1.4 Zufallsvariablen und ihre Verteilung

1.4.1 Diskrete Zufallsvariablen

- Ein Zufallsexperiment wird beschrieben durch einen Grundraum Ω und eine Wahrscheinlichkeit P auf Ω.
- Häufig interessieren nicht die Ergebnisse an sich, sondern bestimmte abgeleitete Eigenschaften/Konsequenzen.

Bsp. 1.44. [Würfelwurf mit fairem Würfel; Spiel mit Auszahlungsregel]

Bem. 1.45.

- Gegeben seien ein diskreter, d.h. höchstens abzählbarer, Ergebnisraum Ω und die Wahrscheinlichkeit P auf Ω. "Jede" Abbildung

$$\begin{array}{rcl} X : \Omega & \mapsto & \Omega_X \\ \omega & \mapsto & X(\omega) \end{array}$$

heißt Zufalls*element*. Setzt man für jede *Realisation* $x \in \Omega_X$

$$P_X(\{x\}) := P(\{X = x\}) := P(\{\omega | X(\omega) = x\}),$$

so erhält man eine Wahrscheinlichkeit auf Ω_X. (Oft wird auch $P(X = x)$ statt $P(\{X = x\})$ geschrieben.)

- P_X heißt *Wahrscheinlichkeitsverteilung* von X.
- X (als Variable) beschreibt den Ausgang eines Zufallsexperiments *vor der Durchführung* (Auszahlungs*regel* beim Würfelspiel: wenn 3 dann 10 Euro, wenn . . . , dann . . .).

- x (als Realisation) gibt den Wert der Variablen nach Durchführung des Zufallsexperiments an (daher „Realisation“, konkreter Auszahlungs$betrag$).
- Weiteres Beispiel:
  * X Größe der nächsten eintretenden Person (als Messvorschrift)
  * x Wert, z.B. 167 cm
- In der Verwendung analog zur Unterscheidung Merkmal / Merkmalsausprägung in Statistik I (siehe später).
- Es ist häufig üblich, bei P_X den Index wegzulassen, also $P(\{x\})$ statt $P_X(\{x\})$ zu schreiben.
- Ist $\Omega_X \subset \mathbb{R}$, so bezeichnet man das Zufallselement X als $Zufallsvariable$. In der Literatur wird der Begriff Zufalls$element$ relativ selten verwendet, gerade aber in den Sozialwissenschaften sind oft nicht reelle Zahlen im Sinne einer metrischen Skala gegeben: Zufallselemente umfassen auch nominal- und ordinalskalierte Merkmale.

Definition 1.46.

Gegeben sei eine diskrete Zufallsvariable X mit der Wahrscheinlichkeitsverteilung P. Die Menge

$$\mathcal{X} := \{x \in \mathbb{R} | P(\{x\}) > 0\}$$

heißt $Träger\ von\ X$.

Bem. 1.47.

- Die *Wahrscheinlichkeitsfunktion* $f(x)$ einer diskreten Zufallsvariable X ist für $x \in \mathbb{R}$ definiert durch

$$f(x) := \begin{cases} P(X = x_i) = p_i, & x = x_i \in \mathcal{X} = \{x_1, x_2, \ldots, x_k, \ldots\} \\ 0, & \text{sonst.} \end{cases}$$

- Man kann $\{X = x_i\}$ mit $x_i \in \mathcal{X}$ als Elementarereignisse sehen. Durch die Wahrscheinlichkeitsfunktion ist also die ganze Wahrscheinlichkeitsverteilung eindeutig bestimmt. Für beliebige Mengen A erhält man:

$$P(X \in A) = \sum_{x_i \in A \cap \mathcal{X}} P(X = x_i) = \sum_{x_i \in A \cap \mathcal{X}} f(x_i)$$

Bsp. 1.48. [Zum Rechnen mit Zufallsvariablen]

Sei X die Zufallsvariable $Anzahl\ der\ Haushaltsmitglieder$ mit der Verteilung

$$P(\{X=1\}) = 0.4 = f(1)$$
$$P(\{X=2\}) = 0.3 = f(2)$$
$$P(\{X=3\}) = 0.2 = f(3)$$
$$P(\{X=4\}) = 0.1 = f(4)$$

(Annahme: Nur bis zu 4-Personen-Haushalte). Man berechne die Wahrscheinlichkeit, bei reiner Zufallsauswahl vom Umfang 1 einen Mehrpersonenhaushalt zu erhalten und die Wahrscheinlichkeit des Ereignisses „Die Zahl der Haushaltsmitglieder ist gerade“.

$$\begin{aligned} P(\{X>1\}) &= P(\{X=2\}) + P(\{X=3\}) + P(\{X=4\}) \\ &= f(2) + f(3) + f(4) \\ &= 0.3 + 0.2 + 0.1 \\ &= 0.6 \end{aligned}$$

alternativ:

$$\begin{aligned} P(\{X > 1\}) &= 1 - P(\{X \leq 1\}) \\ &= 1 - P(\{X = 1\}) \\ &= 0.6 \end{aligned}$$

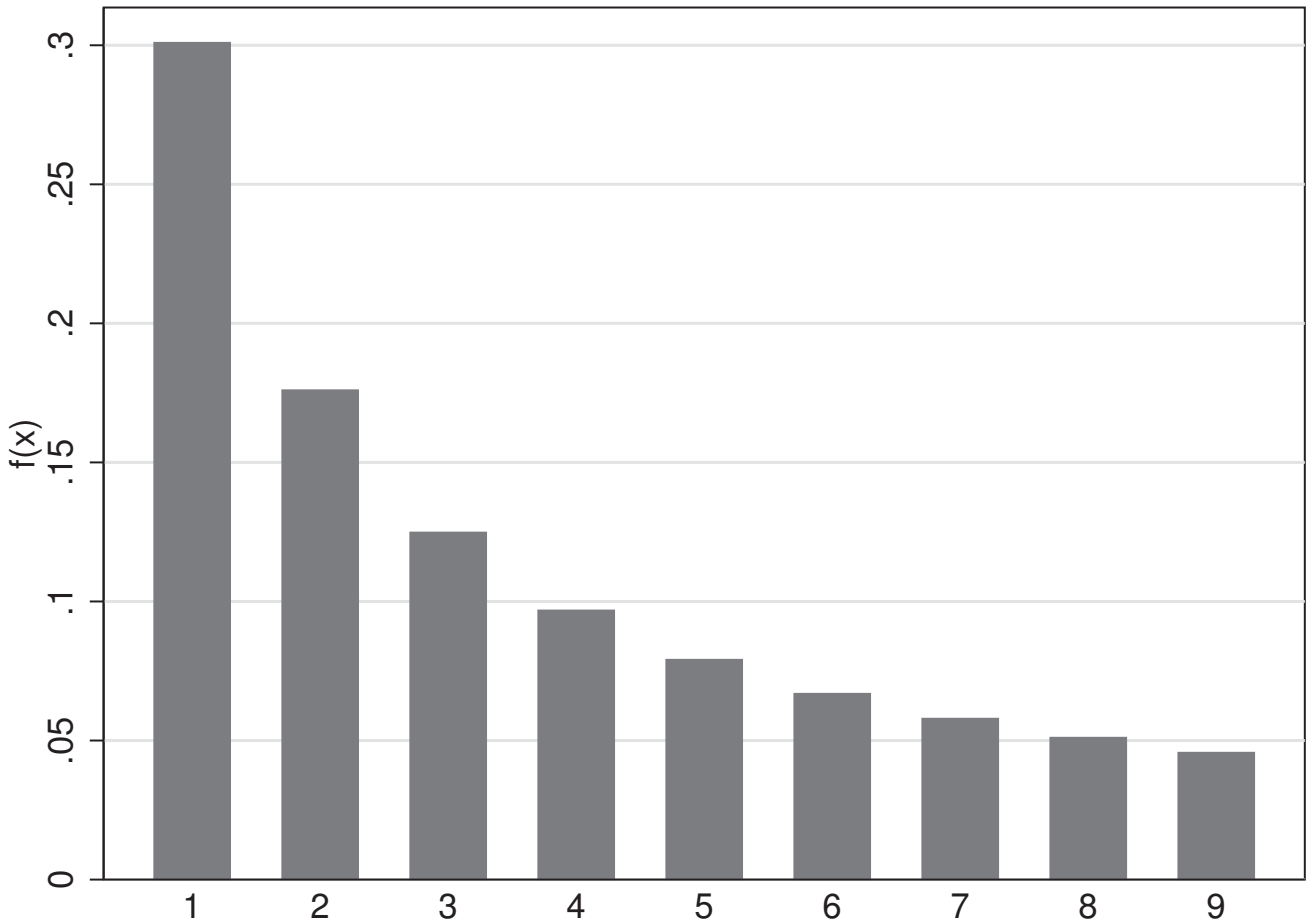
$$\begin{aligned} P(\{X = 2\} \cup \{X = 4\}) &\overset{\text{disjunkt}}{=} P(\{X = 2\}) + P(\{X = 4\}) \\ &= f(2) + f(4) \\ &= 0.3 + 0.1 \\ &= 0.4 \end{aligned}$$

Bsp. 1.49. [Benfords Gesetz]

Simon Newcomb (1835–1909) und später Frank Benford (1883–1948) machten die (zunächst) verblüffende Entdeckung, dass die Anfangsziffern 1–9 von ganzen Zahlen in vielen Fällen nicht gleich häufig vorkommen. Am häufigsten ist die Anfangsziffer 1, am zweithäufigsten die Anfangsziffer 2 usw.

Beispiele sind

- die Häufigkeit der Anfangsziffern von Zahlen in Zeitungsartikeln
- die Häufigkeit der Anfangsziffern von in Steuererklärungen angebenen Beträgen
- die Häufigkeit der ersten Ziffer der Dateigröße von gespeicherten Dateien.
- in regionalen, nicht bevölkerungsproportional organisierten Stimmkreisen die Anzahl der abgegebenen Stimmen bei einer Wahl


f(x)
0
.05
.1
.15
.2
.25
.3
1
2
3
4
5
6
7
8
9

Benford postulierte für die Zufallsvariable

$$X = \text{„Anfangsziffer von Zahlen“}$$

die Wahrscheinlichkeitsfunktion

$$f(x) = P(X = x) = \begin{cases} \log_{10}\left(\dfrac{x+1}{x}\right), & x = 1, \ldots, 9 \\ 0, & \text{sonst} \end{cases}$$

Bem. 1.50. [Induktive Brücke II] *:*

Gegeben sei eine Grundgesamtheit $\mathcal{G}$ (z.B. alle Wähler(innen)). Wir betrachten eine reine Zufallsauswahl mit Ergebnisraum

$$\mathcal{S} = \mathcal{G} \times \mathcal{G} \times \ldots \times \mathcal{G}$$

mit Ergebnissen $s = (s_1, s_2, \ldots, s_n)$, wobei s_i der beim i-ten Zug gezogenen Einheit, also z.B. dem i-ten gezogenen Wähler entspricht.

Betrachtet man nun ein Merkmal

$$X : \mathcal{G} \longrightarrow \{a_1, \ldots a_k\}$$

und die Ereignisse $A_{ij} =$„i-te gezogene Person hat Merkmalsausprägung a_j“.
Stehen z.B. a_1, a_2, . . . für bestimmte Parteien, so gibt $X : \mathcal{G} \longrightarrow \{$CDU/CSU, SPD, . . . $\}$ für jede(n) Wähler(in) g$\in\mathcal{G}$ ihre/seine Wahlentscheidung X(g) an. Die Ereignisse A_{ij} sind nun durch Zufallselemente beschreibbar:

Sei X_i die „Auswertung des Merkmals X an der i-ten zufällig ausgewählten Person“, d.h. an s_i, so ist X_i ein Zufallselement

$$\begin{aligned} X_i \colon \mathcal{G} &\longrightarrow \Omega_X = \{a_1, \ldots, a_k\} \\ \omega &\longmapsto X(\omega_i) \end{aligned}$$

Das Ereignis A_{ij} lässt sich dann schreiben als

$$\{X_i = a_j\}\,.$$

Es gilt also für jedes i (z.B. Nummer des Wählenden in der Stichprobe) und j (z.B. Name der Partei)

$$P_{X_i}(\{a_j\}) = P(\{X_i = a_j\}) = P(A_{ij}) = f_j.$$

Die Wahrscheinlichkeitsverteilung des Zufallselements X_i (Stichprobe!) spiegelt also genau die Häufigkeitsverteilung des Merkmals X (Grundgesamtheit!) wider.

Fasst man man die einzelnen X_i zusammen, so bezeichnet man den Vektor $(X_1, X_2, \ldots, X_n)$ als $i.i.d.\ Stichprobe$ oder $reine\ Zufallsstichprobe$ des Merkmals $\widetilde{X}$. Die Abkürzung $i.i.d.$ steht für

- independently (die einzelnen Ziehungen sind stochastisch unabhängig)

- identically distributed (jedes X_i besitzt dieselbe Wahrscheinlichkeitsverteilung)

Nach dem Durchführen des Zufallsexperiments und der Auswertung von X erhält man die Realisationen $x_1 := X_1(s_1), x_2 := X_2(s_2), \ldots, x_n := X_n(s_n)$, also einen Vektor $(x_1, x_2, \ldots, x_n)$, der formal korrekt als Realisation oder $Stichprobenrealisation$ der $i.i.d.$ Stichprobe $(X_1, X_2, \ldots, X_n)$ bezeichnet werden würde, allgemein üblich aber einfach auch als $Stichprobe$ bezeichnet wird.

Man nimmt diese Stichprobe als Realisation der Stichprobe $X_1, \ldots, X_n$ und versucht jetzt auf die Grundgesamtheit, genauer auf die $f_1, \ldots, f_n$, zu schließen.

Koppelt man die einzelnen Zufallsexperimente, so kann man die sogenannte $gemeinsame\ Verteilung$ der $X_1, X_2, \ldots, X_n$ berechnen.

$$\begin{aligned} & P(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ldots \cap \{X_n = x_n\}) \\ = \quad & P(\{X_1 = x_1\}) \cdot P(\{X_2 = x_2\}) \cdot \ldots \cdot P(\{X_n = x_n\}) \end{aligned}$$

und damit für jede potentielle Stichprobe(nrealisation) die Wahrscheinlichkeit, genau sie zu erhalten. In ihr finden sich wieder die Häufigkeitsverhältnisse der Grundgesamtheit wieder ("Repräsentativität").