Dr. Katharina Best
Sommersemester 2011
14. April 2011

ulm university universität **uulm**

Reader – Teil 1: Beschreibende Statistik

WiMa-Praktikum

Um Daten darzustellen und eine Übersicht über die Struktur der Daten zu erstellen, stellt die beschreibende Statistik sowohl graphische Methoden als auch den Zugriff auf Kennzahlen bereit, welche wir im folgenden näher kennenlernen wollen.

Bezeichnungen und Begriffe

Unter einer *Stichprobe* verstehen wir bei der Datenanalyse eine konkrete Stichprobe, also eine Realisierung $x = (x_1, \ldots, x_n)$ der Zufallsstichprobe $X = (X_1, \ldots, X_n)$. Die einzelnen Einträge der Stichprobe werden auch als *Beobachtungen* bezeichnet.

Der *Rang (engl. rank)* einer Beobachtung x_i innerhalb einer Stichprobe gibt an, an welcher Stelle diese Beobachtung bei aufsteigender Ordnung vorkommt. Sind mehrere Beobachtungen gleich, so bekommen sie den gemittelten Rang.

Eine geordnete Stichprobe $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ ist die dem Rang nach sortierte Reihung der Beobachtungen, wobei $x_{(k)}$ die k-te Ordnungsgröße[1] ist.[2]

Beispiel 1: Wir betrachten die Stichprobe und ihre Ränge:

i	1	2	3	4	5	6	7	8	9	10
x_i	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4
Rang	4	9	5.5	5.5	1	3	8	2	10	7

Die entsprechende geordnete Stichprobe ist

k	1	2	3	4	5	6	7	8	9	10
$x_{(k)}$	0.0	0.8	1.0	1.2	1.3	1.3	1.4	1.8	2.4	4.6

[1] Als Zufallsvariable sprechen wir von Ordnungsstatistik.
[2] Es ist auch die Notation $x_{k:n}$ bekannt, bei der die stichprobenlänge mitgetragen wird.

Statistische Kennzahlen

Lagemaße

Neben dem *Stichprobenmittel*

$$\overline{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$$

werden auch *Quantile* betrachtet. Sei $\alpha \in (0,1)$. Ein α-Quantil x_α teilt die Stichprobe im Verhältnis α zu $1-\alpha$.

Ein beliebtes Dreiergespann ist die Quartilbildung, also die $0.25, 0.5, 0.75$-Quantile, das mittlere ist dabei der *Median*, der eigenständig als

$$\text{med}(x) = \begin{cases} x_{(n+1/2)} & \text{für } n \text{ ungerade,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(1+n/2)}\right) & \text{für } n \text{ gerade} \end{cases}$$

definiert ist.

Der *Modalwert* ist der am häufigsten vorkommende Stichprobenwert.

Streuungsmaße

Hier steht uns ebenfalls eine große Auswahl zur Verfügung. Zum Einen haben wir die *empirische Varianz*

$$s_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

und das *empirische Stichprobenvarianz* $s_n := +\sqrt{s_n^2}$, aber auch der *Variationskoeffizient (engl. coefficient of variation)* $\frac{s*100}{\overline{x}}$, die *Spannweite (engl. range)* $x_{(n)} - x_{(1)}$ und der *Quartilsabstand (engl. interquartile range)* $d_Q = x_{0.75} - x_{0.25}$ sind interessant.

Formmaße

Formmaße geben Information über die Asymmetrie und Steilheit der Daten Auskunft. Die Asymmetrie wird durch die *empirische Schiefe (engl. skewness)*

$$\hat{g}_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

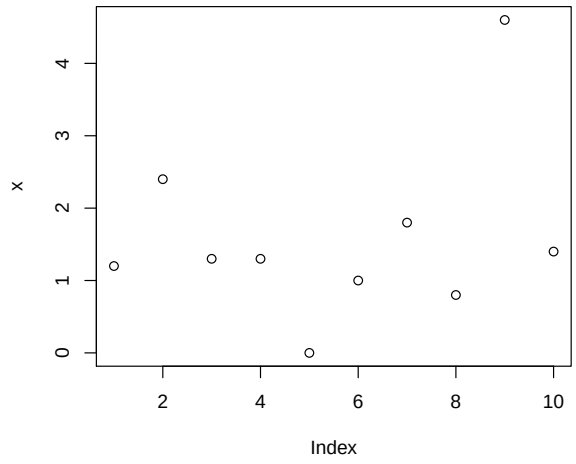
angegeben. Die *Steilheit (engl. kurtosis)* kann mit der

$$\hat{g}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - 3$$

Graphische Darstellungen

Informationserhaltende Diagramme

Für eindimensionale Stichproben können wir mit einem einfachen *Punktediagramm* einen ersten Eindruck über die Daten gewinnen, ohne Information zu verlieren.



Etwas strukturierter macht das die *Stamm-und-Blatt*-Darstellung. In der einfachsten Variante wird der Stamm aus 10 Zeilen gebildet, beginnend mit den Ziffern 0 bis 9. Die einzelnen Beobachtungen werden mit ihrer zweiten Ziffer in die passende Zeile der ersten eingetragen. Je nach Daten kann hier der Stamm entsprechend feiner oder auch länger gewählt werden.

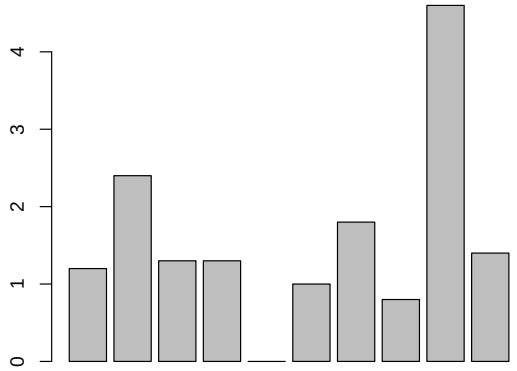
```
0 | 08
1 | 023348
2 | 4
3 |
4 | 6
```

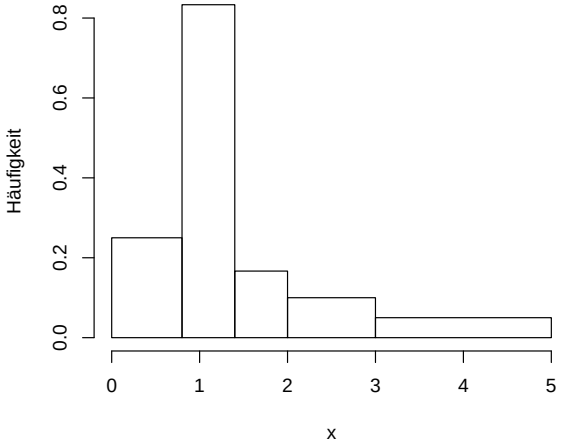
Auf diese Weise bekommen wir Information über die Gestalt der Dichte, können aber auch sofort die numerischen Daten sehen.

Stabdiagramm und Histogramm

Stabdiagramme bilden eine Fortsetzung der Stamm-und-Blatt-Darstellung. Die Informationen der Blatt-Ziffern wird weggelassen. Statt dessen werden Balken entsprechender Höhe angezeigt.

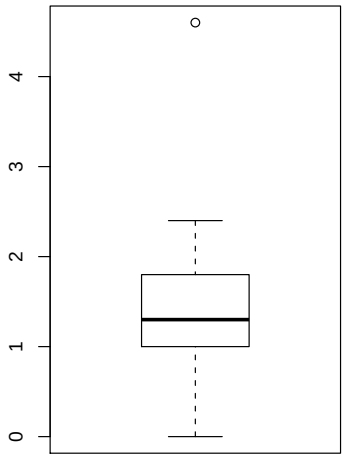
Diese Darstellung ist auch für kategoriale Daten geeignet.





Das Histogramm nähert sich auf dieser Grundlage der Dichte an. Hier sind Balken auch unterschiedlicher Breite möglich. Die Höhe richtet sich dann nicht mehr nach der Anzahl der Beobachtungen allein, sondern die Fläche ist hier der bestimmende Aspekt.

Boxplot



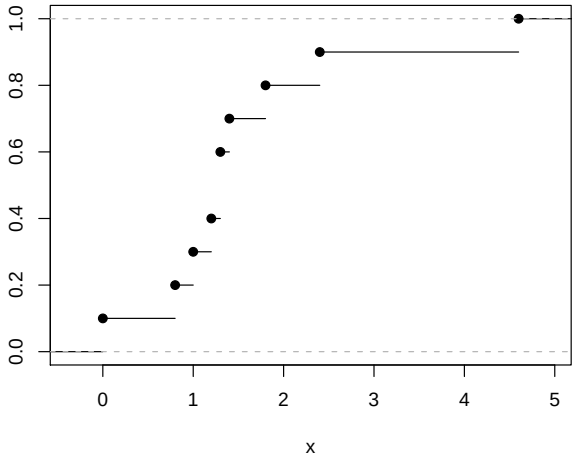
Ein Boxplot ist ein Zwitter zwischen graphischer Darstellung und Lokationsmaß. Er zeigt sowohl den Median, den Quartilsabstand wie auch Ausreißer. Darüber hinaus gibt er Hinweise auf Schiefe der Daten.

Empirische Verteilungsfunktion

Die empirische Verteilungsfunktion ist definiert als

$$F(x) := \frac{1}{n} \left| \{ i : x_i \leq x \} \right|$$

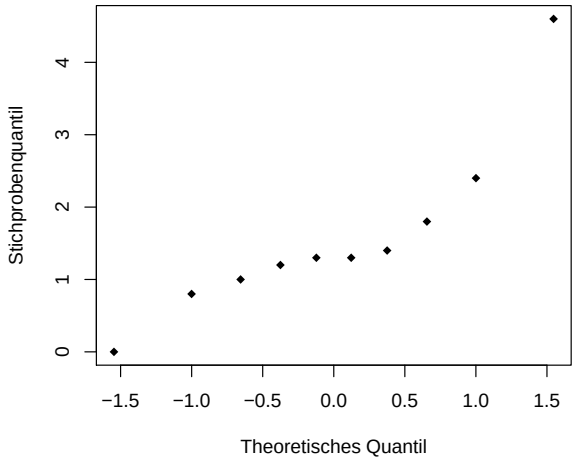
und bildet eine Treppenfunktion, welche mit steigendem n gegen die Verteilungsfunktion konvergiert.



Probability Plot

Besteht bereits ein Verdacht, welcher Verteilung die Beobachtungen folgen, so kann sich dieser leicht graphisch überprüfen. Die Daten werden gegen die theoretischen Daten geplottet.

Dabei werden für den Plot die Punkte $\left(x_{(j)}, z_{(j)}\right)$ betrachtet, die aus den Werten der geordneten Stichprobe und den $\frac{j-0.5}{n}$-Quantilen der vergleichenden Verteilung gebildet werden.



Datentransformation

Es kommt durchaus vor, dass der Zusammenhang der Daten nicht direkt, sondern nur funktional gegeben ist. Auch kann es sein, dass erst transformierte Daten einer sich durch eine bekannte Verteilung beschreiben lassen.