ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Kurzüberblick: R

Markus Kalisch, Seminar für Statistik, ETHZ

Ursprung von R

- ~1970, Bell-Labs intern: “**An Interactive Environment for Data Analysis and Graphics**” → Interne Software mit Namen “S”, geschrieben in Fortran
- ~1985, neue Version geschrieben in C, wird auch ausserhalb Bell-Labs verteilt; kommerziell “S-Plus”
- ~2000, open-source version von “S” mit dem Namen “R”; **geschrieben in C**

Konzepte in R

- Schnelles, interaktives Auswerten von Daten
- Rechenzeit sekundär; Schnittstellen zu anderen Programmiersprachen vorhanden (z.B. C++ in package ‘Rcpp’)

- In erster Linie **funktionale** Programmiersprache
- **Objekt-orientierte** Programmieren auch unterstützt; z.T. gewöhnungsbedürftig:
  - Functional OOP: Standard in R, unüblich in anderen Sprachen
  - Encapsulated OOP: Relativ neu in R; wie C++, Java, …

R heute

- Open-source Programmiersprache basierend auf C
- **High-level** (wie Matlab, Python, …)
- Fokus auf Datenanalyse / Statistik
- **Sehr viele Pakete** (> 10’000) mit Zusatzfunktionen

- Alle wichtigen numerischen Operationen sind enthalten
- Alle gängigen Betriebssysteme werden unterstützt

- **Grosse Verbreitung** in Industrie und Forschung

- https://www.r-project.org/
  (Manuals, CRAN, Packages, Task Views)

Fragen und Antworten

- Welche Editoren gibt es ?
- Wie kompliziert ist der Installationsprozess ?
- Was kann das Programm ?
- Voraussetzungen bei den Schülern an Know-How ?
- Ist die Syntax an ein System angelehnt das man kennt ?
- Was hat das Programm für Features, die auch grad für den Lehrer sinnvoll sind ?

- Ideen für die Schule

Welche Editoren gibt es ?

- R kommt mit einem minimalistischen Editor
- Viel komfortabler: Editor **Rstudio**
  https://www.rstudio.com/
- Rstudio ist mit Abstand der verbreitetste Editor für R
- Alternativen: Emacs, Tinn-R, Eclipse, …

- Achtung:
  R = Programmiersprache
  Rstudio = Einer von vielen möglichen Editoren
  (wird gerne verwechselt…)

Wie kompliziert ist der Installationsprozess ?

- Ziemlich einfach
- R installieren: https://stat.ethz.ch/CRAN/index.html
- Rstudio installieren (nicht nötig, aber empfohlen): https://www.rstudio.com/

- Beim Starten sollte Rstudio automatisch die installierte Version von R erkennen
  sonst in Rstudio: Tools → Global Options → R version

- Kenne keine brauchbare online Version von R

Was kann das Programm ?

- “base” package + 10’000+ Zusatzpakete → SEHR VIEL
- Fokus auf Datenbearbeitung und Datenanalyse

- Ein paar Beispiele: bsp.R

Voraussetzung an Schüler Know-How

- Installieren können (evtl. schon ein Lernziel an sich)
- Dann kann R sofort als “Taschenrechner” verwendet werden

Ist die Syntax an ein System angelehnt das man kennt ?

- Alle ‘high-level’ Sprachen sind sehr ähnlich:
  R, Python, Matlab, …
- R ‘verzeiht viele Fehler’: Z.B. müssen Variablen nicht deklariert werden
  Vorteil für Schüler: Einfacher Einstieg
  Nachteil: Weniger effizient, mögliche Fehlerquelle

Features für Lehrer

- Paket knitr: Text (Latex, Markdown, …) und Auswertung (in R) in einem Dokument vereinen (knitrBsp.Rmd)

- Paket shiny: Interaktive Datenauswertung (siehe bsp.R)

Fragen und Antworten

- Welche Editoren gibt es ?
- Wie kompliziert ist der Installationsprozess ?
- Was kann das Programm ?
- Voraussetzungen bei den Schülern an Know-How ?
- Ist die Syntax an ein System angelehnt das man kennt ?
- Was hat das Programm für Features, die auch grad für den Lehrer sinnvoll sind ?

- Ideen für die Schule

Mögliche Projekte in der Schule

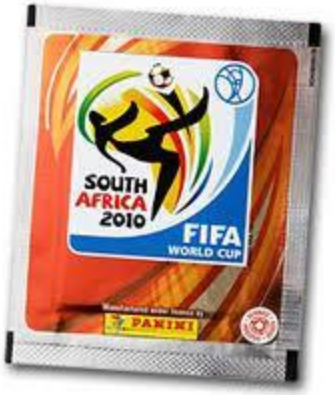
- Permutations Test:
  Werden Panini-Bilder zufällig eingetütet ?

- Runs Test: Ist eine Sequenz von 0/1 zufällig erzeugt worden ?


SOUTH
AFRICA
2010
FIFA
WORLD CUP
OFFICIAL LICENSED
STICKER ALBUM
Manufactured under licence by
PANINI
www.paninigroup.com

661 Bilder

Packung



5 Bilder

Box

100 Packungen =
500 Bilder


C&A
C&A
kiosk
STIMOROL
STIMOROL
Bolt


SOUTH AFRICA 2010
FIFA WORLD CUP
PANINI


FIFA WORLD CUP

Beobachtung von Vorjahren

- Ganze Box: Wenige doppelte Bilder
- Einzelne Packungen an verschiedenen Kiosks: Viele doppelte Bilder

- “Null”hypothese: Bilder werden zufällig verpackt (“Null”, weil kein System hinter dem Verpacken steckt)
- Alternativhypothese: Die Bilder werden systematisch verpackt, sodass man weniger doppelte hat

- Wie könnte man zwischen diesen beiden Hypothesen unterscheiden?

Hypothesentest

- Ich habe eine Box mit 500 Bildern gekauft. In eine leeres Album (661 mögliche Bilder) konnte ich 477 Bilder einkleben.

- Angenommen, die Nullhypothese stimmt:
  Ist es plausibel, dass ich dann 477 Bilder einkleben kann?

- Passen die Nullhypothese “zufällig verpackt” und die Beobachtung “477 Bilder eingeklebt” zusammen?

Problem: Was ist “normal”?

- Wenn wir viel mehr Bilder als “normal” einkleben konnten, wurden die Bilder wohl nicht zufällig verpackt.

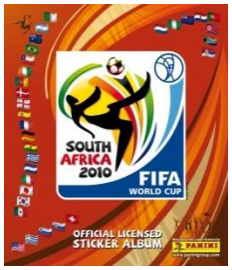
- Angenommen, die Nullhypothese stimmt (Bilder zufällig verpackt):

  Wie viele Bilder kann man normalerweise einkleben?

- Signifikanzniveau α: Wie “abnormal” muss die Beobachtung sein, damit wir der Nullhypothese nicht mehr glauben?
  Z.B.: $\alpha = 1/1.000.000$; wir lehnen die Nullhypothese ab, wenn wir etwas beobachten, das weniger wahrscheinlich als 1/1.000.000 ist.

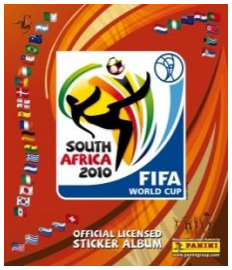
Lösung: Computersimulation

1: 



350 Bilder eingeklebt

2: 



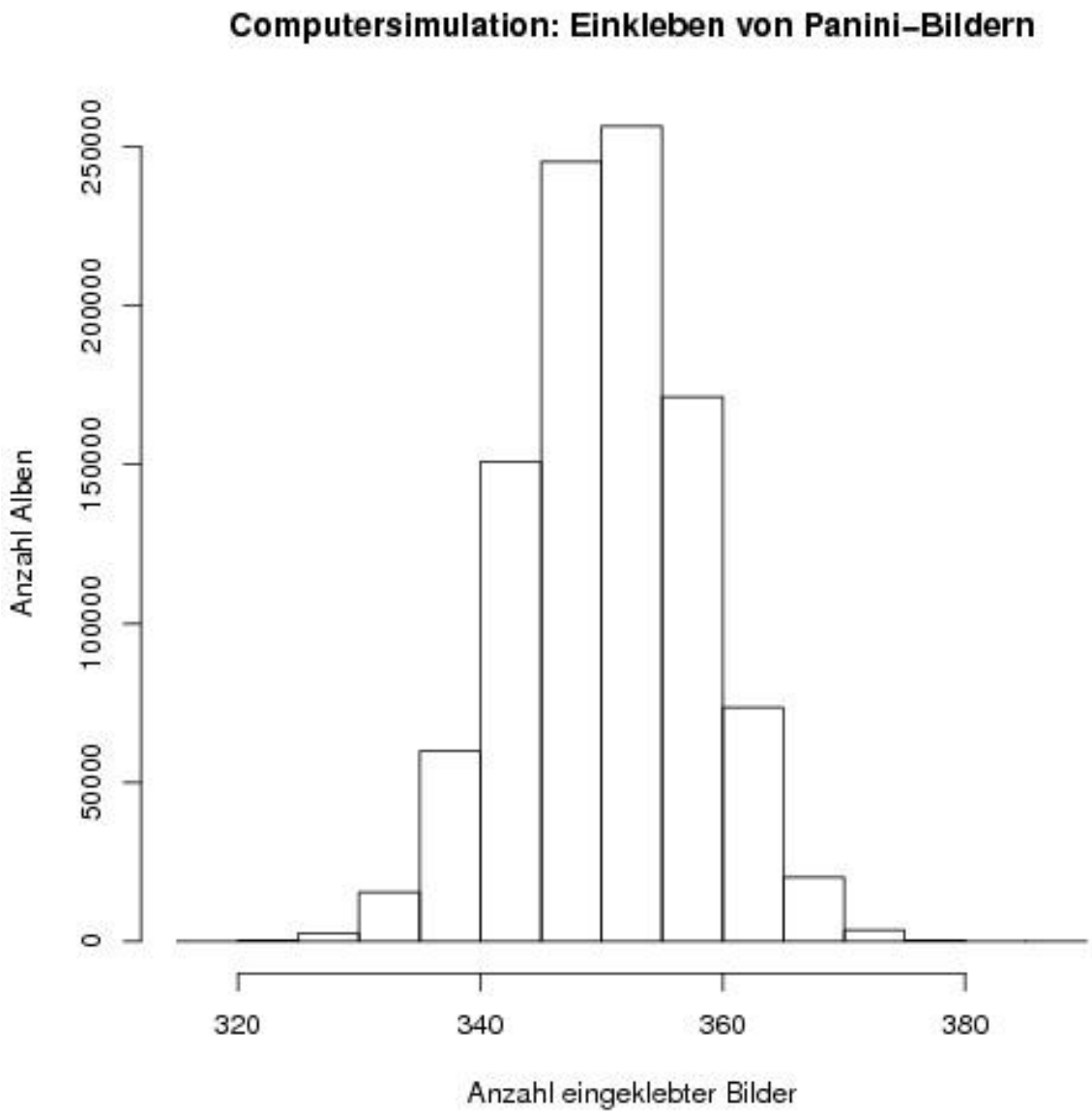
361 Bilder eingeklebt

⋮

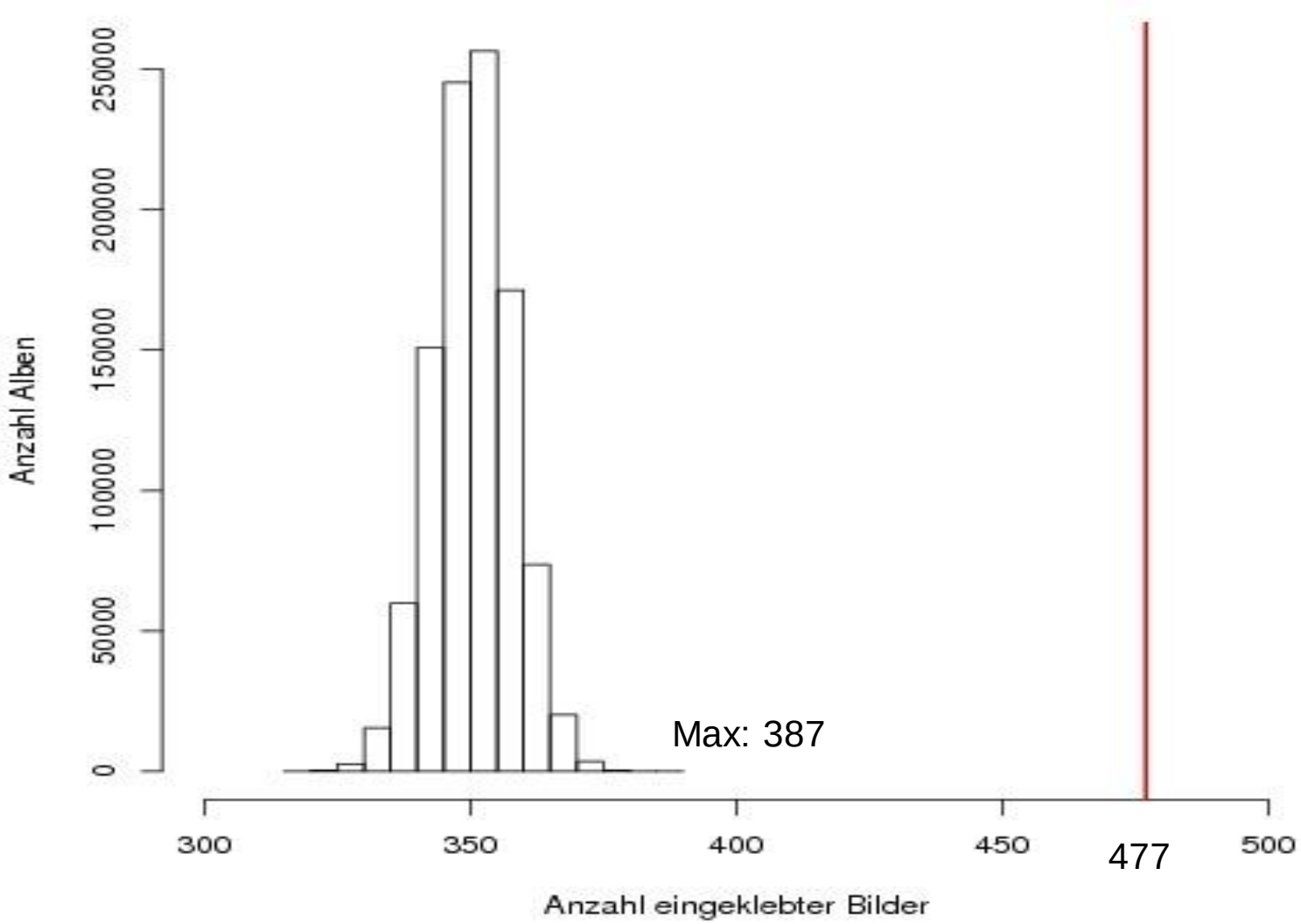
1 Mio: 

358 Bilder eingeklebt

Ergebnis der Computersimulation

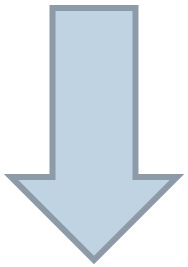


Passt unsere Beobachtung zur Computersimulation?



Schlussfolgerung

- Angenommen, die Bilder werden zufällig verpackt. Die Wa. 477 oder mehr Bilder einkleben zu können ist kleiner als ein Millionstel !



- Beobachtung und Simulation passen nicht zusammen:

  Die Bilder werden wohl NICHT zufällig eingepackt.

Zusammenfassung: Hypothesentest

1. Modell: Ziehen 500 Bilder mit Zurücklegen aus 661 Bildern
2. Nullhypothese: “Panini-Bilder in Kiste zufällig eingepackt”
   Alternative: “Systematisch eingepackt, sodass weniger Doppelte”
3. Teststatistik: Anz. Bilder, die man in eine leeres Album einkleben kann, wenn man eine Kiste mit 500 Bildern hat
   Verteilung der Teststatistik, wenn Nullhypothese stimmt: Computersimulation
4. Signifikanzniveau $\alpha = 1/1.000.000$
5. Verwerfungsbereich der Teststatistik:
   Computer beobachtet bei 1 Mio Simulationen nie mehr als 387 eingeklebte Bilder
   Verwerfungsbereich: K = {388, 389, …, 500}
6. Testentscheid: Der beobachtete Wert (477) liegt im Verwerfungsbereich der Teststatistik. Daher wird die Nullhypothese auf dem Signifikanzniveau 1/1.000.000 verworfen.

Mögliche Projekte in der Schule

- Permutations Test:
  Werden Panini-Bilder zufällig eingetütet ?

- Runs Test: Ist eine Sequenz von 0/1 zufällig erzeugt worden ?

0/1-Zufallssequenz

- Hausaufgabe: 200 mal Münze werfen, damit zufällige 0/1-Sequenz erzeugen
- LANGWEILIG !!! ☹
- Mogeln: Willkürlich auf Tasten 0/1 tippen

- Wie vergleichen sich die Mogelsequenzen mit echten Zufallssequenzen ?
- Möglicher Exkurs: Pseudozufallszahlen

- Test-Batterien für Zufallszahlen: Diehard Tests
  Ein Test dabei ist der sog. “Runs-Test”
  https://en.wikipedia.org/wiki/Diehard_tests

Runs

00111011 0

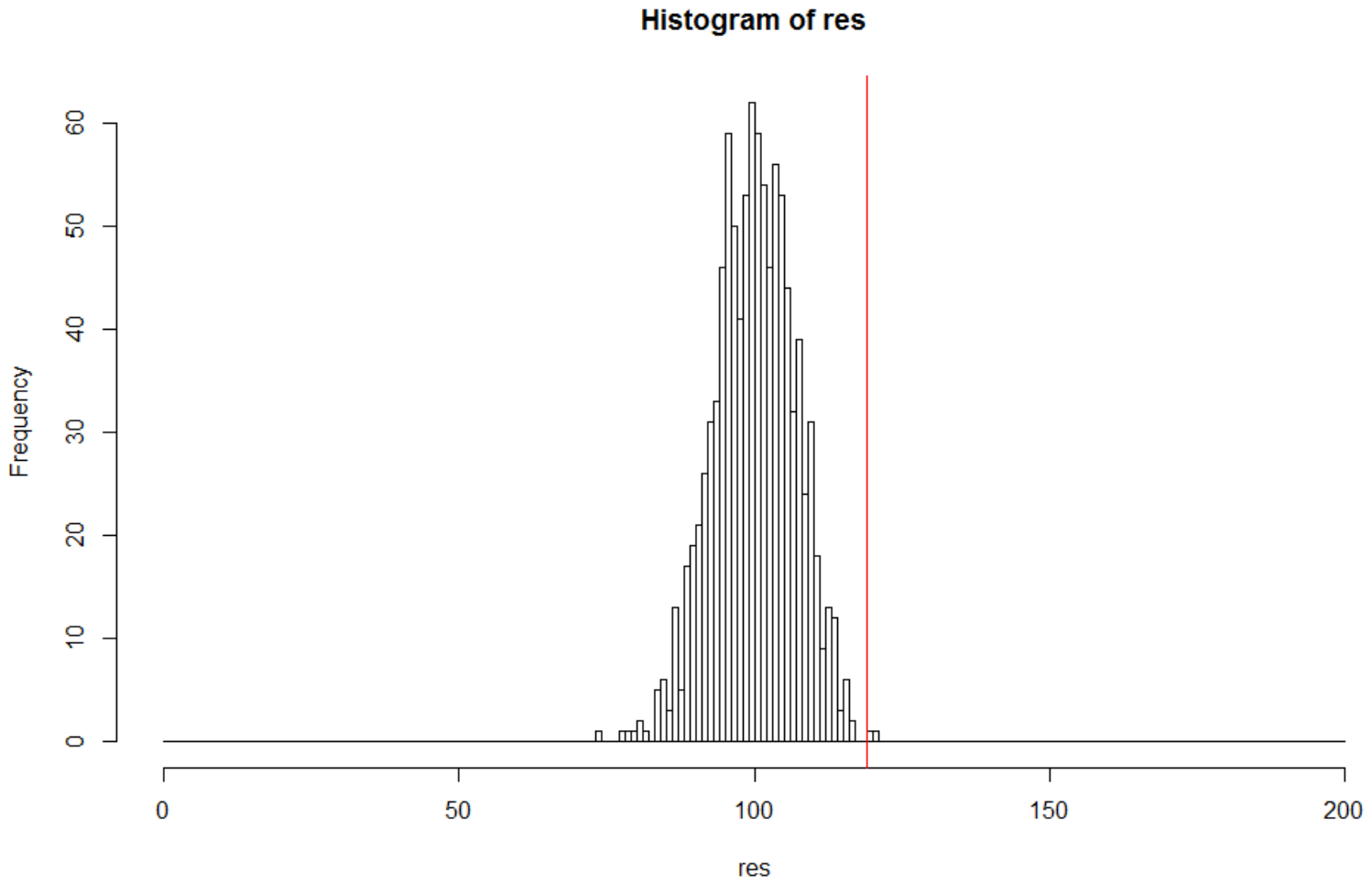
Sequenz hat 5 “runs”

Runs-Test

- Schüler ‘ermogeln’ Zufallssequenz der Länge 200
- Wir zählen die Anzahl Runs

- Wir simulieren 1000 ‘echte’ Zufallssequenzen der Länge 200
- Wir zählen die Anzahl Runs pro Sequenz
- Wir vergleichen

Runs-Test Ergebnis



Schlussfolgerung

- Nullhypothese wurde (knapp) verworfen

- Bei kurzen Sequenzen (<100) ist es relativ leicht eine Sequenz zu erzeugen, die im Runs-Test nicht auffällt.
- Je länger die Sequenz, desto mehr Macht hat der Runs-Test, d.h., desto eher kann er Abweichungen vom echten Zufall entdecken. (Empfehle ≥ 200 für die Schule)
- Kann jemand eine Sequenz der Länge 1000 eintippen, die nicht auffällig ist ?

- Echte Test-Prozeduren für Pseudo-Zufallszahlengeneratoren enthalten mehrere solche Tests und sind daher noch viel schwieriger zu schlagen.

Weitere Ressourcen zu R

- Kostenlose, interaktive Plattformen die Grundlagen in R unterrichten (beides auch für Python):

http://tryr.codeschool.com/
https://www.datacamp.com/

- Quick-R: http://www.statmethods.net/

Anmerkungen nach Diskussionen nach Vortrag

- etutoR: Lernplattform auf Deutsch; begleitet Statistik 1 VL, und behandelt schnell mal **Statistik-Themen**. Daher empfehle ich es nicht für die Schule. Hier eine völlig informelle (Prototyp) Version, die dafür aber sehr leicht zugänglich und inhaltlich praktisch ausgereift ist: http://stat.ethz.ch/~meier/etutoR/