Statistik mit R für Fortgeschrittene (Interne Weiterbildung FOR SS16-08) Block 1: Ausrichtung

Ass.-Prof. Dr. Wolfgang Trutschnig

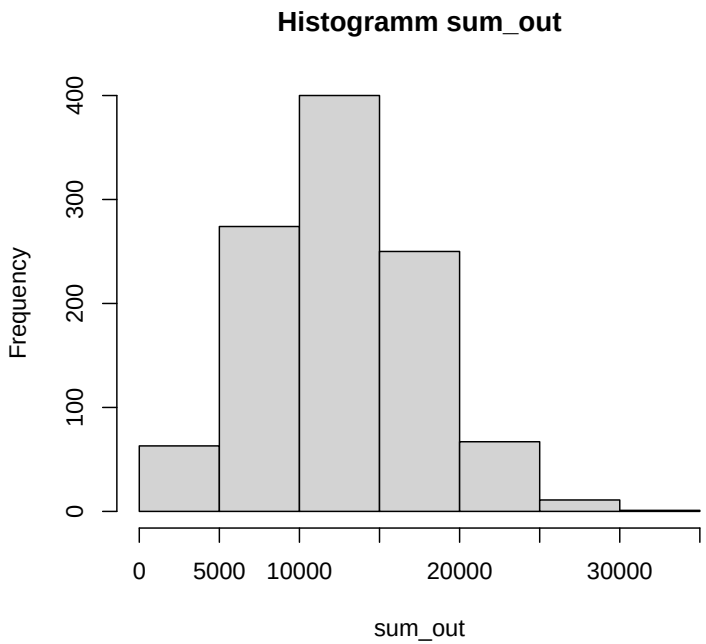
Arbeitsgruppe Stochastik/Statistik
FB Mathematik
Universität Salzburg
www.trutschnig.net

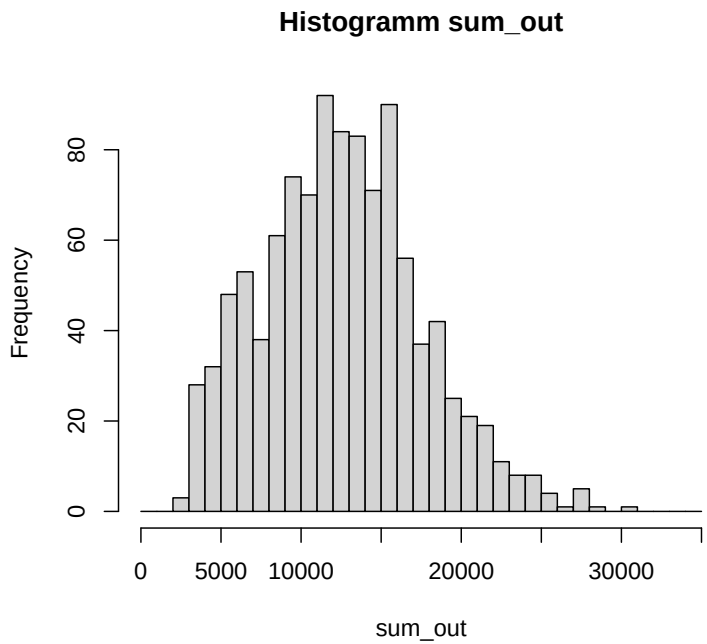
Salzburg, 2016-06-03

Mögliche Inhalte (Diskussionsgrundlage)

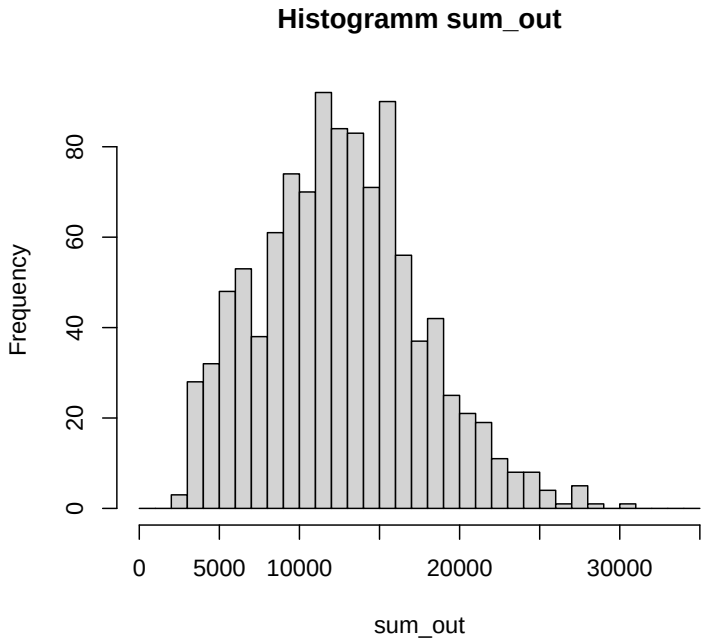
1. Fokus auf Statistik, R als Werkzeug
2. Fokus auf effizientes Programmieren in R, Datenanalyse als Werkzeug
3. R-shiny: interaktive apps mit R (web application framework for R)
   - Sehr nützlich für die Lehre
   - Ermöglicht 'Herumspielen' mit Daten
4. knitR: dynamic reporting mit R
   - Kombination R mit LaTeX (Textverarbeitung)
   - Stichwort: Reproduzierbare Forschung
5. Wünsche?

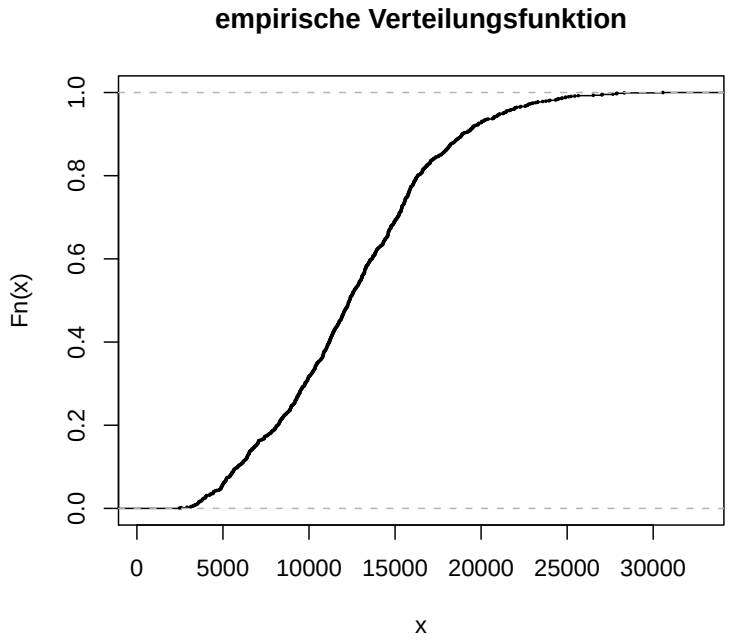
- Gegeben sind numerische Daten (sample) $x_1, \ldots, x_n$ im Intervall $[a, b]$
- **Histogramm:** Übersichtliche, einfache Darstellung der Daten: Zerlege $[a, b]$ in Intervalle $I_1, \ldots, I_k$ und **zähle** wie viele Werte in welchem Intervall liegen, d.h. $h_n(I_j) := \#\{m : x_m \in I_j\}$, $r_n(I_j) := h_n(I_j)/n$



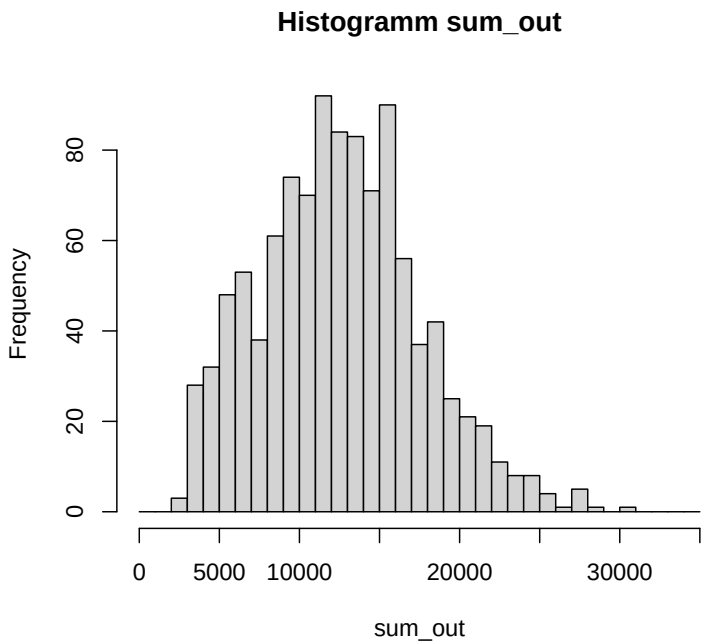


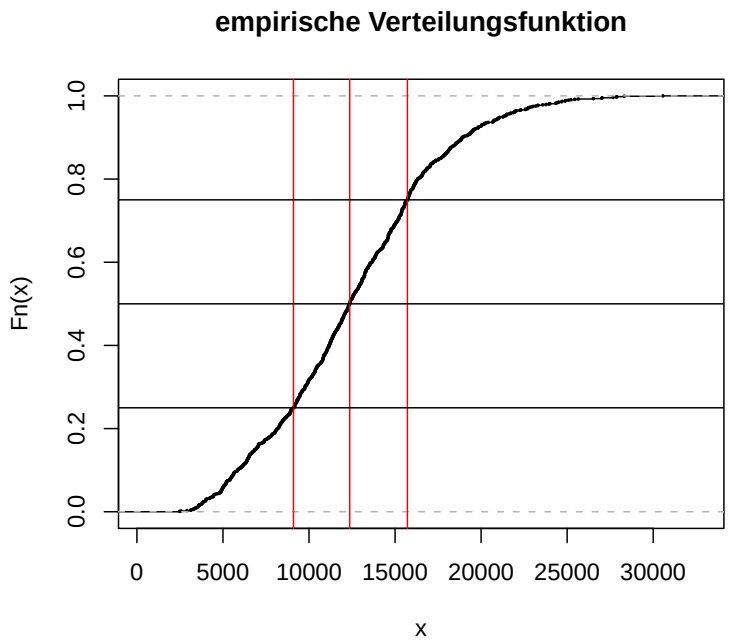
- Gegeben sind numerische Daten (sample) $x_1, \ldots, x_n$ im Intervall $[a, b]$
- **Empirische Verteilungsfunktion:** Übersichtliche, einfache Darstellung der Daten: für jedes $x \in [a, b]$ zähle wie viele Werte $\leq x$ sind, d.h. $F_n(x) := \#\{i : x_i \leq x\}/n$



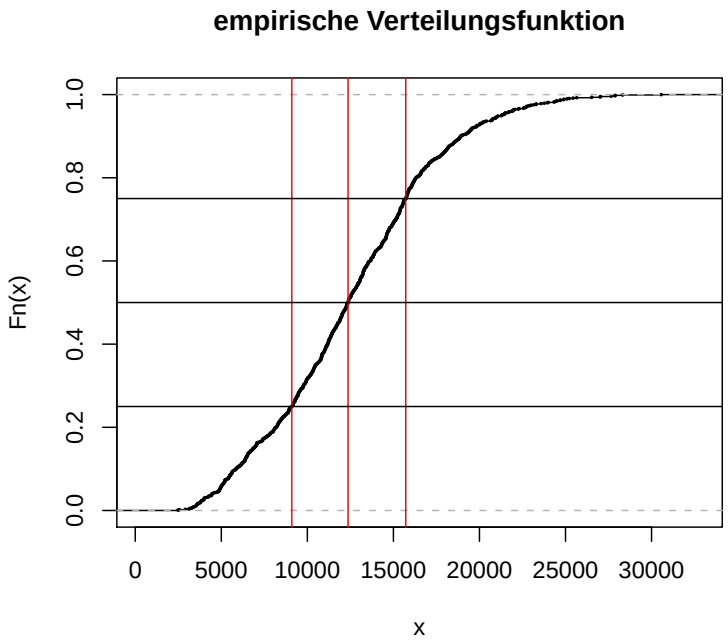


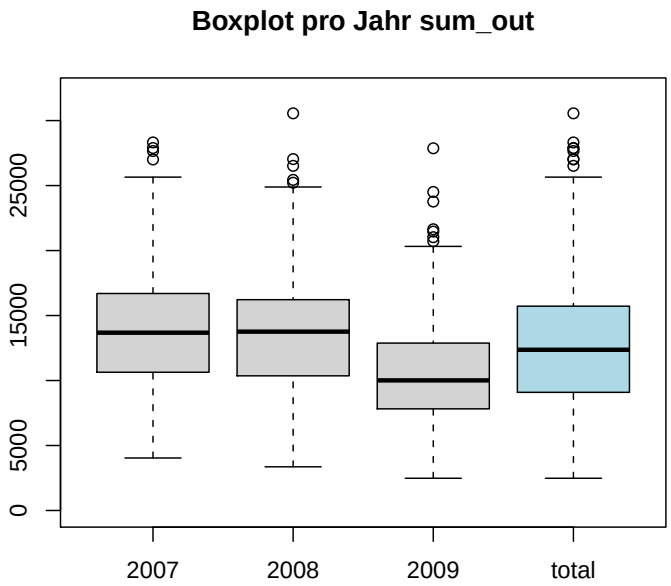
- Gegeben sind numerische Daten (sample) $x_1, \ldots, x_n$ im Intervall $[a, b]$, F_n bezeichne die empirische Verteilungsfunktion.
- Für jedes $p \in [0, 1]$ heisst $F^{(-1)}(p) := \min\{x : F_n(x) \geq p\}$ **p-Quantil** der Stichprobe.



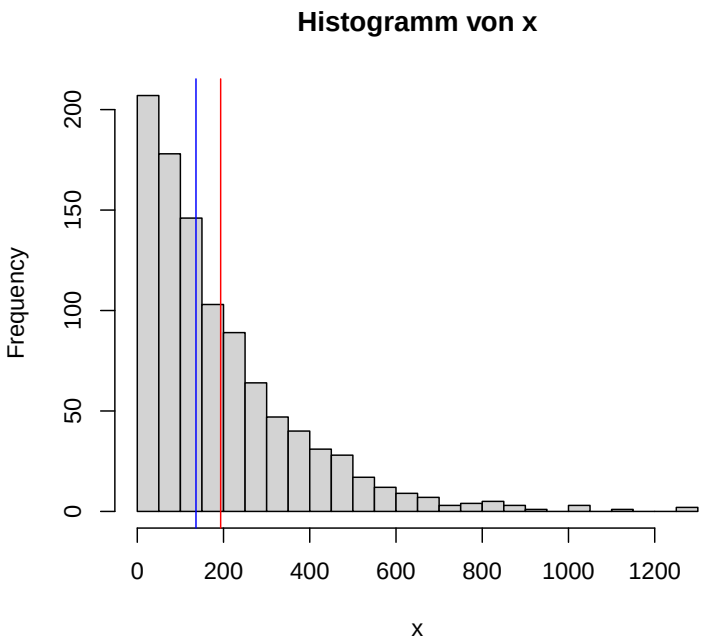


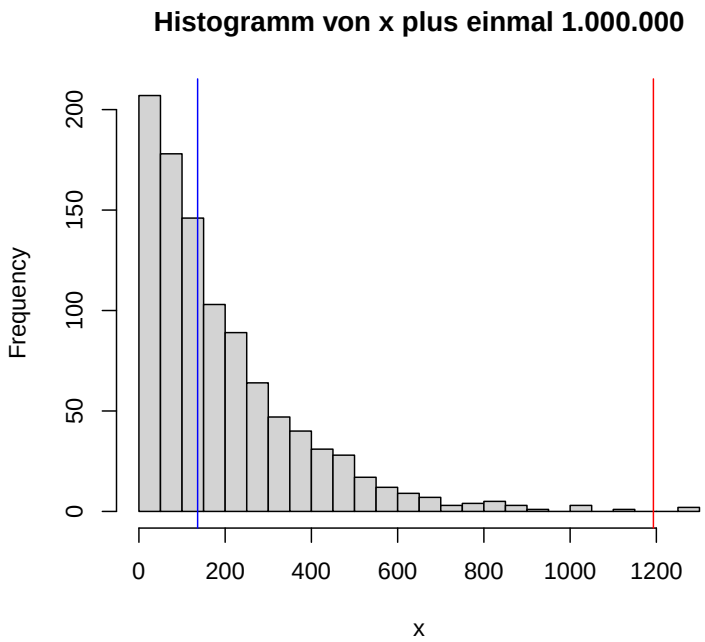
- Gegeben sind numerische Daten (sample) $x_1, \ldots, x_n$ im Intervall $[a, b]$
- Ein **boxplot** ist eine zusammenfassende Darstellung basierend auf den 0.25-, 0.5-, 0.75-Quantilen und Ausreißern:





- Warum bisher **Mittelwert** nicht einmal erwähnt?
- OeNB: "Der durchschnittliche österreichische Haushalt verfügte 2004 über ein Geldvermögen von rund 55.000 Euro".
- Informationsgehalt ?
- Mittelwert ist sehr sensitiv auf Ausreißer - Veränderung eines einzigen Wertes kann Mittelwert extrem verändern - **nicht robust !**





Learning by doing

- Verwendung der deskriptiven tools zur Analyse eines ersten realen Datensatzes:

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).
- Komplettes Skript unter www.trutschnig.net/courses