A bag full of tricks…

Fahrplan

- Pseudozufallszahlen
- Fehlende Werte (NA)
- Profigrafik: ggplot2
- Reproduzierbare Auswertungen: knitr

Pseudozufallszahlen

- Echte Zufallszahlen: Radioaktiver Zerfall etc.

  → aufwändig
- Pseudozufallszahlen: Deterministische Folge von Zahlen; kein echter Zufall
- Bzgl. bekannten Tests nicht von «echten Zufallszahlen» zu unterscheiden
- Folge ist durch Startzahl (“seed”) eindeutig bestimmt

- Falls seed nicht explizit gesetzt: seed wird in Abhängigkeit von der aktuellen Uhrzeit gesetzt

Pseudozufallszahlen und Reproduzierbarkeit

Bei welchen Methoden muss man den random seed setzen, wenn man die Ergebnisse reproduzieren will?

A) 1-weg ANOVA
B) Mixed Effects Models
C) K-means clustering
D) PCA
E) Simulation der Stichprobengrösse

Fahrplan

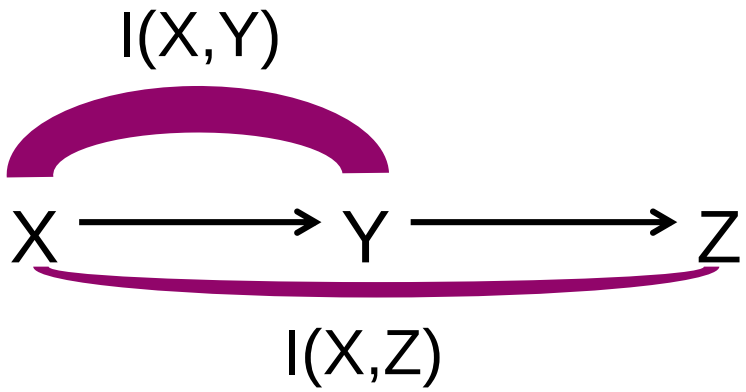
- Pseudozufallszahlen
- Fehlende Werte (NA)
- Profigrafik: ggplot2
- Reproduzierbare Auswertungen: knitr

Fehlende Werte: NA

- Gute Codierung: “NA” - Not available
  In R: Funktion “is.na()”
- Schlecht Codierung: 999, 0, etc.
  Könnte man mit realen Messungen verwechseln

- Überblick über Problematik und bestehende Methoden
- **Hauptbotschaft**:
  Es gibt keine Methode, die zuverlässig mit NAs umgehen kann
  → vermeiden, wenn es geht
  → pragmatische Methoden, wenn es nicht geht

Informationstheorie 101: Data Processing Inequality



«Informationsgehalt»

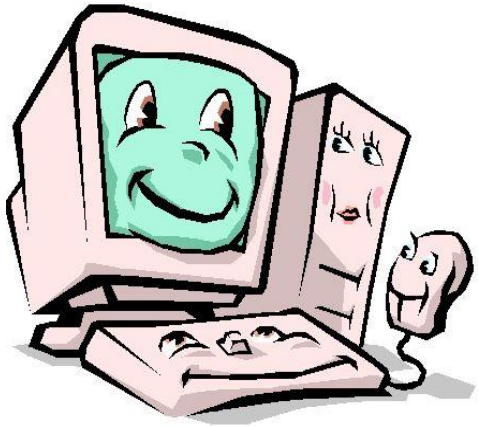
$$I(X,Y) \geq I(X,Z)$$

Nachbearbeitung kann keine Informationen hinzufügen

Natur

.raw

.jpg



Nachbearbeitung kann keine Informationen hinzufügen

Natur

Daten mit NAs

A	B	C
1.3	5.4	7.2
3.2	?	?
?	8.3	?

NAs irgendwie bearbeitet

A	B	C
1.3	5.4	7.2
3.2	7.2	5.6
8.1	8.3	8.2

Informationstheorie bzgl. NAs

- Die Information ist verloren !
  Man kann sie aus den Daten nicht wiedergewinnen.
- Daher: NAs vermeiden, soweit möglich.
- Nicht noch mehr Information verlieren:
  Clevere Methoden verwenden

Überblick über fehlende Werte in den Daten

- R: Funktion “md.pattern” in package “mice”

```
> md.pattern(boys) ## 223 rows are complete
     age reg wgt hgt bmi hc gen phb  tv
223    1   1   1   1   1  1   1   1   1    0
  1    1   1   1   1   1  1   1   0   1    1
 19    1   1   1   1   1  1   1   1   0    1
  1    1   1   1   1   1  1   1   0   1    2
  1    1   1   1   0   0  0   1   1   1    3
437    1   1   1   1   1  1   0   0   0    3
  1    1   1   1   0   0  0   0   1   1    4
 43    1   1   1   1   1  0   0   0   0    4
  3    1   0   1   1   1  1   0   0   0    4
 16    1   1   1   1   0  0   1   0   0    5
  1    1   1   1   0   1  0   1   0   0    5
  1    1   1   1   1   0  0   0   0   0    6
  1    1   1   1   0   0  0   0   0   0    7
       0   3   4  20  21 46 503 503 522 1622
```

Auftretende Muster:
- Keine NAs: 223 Zeilen
…
- Alles ausser age und reg sind NAs: 1 Zeile

Dieses Muster hat 5 NA pro Zeile

1622 NA

20 NA in Spalte “hgt”

Arten von NAs

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

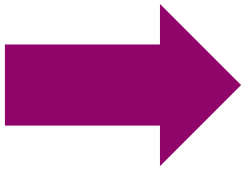
OK

PROBLEM

Verteilung der NAs

Complete data Y_{com}

A	B	C
1.3	2.5	6.3
2.0	3.6	5.4
1.6	2.3	4.3



Y_{obs}

A	B	C
1.3	2.5	
2.0		5.4
1.6		4.3

Y_{mis}

A	B	C
		6.3
	3.6	
	2.3	

R

A	B	C
1	1	0
1	0	1
1	0	1

Einige Einträge fehlen

Bsp: Blutdruck

- 30 TN im Januar(X) und Februar(Y)
- MCAR: Lösche 23 Y-Werte zufällig
- MAR: Behalte nur Y-Werte, bei denen X > 140 (follow-up)
- MNAR: Behalte nur Y-Werte, bei denen Y > 140 (teste jeden, aber behalte nur kritische Werte)

	Y			
X	Complete	MCAR	MAR	MNAR
	Data for individual participants			
169	148	148	148	148
126	123	—	—	—
132	149	—	—	149
160	169	—	169	169
105	138	—	—	—
116	102	—	—	—
125	88	—	—	—
112	100	—	—	—
133	150	—	—	150
94	113	—	—	—
109	96	—	—	—
109	78	—	—	—
106	148	—	—	148
176	137	—	137	—
128	155	—	—	155
131	131	—	—	—
130	101	101	—	—
145	155	—	155	155
136	140	—	—	—
146	134	—	134	—
111	129	—	—	—
97	85	85	—	—
134	124	124	—	—
153	112	—	112	—
118	118	—	—	—
137	122	122	—	—
101	119	—	—	—
103	106	106	—	—
78	74	74	—	—
151	113	—	113	—

Praktisches Problem

- Typ des NA kann nicht getestet warden
- Viele Methoden gehen nur, wenn man keine fehlenden Werte hat
- “Imputation”: Fehlende Werte beseitigen

- Pragmatisch:
  - Methoden, die unter MAR gelten bevorzugen
  - Methoden, die nur unter MCAR gelten möglichst meiden

Imputation – mit NAs arbeiten

- Complete-case analysis - valid for MCAR
- Single Imputation - valid for MAR
- (Multiple Imputation – valid for MAR)

Complete-case analysis

- Lösche alle Zeilen, die mind. ein NA haben
- Problem:
  - Verlust von Informationen; ineffizient
  - erzeugt systematische Fehler, falls MAR
- OK, falls $\leq 5\%$ NAs
- R: Function “na.omit()”

A	B	C	D
NA	3	4	6
3	2	3	NA
2	NA	5	4
5	7	NA	5
6	NA	9	2

- 25% der Werte fehlen
- KEINE vollständige Zeile
  Complete-case analysis ist nutzlos

Single Imputation

- Unconditional Mean
- Unconditional Distribution
- Conditional Mean
- Conditional Distribution

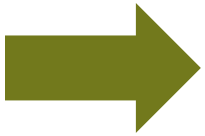
Easy / Inaccurate

Hard / Accurate

Unconditional Mean: Idee

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

Mean = 4.75

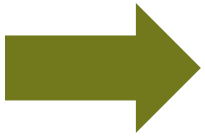


A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	4.75

Unconditional Distribution: Hot Deck Imputation

Wähle zufällig einen beobachteten Wert in der Spalte

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA



A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	6.3

Conditional Mean: Z.B. Linear Regression

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

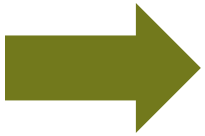
Schätze lm(C ~ A + B)
oder ähnliche Methode

Verwende um C vorherzusagen

Conditional Mean: Z.B. Linear Regression

Vorhersge der lin. Regression

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA



A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	8

Conditional Distribution: Z.B. Linear Regression

- Starte mit Conditional Mean
- Füge Zufallsstreuung hinzu

Vorhersage mit linearer Regression
PLUS ZUFÄLLIGER FEHLER

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	NA

A	B	C
2.1	6.2	3.2
3.4	3.7	6.3
4.1	4.5	8.3

Pragmatisch:
Conditional Mean Imputation mit missForest

- Verwende Random Forest statt Linearer Regression
- Guter Kompromiss: Einfach und doch genau
- Kann mit gemischten Variablentypen arbeiten (kontinuierlich, kategorisch, gemischt)
- Schätzt Güte der Imputation
  OOBerror: Fehler der Imputation als Anteil der totalen Variation
  nahe bei 0 - gut
  nahe bei 1 - schlecht

Measuring quality of imputation

- Normalized Root Mean Squared Error (NRMSE):

$$NRMSE = \sqrt{\frac{\left(mean\left(Y_{com} - Y_{imp}\right)^2\right)}{var(Y_{com})}}$$

- Proportion of falsely classified entries (PFC) over all categorical values

$$PFC = \frac{(nmb.missclassified)}{nmb.categorical\ values}$$

Vor- und Nachteile von missForest

- Akzeptable Methode, falls MAR
- Leicht anzuwenden: Funktion “missForest” in package “missForest”
- Schätzt Genauigkeit der Imputation
- Genauigkeit ist tendenziell zu optimistisch, weil
  - imputierte Werte haben keine Streuung
  - es wird so getan, als wäre das Modell für die Imputation das wahre Modell

- Bessere Lösung: Multiple Imputation

Fazit: Fehlende Werte

- Problem ist nicht abschliessend gelöst
- Es gibt pragmatische, plausible Methoden
- Single Imputation ist tendenziell besser als Complete Case Analysis

Fahrplan

- Pseudozufallszahlen
- Fehlende Werte (NA)
- Profigrafik: ggplot2
- Reproduzierbare Auswertungen: knitr

Grafik in R

- Standardbefehle:
  plot, points, lines, boxplot, hist

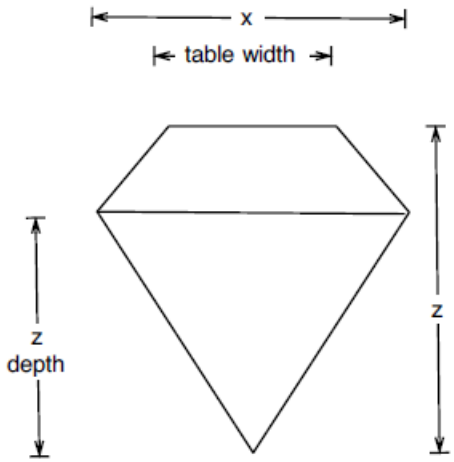
- ggplot2: Sehr mächtiges Grafikpaket
- Einfacher Einstieg: Funktion qplot (“quick plot”)

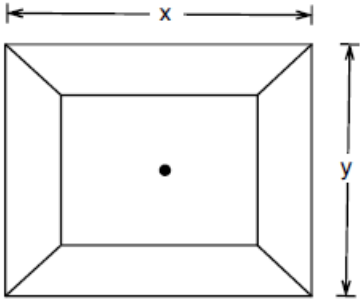
Bsp: Diamanten

carat	cut	color	clarity	depth	table	price	x	y	z
0.2	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.2	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.2	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.3	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.3	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.2	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48

Faktoren

Kontinuierliche Variablen





depth = z depth / z * 100
table = table width / x * 100

qplot: Die wichtigsten Argumente

- Color: Farbe der Plotsymbole (Faktor)
- Shape: Form der Plotsymbole (Faktor)
- Size: Grösse der Plotsymbole (kont. Variable)

Fahrplan

- Pseudozufallszahlen
- Fehlende Werte (NA)
- Profigrafik: ggplot2
- Reproduzierbare Auswertungen: knitr

«Trouble at the lab»: Economist, Oktober 2013

http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

Wissenschaftliche Ergebnisse sind überraschend schlecht reproduzierbar

Krebsforschung: Probleme beim Reproduzieren

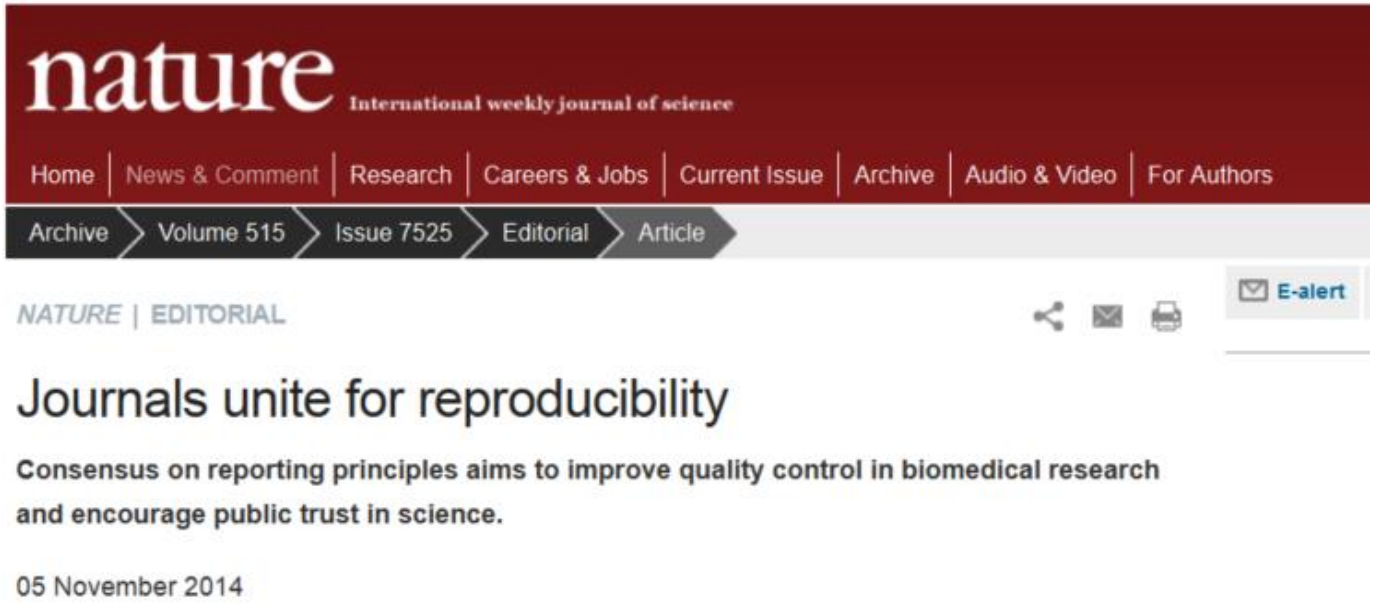
A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic

Aaron Mobley, Suzanne K. Linder, Russell Braeuer, Lee M. Ellis, Leonard Zwelling

Published: May 15, 2013 • DOI: 10.1371/journal.pone.0063221

To examine a microcosm of the academic experience with data reproducibility, we surveyed the faculty and trainees at MD Anderson Cancer Center using an anonymous computerized questionnaire; we sought to ascertain the frequency and potential causes of non-reproducible data. We found that ~50% of respondents had experienced at least one episode of the inability to reproduce published data; many who pursued this issue with the original authors were never able to identify the reason for the lack of reproducibility; some were even met with a less than "collegial" interaction.

Handlungsbedarf



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 515 > Issue 7525 > Editorial > Article

NATURE | EDITORIAL

E-alert

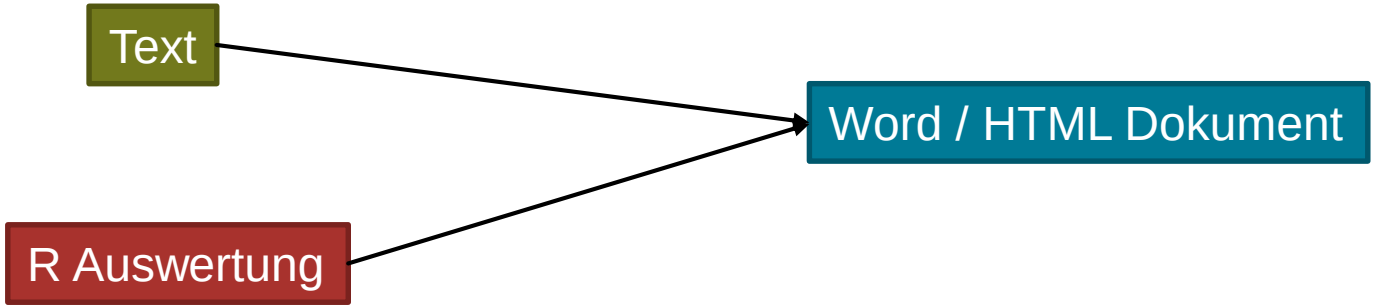
Journals unite for reproducibility

Consensus on reporting principles aims to improve quality control in biomedical research and encourage public trust in science.

05 November 2014

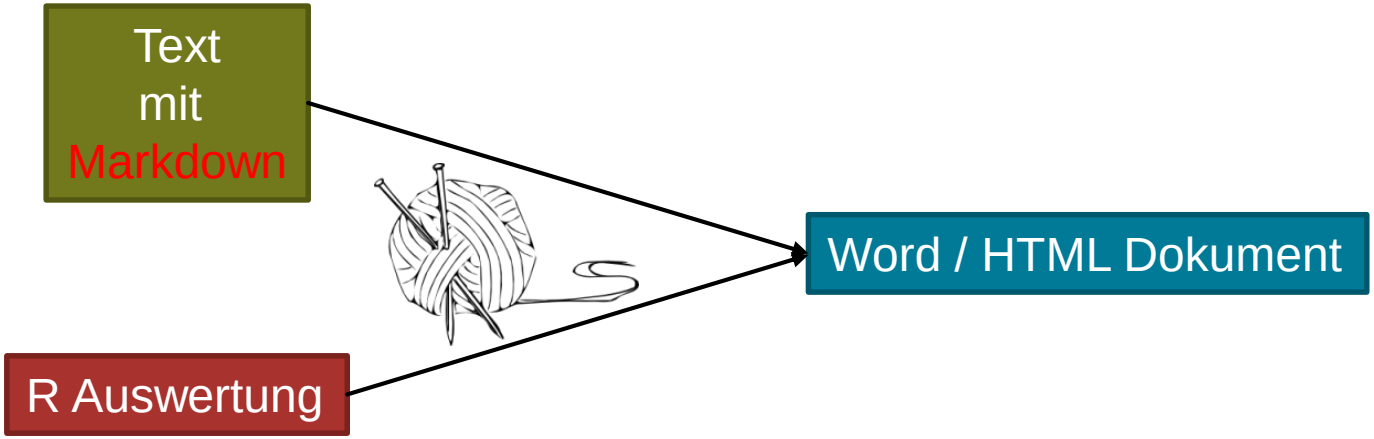
Ein kleiner Beitrag: Dynamsicher Bericht

- Auswertung und Text verknüpfen
  → weniger Fehler
  → reproduzierbarer
- R Studio unterstützt «Notebook»
- R File: File – Compile Notebook – pdf / word / html



Etwas komplexer: knitr

- Verknüpfe z.B. Word oder HTML und R
- knitr: Dynamischer Bericht
  (to knit = stricken)
- Von R-Studio unterstützt: Package ‘knitr’



Für Perfektionisten

- Latex & R geht auch mit knitr: Aufwändiger, aber dafür mehr Kontrolle
- Alternative zu kintr: Sweave
- Latex muss installiert sein