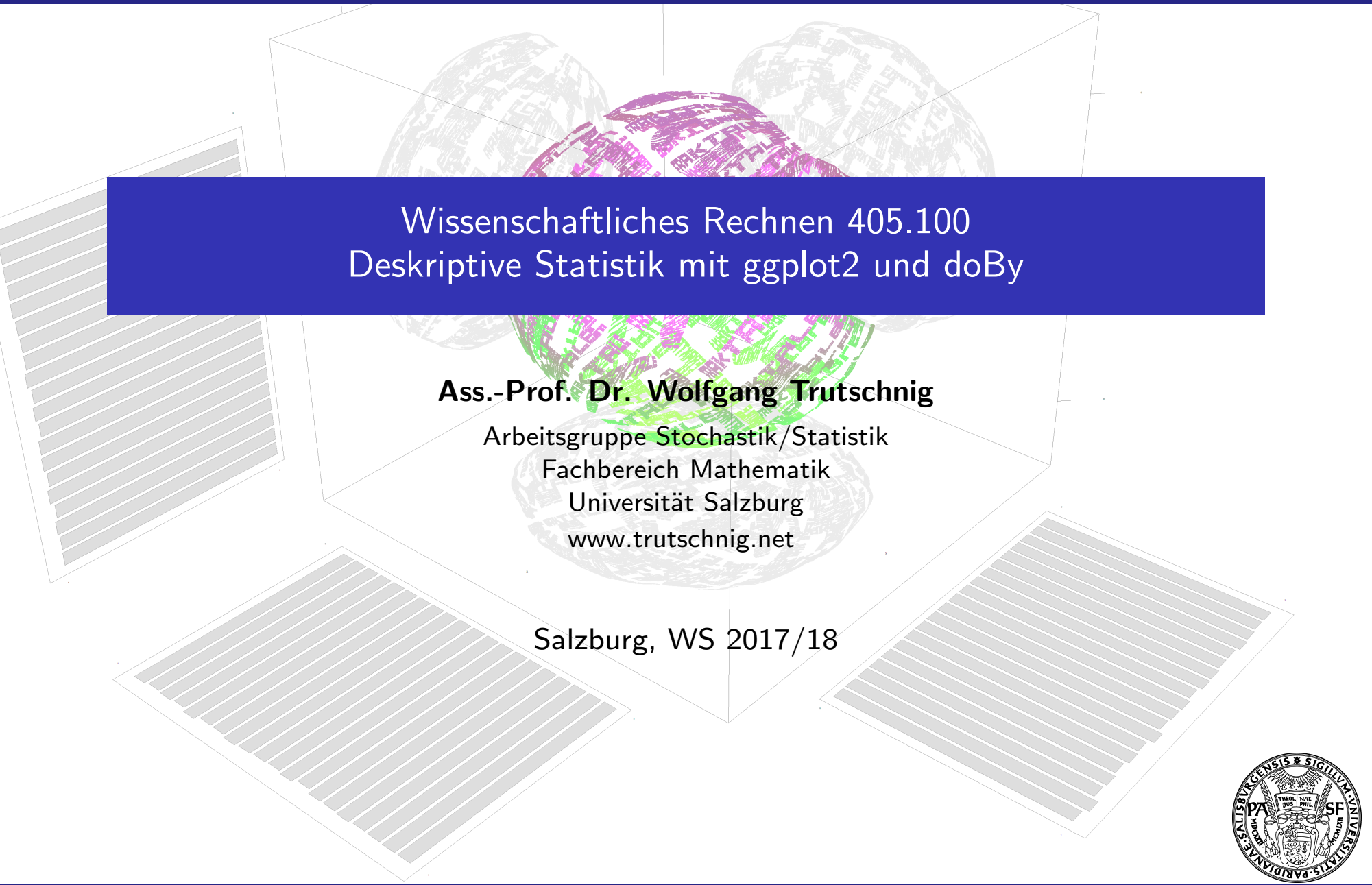


Wissenschaftliches Rechnen 405.100
Deskriptive Statistik mit ggplot2 und doBy

Ass.-Prof. Dr. Wolfgang Trutschnig

Arbeitsgruppe Stochastik/Statistik
Fachbereich Mathematik
Universität Salzburg
www.trutschnig.net

Salzburg, WS 2017/18

Worum geht's?

- Die einfache **Aggregation** und die Erstellung aussagekräftiger (und schön anzusehender) **Grafiken** sind in der Regel die ersten Schritte in der Analyse von Daten.
- Wir verwenden die folgenden zwei Pakete:
- **ggplot2**: Extrem leistungsstarkes und flexibles Paket zur einfachen Erstellung von eleganten Grafiken. Autor: Hadley Wickham (Entwicklung seit 2005, derzeit chief scientist @R-Studio), 'ggplot2: Elegant Graphics for Data Analysis' (Springer)
- **doBy**: (Ein mögliches) Paket zur einfachen Berechnung von 'groupwise summary statistics' (ohne Schleifen). '**do** something on data which is grouped **By** some variables'. Autor: Soren Hojsgaard

Learning by doing

- Um die Funktionalität beider Pakete kennenzulernen wird ein einziger (fast realer) Datensatz analysiert:
- ATM.txt, zu finden unter www.trutschnig.net/courses

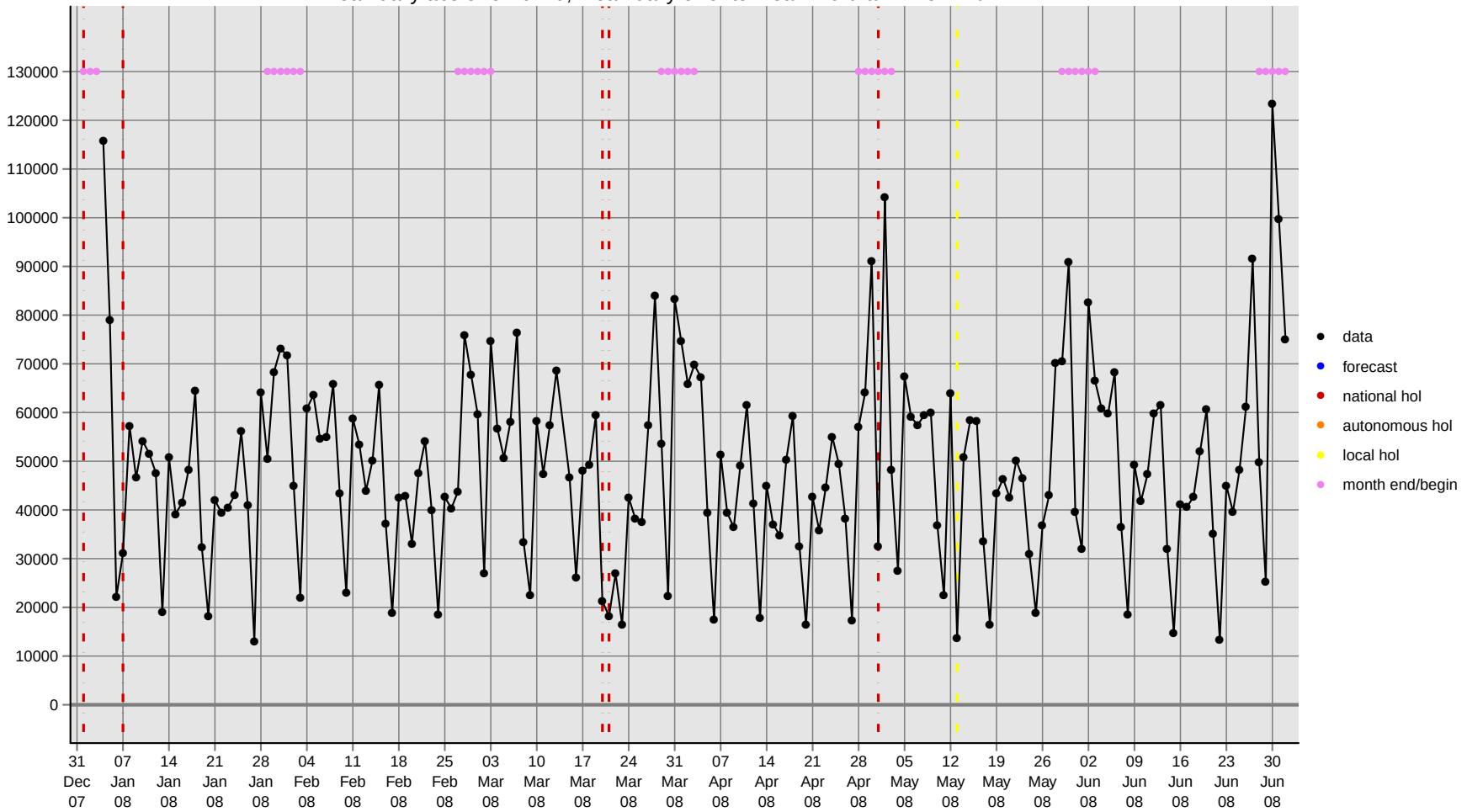
ymd	weekday	nr_weekday	sum_out	holiday
2007-01-01	Mon	1	4040	1.00
2007-01-02	Tue	2	22760	1.50
2007-01-03	Wed	3	18810	0.00
2007-01-04	Thu	4	24910	0.00
2007-01-05	Fri	5	25650	0.50
2007-01-06	Sat	6	5650	1.00

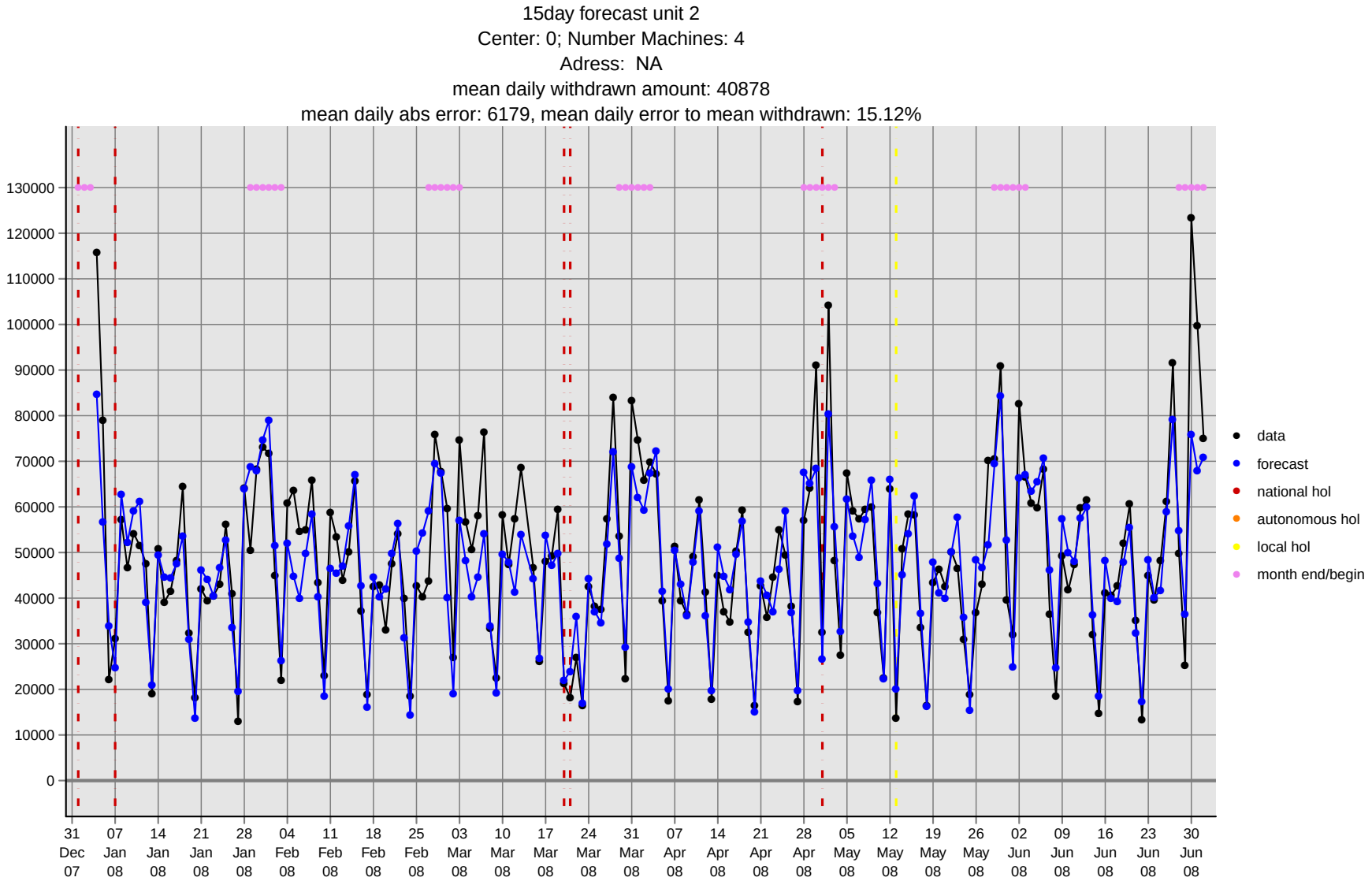
- Der Datensatz enthält die Zeitreihe der bei einem Bankomaten (einer Filiale einer Bank) abgehobenen täglichen Geldmenge.
- Ursprüngliche Problemstellung: Entwicklung von zuverlässigen forecasts für die abgehobenen täglichen Geldmenge zum Zwecke der Optimierung des Zuliefersystems (500 verschiedene Filialen, Zeitreihen von 3 Jahren).

Wir analysieren den Datensatz mit Hilfe von ggplot2 und doBy, und gehen wie folgt vor (siehe R-Codes_WR04.R):

1. Importieren der Daten und ein erster Überblick.
2. Übersichtlicher plot der Zeitreihe - welche Muster sind zu erkennen?
3. Jährliche Histogramme und boxplots - welche Muster sind zu erkennen?
4. Verfeinerte boxplots (violins) je Wochentag und Tag des Monats - welche Muster sind zu erkennen?
5. Was passiert an Feiertagen?
6. Exportieren schöner plots.
7. Forecasting der täglich abgehobenen Geldmenge?
8. Berechnen einfacher summary statistics via doBy, Vergleich mit den Grafiken.

15day forecast unit 2
Center: 0; Number Machines: 4
Adress: NA
mean daily withdrawn amount: 40878
mean daily abs error: 6179, mean daily error to mean withdrawn: 15.12%




15day forecast unit 2
Center: 0; Number Machines: 4
Adress: NA
mean daily withdrawn amount: 40878
mean daily abs error: 6179, mean daily error to mean withdrawn: 15.12%
130000
120000
110000
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
31 Dec 07
07 Jan 08
14 Jan 08
21 Jan 08
28 Jan 08
04 Feb 08
11 Feb 08
18 Feb 08
25 Feb 08
03 Mar 08
10 Mar 08
17 Mar 08
24 Mar 08
31 Mar 08
07 Apr 08
14 Apr 08
21 Apr 08
28 Apr 08
05 May 08
12 May 08
19 May 08
26 May 08
02 Jun 08
09 Jun 08
16 Jun 08
23 Jun 08
30 Jun 08
data
forecast
national hol
autonomous hol
local hol
month end/begin

- Gegeben eine Stichproben $x_1, x_2, \ldots, x_n \in \mathbb{R}$
- Grundidee der empirischen Verteilungsfunktion $\hat{F}_n$:
  Für jedes $x \in \mathbb{R}$ berechne den Anteil der Elemente x_i der Stichproben die $\leq x$ sind, i.e.

$$\hat{F}_n(x) := h_n((-\infty, x]) = \frac{1}{n} \# \Big\{ i : 1 \leq i \leq n : x_i \leq x \Big\}$$

- Damit erhalten wir eine Funktion $\hat{F}_n : \mathbb{R} \to \mathbb{R}$ mit folgenden Eigenschaften:

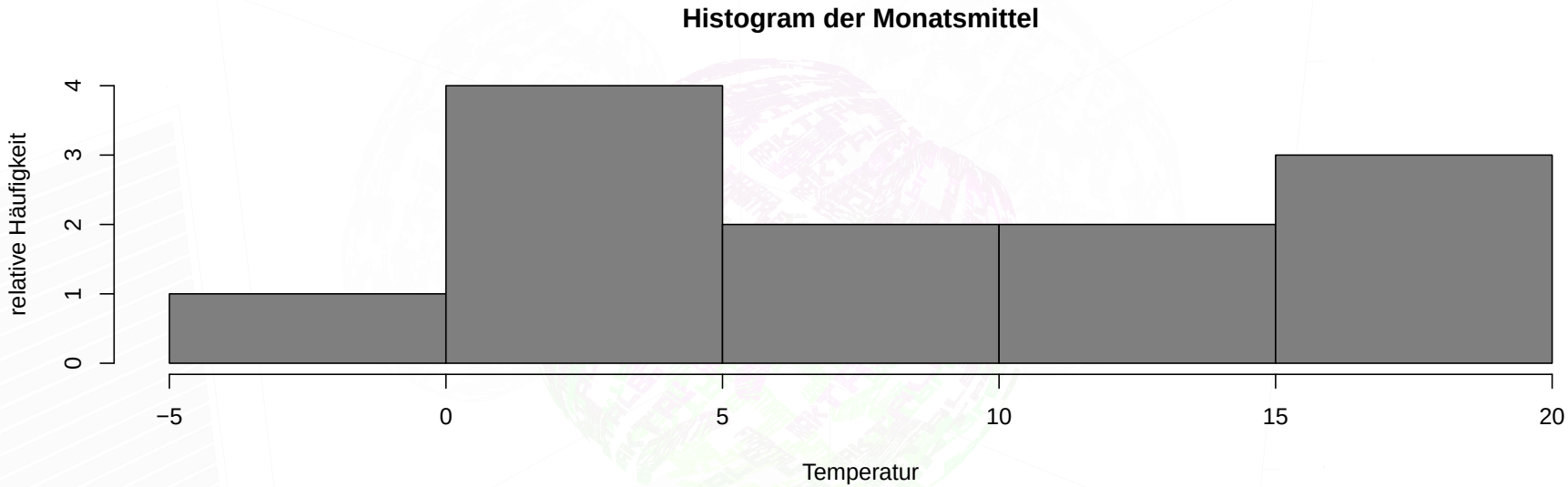
  (F1) $\hat{F}_n(x) \in [0, 1]$ für jedes $x \in \mathbb{R}$.

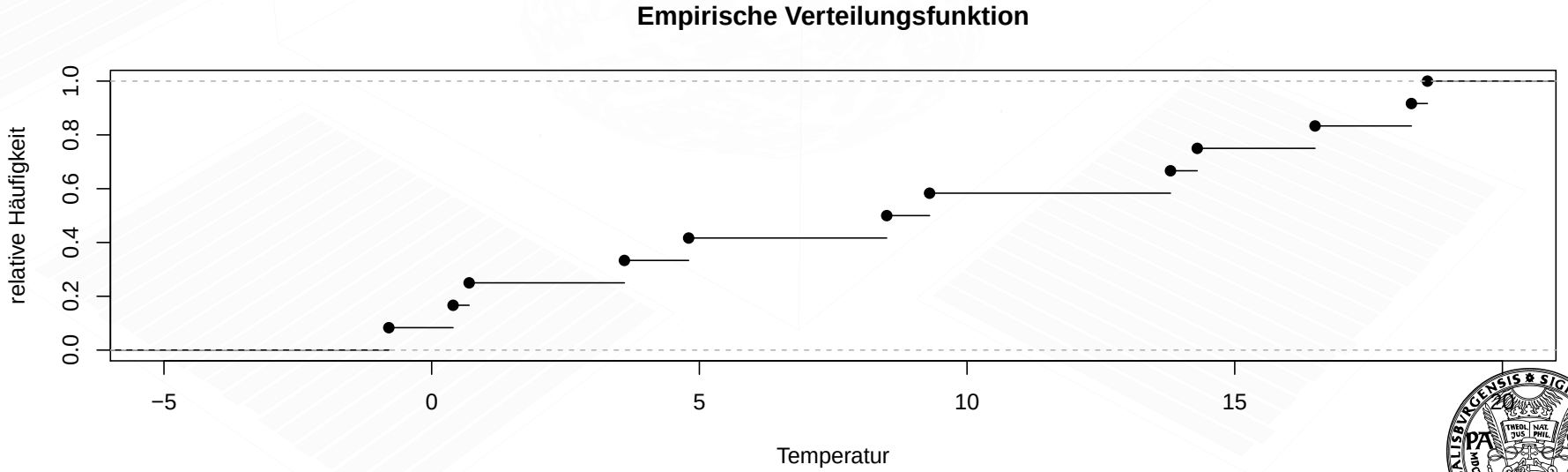
  (F2) Aus $x \leq y$ folgt $\hat{F}_n(x) \leq \hat{F}_n(y)$.

  (F3) $\hat{F}_n(x) = 0$ für $x < \min\{x_i : 1 \leq i \leq n\}$.

  (F4) $\hat{F}_n(x) = 1$ für $x \geq \max\{x_i : 1 \leq i \leq n\}$.

  (F5) Jedes x_i ist eine Sprungstelle von $\hat{F}_n$.


Histogram der Monatsmittel
relative Häufigkeit
0
1
2
3
4
−5
0
5
10
15
20
Temperatur


Empirische Verteilungsfunktion
relative Häufigkeit
0.0
0.2
0.4
0.6
0.8
1.0
−5
0
5
10
15
20
Temperatur

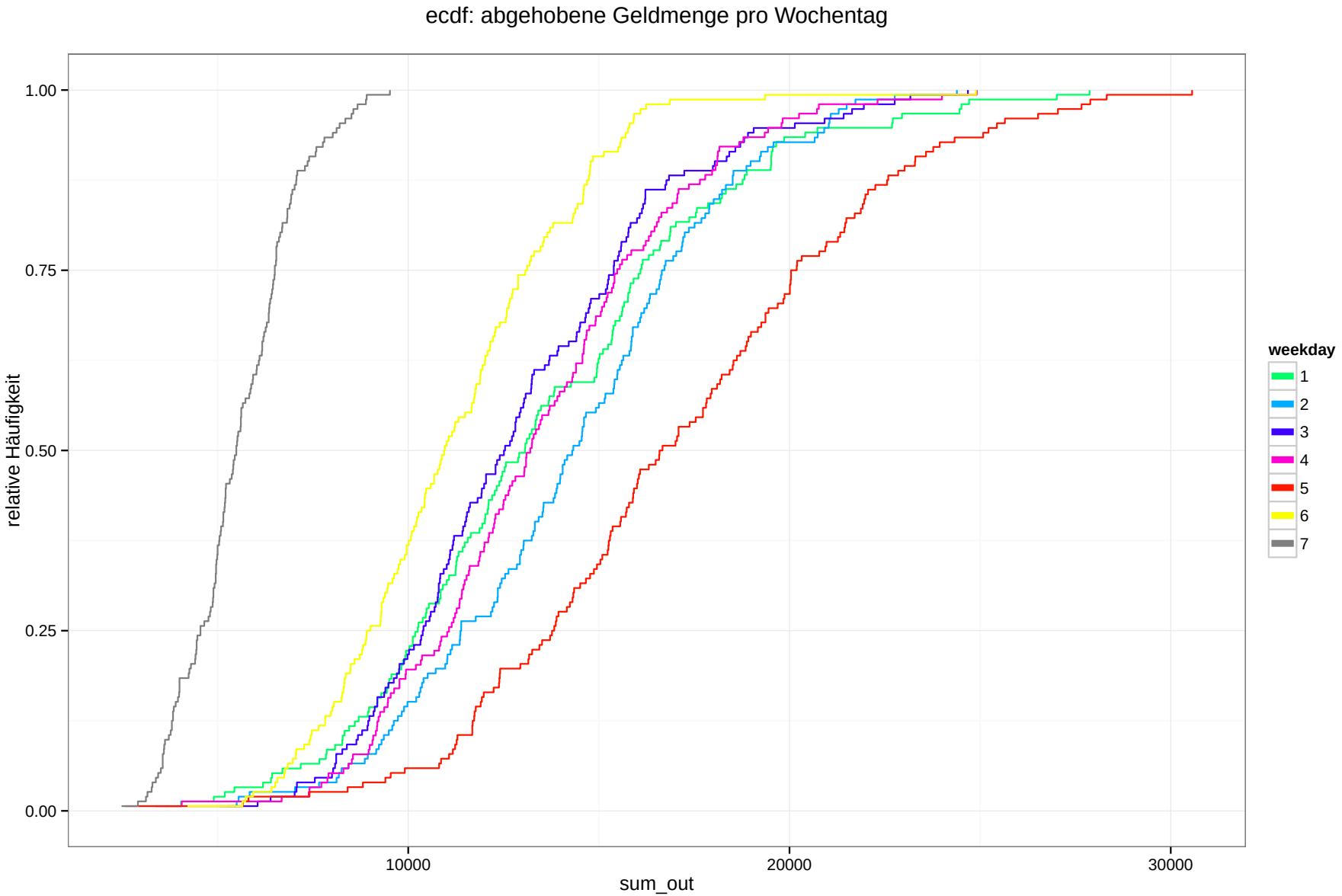
Definition

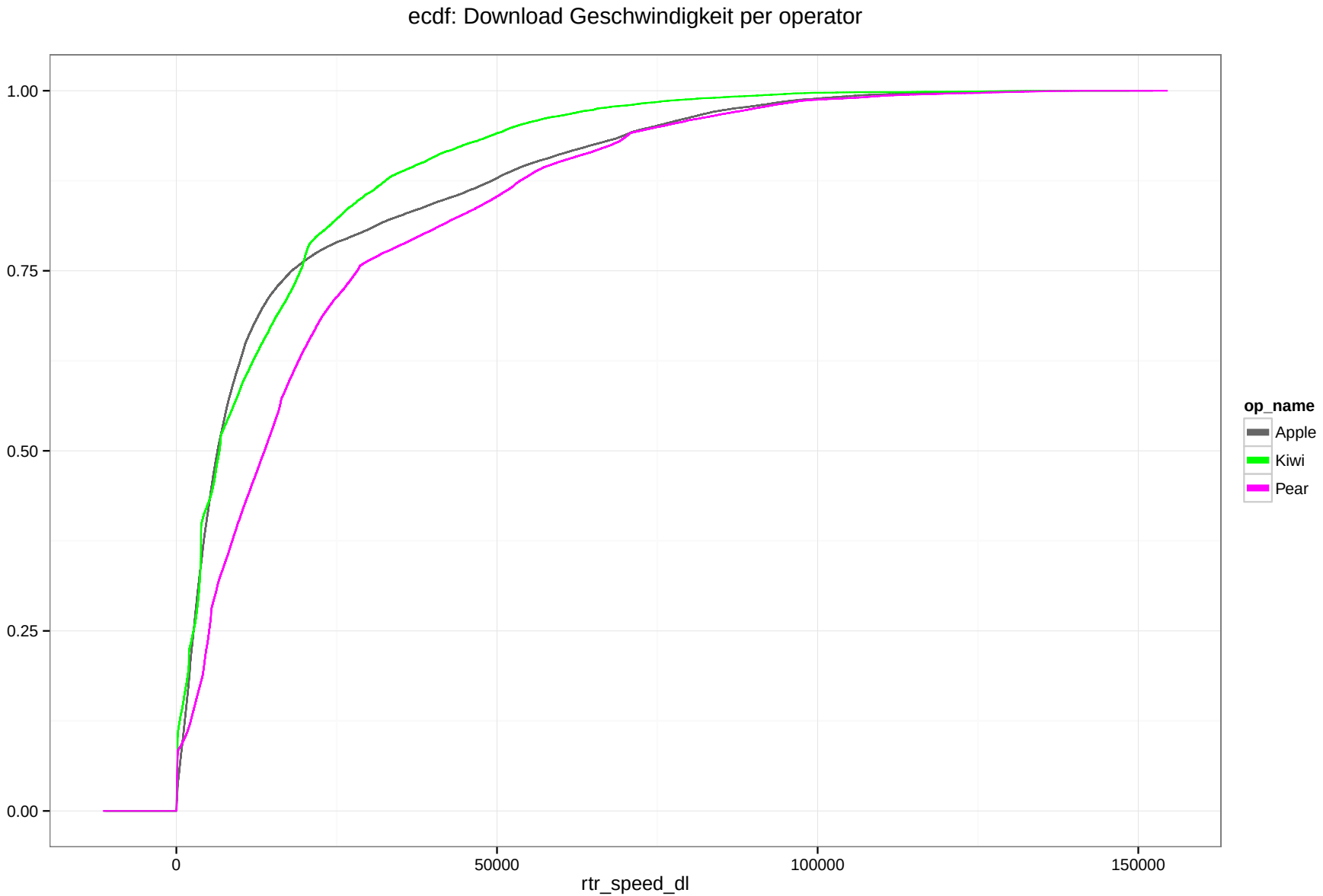
Gegeben eine Stichproben $x_1, x_2, \ldots, x_n \in \mathbb{R}$. *Dann heisst die durch*

$$\hat{F}_n(x) := h_n((-\infty, x]) = \frac{1}{n}\#\Big\{i : 1 \leq i \leq n : x_i \leq x\Big\}$$

definierte Funktion $\hat{F}_n : \mathbb{R} \to \mathbb{R}$ *die empirische Verteilungsfunktion (empirical cumulative distribution function, kurz: ecdf) der Stichprobe* $x_1, x_2, \ldots, x_n$.

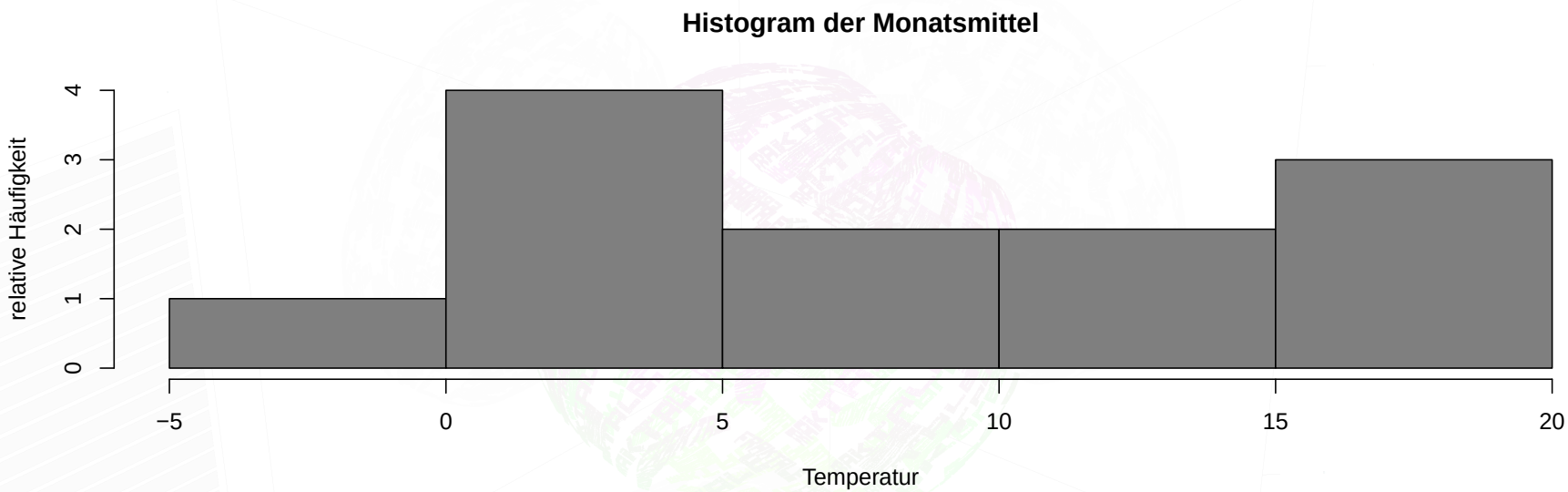
- Lässt sich die komplette Stichprobe $x_1, \ldots, x_n$ rekonstruieren, wenn $\hat{F}_n$ bekannt ist?
- Anders formuliert: Enthält die ecdf genau so viel Information wie die Stichprobe $x_1, x_2, \ldots, x_n$ oder geht - wie bei Histogrammen - Information verloren?

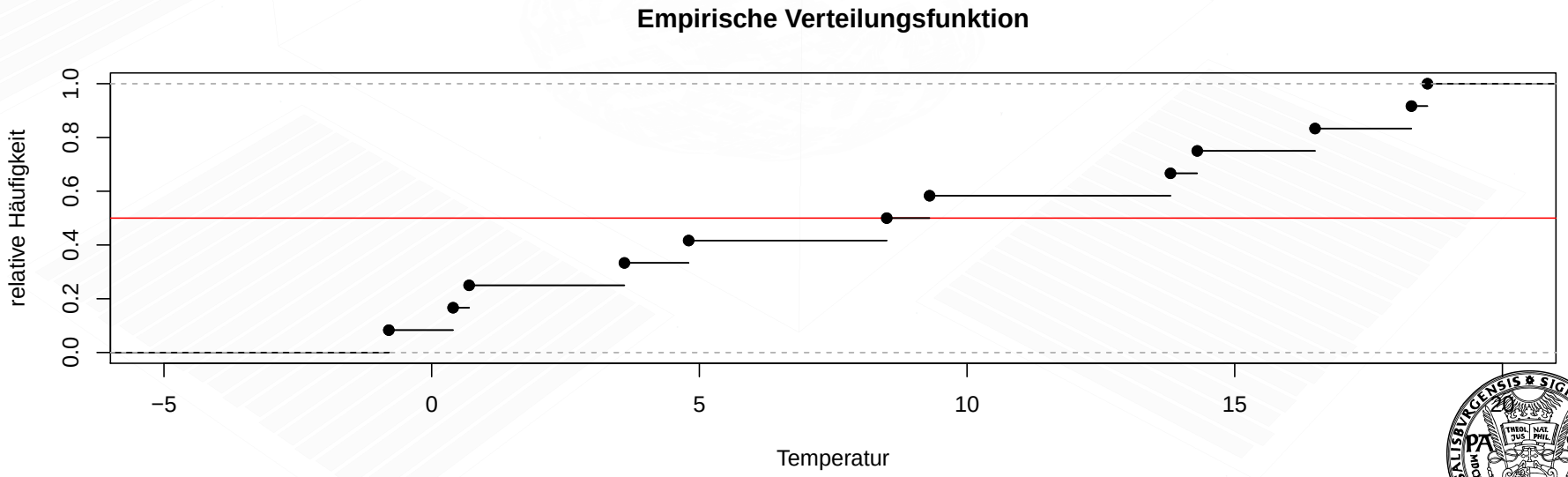

ecdf: abgehobene Geldmenge pro Wochentag
relative Häufigkeit
1.00
0.75
0.50
0.25
0.00
10000
20000
30000
sum_out
weekday
1
2
3
4
5
6
7

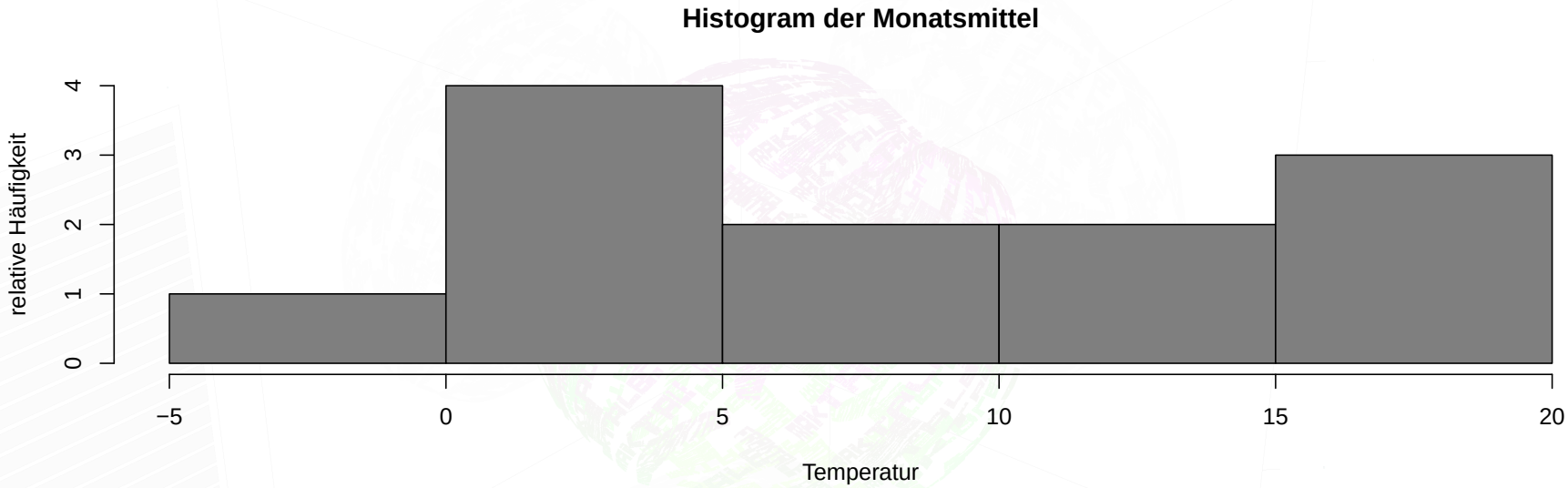

ecdf: Download Geschwindigkeit per operator
1.00
0.75
0.50
0.25
0.00
0
50000
100000
150000
rtr_speed_dl
op_name
Apple
Kiwi
Pear

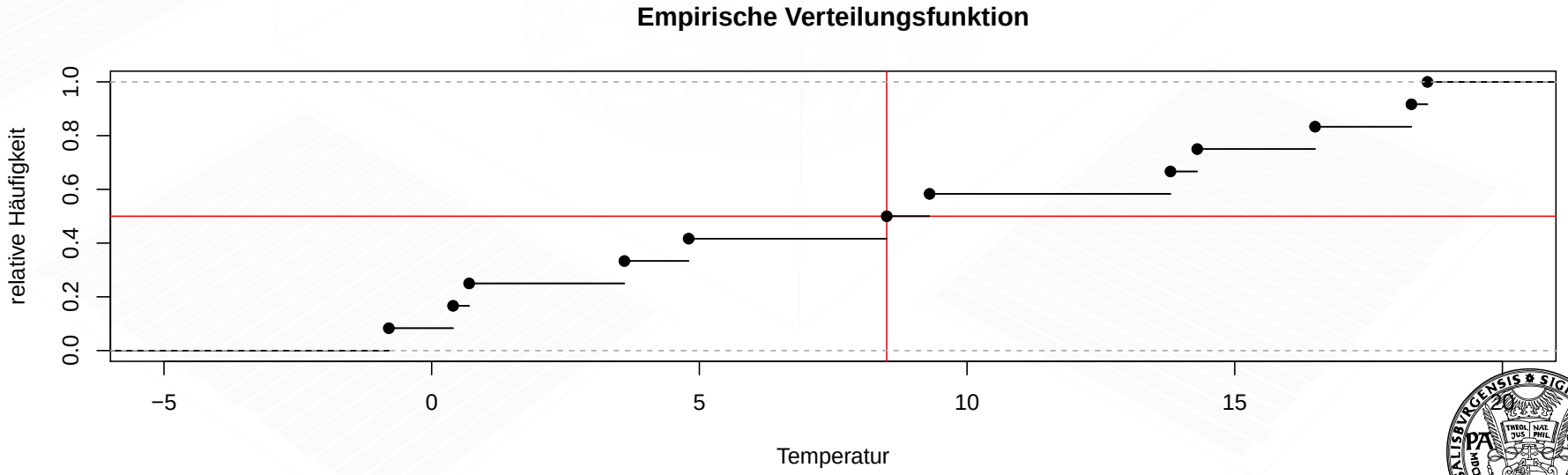
- Gegeben eine Stichproben $x_1, x_2, \ldots, x_n \in \mathbb{R}$.
- Oft betrachtet man nicht die komplette empirische Verteilungsfunktion $\hat{F}_n$, sondern nur spezielle Punkte (Schwellenwerte) der Verteilungsfunktion.
- Angenommen, wir suchen den kleinsten Wert z, sodass $\hat{F}(z) \geq 0.5$ gilt.
- Wie kann dieser Wert $z_{0.5}$ ermittelt werden - rechnerisch und/oder graphisch?
- Welche Eigenschaften hat $z_{0.5}$?
- Wie viele Werte x_i erfüllen $x_i \leq z_{0.5}$, wie viele $x_i \geq z_{0.5}$?
- $z_{0.5}$ heisst *Median* von $x_1, \ldots, x_n$.

Quantile und Boxplots






Histogram der Monatsmittel
relative Häufigkeit
0
1
2
3
4
−5
0
5
10
15
20
Temperatur


Empirische Verteilungsfunktion
relative Häufigkeit
0.0
0.2
0.4
0.6
0.8
1.0
−5
0
5
10
15
20
Temperatur

- Starten wir mit $p = 0.75$ statt mit 0.5 und definieren $z_{0.75}$ als kleinsten Wert z, sodass $\hat{F}_n(z) \geq 0.75$, dann können wir vollkommen analog vorgehen.
- Wir landen bei folgender allgemeinen Definition:

Definition (Quantil)

Gegeben eine Stichproben $x_1, x_2, \ldots, x_n \in \mathbb{R}$, deren empirische Verteilungsfunktion $\hat{F}_n$ und $p \in (0, 1]$. Dann heisst z_p, definiert durch

$$z_p := \min\{x \in \mathbb{R} : \hat{F}_n(x) \geq p\},$$

das p-Quantil (p-quantile) der Stichprobe $x_1, x_2, \ldots, x_n$.
$z_{0.5}$ heisst auch Median, $z_{0.25}$ unteres Quartil und $z_{0.75}$ oberes Quartil.

- Gilt immer $z_p \in \{x_1, \ldots, x_n\}$?
- Oft werden Quantile auch ausgehend von der geglätteten empirischen Verteilungsfunktion $\overline{F}_n$ berechnet.
- In R kann der user aussuchen, welche Art der Berechnung er will $\rightarrow$ siehe Aufgabe am Übungsblatt.
- Median, unteres und oberes Quartil bilden die Grundlagen der sogenannten *boxplots*.
- Boxplots sind ein einfaches, aber sehr nützliches Werkzeug, um sich einen ersten Überblick über numerische Daten zu verschaffen.
- Erlaubt übersichtliche Gegenüberstellung der Werte einer Variable für verschiedene Gruppen.