

## House Votes (cont.)

Ziel ist es Abgeordnete zu finden, die sehr oft gegen die eigene Mehrheit stimmen.

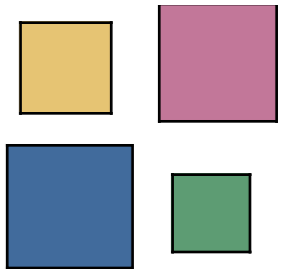
Funktion, die zu einer Abstimmung (Spalte) liefert, ob die Demokraten dafür gestimmt haben.

```
demYes <- function(i) {  
  t1 <- table(HV1[HV$Party=="democrat",i])  
  t1[1] < t1[2]  
}
```

Bereitstellung dieser Information in einem Array

```
n <- dim(HV1)[1]  
k <- dim(HV1)[2]  
Votes <- matrix(0, n, k)  
demYesA <- rep(0, k)  
  
for(j in 1:k) {  
  demYesA[j] <- demYes(j)  
}
```

Markieren, ob ein Abgeordneter nicht für das “Ja” der Demokraten gestimmt hat.



## House Votes (cont.)

```
for(i in 1:n) {
  for(j in 1:k) {
    if( !is.na(HV1[i,j])) {
      if(HV1[i,j] == "n" && demYesA[j])
        Votes[i,j] <- 1
    }
  }
}
```

Wie oft stimmten wie viele Abgeordnete gegen die Demokraten?

```
barplot(table(apply(Votes, 1, sum)))
```

Das gleiche mit iPlots gelinkt mit der Partei

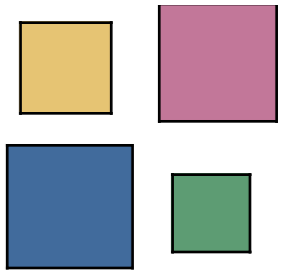
```
ibar(apply(Votes, 1, sum)); ibar(HV[,2])
```

Wie oft enthielten sich einzelne Abgeordnete?

```
barplot(table(apply(HV1, 1, function(x){sum(is.na(x))})))
```

Gab es einen Abgeordneten der nie abstimmte?

```
> HV1[apply(HV1, 1, function(x){sum(is.na(x))})==16,]
  HandicapInf WaterProj Budget PhysicianFee ElSalvador ReligiousInSchool
249      <NA>      <NA>   <NA>          <NA>          <NA>          <NA>
  AntiSatelliteTest ContrAsAid MxMissile Immigration SynfuelsCut EducationSpend
249      <NA>      <NA>   <NA>          <NA>          <NA>          <NA>
  SuperfundSue Crime DutyFreeExport SAfricaExports
249      <NA> <NA>          <NA>          <NA>
```



## EDA: Was es NICHT ist

- **Statistik ohne Mathematik**

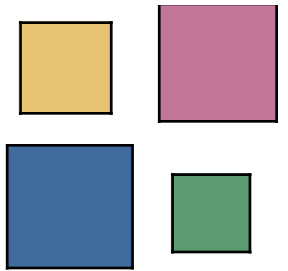
Strikte Verteilungsannahmen werden in der EDA nicht gemacht, **aber** das ganze Instrumentarium der math. Statistik steht der EDA zur Verfügung!

- **Deskriptive Statistik**

Die EDA bedient sich vieler Daten beschreibenden Hilfsmittel, **aber** der Bereich der deskriptiven Statistik alleine ist zu simplifizierend.

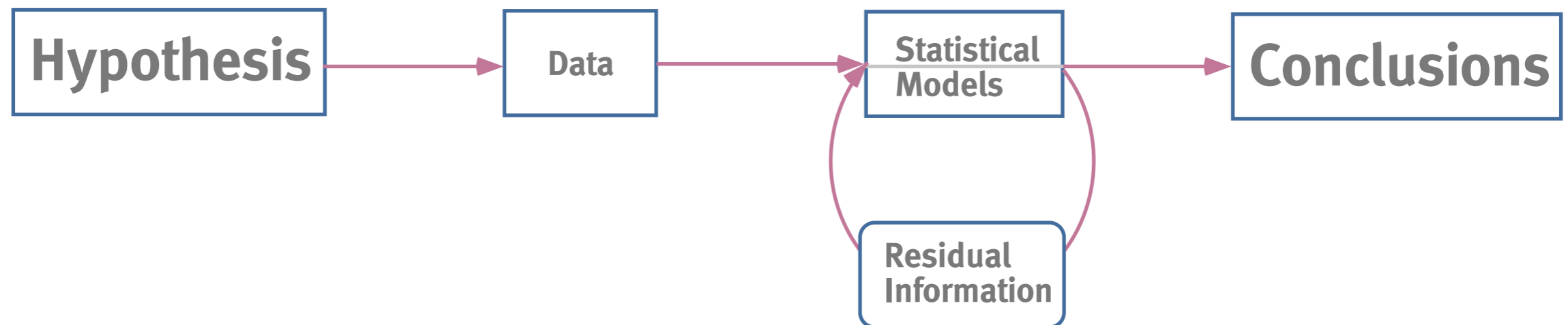
- **Daten Visualisierung**

Graphische Darstellungen sind ein Kernstück der EDA, **aber** um statistische Sicherheit für eine Aussage zu bekommen sind numerische, quantitative Methoden unerlässlich.

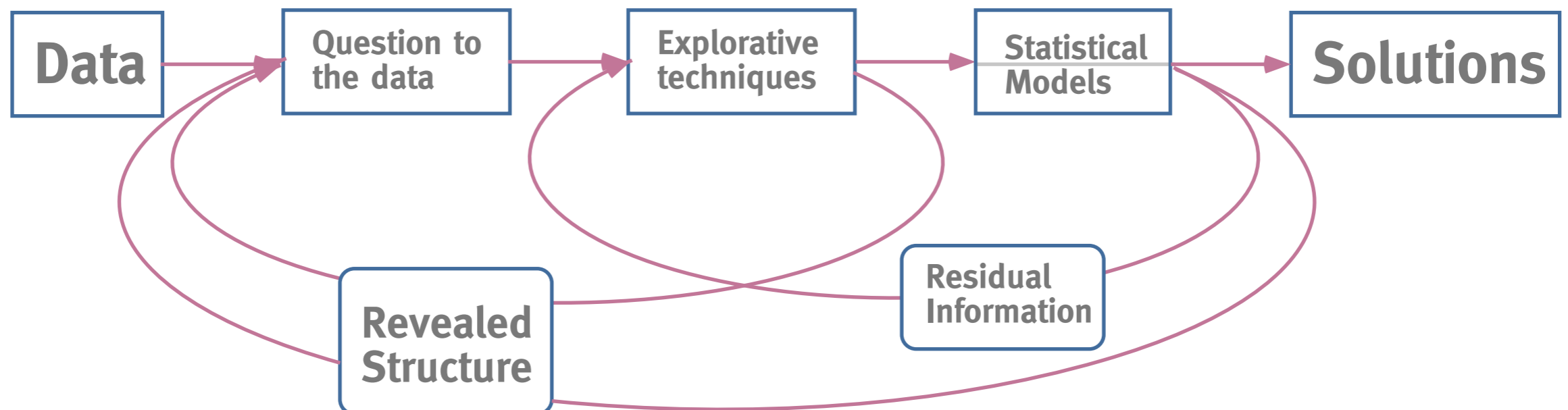


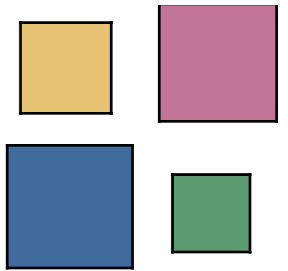
# EDA vs. klassische Statistik

## Classical



## Exploratory





## EDA: Zitate von Tukey

- **1964**

“... must be considered as an open-ended, highly interactive, iterative process, whose actual steps are segments of a stubbornly branching, tree-like pattern of possible actions.”

- **1977**

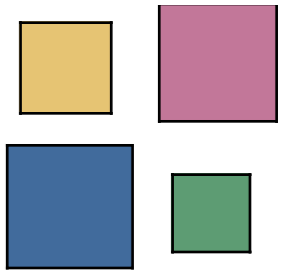
“There is no excuse for failing to plot and look.”

- **1980**

“Finding the question is often more important than finding the answer.”

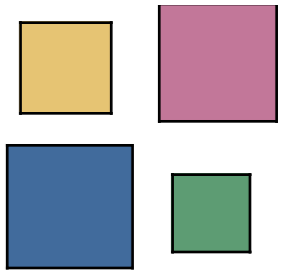
“EDA is

- an attitude, AND
- a flexibility, AND
- some graph paper (or transparencies, or both )”



## Tukey: EDA in der Lehre (1980)

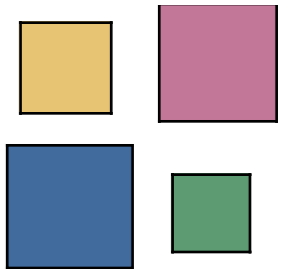
1. There is NO question of teaching confirmatory OR exploratory – we need both.
2. We need to think about science and engineering more broadly than the narrow, inadequate paradigm of a straight line from question to answer.
3. When we want to do careful confirmation on important questions we need to be very careful – randomizing and avoiding multiplicity
4. We need to teach exploratory as an attitude, as well as some helpful techniques, and we probably need to teach it before confirmatory.



## EDA: Tukeys Buch (1977)

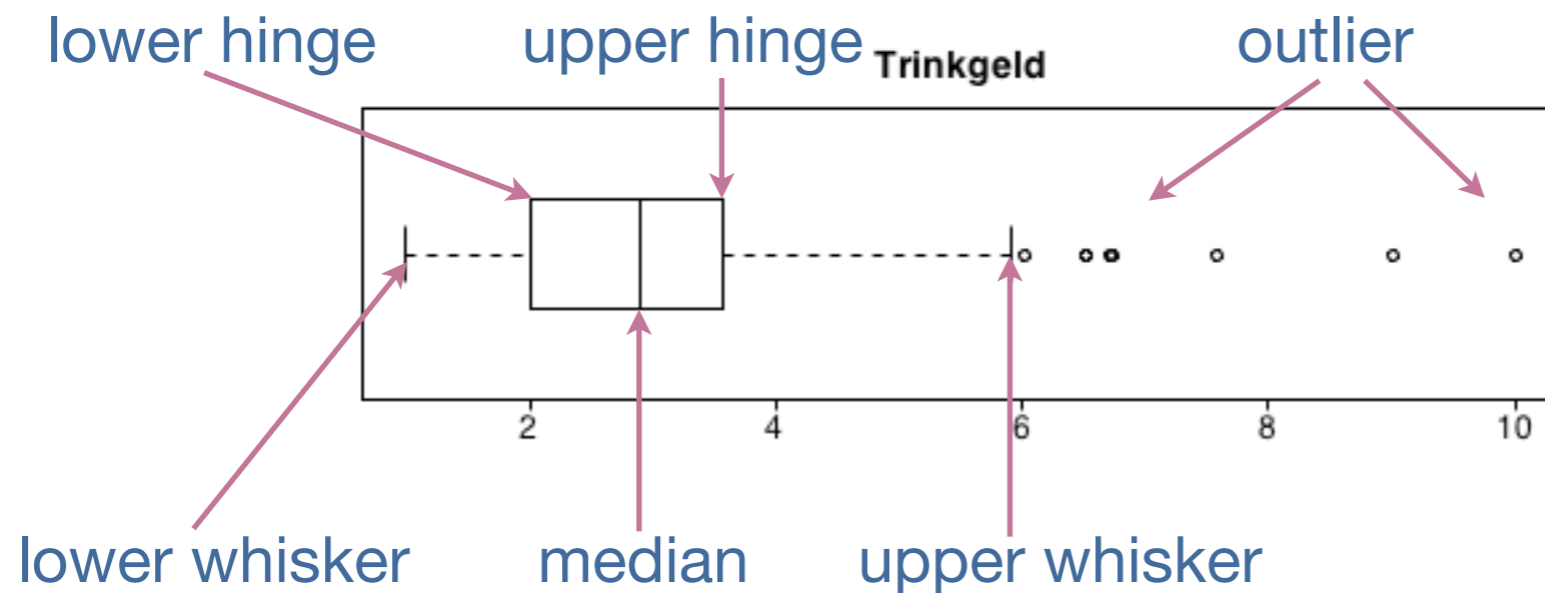
- Headlines
  - SCRATCHING DOWN NUMBERS (stem-and-leaf)
  - SCHEMATIC SUMMARIES (pictures and numbers)
  - EASY RE-EXPRESSION
  - EFFECTIVE COMPARISON (including well-chosen expression)
    - ...
  - SHAPES OF DISTRIBUTION
  - MATHEMATICAL DISTRIBUTIONS
  - POSTSCRIPT
    - A Our relationship to the computer



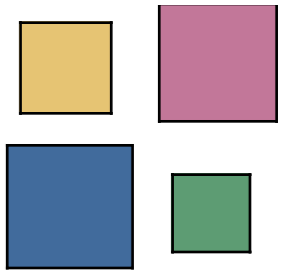


# EDA: Schematic Summaries

- Boxplots



- median: mittlere Beobachtung
- hinges: median der Werte größer/kleiner des Median
- h-spread: Differenz zwischen den beiden hinges
- whiskers: Linien bis zum ersten Wert kleiner/größer als das 1.5 fache des h-spread (jeweils ab dem dazugehörigen hinge gemessen)
- outlier: alles was weiter entfernt ist
- extreme outlier: weiter als das 3-fache des h-spread entfernt



## EDA: Easy re-expression

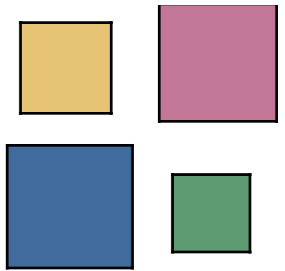
- Jede Art von dynamischer Transformation:

- Box-Cox

$$x_{BC}^{\lambda} := \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{für } x \neq 0 \\ \ln(x) & \text{für } x = 0. \end{cases}$$

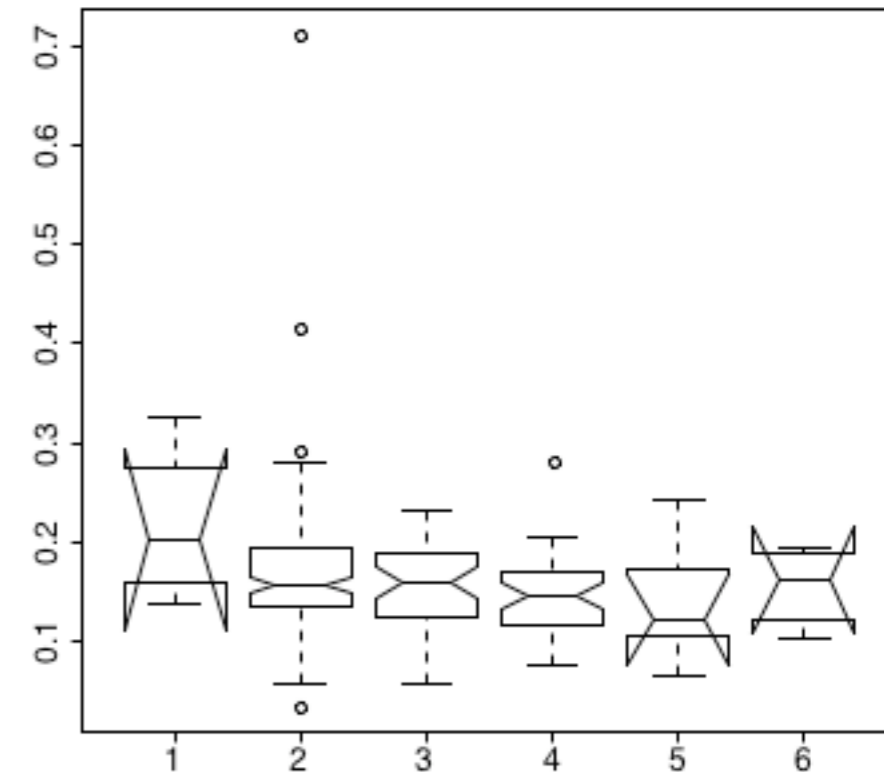
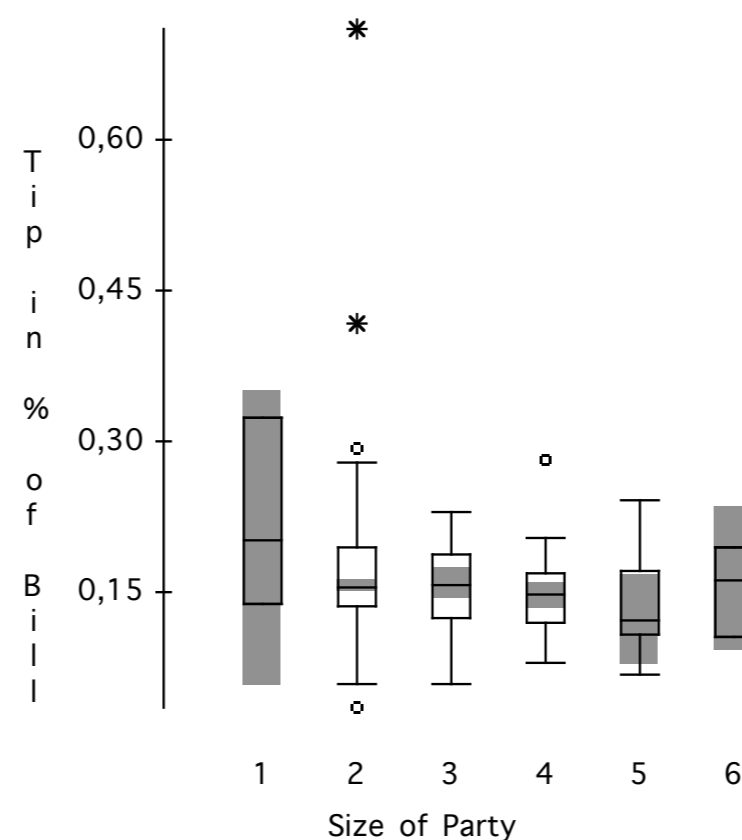
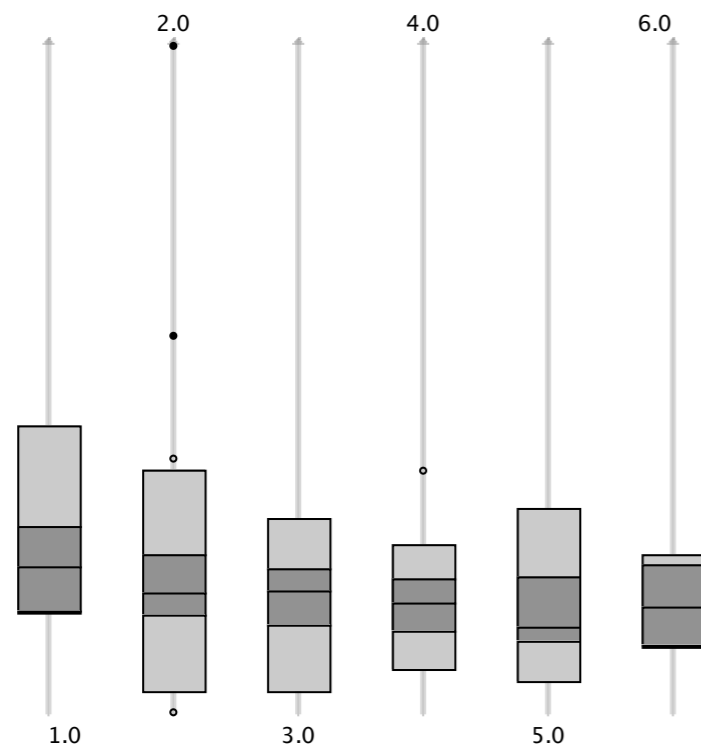
- Ausreißerelimination

- Oft wird versucht Daten “normal” zu machen, damit klassische Methoden angewendet werden können.
- Eine log-Transformation ist aber auch oft nötig, um die Daten überhaupt sinnvoll darzustellen.
- Oft ist eine Behandlung von Ausreißern aber ausreichend, und eine Transformation ist dann nicht mehr nötig.

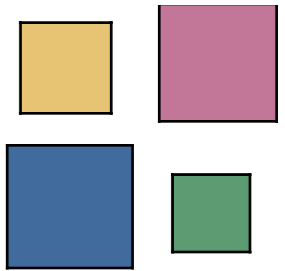


## EDA: Effective comparison

- Box-plots sind excellent für Gruppenvergleiche geeignet:



- Nachteil:  
Gruppengröße kann nicht abgelesen werden  
⇒ notched Boxplots
- Tukeys HSD Test (Honest Significant Differences)

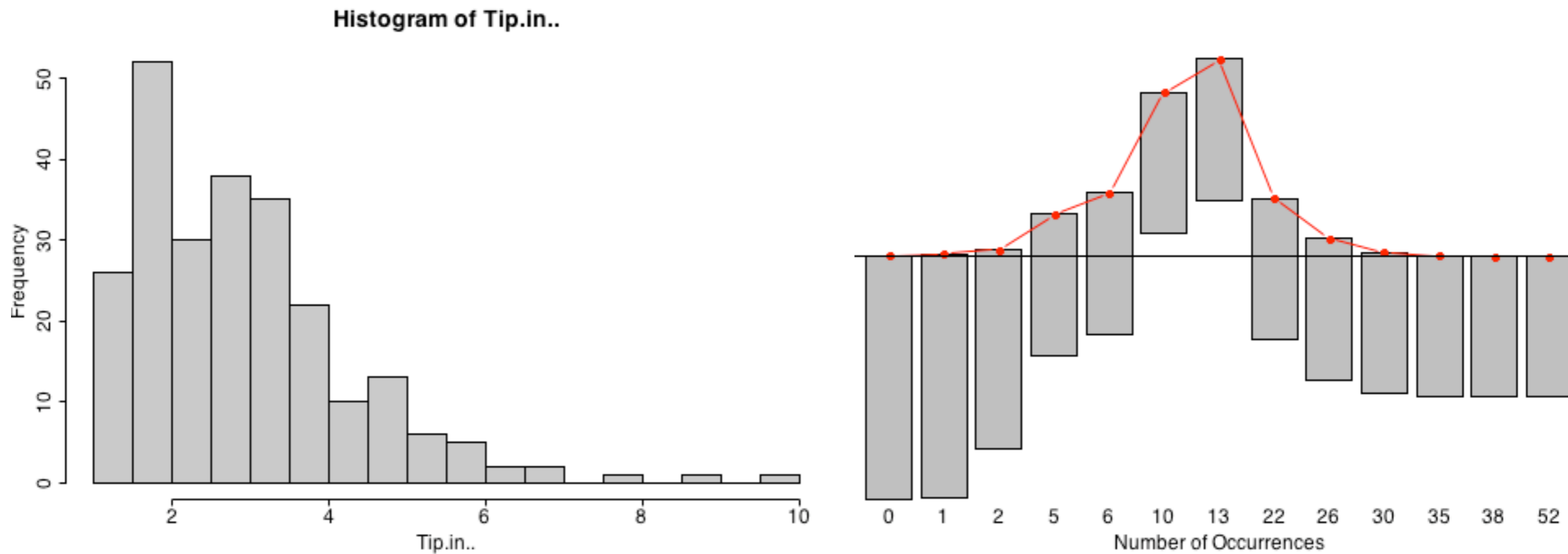


## EDA: Verteilungen

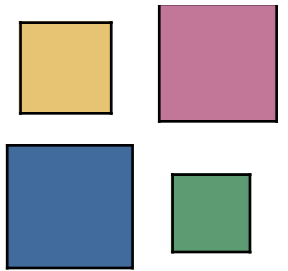
- Rootogram

Idee:

Visualisiere die Abweichung von einer angenommenen Verteilung

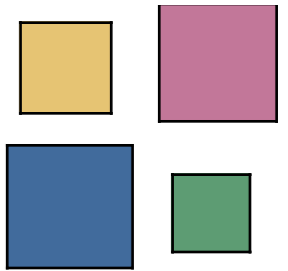


- qq-Plot wird im Allgemeinen bevorzugt.



## EDA: Robustheit

- EDA trifft weniger strenge Verteilungsannahmen als die klassische Statistik
- $\Rightarrow$  die EDA legt(e) ein neues Gewicht auf Robustheit, d.h. die Behandlung von Ausreißern
- Ausreißer
  - unidimensional
  - zweidimensional
  - p-variabel
  - nach Modell (Residuum)
  - ...
- Ausreißerbestimmung und Behandlung ist eine sehr individuelle Sache, und kann daher kaum systematisiert oder automatisiert werden.



## EDA: Der Computer

- Tukey sah schon sehr früh, dass “graph paper” nicht mehr lange aktuell sein wird.
- Schon 1971 erstes Projekt zu interaktiver, dynamischer Graphik.
- Nur Methoden, die wirklich implementiert werden können auch in der EDA verwendet werden.
- **Aber**  
Die meiste Software ist immer noch auf dem Stand der Graphiken der 70er und 80er Jahre  
Die meisten Pakete unterstützen eine explorative Arbeitsweise nur sehr wenig.