§3 Statistische Maßzahlen pfad

▼

```
SetDirectory[$UserDocumentsDirectory] ;

<< MultivariateStatistics`;

Ordnungsstatistik[stich_, k_] := If[Length[stich] ≥ k,
    Part[Sort[stich], k], Print["diese Stichprobe besitzt weniger als ", k, " Elemente"]]

Rang[stich_, element_] := Module[{pos, anz},
    pos = Flatten[Position[Sort[stich], element]];
    anz = Length[pos];
    If[MemberQ[stich, element], Apply[Plus, pos] / anz // N,
      Print[element, " ist kein Element dieser Stichprobe"]]]
```

Neben der Möglichkeit, umfangreiche statistische Daten graphisch zu veranschaulichen, lassen sich quantitative Merkmale auch durch aussagekräftige statistische Maßzahlen charakterisieren. Wir werden in diesem Abschnitt einige wichtige statistische Maßzahlen für univariate bzw bivariate quantitative Daten definieren und zeigen, wie sich diese statistischen Maßzahlen mit Hilfe von Mathematica berechnen lassen.

3.1 Statistische Maßzahlen univariater Daten

Wir gehen davon aus, dass unser statistisches Datenmaterial bereits in Form einer Datenmatrix, welche im Datenfile *datenfile* unseres Datenordners abgelegt ist, vorliegt und befassen uns in diesem Abschnitt mit der Frage, durch welche aussagekräftigen Maßzahlen sich die zu einem *quantitativen* Merkmal gehörende Stichprobe charakterisieren lässt. Dabei bezeichne $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ stets eine Stichprobe des zur Diskussion stehenden *quantitativen* Merkmals und $\vec{x}^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$ die mit Hilfe von Sort der Größe nach geordnete Stichprobe.

■ Maßzahlen für das "Zentrum" einer Stichprobe

Mit den folgenden Maßzahlen werden Werte beschreiben, die als "Zentrum" der Stichprobe angesehen werden:

3.1.1 Definition:

a) Unter dem Mittelwert $\bar{x}$ der Stichprobe $\vec{x}$ versteht man die Zahl

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

Der Mittelwert ist die in der Praxis am häufigsten verwendete Maßzahl für das Zentrum einer Stichprobe.

b) Unter dem getrimmten Mittelwert $\bar{x}_{p,q}$ der Stichprobe $\vec{x}$ versteht man den Mittelwert jener Liste von Daten, welche man erhält, wenn man von der ursprünglichen Stichprobe $100 \times p$ Prozent der kleinsten und $100 \times q$ Prozent der größten Werte weglässt. Den getrimmten Mittelwert verwendet man in der Praxis dann als Maßzahl für das Zentrum einer Stichprobe, wenn Ausreißer nicht berücksichtigt werden sollen.

c) Unter dem Median $\tilde{x}$ der Stichprobe $\vec{x}$ versteht man die Zahl

$$\tilde{x} = \begin{cases} x^*_{(n+1)/2} & \text{falls } n \text{ ungerade ist} \\ (x^*_{n/2} + x^*_{(n+2)/2})/2 & \text{falls } n \text{ gerade ist} \end{cases}$$

Den Median verwendet man in der Praxis oft dann als Maßzahl für das Zentrum einer Stichprobe, wenn diese Stichprobe ein stark asymmetrisches Histogramm aufweist.

Diese drei Maßzahlen für das Zentrum der Stichprobe $\vec{x}$ lassen sich mit den Befehlen Mean, TrimmedMean bzw Median ermitteln:

■ Mean[*stich*]

berechnet den Mittelwert $\bar{x}$ der Stichprobe *stich*.

■ TrimmedMean[*stich*, {*p*, *q*}]

berechnet den getrimmten Mittelwerte $\bar{x}_{p,q}$ der Stichprobe *stich*.

■ Median[*stich*]

berechnet den Median $\tilde{x}$ der Stichprobe *stich*.

3.1.2 Beispiel: Vom Merkmal Zugfestigkeit (dritte Spalte) des Datenmaterials *stahl* berechne man den Mittelwert $\bar{x}$, den getrimmten Mittelwert $\bar{x}_{p,q}$ sowie den Median $\tilde{x}$.

▼

Lösung: Wir lesen dazu das im Datenordner abgelegte Datenfile *stahlfile* ein, rufen die dritte Spalte dieser Datenmatrix auf und bezeichnen die zugehörige Stichprobe mit "zugfestigkeit". Von dieser Stichprobe berechnen wir die gesuchten Maßzahlen und geben diese in übersichtlicher Weise in Form einer Tabelle aus:

```
p = 0.02; q = 0.03;
zugfestigkeit = Rest[Part[<< stahlfile, All, 3]];
TableForm[{
    {"Mittelwert:", Mean[zugfestigkeit]},
    {"getrimmter Mittelwert:", TrimmedMean[zugfestigkeit, {p, q}]},
    {"Median:", Median[zugfestigkeit]}},
  TableSpacing → {1, 10}]
Clear[zugfestigkeit, p, q]
```

```
Mittelwert:                    715.779
getrimmter Mittelwert:         714.566
Median:                        717.
```

■ Maßzahlen für die "Ordnungsstruktur" einer Stichprobe

Mit den folgenden Maßzahlen lassen sich einige Aspekte der "Ordnungsstruktur" der Stichprobe beschreiben:

3.1.3 Definition:

a) Den k-ten Eintrag x_k^* der geordneten Stichprobe $\vec{x}^*$ nennt man k-te Ordnungsstatistik der Stichprobe $\vec{x}$.

b) Gelangt beim Anordnen der Stichprobe $\vec{x}$ das Element *element* dieser Stichprobe an die k-te Stelle der geordneten Stichprobe $\vec{x}^*$ (gilt also $x_k^* = \text{element}$), so nennt man diesen Index k den Rang des Elements element und bezeichnet ihn mit rg[$\vec{x}$, *element*]. Man beachte dabei: Sind mehrere Elemente der Stichprobe $\vec{x}$ gleich, so erhalten diese als Rang das arithmetische Mittel der ihnen an sich zustehenden Ränge zugeordnet. So besitzen beispielsweise die Elemente 1, 2, 3, 5 der Stichprobe $\vec{x} = \{1, 3, 5, 3, 2, 3, 2\}$ die Ränge 1, 2.5, 5, 7.

Die k-te Ordnungsstatistik x_k^* der Stichprobe $\vec{x}$ sowie der Rang $\text{rg}\left[\vec{x}, x_i\right]$ des i-ten Elements der Stichprobe $\vec{x}$ lassen sich mit den Befehlen Ordnungsstatistik bzw Rang berechnen:

■ Ordnungsstatistik[*stich*, *k*]

ermittelt die k-te Ordnungsstatistik der Stichprobe *stich*.

■ Rang[*stich*, *element*]

ermittelt den Rang des Elements *element* der Stichprobe *stich*.

3.1.4 Beispiel: Vom Merkmal Zugfestigkeit (dritte Spalte) des Datenmaterials *stahl* berechne man die k-te Ordnungsstatistik sowie den Rang des i-ten Elements dieser Stichprobe.

▼

Lösung: Wir lesen dazu das im Datenordner abgelegte Datenfile *stahlfile* ein, rufen die dritte Spalte dieser Datenmatrix auf und bezeichnen die zugehörige Stichprobe mit "zugfestigkeit". Von dieser Stichprobe berechnen wir die gesuchten Maßzahlen und geben diese in übersichtlicher Weise in Form einer Tabelle aus:

```
k = 25; i = 10;
zugfestigkeit = Rest[Part[<< stahlfile, All, 3]];
TableForm[{
    {"k-te Ordnungsstatistik:", Ordnungsstatistik[zugfestigkeit, k]},
    {"Rang des i-ten Elements:", Rang[zugfestigkeit, zugfestigkeit[[i]]]},
  TableSpacing → {1, 5}]
Clear[zugfestigkeit, k, i]
```

```
k-te Ordnungsstatistik:      690.
Rang des i-ten Elements:     84.
```

Mit den folgenden Maßzahlen lassen sich weitere Aspekte der "Ordnungsstruktur" der Stichprobe beschreiben:

3.1.5 Definition:

a) Unter dem *p*-Quantil $\hat{x}_p$ der Stichprobe $\vec{x}$ versteht man jenen Wert $\hat{x}_p$ der Stichprobe, für den gilt

$$\frac{1}{n}\,|\{i \,|\, x_i < \hat{x}_p\}| < p \quad \text{und} \quad \frac{1}{n}\,|\{i \,|\, x_i \le \hat{x}_p\}| \ge p$$

b) Das interpolierte *p*-Quantil $\hat{x}^*_p$ der Stichprobe $\vec{x}$ berechnet man folgendermaßen: Zuerst setzt man

$$\hat{x}^*_0 := x^*_1, \quad \hat{x}^*_{1/(2\,n)} := x^*_1, \quad \hat{x}^*_{3/(2\,n)} := x^*_2, \quad \hat{x}^*_{5/(2\,n)} := x^*_3, \quad \ldots, \quad \hat{x}^*_{(2\,n-1)/(2\,n)} := x^*_n, \quad \hat{x}^*_1 = x^*_n$$

und ermittelt anschließend den Wert $\hat{x}^*_p$ durch lineare Interpolation.

c) Die interpolierten Quantile $\hat{x}^*_{0.25}$, $\hat{x}^*_{0.5}$ und $\hat{x}^*_{0.75}$ nennt man Quartile der Stichprobe $\vec{x}$ (wobei man beachte, dass das interpolierte 0.5-Quartil $\hat{x}^*_{0.5}$ dem Median $\tilde{x}$ entspricht).

▼

Während die Quantile $\hat{x}_p$ einer Stichprobe stets Elemente dieser Stichprobe sind, müssen die interpolierten Quantile $\hat{x}^*_p$ (und damit auch die Quartile) keineswegs Elemente dieser Stichprobe sein.

Zur Berechnung der Quantile und Quartile der Stichprobe $\vec{x}$ dienen die Befehle Quantile und Quartiles:

■ Quantile[*stich*, *p*] bzw Quantile[*stich*, *p*, {{1/2, 0}, {0, 1}}]

berechnet das *p*-Quantil $\hat{x}_p$ bzw das interpolierte *p*-Quantil $\hat{x}^*_p$ der Stichprobe *stich*.

■ Quartiles[*stich*]

gibt die drei interpolierten Quantile $\hat{x}^*_{0.25}$, $\hat{x}^*_{0.5}$, $\hat{x}^*_{0.75}$ der Stichprobe *stich* in Form einer Liste aus.

3.1.6 Beispiel: Vom Merkmal Verbrauch (siebente Spalte) des Datenmaterials *fahrzeuge* berechne man die 20%, 50% und 80% Quantile sowie die entsprechenden interpolierten Quantile.

▼

Lösung: Wir lesen dazu das im Datenordner abgelegte Datenfile *fahrzeugefile* ein, rufen die siebente Spalte dieser Datenmatrix auf und bezeichnen die zugehörige Stichprobe mit "verbrauch". Von dieser Stichprobe berechnen wir die gesuchten Quantile und geben diese in übersichtlicher Weise in Form einer Tabelle aus:

```
verbrauch = Rest[Part[<< fahrzeugefile, All, 7]];
TableForm[{Table[Quantile[verbrauch, p], {p, 0.2, 0.8, 0.3}],
    Table[Quantile[verbrauch, p, {{1/2, 0}, {0, 1}}], {p, 0.2, 0.8, 0.3}]},
  TableHeadings → {{Quantile, interpolierte Quantile}, {"20%", "50%", "80%"}},
  TableSpacing –> {2, 5}]
Clear[verbrauch]
```

	20%	50%	80%
Quantile	15.22	16.15	17.47
interpolierte Quantile	15.255	16.155	17.48

■ Maßzahlen für die "Schwankung" einer Stichprobe

Mit den folgenden Maßzahlen lässt sich beschreiben, wie stark die Stichprobe um ihr Zentrum "schwankt":

3.1.7 Definition:

a) Unter der Varianz v_x der Stichprobe $\vec{x}$ versteht man die Zahl

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

b) Unter der Standardabweichung s_x der Stichprobe $\vec{x}$ versteht man die Wurzel ihrer Varianz v_x, also

$$s_x = \sqrt{v_x}$$

Die Standardabweichung ist die in der Praxis am häufigsten verwendeten Maßzahlen, mit der die Schwankung einer Stichprobe beschrieben wird.

c) Unter der Spannweite r_x der Stichprobe $\vec{x}$ versteht man die Zahl

$$r_x = x_n^* - x_1^*$$

Die Spannweite dient in der Praxis dann als Maßzahl für den Schwankungsbereich einer Stichprobe, wenn der Einfluss von Ausreißern berücksichtigt werden soll.

d) Unter dem Interquartile-Range $\hat{r}_x$ der Stichprobe $\vec{x}$ versteht man die Zahl

$$\hat{r}_x = \hat{x}_{0.75}^* - \hat{x}_{0.25}^*$$

Der Interquartile-Range dient in der Praxis dann als Maßzahl für den Schwankungsbereich einer Stichprobe, wenn der Einfluss von Ausreißern nicht berücksichtigt werden soll.

▼

Warum man bei der Definition der Varianz v_x durch $n-1$ und nicht durch n dividiert, wird im Zusammenhang mit dem Begriff der Erwartungstreue von Schätzern klar werden.

Abgesehen von der Spannweite r_x, welche sich unter Verwendung von Max und Min leicht berechnen lässt, können diese Maßzahlen für die Schwankung der Stichprobe $\vec{x}$ mit den Befehlen Variance, StandardDeviation und InterquartileRange berechnet werden:

■ Variance[*stich*]

berechnet die Varianz v_x der Stichprobe *stich*.

■ StandardDeviation[*stich*]

berechnet die Standardabweichung s_x der Stichprobe *stich*.

■ InterquartileRange[*stich*]

berechnet den Interquartile-Range $\hat{r}_x$ der Stichprobe *stich*.

3.1.8 Beispiel: Vom Merkmal Gewicht (zweite Spalte) des Datenmaterials *walzzeit* berechne man die Varianz v_x, die Standardabweichung s_x, die Spannweite r_x und den Interquartile-Range $\hat{r}_x$.

▼

Lösung: Wir lesen dazu das im Datenordner abgelegte Datenfile *walzzeitfile* ein, rufen die zweite Spalte dieser Datenmatrix auf und bezeichnen die zugehörige Stichprobe mit "gewicht". Von dieser Stichprobe berechnen wir die gesuchten Maßzahlen und geben diese in übersichtlicher Weise in Form einer Tabelle aus:

```
gewicht = Rest[Part[<< walzzeitfile, All, 2]];
TableForm[{{"Varianz:", Variance[gewicht]},
    {"Standardabweichung:", StandardDeviation[gewicht]},
    {"Spannweite:", Max[gewicht] – Min[gewicht]},
    {"Interquartile–Range:", InterquartileRange[gewicht]}},
  TableSpacing –> {1, 5}]
Clear[gewicht]
```

```
Varianz:                1.96218
Standardabweichung:     1.40078
Spannweite:             3.72
Interquartile-Range:    2.78
```

■ Maßzahlen für die "Form" des Histogramms einer Stichprobe

Mit den folgenden Maßzahlen werden einige Aspekte der "Form" des Histogramms der Stichprobe beschrieben:

3.1.9 Definition:

a) Unter dem *r*-ten Moment ${}^{(r)}\overline{x}$ der Stichprobe $\vec{x}$ versteht man die Zahl

$${}^{(r)}\overline{x} = \frac{1}{n}\sum_{i=1}^{n} {x_i}^r$$

Man beachte, dass das erste Moment ${}^{(1)}\overline{x}$ mit dem Mittelwert $\overline{x}$ übereinstimmt.

b) Unter dem *r*-ten zentralen Moment ${}^{(r)}\tilde{x}$ der Stichprobe $\vec{x}$ versteht man die Zahl

$${}^{(r)}\tilde{x} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^r$$

Man beachte, dass das zweite zentrale Moment ${}^{(2)}\tilde{x}$ mit der Varianz v_x nicht! übereinstimmt; im ersten Fall wird nämlich durch n, im zweiten Fall wird durch $n - 1$ dividiert.)

c) Unter der Schiefe φ_x der Stichprobe $\vec{x}$ versteht man die Zahl

$$\varphi_x = \frac{{}^{(3)}\tilde{x}}{({}^{(2)}\tilde{x})^{3/2}}$$

Die Schiefe dient in der Praxis als Maßzahl für die Asymmetrie des Histogramms der Stichprobe $\vec{x}$. Besitzt die Stichprobe einen deutlich ausgeprägten linken "Schwanz", so ist die Schiefe negativ; besitzt die Stichprobe einen deutlich ausgeprägten rechten "Schwanz", so ist die Schiefe positiv.

d) Unter der Kurtosis ψ_x der Stichprobe $\vec{x}$ versteht man die Zahl

$$\psi_x = \frac{{}^{(4)}\tilde{x}}{({}^{(2)}\tilde{x})^2}$$

Die Kurtosis dient in der Praxis als Maßzahl für die Wölbung des Histogramms der Stichprobe $\vec{x}$. Für eine normalverteilte Stichprobe ist die Kurtosis annähernd gleich 3; besitzt die Verteilung der Stichprobe einen deutlichen "Gipfel" und ausgeprägte "Schwänze", so ist die Kurtosis größer als 3; besitzt die Verteilung der Stichprobe hingegen steile "Flanken", so ist die Kurtosis kleiner als 3.

Diese Maßzahlen, mit denen sich die Form des Histogramms einer Stichprobe beschreiben lässt, können mit den Befehlen ExpectedValue, CentralMoment, Skewness bzw Kurtosis berechnet werden:

■ ExpectedValue[x^r, *stich*, *x*]

berechnet das *r*-te Moment ${}^{(r)}\overline{x}$ der Stichprobe *stich*.

■ CentralMoment[*stich*, *r*]

berechnet das *r*-te zentrale Moment ${}^{(r)}\tilde{\overline{x}}$ der Stichprobe *stich*.

■ Skewness[*stich*]

berechnet die Schiefe φ_x der Stichprobe *stich*.

■ Kurtosis[*stich*]

berechnet die Kurtosis ψ_x der Stichprobe *stich*.

3.1.10 Beispiel: Durch Simulation erzeuge man eine Liste von $n = 100$ mit dem Parameter $\lambda = 2$ exponentialverteilten Zufallszahlen und berechne von dieser Stichprobe das 2-te, 3-te und 4-te zentrale Moment sowie die Schiefe und den Excess.

▼

Lösung: Wir erzeugen zuerst in der bekannten Weise eine Liste von $n = 100$ mit dem Parameter $\lambda = 2$ exponentialverteilten Zufallszahlen und geben dieser Stichprobe den Namen "stich". Anschließend berechnen wir von dieser Stichprobe die gesuchten Maßzahlen und geben diese in übersichtlicher Weise in Form einer Tabelle aus:

```
n = 100; λ = 2;
stich = RandomReal[ExponentialDistribution[λ], {n}];
TableForm[{{"2−tes zentrales Moment:", CentralMoment[stich, 2]},
    {"3−tes zentrales Moment:", CentralMoment[stich, 3]},
    {"4−tes zentrales Moment:", CentralMoment[stich, 4]},
    {"Schiefe:", Skewness[stich]},
    {"Kurtosis:", Kurtosis[stich]}},
  TableSpacing → {1, 5}]
Clear[n, λ, stich]
```

```
2-tes zentrales Moment:      0.171279
3-tes zentrales Moment:      0.123129
4-tes zentrales Moment:      0.186424
Schiefe:                     1.73701
Kurtosis:                    6.35465
```

Man beachte, wie stark die Schiefe und vor allem die Kurtosis von Stichprobe zu Stichprobe variieren. Dies zeigt, dass diese beiden Maßzahlen nur von qualitativem Interesse sind.

3.2 Statistische Maßzahlen bivariater Daten

Wir gehen davon aus, dass unser statistisches Datenmaterial bereits in Form einer Datenmatrix, welche im Datenfile *datenfile* unseres Datenordners abgelegt ist, vorliegt und befassen uns in diesem Abschnitt mit der Frage, durch welche aussagekräftigen Maßzahlen sich die gegenseitige Abhängigkeit zwischen zwei quantitativen Merkmalen charakterisieren lässt.

■ Maßzahlen für die gegenseitige "Abhängigkeit" zweier Stichproben

Gegeben seien die beiden Stichproben $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ und $\vec{y} = \{y_1, y_2, \ldots, y_n\}$ zweier *quantitativer* Merkmale. Mit den folgenden Maßzahlen lässt sich die gegenseitige "Abhängigkeit" dieser beiden Stichproben beschreiben:

3.2.1 Definition:

a) Unter der Kovarianz $k_{x,y}$ der Stichproben $\vec{x}$ und $\vec{y}$ versteht man die Zahl

$$k_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})$$

Die Kovarianz ist positiv bei gleichsinnigem Zusammenhang (zu großen x-Werten gehören große y-Werte) und negativ bei gegensinnigem Zusammenhang (zu großen x-Werten gehören kleine y-Werte). Da der Zahlenwert $k_{x,y}$ aber wenig Aussagekraft hat, wird die Kovarianz in der Praxis eher selten als Maßzahl für die Abhängigkeit der beiden Stichproben $\vec{x}$ und $\vec{y}$ verwendet. Man beachte übrigens, dass $k_{x,x} = v_x$ ist.

b) Unter dem Pearsonschen Korrelationskoeffizient $r_{x,y}$ der Stichproben $\vec{x}$ und $\vec{y}$ versteht man die Zahl

$$r_{x,y} = \frac{k_{x,y}}{s_x\, s_y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}\ \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

Analog zur Kovarianz ist auch der Pearsonsche Korrelationskoeffizient bei gleichsinnigem Zusammenhang positiv und bei ungleichsinnigem Zusammenhang negativ. Da wegen der Cauchy-Schwarz-Ungleichung stets $-1 \le r_{x,y} \le 1$ gilt, lässt sich das Ausmaß der gegenseitigen Abhängigkeit der beiden Stichproben $\vec{x}$ und $\vec{y}$ mit $r_{x,y}$ wesentlich besser erfassen als mit $k_{x,y}$. Außerdem gilt: Ist $r_{x,y} = \pm 1$, so liegen die Wertepaare $\{x_i, y_i\}$ auf einer Geraden; ist $r_{x,y} = 0$, so besteht zwischen den beiden Stichproben kein (linearer) Zusammenhang.

c) Unter dem Spearmanschen Rangkorrelationskoeffizient $\rho_{x,y}$ der Stichproben $\vec{x}$ und $\vec{y}$ versteht man die Zahl

$$\rho_{x,y} = \frac{\sum_{i=1}^{n} (\mathrm{rg}[x,\, x_i] - (n+1)/2)\,(\mathrm{rg}[y,\, y_i] - (n+1)/2)}{\sqrt{\sum_{i=1}^{n} (\mathrm{rg}[x,\, x_1] - (n+1)/2)^2}\ \sqrt{\sum_{i=1}^{n} (\mathrm{rg}[y,\, y_i] - (n+1)/2)^2}}$$

Beim Spearmanschen Rangkorrelationskoeffizient handelt es sich offenbar um den Korrelationskoeffizient der Ränge der Stichproben $\vec{x}$ und $\vec{y}$. Damit gilt ebenfalls stets $-1 \le \rho_{x,y} \le 1$. Ähnlich wie der Pearsonsche Korrelationskoeffizient beschreibt auch der Spearmansche Rangkorrelationskoeffizient das Ausmaß der gegenseitigen Abhängigkeit der beiden Stichproben $\vec{x}$ und $\vec{y}$. In der Praxis wird der Spearmansche Rangkorrelations-koeffizient dann verwendet, wenn nicht die konkreten Werte der beiden Stichproben sondern nur ihre Größenbeziehungen relevant sind.

d) Unter dem Kendallschen Rangkorrelationskoeffizient $\tau_{x,y}$ der Stichproben $\vec{x}$ und $\vec{y}$ versteht man die Zahl

$$\tau_{x,y} = \frac{2}{n\,(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathrm{sign}[\mathrm{rg}[x,\, x_i] - \mathrm{rg}[x,\, x_j]]\ \mathrm{sign}[\mathrm{rg}[y,\, y_i] - \mathrm{rg}[y,\, y_j]]$$

Der Kendallsche Rangkorrelationskoeffizient ist eine Maßzahl für den Grad der Monotonie und beschreibt damit ebenfalls in gewisser Weise die Abhängigkeit der beiden Stichproben. Auch für den Kendallschen Rangkorrelationskoeffizient gilt offenbar stets die Beziehung $-1 \le \tau_{x,y} \le 1$, wobei Werte von $\tau_{x,y}$ in der Nähe von ± 1 wieder eine starke gleichartige Monotonie der beiden Stichproben signalisieren. In der Praxis verwendet man den Kendallschen Rangkorrelationskoeffizient in den gleichen Situationen wie den Spearmanschen Rangkorrelationskoeffizient. Man berechnet übrigens meistens alle drei Korrelationskoeffizienten gemeinsam.

Diese vier Maßzahlen, mit denen sich die gegenseitige Abhängigkeit zweier Stichproben beschreiben lässt, lassen sich mit den Befehlen Covariance und Correlation sowie den im Paket MultivariateStatistics` implementierten Befehlen SpearmanRankCorrelation und KendallRankCorrelation berechnen:

■ Covariance[*xstich, ystich*]

berechnet die Covariance $k_{x,y}$ der beiden Stichproben *xstich* und *ystich*.

■ Correlation[*xstich, ystich*]

berechnet den Pearsonschen Korrelationskoeffizient $r_{x,y}$ der beiden Stichproben *xstich* und *ystich*.

■ SpearmanRankCorrelation[*xstich, ystich*]

berechnet den Spearmanschen Rangkorrelationskoeffizient $\rho_{x,y}$ der beiden Stichproben *xstich* und *ystich*.

■ KendallRankCorrelation[*xstich, ystich*]

berechnet den Kendallschen Rangkorrelationskoeffizient $\tau_{x,y}$ der beiden Stichproben *xstich* und *ystich*.

3.2.2 Beispiel: Von den beiden quantitativen Merkmalen Kohlenstoff (zweite Spalte) und Zugfestigkeit (dritte Spalte) des Datenmaterials *stahl* berechne man die Kovarianz $c_{x,y}$ sowie alle drei Korrelationskoeffizienten $r_{x,y}$, $\rho_{x,y}$ und $\tau_{x,y}$.

▼

Lösung: Wir lesen dazu das im Datenordner abgelegte Datenfile *stahlfile* ein, rufen die zweite und dritte Spalte dieser Datenmatrix auf und bezeichnen die zugehörigen Stichproben mit "kohlenstoff" bzw "zugfestigkeit". Die gesuchten Maßzahlen geben wir in übersichtlicher Weise in Form einer Tabelle aus:

```
kohlenstoff = Rest[Part[<< stahlfile, All, 2]];
zugfestigkeit = Rest[Part[<< stahlfile, All, 3]];
TableForm[{
    {"Kovarianz: ", Covariance[kohlenstoff, zugfestigkeit]},
    {"Pearson: ", Correlation[kohlenstoff, zugfestigkeit]},
    {"Spearman: ", SpearmanRankCorrelation[kohlenstoff, zugfestigkeit] // N},
    {"Kendall: ", KendallRankCorrelation[kohlenstoff, zugfestigkeit] // N}},
  TableSpacing → {1, 5}]
Clear[kohlenstoff, zugfestigkeit]
```

```
Kovarianz:      71.2524
Pearson:        0.62625
Spearman:       0.610589
Kendall:        0.47059
```

Wie erkennen, dass alle drei Korrelationskoeffizienten zwar deutlich positiv aber doch von 1 relativ weit entfernt sind. Die beiden Merkmale Kohlenstoff und Zugfestigkeit hängen daher zwar deutlich in gleichsinniger Weise voneinander ab, es besteht jedoch kein exakter linearer Zusammenhang zwischen diesen beiden Merkmalen (man vergleiche dazu die in Beispiel 2.2.2 durch ein Scatter-Plot graphisch dargestellten Abhängigkeit dieser beiden Merkmale).