

## Statistische Software (R)

Paul Fink, M.Sc.

Institut für Statistik  
Ludwig-Maximilians-Universität München

*Daten einlesen*



Daten als eine Zusammenstellung von Informationen zu einem bestimmten Thema.

Am Sonntag 11.05.2014 hatte es in München um Punkt 12 Uhr 15°C, während es am Samstag 11.05.2014 in München um Punkt 12 Uhr noch 20°C waren. Am Freitag zuvor waren es zur selben Zeit am selben Ort allerdings erst 12°C. ...

Viel Irrelevantes und Informationen doppelt  
⇒ **DRY-Prinzip**

## (Meta)Daten

Es geht hier um Temperaturbeobachtungen (gemessen in °C) an einem bestimmten Ort (irgendwo in München) zu jeweils einer bestimmten Zeit (12 Uhr)

Diese Informationen sind fest und ändern sich nicht mit der Beobachtung. Solche Art von Informationen heißen **Metadaten**

Die eigentlichen variablen Daten lassen sich in einer Tabelle darstellen, mit jeder Beobachtung in einer Zeile

Wochentag	Datum	Temperatur
Sonntag	11.05.2014	15
Samstag	10.05.2014	20
Freitag	09.05.2014	12
⋮	⋮	⋮

## Datenformate - Textformat

Jedes Zeichen der Information wird als Text gespeichert.

Spezielle Zeichen trennen die Spalten innerhalb einer Zeile

**Vorteil:** Sehr einfaches Format, einfach zu editieren

**Nachteil:** Geht in der Regel nur für Datensätze im sog. Rechteck-Schema

Beispiel: CSV- oder Fixed-Width-Format

R: Einlesen wird direkt unterstützt

Zeichen werden unterschiedlich gespeichert, je nach Typ:  
Ganze Zahl, reelle Zahl, Text, ...

⇒ *Intelligentes* Speichern

**Vorteil:** Spart Platz, kann zur Vermeidung von Redundanz verwendet werden

**Nachteil:** Man braucht spezielle Software dazu, die auflöst, was als was gespeichert ist.

Beispiel: SPSS sav-Dateien, Excel-Spreadsheets

R: Einlesen wird durch Zusatzpakete unterstützt

Jede Datei auf dem Computer liegt an einem Ort in der Ordner-Baum-Struktur, identifiziert über den sogenannten *Pfad*.

Jeder Ordner Ebene wird über Pfadtrenner verbunden, bei MS Windows \, bei Mac und \*nix-Systemen /

```
> # Geht auch unter Windows
> (pfad <- "Der/Pfad/zu/meiner/Datei")
[1] "Der/Pfad/zu/meiner/Datei"
> # Nur unter Windows
> (pfad_ms <- "Der\\Pfad\\zu\\meiner\\Datei")
[1] "Der\\Pfad\\zu\\meiner\\Datei"
> # Pfadtrenner automatisch eingefuegt
> (pfad_r <- file.path("Der", "Pfad", "zu", "meiner", "Datei"))
[1] "Der/Pfad/zu/meiner/Datei"
```

## Relative vs. Absolute Pfadangabe

Bei der absoluten Angabe muss man immer im OS-Wurzelverzeichnis starten (Laufwerksbuchstabe bei Windows)

```
> "C:/Der/absoulte/Pfad/zu/meiner/Datei"
[1] "C:/Der/absoulte/Pfad/zu/meiner/Datei"
```

Bei der relativen Angabe nimmt R das aktuelle Arbeitsverzeichnis und geht dann herunter

```
> "Der/relative/Pfad/zu/meiner/Datei"
[1] "Der/relative/Pfad/zu/meiner/Datei"
> # entspricht
> # file.path(getwd(),"Der/relative/Pfad/zu/meiner/Datei")
```

Spezielle Verzeichnisnamen:

".." Verzeichnisebene oberhalb der aktuellen  
"." Aktueller Ordner Ebene

## Arbeitsverzeichnis

Verzeichnis, in dem man mit R arbeitet:

- Speicherort für History und Workspace
- Wurzelverzeichnis für relative Pfadangaben

```
> getwd() # Abfragen
> setwd("Pfad/zum/arbeitsverzeichnis")
> ?read.table
```

**DIE** grundlegende Funktion in R zum Einlesen von Textformaten ist `read.table()`

Funktionen zum einlesen von bestimmten Textformaten, rufen meist nur die Funktion `read.table()` mit anderen vorgegeben Argumenten auf, zum Beispiel `read.csv2()`

Hilfe: `?read.table`

Die Funktion liefert nach dem Einlesen ein Objekt von Typ `data.frame` zurück

## Einlesen in R - Binärformat

Daten in Excel-Format `xls(x)` :

- Im csv-Format speichern und als Textformat einlesen
- Einlesen mit Funktion `read.xlsx()` aus Paket `openxlsx`

Daten in SPSS-Format `sav` :

- Einlesen mit Funktion `read.spss()` aus Paket `foreign`

**TIPP:** Einlesen aus Binärformaten vermeiden!

Standardmäßig wandelt R beim Einlesen die Spalten in geeignete Formate um:

- Spalten mit Zahlen werden als `numeric` eingelesen
- Spalten mit Text werden als Faktor eingelesen

Wichtige Argumente für `read.table()`:

<code>header:</code>	Erste Zeile enthält Variablennamen
<code>sep:</code>	Trennzeichen zwischen den Spalten
<code>dec:</code>	Dezimaltrennzeichen (1.3 vs. 1,3)
<code>as.is:</code>	Keine automatische Umwandlung
<code>col.classes:</code>	Typ der Spalten vorgeben