§4 Stichproben und Statistiken

Will man sich nicht nur auf die beschreibende Statistik - also die graphische Darstellung von statistischen Daten und die Berechnung von gewissen Maßzahlen - beschränken, so muss zunächst der Begriff "Stichprobe" genauer definiert werden, wobei vor allem die Zufälligkeit der Stichprobe in adäquater Weise in Betracht gezogen werden muss.

Sobald aber eine Stichprobe als zufällig angesehen wird, werden davon abhängige statistische Maßzahlen notgedrungen zu Zufallsvariablen - man spricht von Statistiken. Wir werden uns in diesem Kapitel mit einigen wichtigen Statistiken im Detail befassen. Weitere wichtige Statistiken werden wir in den folgenden Kapiteln bei Bedarf definieren und ihre Eigenschaften besprechen.

4.1 Die Zufälligkeit einer Stichprobe

Bisher sind wir stets davon ausgegangen, dass ein statistisches Datenmaterial in Form einer Datenmatrix vorliegt, wobei wir uns über die Entstehung dieses Datenmaterials keine Gedanken gemacht haben. In vielen Fällen entsteht ein derartiges Datenmaterial aber durch eine zufällige Auswahl von einzelnen Objekten, an denen gewisse Merkmale beobachtet werden. Beim Datenmaterial *stahl* wurden beispielsweise einzelne Stahlbleche der Produktion zufällig entnommen und von diesen Blechen der Kohlenstoffgehalt und die Zugfestigkeit ermittelt.

Mit den folgenden Begriffsbildungen wird der Zufälligkeit einer Stichprobe Rechnung getragen.

4.1.1 Definition:
a) Eine Zufallsvariable X zusammen mit ihrer Verteilung $\mathbb{P}_X$ nennt man Grundgesamtheit.

b) Eine Liste $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ von jeweils n durch unabhängiges Experimentieren gewonnene Realisierungen der Zufallsvariablen X nennt man eine nach dem Verteilungsgesetz $\mathbb{P}_X$ ausgewählte konkrete Stichprobe vom Umfang n. Man spricht in diesem Zusammenhang auch davon, dass man "aus einer $\mathbb{P}_X$-verteilten Grundgesamtheit eine Stichprobe vom Umfang n zieht".

c) Wir fassen nun die Zahl x_k - also die Realisierung der Zufallsvariablen X beim k-ten Versuch - als Realisierung einer Zufallsvariablen X_k auf und nennen die Liste $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ dieser vollständig unabhängigen und identisch $\mathbb{P}_X$-verteilten Zufallsvariablen eine nach dem Verteilungsgesetz $\mathbb{P}_X$ ausgewählte mathematische Stichprobe vom Umfang n.

▼

Die Begriffe $\mathbb{E}$-wertige Grundgesamtheit (darunter versteht man eine $\mathbb{E}$-wertige Zufallsvariable zusammen mit ihrer Verteilung), $\mathbb{E}$-wertige konkrete Stichprobe von Umfang n und $\mathbb{E}$-wertige mathematische Stichprobe vom Umfang n werden analog definiert.

Sobald man eine Stichprobe als zufällig ansieht, wird jede davon abhängige Maßzahl automatisch zu einer Zufallsvariablen (symbolisch drücken wir dies dadurch aus, dass aus den kleinen Buchstaben, mit denen wir Maßzahlen bezeichnet haben, große Buchstaben werden). Wir definieren in diesem Zusammenhang:

4.1.2 Definition: Ist $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ eine nach dem Verteilungsgesetz $\mathbb{P}_X$ ausgewählte mathematische Stichprobe vom Umfang n und ist $s : \mathbb{R}^n \to \mathbb{R}$ eine (stückweise stetige) Abbildung, so nennt man

$$S^{(n)} = s[\vec{X}]$$

eine Statistik. Geht aus dem Kontext der Stichprobenumfang n klar hervor, so verzichten wir der Einfachheit halber auf den oberen Index, schreiben statt $S^{(n)}$ also einfach S.

Die Verteilung einer Statistik S und damit auch ihr Erwartungswert, ihre Streuung, ihre Varianz sowie die Wahrscheinlichkeit von Ereignissen der Form $\{S \in A\}$ hängen von der Verteilung $\mathbb{P}_X$ der Zufallsvariablen X

ab. Wir bringen diese Tatsache dadurch zum Ausdruck, indem wir in Zukunft die Schreibweise $\mathbb{E}[S; \mathbb{P}_X]$, $\mathbb{S}[S; \mathbb{P}_X]$, $\mathbb{V}[S; \mathbb{P}_X]$ sowie $\mathbb{P}[\{S \in A\}; \mathbb{P}_X]$ verwenden.

Unter dem Wissensgebiet "Statistik" versteht man die "Lehre von den Statistiken". Das Wissensgebiet der Statistik befasst sich also mit den wahrscheinlichkeitstheoretischen Eigenschaften der verschiedenen Statistiken zusammen mit ihren äußerst zahlreichen Anwendungen.

4.2 Einige wichtige Statistiken

Im Rahmen dieses Lehrgangs werden wir vielen Statistiken begegnen. Abhängig von der jeweiligen Fragestellung werden von diesen Statistiken der Erwartungswert, die Varianz, die Verteilung bzw die asymptotische Verteilung (darunter versteht man die Verteilung dieser Statistik für den Fall, dass der Stichprobenumfang n gegen ∞ geht) benötigt. Die Ermittlung dieser Größen kann mitunter sehr aufwändig (ja oft sogar unmöglich) sein und wird im Rahmen von Sätzen (wobei wir oft auf Beweise verzichten müssen), Bemerkungen und Beispielen erfolgen, die über den ganzen Lehrgang verstreut sind.

Um zu zeigen, wie mit Statistiken gearbeitet wird, werden wir in diesem Abschnitt exemplarisch einige besonders wichtige Statistiken ausführlich behandeln.

Mittelwert, Varianz und Standardabweichung

Im Rahmen der Statistik treten die Statistiken Mittelwert, Varianz und Standardabweichung besonders häufig auf. Wir beginnen daher mit einer genauen Analyse dieser drei Statistiken:

4.2.1 Definition: Ist $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ eine nach dem Verteilungsgesetz $\mathbb{P}_X$ ausgewählte mathematische Stichprobe vom Umfang n, so nennt man die Statistiken

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad V_X = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \qquad S_X = \sqrt{V_X}$$

den Mittelwert, die Varianz und die Standardabweichung von $\vec{X}$ (man vergleiche dazu die Maßzahlen $\overline{x}$, v_x und s_x aus Definition 3.1.1 bzw Definition 3.1.7 sowie die Begriffsbildung in Satz 23.4.2; die in diesem Satz verwendete Beifügung "empirisch" lassen wir in Hinkunft der Einfachheit halber weg).

4.2.2 Satz: Ist die mathematische Stichprobe $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ vom Umfang n nach dem Verteilungsgesetz $\mathbb{P}_X$ ausgewählt, so gilt

$$\mathbb{E}[\overline{X}; \mathbb{P}_X] = \mathbb{E}[X] \qquad \mathbb{E}[V_X, \mathbb{P}_X] = \mathbb{V}[X]$$

$$\mathbb{V}[\overline{X}; \mathbb{P}_X] = \frac{1}{n}\,\mathbb{V}[X] \qquad \mathbb{V}[V_X, \mathbb{P}_X] = \frac{1}{n}\,\mathbb{E}[(X - \mathbb{E}[X])^4] - \frac{n-3}{n\,(n-1)}\,\mathbb{V}[X]^2$$

▼

Beweis: a) Aus der Linearität des Erwartungswerts folgt unmittelbar

$$\mathbb{E}[\overline{X}; \mathbb{P}_X] = \mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} X_i] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \mathbb{E}[X]$$

b) Berücksichtigt man die Tatsache, dass für $i \neq k$ die beiden Zufallsvariablen X_i und X_k unabhängig sind und damit $\mathbb{E}[X_i\, X_k] = \mathbb{E}[X_i]\,\mathbb{E}[X_k] = \mathbb{E}[X]^2$ gilt, so erhält man durch einfaches Umformen

$$\mathbb{E}[V_X, \mathbb{P}_X] = \mathbb{E}[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \frac{1}{n} \sum_{j=1}^{n} X_j)^2] =$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}[X_i^2 - \frac{2}{n} X_i^2 - \frac{2}{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} X_i X_j + \frac{1}{n^2} \sum_{j=1}^{n} X_j^2 + \frac{1}{n^2} \sum_{\substack{j,k=1 \\ j\neq k}}^{n} X_j X_k] =$$

$$= \frac{1}{n-1} (n\,\mathbb{E}[X^2] - 2\,\mathbb{E}[X^2] - 2\,(n-1)\,\mathbb{E}[X]^2 + \mathbb{E}[X^2] + (n-1)\,\mathbb{E}[X]^2) = \mathbb{V}[X]$$

c) Offenbar gilt

$$\mathbb{E}[(\overline{X})^2; \mathbb{P}_X] = \mathbb{E}[(\frac{1}{n} \sum_{i=1}^{n} X_i)^2] = \frac{1}{n^2}\,\mathbb{E}[\sum_{i=1}^{n} X_i^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} X_i X_j] = \frac{1}{n}\,\mathbb{E}[X^2] + \frac{n-1}{n}\,\mathbb{E}[X]^2$$

Zusammen mit a) erhält man damit

$$\mathbb{V}[\overline{X}; \mathbb{P}_X] = \mathbb{E}[(\overline{X})^2; \mathbb{P}_X] - \mathbb{E}[\overline{X}; \mathbb{P}_X]^2 = \frac{1}{n}\,\mathbb{E}[X^2] + \frac{n-1}{n}\,\mathbb{E}[X]^2 - \mathbb{E}[X]^2 = \frac{1}{n}\,\mathbb{V}[X]$$

d) Der Beweis der letzten Aussage ist umfangreich, aber im Prinzip ähnlich den oben angeführten Beweisen.

Der folgende Satz ist für zahlreiche Tests von zentraler Bedeutung.

4.2.3 Satz: Ist die mathematische Stichprobe $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ vom Umfang n nach dem Verteilungsgesetz $\mathbb{P}_X = \mathcal{N}[\mu, \sigma]$ ausgewählt, so gilt speziell

a) der Mittelwert $\overline{X}$ und die Varianz V_X von $\vec{X}$ sind unabhängig;

b) der Mittelwert $\overline{X}$ von $\vec{X}$ ist $\mathcal{N}[\mu, \sigma/\sqrt{n}\,]$-verteilt;

c) die Statistik $(n-1)\,V_X/\sigma^2$ ist $Chi[n-1]$-verteilt;

d) die Statistik $T_X = (\overline{X} - \mu)\sqrt{n}\,/S_X$ ist $\mathcal{T}[n-1]$ verteilt.

▼

Beweis: Die Zufallsvariablen $Z_1, Z_2, \ldots, Z_n$ mit $Z_i = (X_i - \mu)/\sigma$ sind offenbar vollständig unabhängig und identisch $\mathcal{N}[0, 1]$-verteilt, erfüllen also die Voraussetzungen von Satz 23.3.8 und Satz 23.4.2. Damit gilt:

- die beiden Zufallsvariablen $\overline{Z}$ und V_Z sind unabhängig (Satz 23.3.8);
- die Zufallsvariable $\overline{Z}$ ist $\mathcal{N}[0, 1/\sqrt{n}\,]$-verteilt (Faltungsformeln und Satz über die affine Transformation);
- die Zufallsvariable $(n-1)\,V_Z$ ist $Chi[n-1]$-verteilt (Satz 23.3.8);
- die Zufallsvariable $T_Z = \overline{Z}\sqrt{n}\,/S_Z$ ist $\mathcal{T}[n-1]$ verteilt (Satz 23.4.2).

Unsere Behauptungen ergeben sich aus diesen vier Eigenschaften unter Berücksichtigung der Tatsache, dass zwischen den Zufallsvariablen $\overline{Z}$, V_Z, T_Z und $\overline{X}$, V_X, T_X offensichtlich die folgenden Beziehungen bestehen:

$$\overline{X} = \sigma\,\overline{Z} + \mu \qquad \text{bzw} \qquad V_X = \sigma^2\,V_Z \qquad \text{bzw} \qquad T_X = T_Z$$

Ordnungsstatistiken

Auch die Ordnungsstatistiken und ihre Verwandten besitzen im Rahmen der Statistik eine besondere Bedeutung:

4.2.4 Definition: Ist $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ eine nach dem Verteilungsgesetz $\mathbb{P}_X$ ausgewählte mathematische Stichprobe vom Umfang n, so nennt man die Statistiken (man vergleiche dazu Definition 19.2.5)

$$X_k^* = \mathrm{Min}[\mathrm{Max}[X_{i_1}, X_{i_2}, \ldots, X_{i_k}] \mid i_1 < i_2 < \ldots < i_k \in \{1, 2, \ldots, n\}]$$

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

$$R_X = X_n^* - X_1^*$$

die k-te Ordnungsstatistik, den Median bzw die Spannweite von $\vec{X}$ (man vergleiche dazu die Maßzahlen x_k^*, $\tilde{x}$ und r_X aus Definition 3.1.3 bzw Definition 3.1.1 bzw Definition 3.1.7).

4.2.5 Satz: Ist die mathematische Stichprobe $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ vom Umfang n nach einem stetigen Verteilungsgesetz $\mathbb{P}_X$ mit der Verteilungsfunktion $\mathbb{F}$ und der Verteilungsdichte $\mathbb{f}$ ausgewählt, so gilt für die Verteilungsdichte $\mathbb{f}_{X_k^*}$ der k-ten Ordnungsstatistik

$$\mathbb{f}_{X_k^*}[x] = k \binom{n}{k} \mathbb{f}[x]\, \mathbb{F}[x]^{k-1} (1 - \mathbb{F}[x])^{n-k}$$

und die gemeinsame Verteilungsdichte $\mathbb{f}_{X_i^*, X_k^*}$ der i-ten und k-ten Ordnungsstatistik im Fall $i < k$

$$\mathbb{f}_{X_i^*, X_k^*}[x, y] = i\,(k-i)\, \frac{n!}{i!\,(k-i)!\,(n-k)!}\, \mathbb{f}[x]\, \mathbb{f}[y]\, \mathbb{F}[x]^{i-1} (\mathbb{F}[y] - \mathbb{F}[x])^{k-i-1} (1 - \mathbb{F}[y])^{n-k}$$

▼

Beweis: Die erste Aussage folgt unmittelbar aus Beispiel 19.2.6.

Im Fall $i < k$ gilt für alle $x < y$

$$\mathbb{f}_{X_i^*, X_k^*}[x, y] = \frac{1}{dx\,dy}\, \mathbb{P}[\{X_i^* \in [x, x + dx]\} \cap \{X_k^* \in [y, y + dy]\}] =$$

$$= \frac{1}{dx\,dy}\, \mathbb{P}[\left\{\begin{array}{l}\text{Von den Zufallsvariablen } X_1, X_2, \ldots, X_n \text{ liegt je eine im Intervall} \\ {[x, x + dx]} \text{ bzw } [y, y + dy],\ \ i - 1 \text{ sind kleiner als } x,\ \ n - k \text{ sind} \\ \text{größer als } y \text{ und die restlichen } k - i - 1 \text{ liegen im Intervall } [x, y]\end{array}\right\}] =$$

$$= n\, \mathbb{f}[x]\, (n-1)\, \mathbb{f}[y] \binom{n-2}{i-1} \mathbb{F}[x]^{i-1} \binom{n-i-1}{n-k} (1 - \mathbb{F}[y])^{n-k} (\mathbb{F}[y] - \mathbb{F}[x])^{k-i-1}$$

4.2.6 Beispiel: Man entwickle eine Prozedur, mit der sich die Verteilungsdichte der k-ten Ordnungsstatistik X_k^* von $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ für eine beliebige in *Mathematica* implementierte, stetige Verteilung $\mathbb{P}_X$ zeichnen lässt.

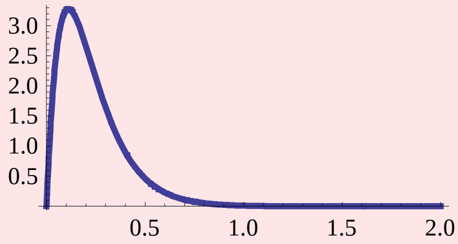
▼

Lösung: Die einzugebenden Parameter sind der Stichprobenumfang n, der Parameter k, das Verteilungsgesetz dist, der Bereich, in dem die Verteilungsdichte der k-ten Ordnungsstatistik gezeichnet werden soll, sowie optionale Angaben betreffend den Stil der Zeichnung:

```
PlotOrdnungsStatistik[n_, k_, dist_, bereich_, options___] := Module[{x = bereich[[1]]},
   Plot[k Binomial[n, k] PDF[dist, x] CDF[dist, x]^(k-1) (1 - CDF[dist, x])^(n-k), bereich, options]]
```

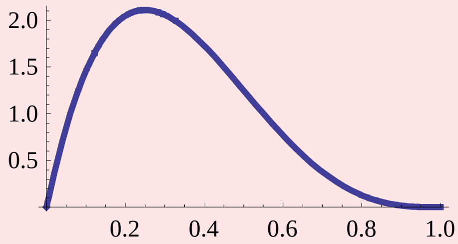
Beispielsweise gilt

```
PlotOrdnungsStatistik[5, 2, ExponentialDistribution[2], {x, 0, 2}, PlotStyle → Thickness[0.015],
  AspectRatio → 0.5]
```



sowie

```
PlotOrdnungsStatistik[5, 2, UniformDistribution[{0, 1}], {x, 0, 1}, PlotStyle → Thickness[0.015],
  AspectRatio → 0.5]
```



4.2.7 Beispiel: Für den Spezialfall $\mathbb{P}_X = \mathcal{U}[\{a, b\}]$ berechne man den Erwartungswert und die Varianz der k-ten Ordnungsstatistik X_k^* von $\vec{X} = \{X_1, X_2, \ldots, X_n\}$.

▼

Lösung: Für alle $a \le x \le b$ gilt für die Verteilungsdichte $\mathbb{f}[x]$ bzw die Verteilungsfunktion $\mathbb{F}[x]$ der Gleichverteilung $\mathcal{U}\{[a, b]\}$ bekanntlich $\mathbb{f}[x] = 1/(b-a)$ bzw $\mathbb{F}[x] = (x-a)/(b-a)$. Mit Hilfe von *Mathematica* lassen sich damit der Erwartungswert $\mathbb{E}[X_k^*; \mathcal{U}[\{a, b\}]]$ und die Varianz $\mathbb{V}[X_k^*; \mathcal{U}[\{a, b\}]]$ der k-ten Ordnungsstatistik X_k^* leicht berechnen.

```
moment1 = k Binomial[n, k] (b - a)^-n Integrate[ x (x - a)^(k-1) (b - x)^(n-k), {x, a, b}];
moment2 = k Binomial[n, k] (b - a)^-n Integrate[ x^2 (x - a)^(k-1) (b - x)^(n-k), {x, a, b}];
FullSimplify[moment1, {0 < k ≤ n, a < b}]
FullSimplify[moment2 - moment1^2, {0 < k ≤ n, a < b}]
Clear[moment1, moment2]
```

$$a + \frac{(-a + b)\, k}{1 + n}$$

$$-\frac{(a - b)^2\, k\, (-1 + k - n)}{(1 + n)^2\, (2 + n)}$$

Es gilt somit

$$\mathbb{E}[X_k^*; \mathcal{U}[\{a, b\}]] = a + \frac{k\,(b-a)}{n+1} \qquad \text{und} \qquad \mathbb{V}[X_k^*; \mathcal{U}[\{a, b\}]] = \frac{k\,(b-a)^2\,(n-k+1)}{(n+1)^2\,(n+2)}$$

4.2.8 Satz: Ist die mathematische Stichprobe $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ vom Umfang n nach einem stetigen

Verteilungsgesetz $\mathbb{P}_X$ mit der Verteilungsfunktion $\mathbb{F}$ und der Verteilungsdichte $\mathbb{f}$ ausgewählt, so gilt für die Verteilungsdichte $\mathbb{f}_{\tilde{X}}$ des Medians im Fall $n = 2\,m$ (im Fall $n = 2\,m + 1$ stimmt der Median $\tilde{X}$ mit der Ordnungsstatistik X^*_{m+1} überein, die Verteilungsdichte von $\mathbb{f}_{\tilde{X}}$ lässt sich in diesem Fall mit Satz 4.2.5 ermitteln)

$$\mathbb{f}_{\tilde{X}}[x] = \frac{2\,(2\,m)!}{(m-1)!\,(m-1)!} \int_{-\infty}^{x} \mathbb{f}[u]\,\mathbb{f}[2\,x-u]\,\mathbb{F}[u]^{m-1}\,(1-\mathbb{F}[2\,x-u])^{m-1}\,d u$$

und für die Verteilungsdichte $\mathbb{f}_{R_X}$ der Spannweite

$$\mathbb{f}_{R_X}[x] = n\,(n-1) \int_{-\infty}^{\infty} \mathbb{f}[u]\,\mathbb{f}[u+x]\,(\mathbb{F}[u+x]-\mathbb{F}[u])^{n-2}\,d u$$

▼

Beweis: Für alle $x \in \mathbb{R}$ gilt unter Verwendung des Satzes von der totalen Wahrscheinlichkeit in differenzieller Form unter Berücksichtigung von Bemerkung 18.2.3

$$\mathbb{f}_{\tilde{X}}[x] = \frac{1}{d x}\,\mathbb{P}[\{(X^*_m + X^*_{m+1})/2 \in [x,\, x + d x]\}] =$$

$$= \frac{1}{d x} \int_{-\infty}^{x} \mathbb{P}[\{X^*_m + X^*_{m+1} \in [2\,x,\, 2\,x + 2\,d x]\} \mid \{X^*_m = u\}]\,\mathbb{f}_{X^*_m}[u]\,d u =$$

$$= 2 \int_{-\infty}^{x} \mathbb{f}_{X^*_{m+1} | \{X^*_m = u\}}[2\,x-u]\,\mathbb{f}_{X^*_m}[u]\,d u = 2 \int_{-\infty}^{x} \mathbb{f}_{X^*_m, X^*_{m+1}}[u,\, 2\,x-u]\,d u =$$

$$= \frac{2\,(2\,m)!}{(m-1)!\,(m-1)!} \int_{-\infty}^{x} \mathbb{f}[u]\,\mathbb{f}[2\,x-u]\,\mathbb{F}[u]^{m-1}\,(1-\mathbb{F}[2\,x-u])^{m-1}\,d u$$

Hinsichtlich der Verteilungsdichte der Spannweite R_X der Stichprobe $\vec{X}$ vergleiche man Beispiel 19.2.7.

4.2.9 Beispiel: Man entwickle eine Prozedur, mit der sich die Verteilungsdichte der Spannweite R_X von $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ für beliebige in *Mathematica* implementierte, stetige Verteilungen $\mathbb{P}_X$ zeichnen lässt.

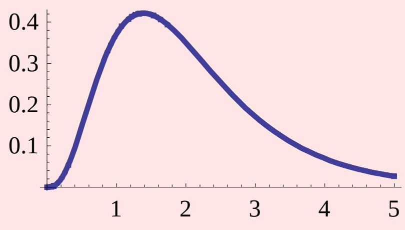
▼

Lösung: Die einzugebenden Parameter sind der Stichprobenumfang n, das Verteilungsgesetz dist, der Bereich, in dem die Verteilungsdichte der Spannweite gezeichnet werden soll, sowie optionale Angaben betreffend den Stil der Zeichnung:

```
PlotSpannweite[n_, dist_, bereich_, options___] := Module[{x = bereich[[1]], umin, umax},
    umin = DistributionDomain[dist][[1, 1]];
    umax = DistributionDomain[dist][[1, 2]];
    Plot[n (n − 1) NIntegrate[PDF[dist, u] PDF[dist, u + x] (CDF[dist, u + x] − CDF[dist, u])^(n−2),
        {u, umin, umax}], bereich, options]]
```

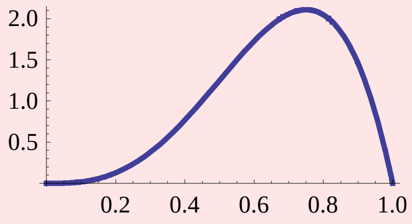
Beispielsweise gilt

```
PlotSpannweite[5, ExponentialDistribution[1], {x, 0, 5}, PlotStyle → Thickness[0.015], AspectRatio → 0.5,
  ImageSize → {150, 80}]
```



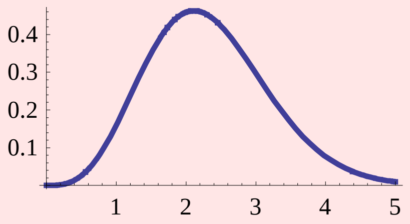
sowie

```
PlotSpannweite[5, UniformDistribution[{0, 1}], {x, 0, 1}, PlotStyle → Thickness[0.015], AspectRatio → 0.5,
  ImageSize → {150, 80}]
```



und

```
PlotSpannweite[5, NormalDistribution[0, 1], {x, 0, 5}, PlotStyle → Thickness[0.015], AspectRatio → 0.5,
  ImageSize → {150, 80}]
```



4.2.10 Beispiel: Für den Spezialfall $\mathbb{P}_X = \mathcal{U}[\{a, b\}]$ berechne man den Erwartungswert und die Varianz des Medians $\tilde{X}$, der Spannweite R_X und der Statistik $(X_1^* + X_n^*)/2$ von $\vec{X} = \{X_1, X_2, \ldots, X_n\}$.

▼

Lösung: a) Die Erwartungswerte $\mathbb{E}[\tilde{X}; \mathcal{U}[\{a, b\}]]$, $\mathbb{E}[R_X; \mathcal{U}[\{a, b\}]]$ und $\mathbb{E}[(X_1^* + X_n^*)/2; \mathcal{U}[\{a, b\}]]$ ergeben sich unmittelbar aus der in Beispiel 4.2.7 ermittelten Formel für den Erwartungswert der k-ten Ordnungsstatistik:

$$\mathbb{E}[\tilde{X}; \mathcal{U}[\{a, b\}]] = \begin{cases} \mathbb{E}[X_{m+1}^*; \mathcal{U}[\{a, b\}]] = (a+b)/2 & \text{für } n = 2\,m+1 \\ (\mathbb{E}[X_m^*; \mathcal{U}[\{a, b\}]] + \mathbb{E}[X_{m+1}^*; \mathcal{U}[\{a, b\}]])/2 = (a+b)/2 & \text{für } n = 2\,m \end{cases}$$

$$\mathbb{E}[R_X; \mathcal{U}[\{a, b\}]] = (\mathbb{E}[X_n^*; \mathcal{U}[\{a, b\}\,\{]] - \mathbb{E}[X_1^*; \mathcal{U}[a, b]]])/2 = \frac{n-1}{n+1}\,(b-a)$$

$$\mathbb{E}[(X_1^* + X_n^*)/2; \mathcal{U}[\{a, b\}]] = (\mathbb{E}[X_n^*; \mathcal{U}[\{a, b\}]] + \mathbb{E}[X_1^*; \mathcal{U}[\{a, b\}]])/2 = \frac{1}{2}\,(a+b)$$

b) Zur Berechnung der Varianzen $\mathbb{V}[\tilde{X}; \mathcal{U}[\{a, b\}]]$, $\mathbb{V}[R_X; \mathcal{U}[\{a, b\}]]$ und $\mathbb{V}[(X_1^* + X_n^*)/2; \mathcal{U}[\{a, b\}]]$ verwenden wir Satz 17.5.6, die in Beispiel 4.2.7 ermittelte Formel für die Varianz der k-ten Ordnungsstatistik sowie die Formel

$$\mathbb{K}[X_i^*, X_k^*; \mathcal{U}[a, b]] = \frac{i\,(n-k+1)}{(n+1)^2\,(n+2)}\,(b-a)^2$$

für die Kovarianz $\mathbb{K}[X_i^*, X_k^*; \mathcal{U}[\{a, b\}]]$ der beiden Ordnungsstatistiken X_i^* und X_k^* mit $i < k$, welche sich unter Verwendung von Satz 4.2.5 analog zu Beispiel 4.2.7 mit Hilfe von *Mathematica* leicht beweisen lässt:

```
FullSimplify[i (k - i) n! / (i! (k - i)! (n - k)! (b - a)^n)
    Integrate[y (b - y)^(n-k) Integrate[x (x - a)^(i-1) (y - x)^(k-i-1), {x, a, y}], {y, a, b},
     Assumptions -> {a < x < y < b, 1 ≤ i < k ≤ n}]
   - (a + i (b - a) / (n + 1)) (a + k (b - a) / (n + 1))]
```

$$\frac{(a-b)^2\, i\,(1-k+n)}{(1+n)^2\,(2+n)}$$

Damit ergibt sich nach elementaren Umformungen

$$\mathbb{V}[\tilde{X}; \mathcal{U}[\{a,b\}]] = \begin{cases} \mathbb{V}[X^*_{m+1}; \mathcal{U}[\{a,b\}]] = \dfrac{1}{4\,(n+2)}\,(b-a)^2 & \text{für } n = 2\,m+1 \\ \mathbb{V}[(X^*_m + X^*_{m+1})/2; \mathcal{U}[\{a,b\}]] = \dfrac{n}{4\,(n+1)\,(n+2)}\,(b-a)^2 & \text{für } n = 2\,m \end{cases}$$

$$\mathbb{V}[R_X; \mathcal{U}[\{a,b\}]] = \mathbb{V}[X^*_n - X^*_1; \mathcal{U}[\{a,b\}]] = \frac{2\,(n-1)}{(n+1)^2\,(n+2)}\,(b-a)^2$$

$$\mathbb{V}[(X^*_1 + X^*_n)/2; \mathcal{U}[\{a,b\}]] = \frac{1}{2\,(n+1)\,(n+2)}\,(b-a)^2$$

Rangstatistiken

Abschließend befassen wir uns noch mit den sogenannten Rangstatistiken:

4.2.11 Definition: Ist $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ eine nach dem Verteilungsgesetz $\mathbb{P}_X$ ausgewählte mathemati-sche Stichprobe vom Umfang n, so nennt man die Statistik $\text{Rg}[\vec{X}, X_i]$, welche jedem $\omega \in \Omega$ die Zahl

$$\text{Rg}[\vec{X}, X_i][\omega] = \text{rg}[\{X_1[\omega], X_2[\omega], \ldots, X_n[\omega]\}, X_i[\omega]]$$

zuordnet, den Rang von X_i (man vergleiche dazu die in Definition 3.1.3 eingeführte Maßzahl $\text{rg}[\vec{x}, \textit{element}]$).

4.2.12 Satz: Ist das Verteilungsgesetz $\mathbb{P}_X$ stetig, so handelt es sich bei der gemeinsamen Verteilung der Ränge $\text{Rg}[\vec{X}, X_1]$, $\text{Rg}[\vec{X}, X_2]$, …, $\text{Rg}[\vec{X}, X_n]$ um die Gleichverteilung auf der Menge aller Permutationen ohne Wiederholung von n Dingen. Für alle $i \in \{1, 2, \ldots, n\}$ gilt damit

$$\mathbb{E}[\text{Rg}[\vec{X}, X_i]; \mathbb{P}_X] = \frac{n+1}{2} \qquad \mathbb{V}[\text{Rg}[\vec{X}, X_i]; \mathbb{P}_X] = \frac{n^2-1}{12}$$

▼

Beweis: Da das Verteilungsgesetz $\mathbb{P}_X$ stetig ist, sind die Zufallsvariablen $X_1, X_2, \ldots, X_n$ (fast) sicher voneinander verschieden. Für (fast) jedes $\omega \in \Omega$ entsprechen ihre Ränge

$$\text{Rg}[\vec{X}, X_1][\omega], \text{Rg}[\vec{X}, X_2][\omega], \ldots, \text{Rg}[\vec{X}, X_n][\omega]$$

damit einer Permutation ohne Wiederholung der Zahlen $1, 2, \ldots, n$, wobei jede dieser Permutationen offenbar mit der gleichen Wahrscheinlichkeit auftreten kann. Für alle $i \in \{1, 2, \ldots, n\}$ ist die Zufallsvariable $\text{Rg}[\vec{X}, X_i]$ damit auf der Menge $\{1, 2, \ldots, n\}$ gleichverteilt. Die oben angeführten Formeln für den Erwartungswert und die Varianz von $\text{Rg}[\vec{X}, X_i]$ erhält man damit leicht unter Verwendung von *Mathematica*

```
Mean[DiscreteUniformDistribution[{1, n}]]
Variance[DiscreteUniformDistribution[{1, n}]]
```

$$\frac{1+n}{2}$$

$$\frac{1}{12}\left(-1+n^2\right)$$

4.2.13 Bemerkung: Die gemeinsame Verteilung der Ränge Rg[$\vec{X}$, X_1], Rg[$\vec{X}$, X_2], …, Rg[$\vec{X}$, X_n] und somit auch die Verteilung jeder beliebigen Funktion dieser Ränge hängt vom stetigen Verteilungsgesetz $\mathbb{P}_X$, nach dem die mathematische Stichprobe $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ ausgewählt wurde, nicht ab. Diese Tatsache nützt man bei den sogenannten verteilungsfreien Verfahren.