

# 1 Univariate Statistiken

Im ersten Kapitel berechnen wir zunächst Kenngrößen einer einzelnen Stichprobe bzw. so genannte empirische Kenngrößen, wie beispielsweise den Mittelwert. Diese können, unter gewissen Voraussetzungen, als Schätzer für „theoretische“ Kenngrößen einer Zufallsvariablen verwendet werden, wie beispielsweise dem Erwartungswert. Danach wird gezeigt, wie man diese bei einem Test bezüglich des Erwartungswertes verwenden kann, dem so genannten t-Test. Am Beispiel dieses Tests wird das Prinzip des Testens von Hypothesen erklärt. Neben dem t-Test, der spezielle Verteilungsvoraussetzungen benötigt, stellen wir auch nichtparametrische Verfahren vor.

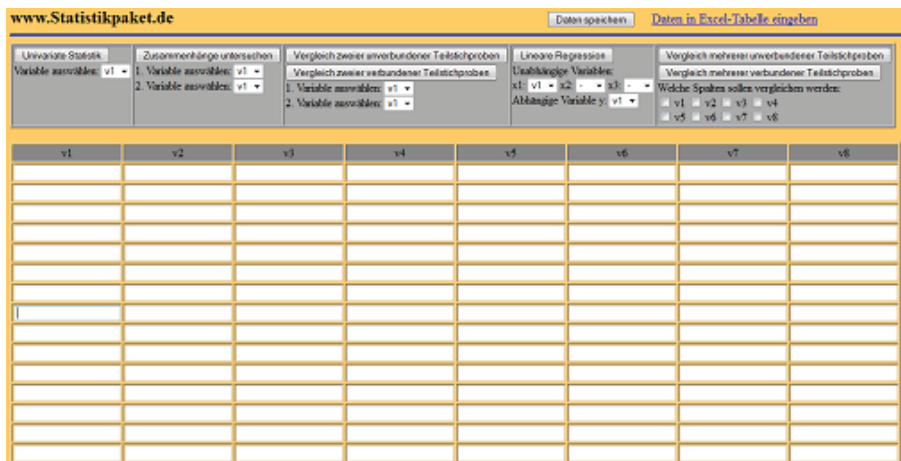
## 1.1 Berechnung von Kenngrößen

Gegeben sei folgende Stichprobe: 167,163,155,167,161,177,173,179. Diese Werte könnten als Körpergrößen von zufällig ausgewählten Schülern einer Schule interpretiert werden. Wir wollen zeigen, wie Statistikpaket.de diese Daten zur Berechnung von Kenngrößen eingeben können. Die Ausgabe wird dann im Anschluss erklärt.

Die folgenden Daten werden zunächst im Browser eingegeben.

v1
167
163
155
167
161
177
173
179

Wir wählen „Wechsel zur Tabellenansicht“ und geben die Daten in der Spalte „v1“ ein:



**Bemerkungen zur Dateneingabe:**

Bei der späteren Auswertung werden die Daten einer Spalte bis zur ersten leeren Zelle gelesen! Es sollten deshalb die fehlenden Werte aus der Stichprobe eliminiert werden, denn die Zellen nach einer leeren Zelle werden nicht berücksichtigt. Browserabhängig können auch nur bestimmte Größen an Datenmengen übertragen werden. Aus diesem Grund sollten auch nicht zu große Datenmengen eingegeben werden. D.h. es sollten keine große Datensätze bestehend aus vielen Spalten und Zeilen, während gleichzeitig in den Zellen Zahlen mit vielen Stellen oder lange Texte stehen, verwendet werden. Im Output werden immer auch noch mal die übertragenen Daten angezeigt.

Nun sollen statistische Kenngrößen berechnet werden. Dazu können Sie →**Univariate Statistik** wählen, womit Sie das folgende Fenster erhalten:

### Univariate Statistik

Statistische Kenngrößen mit Boxplot

Häufigkeitstabellen mit Kreisdiagramm

Histogramm Anzahl Klassen: keine Klassen ▾

t-Test mit  $\mu_0 = 0$  .

Vorzeichentest mit  $\theta_0 = 0$  , gilt auch für Vorzeichenrangtest.

Vorzeichenrangtest

Binomialtest

Kolmogorov-Smirnov-Test

Rangzahlen berechnen

v1

167
163
155
167
161
177
173
179

Hier können Sie nun → **Statistische Kenngrößen** wählen, womit Sie den Output erhalten:

## Statistische Kenngrößen

Hier sind die eingegebenen Daten zu sehen:

167  
163  
155  
167  
161  
177  
173  
179

Stichprobenumfang	8
arithmetisches Mittel	167.75
empirische Varianz	67.357142857143
empirische Standardabweichung	8.2071397488493
Minimum	155
Maximum	179
empirischer Median	167
empirischer Variationskoeffizient	0.048924827116837
empirische Schiefe	-0.044189855841829
empirischer Exzess	-0.88644148039526

Hier noch mal die Abfolge der Auswahl im Menü:

→**Univariate Statistik** →**Statistische Kenngrößen**

**Erläuterung des Outputs:**

Ganz oben ist der Stichprobenumfang zu finden, den wir im Folgenden mit  $n$  bezeichnen. Die Beobachtungen der Stichprobe werden mit  $x_i$  ( $i = 1, 2, \dots, n$ ) bezeichnet. Die Stichprobe ist dann  $x_1, x_2, \dots, x_n$ .

Im Output sind dann folgende Kenngrößen der Stichprobe zu finden:

Das arithmetische Mittel:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Die empirische Varianz:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Die empirische Standardabweichung:

$$s = \sqrt{s^2}$$

Der kleinste und größte Stichprobenwert:

$$\min(x_i) \text{ und } \max(x_i).$$

Der empirische Median:

Hierzu wird zunächst die Stichprobe  $x_1, x_2, \dots, x_n$  geordnet in  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ . Nun kann der empirische Median berechnet werden.

Falls  $n$  gerade ist gilt:  $\tilde{x} = (x_{(n/2)} + x_{(n/2+1)}) / 2$

Falls  $n$  ungerade ist gilt:  $\tilde{x} = x_{((n+1)/2)}$

Weitere Kenngrößen sind der empirische Variationskoeffizient die empirische Schiefe und die empirische Wölbung (engl. skewness & kurtosis):

$$\text{Empirischer Variationskoeffizient} = \frac{s}{\bar{x}}$$

$$\text{Empirische Schiefe} = \frac{n}{(n-1)(n-2)} \frac{1}{s^3} \sum_{i=1}^n (x_i - \bar{x})^3$$

$$\text{Empirische Wölbung} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{1}{s^4} \sum_{i=1}^n (x_i - \bar{x})^4$$

$$\text{Empirischer Exzess} = \text{Empirische Wölbung} - 3 \frac{(n-1)^2}{(n-2)(n-3)}$$

Bei symmetrischen Verteilungen nimmt die Schiefe den Wert 0 an. Da

es sich jeweils um die entsprechenden empirischen Werte, also um Schätzer der theoretischen Kenngrößen handelt, ist der Wert bei Stichproben, die aus Realisierungen von symmetrisch verteilten Zufallsvariablen bestehen, nicht automatisch gleich Null. Ist die Abweichung vom Wert 0 zu groß, so ist dies ein Hinweis darauf, dass die theoretische Verteilung nicht symmetrisch sein könnte. Die Schiefe ist - wie die Wölbung - dimensionslos. Die Wölbung einer normalverteilten Zufallsvariablen hat den Wert 3, während der Exzess hier den Wert 0 annimmt. Im Output finden Sie den empirischen Exzess aus.

**Bemerkung zum Speichern der Daten:**

Die eingegebenen Daten können Sie in Ihrem Browser (Firefox oder Internet-Explorer) speichern. Dazu müssen Sie nur über dem Menü auf den Button **→Daten speichern** klicken und danach im darauf folgenden Fenster auf **→Zum Menü**. Danach werden die Daten oben im Browser in der www-Adresse mit angezeigt. Wenn Sie dann diese Seite als Lesezeichen (oder Favorit) speichern, dann werden die Daten mit gespeichert (nur in Ihrem Browser) und Sie können diese jederzeit wieder aufrufen. Browserabhängig können allerdings nicht beliebig große Datenmengen gespeichert werden.

Nun berechnen wir die statistischen Kenngrößen mit SAS.

### Daten in SAS und Prozeduraufruf:


Nach dem Starten von SAS erhalten Sie den Eröffnungsbildschirm von SAS. Wenn man auf den Editor (Taste F5) klickt, kann man das Programm zum übergeben der Daten in SAS eingeben.

```
data dat1; /*Datensatz dat1 wird erzeugt (temporär)*/  
input x; /*variable x wird erzeugt*/  
cards; /*Karte für Eingabe wird erzeugt,  
alternativ: datalines; */  
167  
163  
155  
167  
161  
177  
173  
179  
run; /*ausführen*/
```

Das Programm kann man mit der Taste F8 abschicken. Fehlerkommentare sieht man in LOG-Fenster (erhält man mit der Taste F6 und mit F7 kommt man zum OUTPUT-Fenster).

Man kann u.a. mit den Prozeduren UNIVARIATE oder MEANS in SAS statistische Kenngrößen berechnen.

```
proc univariate data=dat1;  
var x;  
run;
```

Zum Ausführen drücken Sie einfach wieder die Taste F8 oder das Ikon mit dem rennenden Männchen in der Symbolleiste .

## SAS-Output zur Prozedur Univariante

Die Prozedur UNIVARIATE  
Variable: x

Momente			
<b>N</b>	8	<b>Summe Gewichte</b>	8
<b>Mittelwert</b>	167.75	<b>Summe Beobacht.</b>	1342
<b>Std.abweichung</b>	8.20713975	<b>Varianz</b>	67.3571429
<b>Schiefe</b>	-0.0441899	<b>Kurtosis</b>	-0.8864415
<b>Unkorr. Qu.summe</b>	225592	<b>Korr. Quad.summe</b>	471.5
<b>Variationskoeff.</b>	4.89248271	<b>Stdfeh. Mittelw.</b>	2.90166209

Grundlegende Statistikmaße			
Lage		Streuung	
<b>Mittelwert</b>	167.7500	<b>Std.abweichung</b>	8.20714
<b>Median</b>	167.0000	<b>Varianz</b>	67.35714
<b>Modalwert</b>	167.0000	<b>Spannweite</b>	24.00000
		<b>Interquartilsabstand</b>	13.00000

Tests auf Lageparameter: $\mu_0=0$			
Test	Statistik	p-Wert	
<b>Studentsches t</b>	t 57.81169	<b>Pr &gt;  t </b>	<.0001
<b>Vorzeichen</b>	M 4	<b>Pr &gt;=  M </b>	0.0078
<b>Vorzeichen-Rang</b>	S 18	<b>Pr &gt;=  S </b>	0.0078

Quantile (Definition 5)

Quantil      Schätzwert

<b>Quantile (Definition 5)</b>	
<b>Quantil</b>	<b>Schätzwert</b>
<b>100% Max</b>	179
<b>99%</b>	179
<b>95%</b>	179
<b>90%</b>	179
<b>75% Q3</b>	175
<b>50% Median</b>	167
<b>25% Q1</b>	162
<b>10%</b>	155
<b>5%</b>	155
<b>1%</b>	155
<b>0% Min</b>	155

<b>Extreme Beobachtungen</b>			
<b>Kleinste</b>		<b>Größte</b>	
<b>Wert</b>	<b>Beobachtung</b>	<b>Wert</b>	<b>Beobachtung</b>
155	3	167	1
161	5	167	4
163	2	173	7
167	4	177	6
167	1	179	8

SAS verwendet beim empirischer Variationskoeffizient =  $\frac{S}{x}$  noch der

Faktor 100 verwendet:  $100 \cdot \frac{S}{x}$

**Bemerkung zum Speichern der Daten:**

In unserem obigen Programm werden die Daten nur temporär gespeichert. Das bedeutet, dass der erstellte Datensatz (DATASET) bei der nächsten SAS-Sitzung nicht mehr vorhanden wäre (bspw. falls SAS beendet und neu gestartet wird). Temporär Dateien werden von SAS automatisch unter den Referenzen WORK.

Um die Dateien in einem separaten Verzeichnis speichern zu können, geben Sie nach jedem Neustart (immer) die folgende Programmzeile ein (legen Sie hierzu vorab ein eigenes Verzeichnis an).

```
libname disk 'C:\DATA\';
```

Sobald Sie nun vor einem Datensatzname „disk.“ angeben, z.B. disk.daten, so weiß das SAS-System, dass es auf das oben definierte Verzeichnis zugreifen soll und die Daten werden dort gespeichert.

An unserem Quellcode müsste somit lediglich ‚dat1‘ mit ‚disk.dat1‘ ersetzt werden.