Einführung in R

Michael Zwinkel

Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg
Betreuer: Julian Kunkel

27-04-2016

Gliederung (Agenda)

Was ist R?

- Frei verfügbare Programmiersprache
  - GNU General Public License
- General Purpose Language
  - Nicht nur für spezielle Anwendungsfälle einsetzbar
- Untersuchung statistischer Probleme
  - Einsatz in der Wissenschaft als auch in der Wirtschaft
- Implementierung der Sprache S

Was ist R?

- R basiert auf der Programmiersprache S
- R wurde 1992 von Ross Ihaka und Robert Gentleman entwickelt
- Programmiert in C, Fortran und R selbst
  - 52% in C, 26% in Fortran und 22% in R
- 1993 öffentliche Verbreitung von R
  - Seperate Bekanntmachungen an Personen die in S visiert waren
- 1995 steht R unter der GNU General Public License
  - R war einzige Statistikumgebung für Linux
- 1997 Gründung des "R Development Core Team"
- 1997 startete CRAN
  - Teilen von selbstgeschriebenen Funktionen

Was ist R?

- 2000 R Version 1.0
- 2004 R Version 2.0
  - Lazy Loading
- 2005 R Version 2.1
  - Unterstützung verschiedener Sprachversionen und Zeichenkodierungen
- 2010 R Version 2.11
  - R ist auf 64-Bit Systemen nutzbar
  - R kann bis zu 8 Terabyte Arbeitsspeicher adressieren
- 2013 R Version 3.0
  - R erlaubt Indexwerte von 2^{31} und größer auf 64-Bit Systemen
- 2015 Unternehmen die R nutzen gründeten das "R Consortium"
  - Ziel: R im Unternehmensfeld komfortabler einsetzen

Vor- und Nachteile von R

- Vorteile
  - Open-Source [1]
  - Schnelle Entwicklung [1]
  - Keine Lizenzgebühren [1]
  - Neue statistische Methoden werden sofort in R umgesetzt [1]
  - Probleme werden durch große Nutzergemeinde schnell gelöst
  - R läuft auf allen gängigen Betriebssystemen
- Nachteile
  - Keine komplett grafische Oberfläche [1]
  - Fehlermeldungen helfen oft nicht weiter [1]
  - Funktion von älteren Versionen steht nicht im Vordergrund [1]
  - Qualität der Pakete sollte für den Zweck überprüft werden [1]

Einfache mathematische Rechnungen von R

- R benutzt mathematische Regeln selbstständig
  - Punkt vor Strich
  - Klammer vor Punkt
- Gibt das Ergebnis von Rechnungen in der Konsole aus

```
1 2+2
2 2-1
3 2*2
4 8/2
5 2^3
6 2+3*3*(1+1)
```

Mathematische Rechnungen R

- Mathematische Rechnungen
- Weitergehende mathematische Funktionen

```
1 sqrt(17)#Quadratwurzel
2 log(17) #Logarithmus
3 sin(17) #Sinus
4 cos(17) #Kosinus
5 tan(17) #Tangens
```

- Anzahl der Ziffern ist standardmäßig auf 7 gesetzt
  - sqrt(17) liefert zum Beispiel 4,123105
  - Lässt sich umgehen

```
1 options(digits=22)
2 #Aendert die Anzahl der Ziffern
3 print(sqrt(17), digits=22)
4 #Aendert die Anzahl der Ziffern der Funktion
    ↪ sqrt(17)
```

Statistische Grafiken

- Wichtiger Punkt von R: Erstellen statistischer Grafiken
- Verschiedene Arten von Grafiken können erstellt werden
  - barplot() erstellt ein Säulendiagramm
  - boxplot() erstellt ein Kastendiagramm
  - contour() erstellt Isolinien (z.B. Höhenlinien auf Landkarten)
  - curve() erstellt eine Kurve
  - hist() erstellt ein Histogramm
- Daten werden auf verschiedene Weisen implementiert

Daten für statistische Grafiken

- Manuelle Eingabe der Daten
  - Daten werden manuell hinter der Funktion eingetragen
  - Beispiel: barplot(1,2,3)
- Daten werden aus einem Datensatz ausgelesen
- Die Art der Daten bestimmt wie sie eingelesen werden
  - read.table(): Trennt Daten mit ¨ ¨, Dezimalzahlen mit "."
  - read.csv(): Trennt Daten mit ",", Dezimalzahlen mit "."
  - read.csv2(): Trennt Daten mit ";", Dezimalzahlen mit ","
  - read.delim(): Trennt Daten mit "/t", Dezimalzahlen mit "."
  - read.delim2(): Trennt Daten mit "/t", Dezimalzahlen mit ","
- Excel Tabellen können nicht direkt eingelesen werden

Beispiel

- Wir erstellen Grafiken anhand von Daten einer .txt Datei
- Die Datei umfasst Daten eines Rennens
- Die gesammelten Daten des Rennens sind
  - Startnummer
  - Laufzeit
  - Geschlecht
  - Alter
  - Einkommen

Vorbereitung

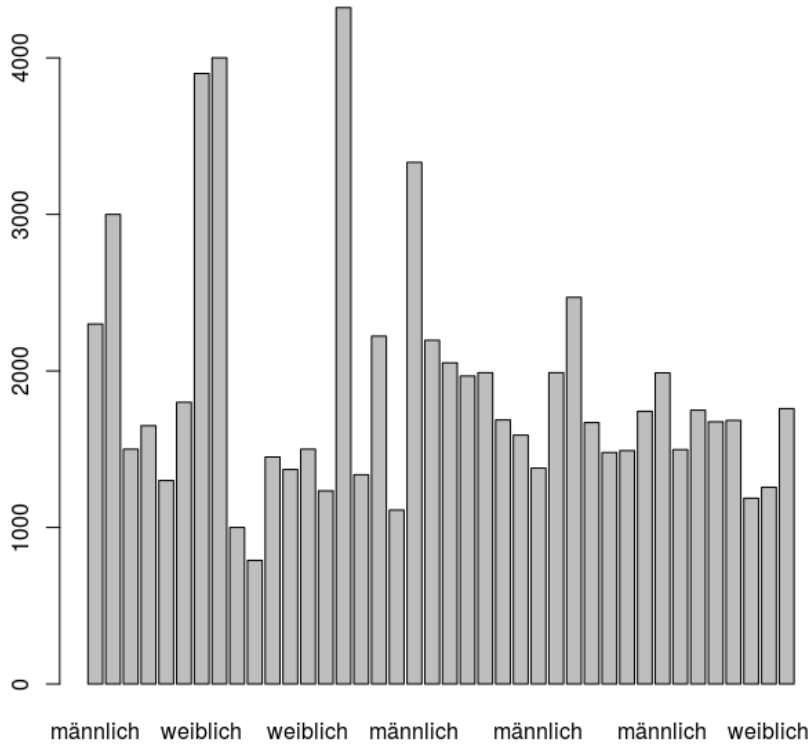
- Zuerst muss der Datensatz in R eingelesen werden

```
1 dat=read.table("/home/user/Documents/r/rennen.txt",
2 header=T)
```

- Gedanken über die Grafik machen
  - Welche Daten machen Sinn?
  - Welche Grafik macht für die ausgewählten Daten Sinn?
  - Was möchte man mit der Grafik zeigen?

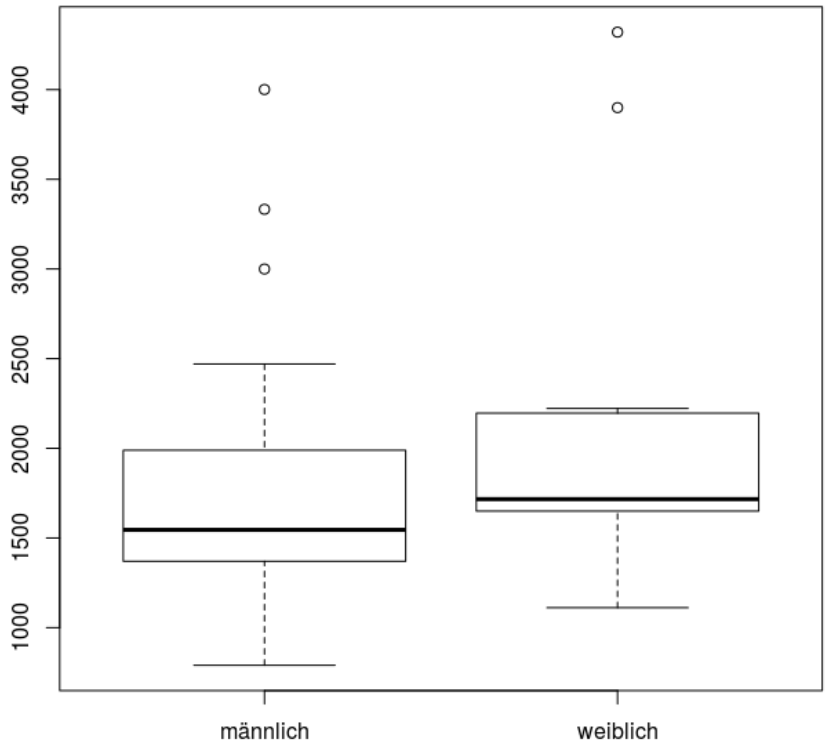
Schlechtes Säulendiagramm

```
1 barplot(Einkommen, width=1, space = NULL, names.arg =
      ↪ Geschlecht)
```



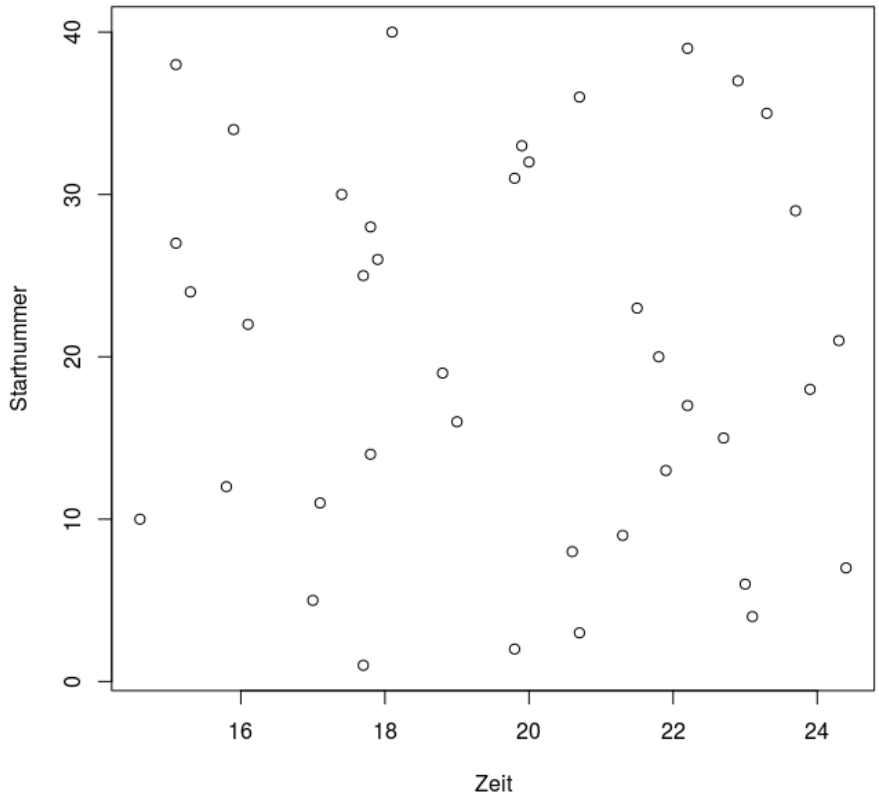
Schlechtes Kastendiagramm

```
1 boxplot(Einkommen~Geschlecht)
```



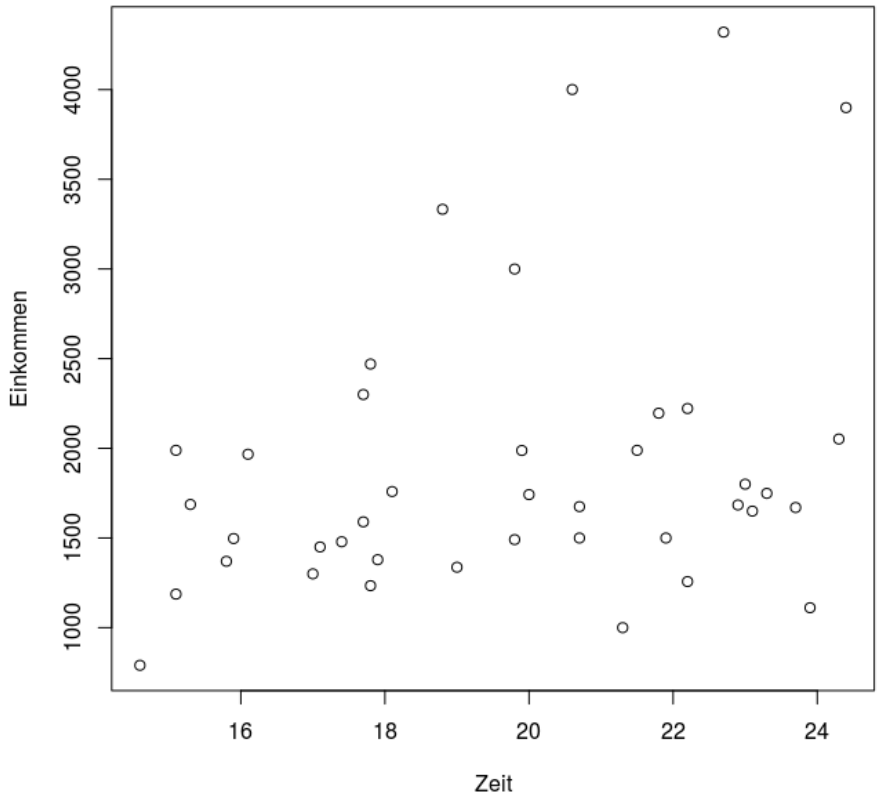
Grafische Darstellung eines Plots

```
1 plot (Startnummer~Zeit)
```



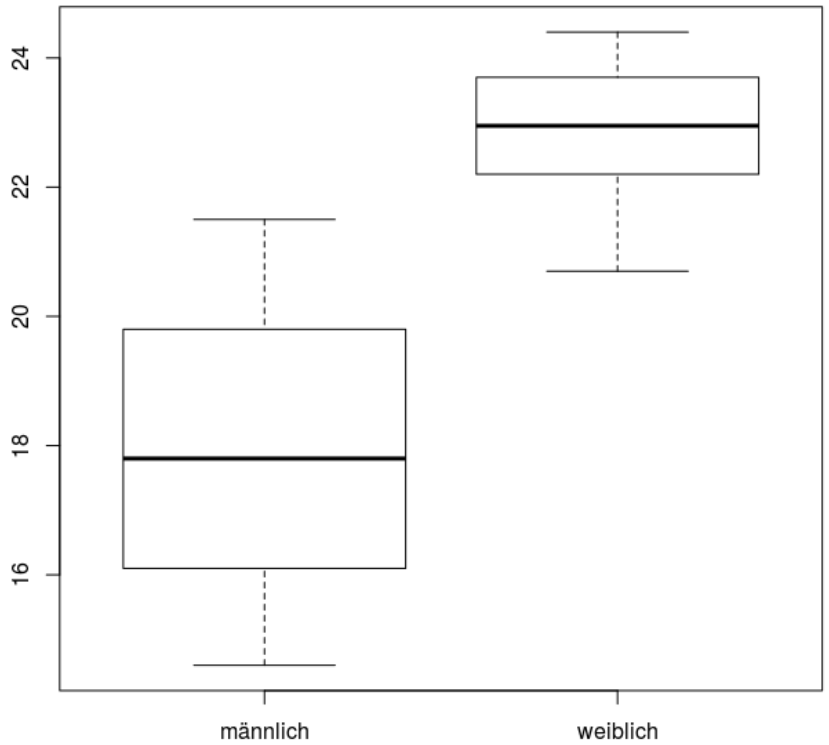
Grafische Darstellung eines Plots

```
1 plot (Einkommen~Zeit)
```



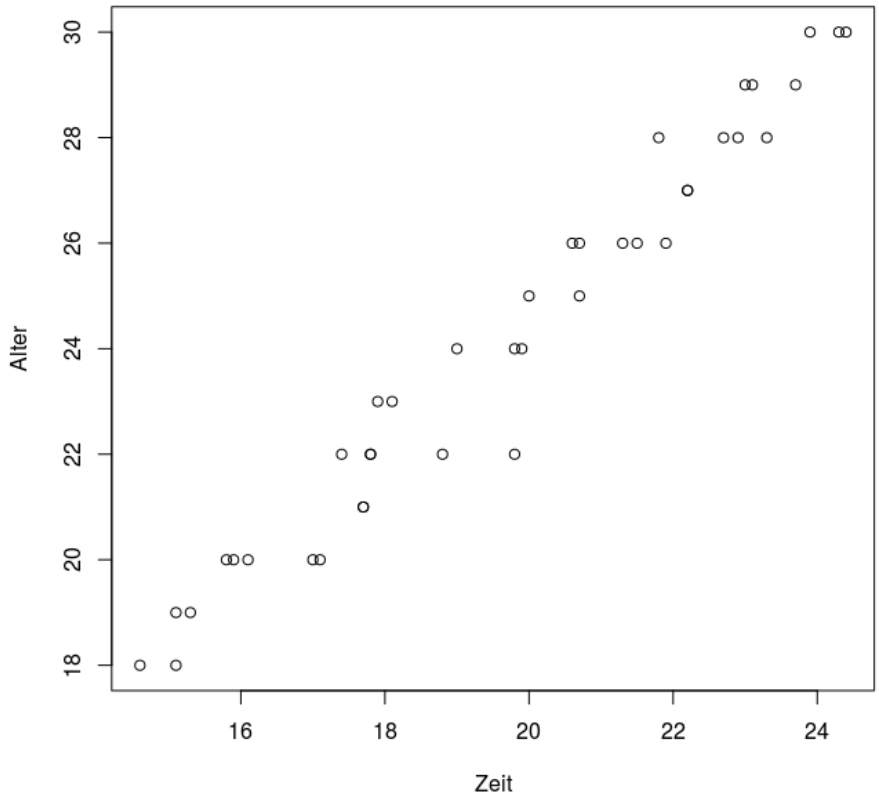
Grafische Darstellung eines Kastendiagramms

```
1 boxplot (Zeit~Geschlecht)
```



Grafische Darstellung eines Plots

```
1 plot (Alter~Zeit)
```



Zusammenfassung

- R ist eine Programmiersprache die auf S basiert
- Wird ständig aktualisiert
- Viele Vorteile
- Weit verbreitet
- R ist in der Lage mathematische Rechnungen durchzuführen
- Statistische Grafiken sind einfach zu erstellen

Literatur

- Zitat [1] Christian Heumann, Vorlesung Programmieren in statistischer Software: R
- https://cran.r-project.org
- https://cran.r-project.org/doc/manuals/R-intro.pdf
- Johannes Hain, Statistik mit R Grundlagen der Datenanalyse