Die Benford-Verteilung
Anwendung auf reale Daten der Marktforschung

1 Wahrscheinlichkeitsraum

Die Mantissenfunktion M_b zur Basis $b \in \mathbb{N} \setminus \{1\}$ ordnet jeder Zahl $x \in \mathbb{R}^+$ ihre Mantisse $M_b(x) = m_b \in [1, b)$ zu. $D_n^{(b)}(x)$ ist die n-te signifikante Ziffer von x zur Basis b.

$(\mathbb{R}^+, \mathcal{M}_b, \tilde{\mathbf{P}})$ ist der Wahrscheinlichkeitsraum mit Grundraum $\mathbb{R}^+$, Mantissen-σ-Algebra $\mathcal{M}_b$ und Benford-Wahrscheinlichkeit $\tilde{\mathbf{P}}$.

Definition 1 (Mantissen-σ-Algebra)
Die von der Menge der Funktionen der signifikanten Ziffern $\{D_n^{(b)}\}_{n\in\mathbb{N}}$ auf $\mathbb{R}^+$ erzeugte σ-Algebra $\mathcal{M}_b = \left\{\bigcup_{e=-\infty}^{\infty} B \cdot b^e \text{ für alle Borelmengen } B \subseteq [1, b)\right\}$ heißt **Mantissen-σ-Algebra** *zur Basis b.*

Definition 2 (Benford-Verteilungsfunktion)
Die Verteilungsfunktion $\tilde{\mathbf{P}}(M_b(x) < m_b) = log_b(m_b)$ für alle $m_b \in [1, b)$ heißt **Benford-Verteilungsfunktion** *zur Basis b.*

Satz 1 (Benford-Wahrscheinlichkeit)
Die **Benford-Wahrscheinlichkeit $\tilde{\mathbf{P}}$** *ist gegeben durch die gemeinsame Verteilung der n ersten signifikanten Ziffern ($n \in \mathbb{N}$) mit $d_1 \in \{1, 2, ..., b-1\}$ und $d_j \in \{0, 1, 2, ..., b-1\}$ für $j = 2, ..., n$*

$$\tilde{\mathbf{P}}(D_1^{(b)} = d_1, ..., D_n^{(b)} = d_n) = \tilde{\mathbf{P}}\Big(\bigcap_{i=1}^{n}\{D_i^{(b)} = d_i\}\Big) = log_b\left(1 + \frac{1}{\sum\limits_{i=1}^{n} d_i \cdot b^{n-i}}\right).$$

2 Spezialfälle von Benfords Gesetz

Satz 2 (Gesetz der ersten Ziffern)
Die Wahrscheinlichkeiten der ersten signifikanten Ziffer $D_1^{(b)}$ ist gegeben durch

$$\tilde{\mathbf{P}}(D_1^{(b)} = d_1) = log_b\Big(1 + \frac{1}{d_1}\Big), \text{ wobei } d_1 = 1, 2, ..., b-1 \text{ ist.}$$

Satz 3 (Gesetz der n-ten Ziffern)
Die Wahrscheinlichkeiten der n-ten signifikanten Ziffer $D_n^{(b)}$ $(n \in \mathbb{N} \setminus \{1\})$ *lauten*

$$\tilde{\mathbf{P}}(D_n^{(b)} = d_n) = \sum_{i=b^{n-2}}^{b^{n-1}-1} log_b\Big(1 + \frac{1}{b \cdot i + d_n}\Big), \text{ wobei } d_n \in \{0, 1, ..., b-1\} \text{ ist.}$$

3 Strukturelle Eigenschaften der Benford-Verteilung

3.1 Skaleninvarianz

Definition 3 (Skaleninvarianz)
Ein Wahrscheinlichkeitsmaß $\mathbf{P}$ *auf* $(\mathbb{R}^+, \mathcal{M}_b)$, *für das* $\mathbf{P}(S) = \mathbf{P}(\alpha S)$ *für alle* $\alpha \in \mathbb{R}^+$ *und alle* $S \in \mathcal{M}_b$ *gilt, heißt* **skaleninvariant**.

Satz 4 (Skaleninvarianz impliziert Benfords Gesetz)
Die Benford-Wahrscheinlichkeit $\tilde{\mathbf{P}}$ *ist das einzige skaleninvariante Wahrscheinlichkeitsmaß auf* $(\mathbb{R}^+, \mathcal{M}_b)$. *Ein Wahrscheinlichkeitsmaß* $\mathbf{P}$ *auf* $(\mathbb{R}^+, \mathcal{M}_b)$ *ist also genau dann skaleninvariant, wenn* $\mathbf{P}$ *die Benford-Wahrscheinlichkeit* $\tilde{\mathbf{P}}$ *ist.*

3.2 Baseninvarianz

Definition 4 (Baseninvarianz)
Ein Wahrscheinlichkeitsmaß $\mathbf{P}$ *auf* $(\mathbb{R}^+, \mathcal{M}_b)$, *für das* $\mathbf{P}(S) = \mathbf{P}(S^{1/n})$ *für alle* $n \in \mathbb{N}^+$ *und alle* $S \in \mathcal{M}_b$ *gilt, heißt* **baseninvariant**.

Satz 5
$\tilde{\mathbf{P}}$ *ist das einzige baseninvariante, atomlose Wahrscheinlichkeitsmaß auf* $(\mathbb{R}^+, \mathcal{M}_b)$.

4 Grenzwertsatz für signifikante Ziffern

"'Werden Wahrscheinlichkeitsverteilungen zufällig gewählt und aus jeder dieser Verteilungen zufällig eine Stichprobe gezogen, dann konvergieren die Häufigkeiten der signifikanten Ziffern der gemeinsamen Stichprobe unter relativ schwachen Bedingungen gegen die Benford-Verteilung."'

Definition 5 (zufälliges Borel-Wahrscheinlichkeitsmaß)
Sei $(\Omega, \mathcal{F}, \mathbf{P})$ *ein Wahrscheinlichkeitsraum,* $\mathbf{P}_{\mathcal{B}}$ *die Menge aller Borel-Wahrschein-*

lichkeitsmaße auf dem Borel-Raum $(\mathbb{R}, \mathcal{B})$ *und* $\mathbf{P}^{(\omega)} \in \mathbf{P}_{\mathcal{B}}$ *ein Borel-Wahrscheinlichkeitsmaß. Eine Abbildung*

$$\begin{aligned} \mathbb{M} : (\Omega, \mathcal{F}, \mathbf{P}) &\to \mathbf{P}_{\mathcal{B}} \\ \omega &\mapsto \mathbb{M}(\omega) = \mathbf{P}^{(\omega)} \end{aligned}$$

für die gilt, dass $\mathbb{M}(\cdot)(B) = \mathbf{P}^{(\cdot)}(B)$ *eine Zufallsvariable für jede Borelmenge* $B \subset \mathbb{R}$ *ist, heißt* **zufälliges Borel-Wahrscheinlichkeitsmaß**.

Definition 6 (erwartetes Wahrscheinlichkeitsmaß)
Sei $E(\cdot)$ *der Erwartungswert bezüglich* $\mathbf{P}$ *auf dem zugrundeliegenden Wahrscheinlichkeitsraum* $(\Omega, \mathcal{F}, \mathbf{P})$. *Das Wahrscheinlichkeitsmaß* $\mathbf{E}\mathbb{M}$, *für das gilt*

$$(\mathbf{E}\mathbb{M})(B) = E(\mathbb{M}(\cdot)(B)) \text{ für alle Borelmengen } B \subset \mathbb{R},$$

heißt **erwartetes Wahrscheinlichkeitsmaß** *des zufälligen Wahrscheinlichkeitsmaßes* $\mathbb{M}$.

Definition 7 (Folge von $\mathbb{M}$-zufälligen k-Stichproben)
Sei $\mathbb{M}$ *ein zufälliges Wahrscheinlichkeitsmaß,* $\mathbb{M}_1, \mathbb{M}_2, \mathbb{M}_3, \ldots$ *eine beliebige i.i.d. Folge von zufälligen Wahrscheinlichkeitsmaßen mit der gleichen Verteilung wie* $\mathbb{M}$, *also* $\mathbb{M}_i(\cdot) \overset{f.s.}{=} \mathbb{M}(\cdot)\ \forall i$, *und* $k \in \mathbb{N}^+$ *fest. Eine Folge von Zufallsvariablen* $X_1, X_2, \ldots$ *auf* $(\Omega, \mathcal{F}, \mathbf{P})$ *heißt* **Folge von $\mathbb{M}$-zufälligen k-Stichproben**, *wenn folgendes gilt:*

(i) Unter der Bedingung, dass für die j*-te Teilstichprobe* $\mathbb{M}_j(\omega) = \mathbf{P}_j^{(\omega)}$ *als Verteilung realisiert wurde, sind die Zufallsvariablen* $X_{(j-1)k+1}, \ldots, X_{jk}$ *i.i.d. mit Verteilungsfunktion* $\mathbf{P}_j^{(\omega)}$ *für alle* $j = 1, 2, \ldots$.

(ii) Die $X_{(j-1)k+1}, \ldots, X_{jk}$ *der* j*-ten Teilstichprobe sind für alle* $j = 1, 2, \ldots$, *alle* $l \neq j$ *und alle* $B \subset \mathbb{R}$ *von* $\{\mathbb{M}_l(\cdot)(B), X_{(l-1)k+1}, \ldots, X_{lk}\}$ *unabhängig.*

Definition 8 (skalenneutrale Mantissenhäufigkeit)
$\#\{\cdots\}$ *bezeichne die Anzahl von Elementen einer Menge* $\{\cdots\}$. *Eine Folge von Zufallsvariablen* $X_1, X_2, \ldots$ *hat eine* **skalenneutrale Mantissenhäufigkeit**,*wenn für alle* $\alpha > 0$ *und alle* $S \in \mathcal{M}$ *gilt*

$$\frac{|\#\{i : i \leq n \wedge X_i \in S\} - \#\{i : i \leq n \wedge X_i \in \alpha S\}|}{n} \longrightarrow 0 \text{ fast sicher.}$$

Definition 9 (basenneutrale Mantissenhäufigkeit)
Eine Folge von Zufallsvariablen $X_1, X_2, ...$ *hat eine* **basenneutrale Mantissenhäufigkeit**, *wenn für alle* $m \in \mathbb{N}$ *und* $S \in \mathcal{M}$ *gilt*

$$\frac{|\#\{i : i \leq n \wedge X_i \in S\} - \#\{i : i \leq n \wedge X_i \in S^{\frac{1}{m}}|}{n} \longrightarrow 0 \text{ fast sicher.}$$

Definition 10 (Skalenunverzerrtheit)
Ein zufälliges Wahrscheinlichkeitsmaß $\mathbb{M}$, *dessen erwartetes Wahrscheinlichkeitsmaß* $\mathbf{E}\mathbb{M}$ *skaleninvariant auf* $(\mathbb{R}^+, \mathcal{M}_b)$ *ist, für das also gilt, dass* $\mathbf{E}\mathbb{M}(S) = \mathbf{E}\mathbb{M}(\alpha S)$ *für* $\alpha > 0$ *und für alle* $S \in \mathcal{M}_b$, *heißt* **skalenunverzerrt**.

Definition 11 (Basenunverzerrtheit)
Ein zufälliges Wahrscheinlichkeitsmaß $\mathbb{M}$, *dessen erwartetes Wahrscheinlichkeitsmaß* $\mathbf{E}\mathbb{M}$ *baseninvariant auf* $(\mathbb{R}^+, \mathcal{M}_b)$ *ist, für das also gilt, dass* $\mathbf{E}\mathbb{M}(S) = \mathbf{E}\mathbb{M}(S^{1/n})$, *für alle* $n \in \mathbb{N}^+$ *und für alle* $S \in \mathcal{M}_b$, *heißt* **basenunverzerrt**.

Satz 6 (Grenzwertsatz für signifikante Ziffern)
Sei $\mathbb{M}$ *ein zufälliges Wahrscheinlichkeitsmaß auf* $(\mathbb{R}^+, \mathcal{M}_b)$ *und* $\hat{t} := \bigcup_{e=-\infty}^{\infty} [1, t) \cdot b^e \in \mathcal{M}_b$ *die Menge der positiven Zahlen mit Mantisse in* $[1, t)$. *Die folgenden Aussagen sind äquivalent:*

(i) $\mathbb{M}$ *ist skalenunverzerrt.*

(ii) $\mathbb{M}$ *ist basenunverzerrt und* $\mathbf{E}\mathbb{M}$ *ist atomlos.*

(iii) $E[\mathbb{M}(\cdot)(\hat{t})] = \log_b t$ *für alle* $t \in [1, b]$.

(iv) Jede $\mathbb{M}$*-zufällige k-Stichprobe hat eine skalenneutrale Mantissenhäufigkeit.*

(v) Jede $\mathbb{M}$*-zufällige k-Stichprobe hat eine basenneutrale Mantissenhäufigkeit und* $\mathbf{E}\mathbb{M}$ *ist atomlos.*

(vi) Für jede $\mathbb{M}$*-zufällige k-Stichprobe* $X_1, X_2, ...$ *gilt*

$$\frac{\#\{i \leq n : M_b(X_i) \in [1, t)\}}{n} \to \log_b t \text{ fast sicher für alle } t \in [1, b).$$