Methoden und Konditionen lernen



# Wozu braucht ein Arzt Statistik?

- **zum Verstehen der medizinischen Fachliteratur**  
(„How to Read a Paper“)  
insbesondere von **Originalarbeiten in Fachzeitschriften**  
über
  - experimentelle
  - klinische
  - epidemiologische
  - sonstige (z. B. gesundheitsökonomische) Studien
- **„Evidence-based Medicine“**  
**Bewertung und Kommunikation von Chancen und Risiken**
- **bei eigenen Untersuchungen**
  - Doktorarbeit
  - Industrie
  - Gesundheitsbehörden

(Source: <http://imsieweb.uni-koeln.de/lehre/epidemiologie>)

# Grundgedanke

- ▶ Frage: ist die Medizin wirksam?
- ▶ **def.** Wahrscheinlichkeit informell: günstige Fälle / alle Fälle
- ▶ Frage: was ist die Wahrscheinlichkeit, dass die Medizin wirkt?

		geheilt	krank geblieben
▶ Antwort: Stichprobe	ohne Medizin	3	7
	mit Medizin	5	5

- ▶ Neue Frage:
  1. Nehmen wir an die Medizin ist nicht wirksam. (*Nullhypothese*)
  2. Was ist die Wahrscheinlichkeit, dass von der Stichprobe wir eine solche oder mehr extreme Ergebnis bekommen? (*p-Wert, Signifikanzwert, p-value*)
- ▶ Wenn der p-Wert gering ist, die Nullhypothese ist wahrscheinlich falsch  $\implies$  der Gegensatz ist richtig (*Alternativhypothese*)
- ▶ Beispiele: später

# Outline

Allgemeine Informationen

Grundgedanke der Biostatistik

## Klinische Studien

Planung

Durchführung

Auswertung

Phasen einer Arzneimittelstudie

R

# Klinische Studien

- ▶ experimentelle Projekte
- ▶ prospektiv geplant und durchgeführt
- ▶ Ziel: Bestimmung optimaler Behandlungen für zukünftige Patienten
- ▶ Behandlungs-, Kontrollgruppe

	geheilt	krank geblieben
ohne Medizin	3	7
mit Medizin	5	5

- ▶ Randomisierung
- ▶ Verblindung

# Outline

Allgemeine Informationen

Grundgedanke der Biostatistik

Klinische Studien

Planung

Durchführung

Auswertung

Phasen einer Arzneimittelstudie

R

# Formalisierung der Fragen

- ▶ Wir beobachten ein bestimmtes Ergebnis (z. B. Response-rate). Was ist die Präzision? (Konfidenzintervall)
- ▶ Ist die Medizin wirksam? Mit welcher Wahrscheinlichkeit? (p-Wert)
- ▶ Prädiktion: Prognose für zukünftige Patienten? (Konfidenzintervall)

# Irrtum

Fehler in Hypothesentests		REALITÄT	
		Hyp. richtig	Hyp. falsch
ENTSCHEIDUNG	Hypothese ablehnen	Fehler 1. Art $\alpha$	✓
	Hypothese annehmen	✓	Fehler 2. Art $\beta$

(Source des Bildes: <http://www.campegym.de/pmwiki/pmwiki.php?n=Ma.HypoTest>)

- ▶ Fehler 1. Art: wenn die in Wirklichkeit richtige Nullhypothese fälschlicherweise abgelehnt wird (siehe auch p-Wert)
- ▶ Fehler 2. Art: wenn die in Wirklichkeit falsche Nullhypothese irrtümlich akzeptiert wird

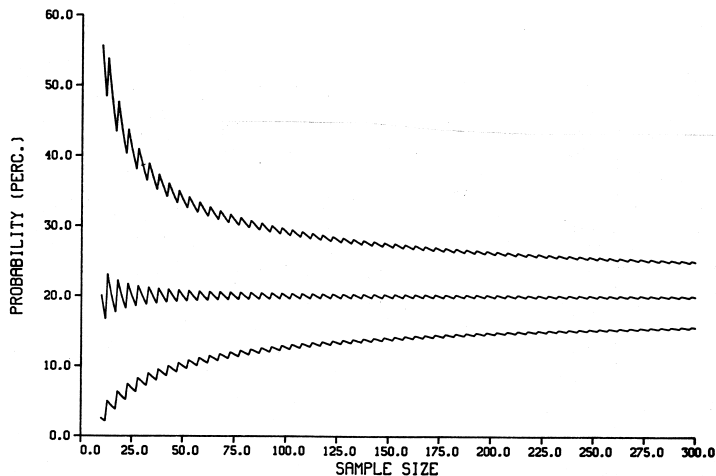
# Fallzahlplanung

- ▶ Eine Pharmafirma hat ein neues Medikament entwickelt, von dem vermutet wird, dass es die Heilungschance bei einer bestimmten Krankheit von 20% (Erfolgschance bei Standardmedikation) auf 30% erhöht.
- ▶ Man plant eine Studie mit  $n$  Patienten.
- ▶ Wie groß ist die Chance, den vermuteten Effekt (10%) nachzuweisen?
- ▶ **def. Power:** Was ist die Wahrscheinlichkeit, dass bei Gültigkeit einer bestimmten Alternativhypothese, diese mit dem Test auch tatsächlich erkannt wird?
- ▶ Bei  $n=30$  Patienten: Power=40%
- ▶ Bei  $n=100$  Patienten: Power=77%



## 95% Konfidenzintervall für Ansprechrates von 20%

P20 ALPHA=.05



# Probenahme

► gemütlich



► systematisch



► geschichtet (Stratifikation)



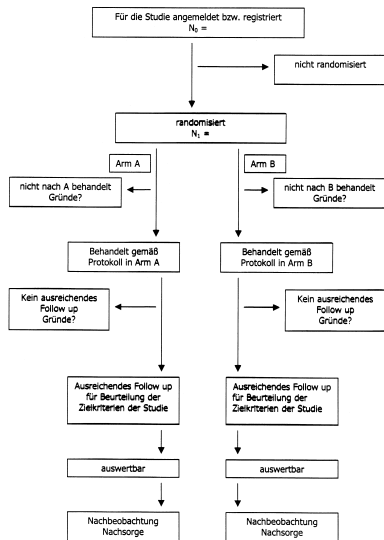
► randomisiert (kann einfach “gemütlich” bekommen)

(Source der Bilder: <http://complex.elte.hu/elemistatiztika>)

# Verblindung

- ▶ Verblindung des Patienten (einfach blind)
- ▶ Verblindung des Patienten und des Arztes (doppelblind)
- ▶ Verblindung spielt bei Zwischenauswertungen:
  - ▶ Independent (Data) Review Board
  - ▶ Studienbiometrie wertet verblindet aus oder gibt die Auswertung an unabhängige Dritte ab

# Patientenfluss



(Source: Jörg Rüdiger Siewert: Praxis Der Viszeralchirurgie: Onkologische Chirurgie, 2. Band, Springer 2006)

# Outline

Allgemeine Informationen

Grundgedanke der Biostatistik

## Klinische Studien

Planung

Durchführung

Auswertung

Phasen einer Arzneimittelstudie

R

# Datenerhebung, Dokumentation

- ▶ das Studienziel bestimmt die Dokumentation und nicht umgekehrt
- ▶ einzigartige Identifizierung für jeden Patient
- ▶ häufige Verwechslungen von Datumsangaben:
  - ▶ Was bedeutet 03.04.10?
  - ▶ Aufnahme in die Studie, Randomisation, Untersuchung vor Behandlungsbeginn, erstes Labor nach Behandlungsbeginn
  - ▶ Abschluss der Therapie und Abschluss der Studie
  - ▶ Ausfülldatum des Arztes und letztem Follow-up Datum des Patienten

# Outline

Allgemeine Informationen

Grundgedanke der Biostatistik

## Klinische Studien

Planung

Durchführung

**Auswertung**

Phasen einer Arzneimittelstudie

R

# Auswertung

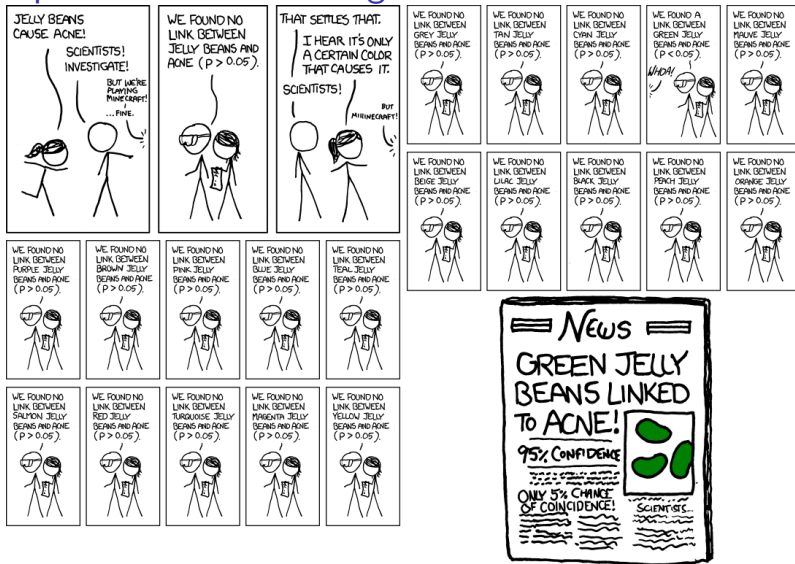
- ▶ “automatisch” nach der Planung
- ▶ verblindet
- ▶ Multiplizität: wendet man Test auf dem Signifikanzniveau 5% zweimal, so ist die Chance von falsch positivem Ergebnis größer, als 5%
  - ▶ mehrfache Endpunkte, Frage
  - ▶ mehrfache Behandlungen
  - ▶ Zwischenauswertungen
  - ▶ Untergruppen
  - ▶ usw.

(*Alphafehler-Kumulierung*, Lösung: z. B. *Bonferroni-Korrektur*)

- ▶ später: statistische Methoden



# Alphafehler-Kumulierung



(Source des Bildes: <http://xkcd.com>)

# Outline

Allgemeine Informationen

Grundgedanke der Biostatistik

Klinische Studien

Planung

Durchführung

Auswertung

Phasen einer Arzneimittelstudie

R

# Phasen

Phase	Personen	Dauer	Hauptziel
0	ca. 10 – 15	Wochen	Pharmakokinetik, Pharmakodynamik, Tests mit subtherapeutischen Dosen, z. B. Microdosing
I	ca. 20 – 80	Wochen bis Monate	Pharmakokinetik, Pharmakodynamik, Verträglichkeit und Sicherheit des Medikaments
II	ca. 50 – 200	Monate	Überprüfung des Therapiekonzepts (Proof of Concept, Phase IIa), Findung der geeigneten Therapie-dosis (Dose Finding, Phase IIb), positive Effekte der Therapie sollten zu beobachten sein
III	ca. 200 – 10.000	Monate bis Jahre	Signifikanter Wirkungsnachweis (Pivotal Study) und Marktzulassung der Therapie; nach Marktzulassung werden laufende Studien dann zu IIIb-Studien
IV	ab ca. 1000 bis Millionen	Jahre	Erfolgen mit bereits zugelassenen Medikamenten in der zugelassenen Indikation. Zulassungsbehörden verlangen oftmals derartige Studien, z. B. zur Feststellung sehr seltener Nebenwirkungen, die erst in großen Patientenkollektiven erkennbar sind. Häufig werden Phase IV Studien aber auch zu Marketing-zwecken verwendet

# Outline

Allgemeine Informationen

Grundgedanke der Biostatistik

Klinische Studien

Planung

Durchführung

Auswertung

Phasen einer Arzneimittelstudie

R

## R

- ▶ R ist eine freie Programmiersprache für statistisches Rechnen und statistische Grafiken



- ▶ R ist auch eine interaktive Entwicklungsumgebung
- ▶ R ist Teil des GNU-Projekts und auf vielen Plattformen verfügbar. Vor allem im akademischen Bereich gilt sie heute als statistische Standardsoftware.
- ▶ open source (Quelltext)

# Grundkenntnisse von Programmierung in R

- a) Kommandozeile, Befehlen, Ergebnis
- b) Mathematische Operators: `+`, `-`, `*`, `/`, `^`
- c) Funktionen: `sqrt()`, `abs()`, `sin()`, `log10()`, `log()`, `exp()`
- d) oben Knopf
- e) Bezeichners: `pi`, eigene Bezeichners, `<-` (Zuweisung)
- f) Datentypen: `class()`, `"numeric"`, `"logical"`, `TRUE`, `FALSE`, `"character"`, Fehlers
- g) Vektor: `c()`, `v[i]`, `length()`, `sum()`, `v[c(i, j, k)]`
- h) Deskriptiv-statistische Funktionen: `mean()`, `sd()`, `var()`, `range()`, `min()`, `max()`, `median()`, `quantile()`
- i) `?`, Funktionen mit mehrere Parameters, `summary()`, `fivenum()`
- j) Abbildungen: `hist()`, `plot()`, Vektoren von Folgen: `1:10`

# Methoden für Datenpresentation in R 1/2

- a) Matrix Datentyp: `matrix()`, 2D-Indexierung: `m[i,j]`, `m[i,]`, `m[,j]`
- b) graphische Darstellung kategorische Daten:  
`barplot(, legend.text=TRUE, beside=TRUE)`, `mosaicplot()`
- c) Statistische Tests für die folgende Typ: abhängige Variable:  
 kategorisch, unabhängige variable: kategorisch: Fisher-test  
`(fisher.test())`, Khiquadrat-test `(chisq.test())`
- d) Tabelle machen von kategorische Daten: `table()`
- e) speichern als CSV in Excel, `read.csv2()`, `head()`
- f) fehlende/spezielle Daten: NA, NaN, Inf
- g) Entscheiden ob eine Variable Normalverteilung hat: `hist()` (sollte symmetrisch sein), `qqnorm()`, `shapiro.test()`
- h) graphische Darstellung numerische Daten: `hist()`, `boxplot()`,  
`plot(density())`, `lines(density())`, Filtrierung der Daten:  
`which()`, mehrere Abbildungen auf eine Abbildung: `par(mfrow=...)`

## Methoden für Datenpresentation in R 2/2

- i) Statistische Tests für die folgende Typ: abhängige Variable: numerisch, Normalverteilung, unabhängige variable: kategorisch, zwei Kategorien: t-test (`t.test()`)
- j) Statistische Tests für die folgende Typ: abhängige Variable: numerisch, nicht-Normalverteilung, unabhängige variable: kategorisch, zwei Kategorien: Wilcoxon-test (`wilcox.test()`)
- k) Statistische Tests für die folgende Typ: abhängige Variable: numerisch, Normalverteilung, unabhängige variable: kategorisch, mehr als zwei Kategorien: ANOVA (`aov()`, `TukeyHSD()`)
- l) Statistische Tests für die folgende Typ: abhängige Variable: numerisch, nicht-Normalverteilung, unabhängige variable: kategorisch, mehr als zwei Kategorien: Kruskal-Wallis-test (`kruskal.test()`)
- m) Statistische Tests für die folgende Typ: abhängige Variable: numerisch, unabhängige variable: numerisch: Korrelation (`cor.test()`), Regression



# Hypothesenprüfungen für unterschiedliche Datentypen

			Abhängige Variable			
			Kategorisch	Numerisch		
				Normalverteilt	Nicht-normalverteilt	Ereigniszeit-data
Unabhängige Variable	Kategorisch	2 Kategorien	Khiquadrat-test <i>chisq.test()</i> , Fisher-test <i>fisher.test()</i>	t-test <i>t.test()</i>	Wilcoxon-test <i>wilcox.test()</i>	Kaplan-Meier <i>survdiff()</i>
		mehr als 2 Kategorien		ANOVA <i>aov()</i>	Kruskal-Wallis-test <i>kruskal.test()</i>	
	numerisch		Logistische Regression	Regression, Korrelationstest <i>cor.test(..., method="pearson")</i>	Robust Regression, Korrelationstest <i>cor.test(..., method="spearman")</i>	

# Graphische Presentation für unterschiedliche Datentypen in R

		Abhängige Variable		
		Kategorisch	Numerisch	
			Nicht-Ereigniszeit	Ereigniszeitdata
Unabhängige Variable	Kategorisch	barplot(), mosaicplot	boxplot, hist(), plot(density())	plot(survfit())
	Numerisch	plot()	plot()	