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. 1.40. [Würfelwurf mit fairem Würfel]

Bem. 1.41.

- 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 $\Omega_X = \mathbb{R}$, 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 1.42.

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*.

Definition 1.43.

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

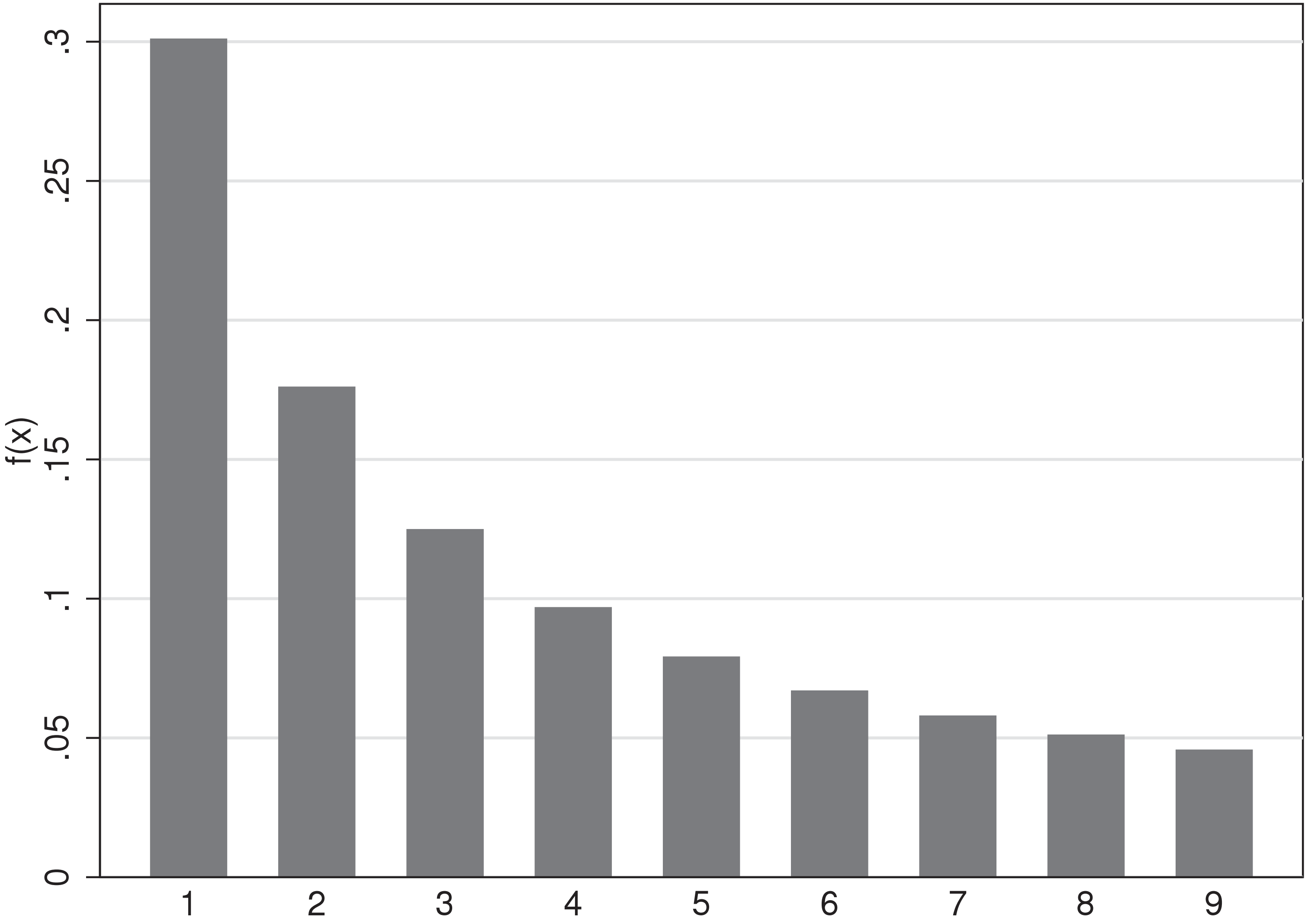
$$f(x) := \begin{cases} P(X = x_i) = p_i, & x = x_i \in \mathcal{X} = \{x_1, x_2, \ldots, x_k, \ldots\} \\ 0, & \text{sonst.} \end{cases}$$

Bsp. 1.44. [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


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

Benford publizierte für die Zufallsvariable

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

die Wahrscheinlichkeitsfunktion

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

Benfords Gesetz findet zum Beispiel Anwendung bei der Fahndung nach Steuerbetrügern, bei der Überprüfung von Wahlergebnissen oder bei der Optimierung von Computerfestplatten.

Bsp. 1.45. [Zum Rechenen mit Zufallsvariablen]

Sei X die Zufallsvariable *Anzahl der Haushaltsmitglieder* mit der Verteilung

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

(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“.

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

alternativ:

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

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

Bem. 1.46. [Induktive Brücke II]

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

$$\Omega = \widetilde{\Omega} \times \widetilde{\Omega} \times \ldots \times \widetilde{\Omega}$$

mit Ergebnissen $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$, wobei ω_i der beim i-ten Zug gezogenen Einheit, also z.B. dem i-ten gezogenen Wähler entspricht.

Dann bezeichnet das Merkmal

$$\widetilde{X} : \widetilde{\Omega} \longrightarrow \{\text{SPD, CDU/CSU, } \ldots\}$$

die individuelle Wahlentscheidung jedes Wählers $\tilde{\omega} \in \widetilde{\Omega}$, $\widetilde{X}(\tilde{\omega})$ von $\tilde{\omega}$ gewählte Partei.

Betrachtet werden die Ereignisse $A_{ij} =$„i-te gezogene Person hat Merkmalsausprägung a_j“. Diese sind nun durch Zufallselemente beschreibbar:

Sei X_i die „Auswertung des Merkmals $\widetilde{X}$ an der i-ten zufällig ausgewählten Person“, d.h. an ω_i, so ist X_i ein Zufallselement

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

Das Ereignis A_{ij} lässt sich dann schreiben als

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

Es gilt also für jedes i (z.B. * des Wählers) 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 $\widetilde{X}$ (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 $\widetilde{X}$ erhält man die Realisationen $x_1 := X_1(\omega_1), x_2 := X_2(\omega_2), \ldots, x_n := X_n(\omega_n)$, 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.

1.4.2 Verteilungsfunktion

Jetzt Zufalls<u>variablen</u> betrachten, also reellwertige Realisationen.
Viele interessierende Ereignisse besitzen folgende Form:

$$\{X \leq a\} \quad \text{oder} \quad \{X \epsilon [a, b]\} = \{a \leq X \leq b\},$$

wobei a und b feste reelle Zahlen sind.

$P(\{X \leq a\})$ für variables a entspricht der empirischen Verteilungsfunktion. In der Tat definiert man:

Definition 1.47.

Sei X eine diskrete Zufallsvariable. Die Funktion

$$\begin{aligned} F : \mathbb{R} &\rightarrow [0;1] \\ x &\mapsto F(x) \end{aligned}$$

$$F(x) := P(X \leq x)$$

heißt *Verteilungsfunktion* von x.

Satz 1.48.

Die Wahrscheinlichkeitsverteilung einer (diskreten) Zufallsvariablen X kann man durch die Verteilungsfunktion eineindeutig erklären.

Die Wahrscheinlichkeit anderer Ereignisse ergibt sich aus dem dritten Kolmogorovschen Axiom. Es gilt zum Beispiel

$$P(a < X \leq b) = F(b) - F(a).$$

Die Ereignisse $\{X \leq a\} = \{\omega | X(\omega) \leq a\}$, $\{a < X \leq b\}$ und $\{X > b\}$ sind disjunkt und ergeben in ihrer Vereinigung Ω. Also gilt

$$\begin{aligned} 1 &= P(\Omega) = P(X \leq a) + P(a < X \leq b) + P(X > b) \\ &\Leftrightarrow \quad 1 - P(X \leq a) - P(X > b) = P(a < X \leq b) \\ &\Leftrightarrow \quad P(X \leq b) - P(X \leq a) = P(a < X \leq b) \end{aligned}$$

Bsp. 1.49. [Fortsetzung von Bsp. 1.45]

Berechne die Verteilungsfunktion und zeichne sie.

Allgemein gilt: $F(x)$ ist eine stückweise konstante Treppenfunktion und $P(X = x)$ ist genau die Sprunghöhe der Verteilungsfunktion im Punkt x.

Bsp. 1.50. [Fortsetzung von Bsp. 1.45]

1.4.3 Stetige Zufallsvariablen

Im Folgenden nur die Grundidee, keine „mathematisch saubere“ Herleitung.

- Eine stetige Zufallsvariable
$$X : \Omega \longrightarrow \Omega_X = \mathbb{R}$$
besitzt überabzählbaren Ergebnisraum Ω_X, d.h. jeder Wert innerhalb eines Intervalls $[a, b]$ ist ein mögliches Ergebnis.

- Vorstellung: Auswertung eines $stetigen$ Merkmals $\tilde{X}$ an zufällig ausgewählter Person aus einer Grundgesamtheit.

- Problem: Den einzelnen Ereignissen kann keine positive Wahrscheinlichkeit mehr zugeordnet werden: „Unendlich oft“ positive Werte aufsummieren ergibt immer eine Zahl $> 1 = P(\Omega)$

Idee: Versuche eine stetige Gleichverteilung auf dem Intervall $[0, 1]$ zu konstruieren. Dazu geben wir uns ein diskretes, gleichabständiges Gitter bestehend aus n Werten aus $[0, 1]$ vor, also die Werte

$$\mathcal{X} = \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\right\}.$$

Die diskrete Gleichverteilung auf diesem Gitter ergibt die Wahrscheinlichkeitsfunktion

$$P(X = x) = \frac{1}{n+1}, \quad x \in \mathcal{X}.$$

$\Rightarrow$ Wahrscheinlichkeit für jeden einzelnen Wert geht gegen Null für $n \to \infty$. Tatsächlich gilt für jede stetige Zufallsvariable

$$P_X(\{x\}) = 0 \quad \text{für jedes } x \in \mathbb{R}.$$

Andererseits bleibt die Wahrscheinlichkeit

$$P(X \leq 0.5) = 0.5$$

im Wesentlichen konstant, unabhängig vom Feinheitsgrad des Gitters. Allgemeiner gilt hier (im Grenzfall) sogar

$$P(X \leq x) = x, \quad \text{für alle } x \in [0, 1]$$

$\Rightarrow$ Verteilungsfunktion betrachten. Im Gegensatz zu diskreten Zufallsvariablen ist die Verteilungsfunktion jetzt allerdings stetig.

Bem. 1.51.

Die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen ist (nicht mehr durch die Wahrscheinlichkeitsbewertung der Elementarereignisse, sondern nur mehr) durch die Verteilungsfunktion

$$F(x) = P(X \leq x)$$

eindeutig festgelegt.

Allgemeiner als zuvor gilt hier

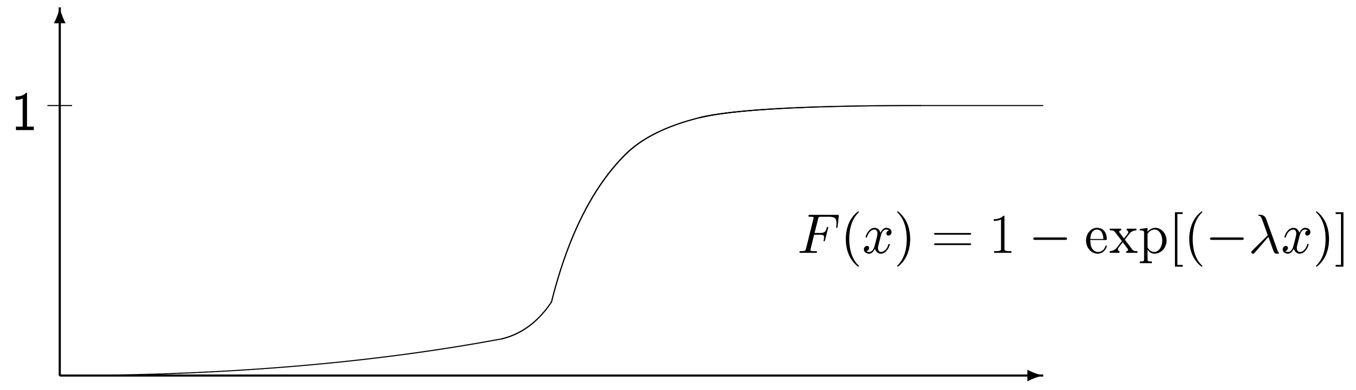
$$\begin{aligned} P(a < X \leq b) &= P(a \leq X \leq b) \\ &= P(a < X < b) = F(b) - F(a) \end{aligned}$$

da $P(X = a) = P(X = b) = 0$.

Bem. 1.52.

- Stetige Zufallsvariablen sind in der Wahrscheinlichkeitsrechnung sehr wichtig. Später wird fast ausschließlich mit stetigen Zufallsvariablen gerechnet.
- Insbesondere ergeben sich Approximationsmöglichkeiten für diskrete durch stetige Zufallsvariablen bei größeren Stichprobenumfängen. Damit lassen sich zahlreiche Berechnungen vereinfachen (auch wenn die stetige Formulierung zunächst komplizierter wirkt).

Typische Verteilungsfunktion: (z.B. zur Beschreibung der Dauer von Arbeitslosigkeit)



Die Kurve ist unterschiedlich steil. Sie hat zwar in keinem Punkt eine Sprungstelle ($P(X = x) = 0$), aber in jedem kleinen Intervall um x ist:

$$P(x - h < X < x + h) = F(x + h) - F(x - h)$$

durchaus unterschiedlich.

Die Steigung

$$\lim_{h \to 0} \frac{F(x+h) - F(x-h)}{h}$$

enthält also wesentliche Information über P. Dies führt zu folgender Definition:

Definition 1.53.

Gegeben sei eine stetige Zufallsvariable X mit differenzierbarer Verteilungsfunktion $F_X(x)$. Dann heißt die Ableitung von $F(x)$ nach x, also

$$f(x) = \frac{dF(x)}{dx}$$

$Dichte$ der Zufallsvariablen X.

Bem. 1.54.

Man kann die Definition allgemein auch anwenden, wenn die Funktion nur „stückweise differenzierbar“ ist.

Umgekehrt erhält man aus der Dichte die Verteilungsfunktion durch Integration:

Satz 1.55.

In der Situation der obigen Definition gilt

$$F(x) = \int_{-\infty}^{x} f(u)\, du$$

und damit für beliebige reelle Zahlen a und b mit $a < b$

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X \leq b) = P(a < X < b) \\ &= P(a \leq X < b) \\ &= \int\limits_a^b f(x)\, dx. \end{aligned}$$

Jede Funktion f auf $\mathbb{R}$ mit $f(x) \geq 0$ für alle x und

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

kann als Dichte verwendet werden.

Alternative Definition stetiger Zufallsvariblen: Eine Zufallsvariable X heißt stetig, wenn es eine Funktion $f(x) \geq 0$ gibt, so dass für jedes Intervall $[a, b]$

$$P(a \leq X \leq b) = \int\limits_a^b f(x)dx \quad (\widehat{=} \text{ Fläche zwischen } a \text{ und } b \text{ unter der Funktion})$$

gilt.

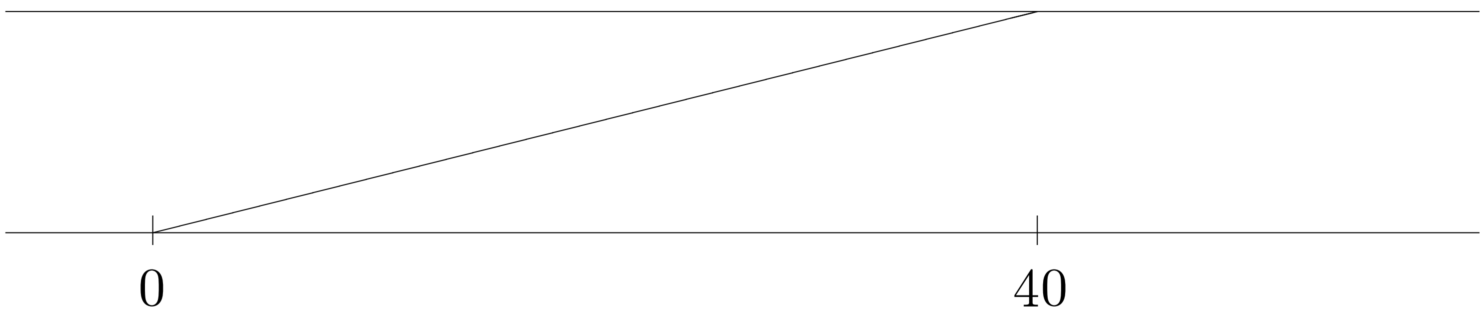
Bsp. 1.56.

Gegeben sei eine Zufallsvariable X mit Verteilungsfunktion

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{40} \cdot x & x \in [0, 40] \\ 1 & x > 40 \end{cases}$$

Bestimmen Sie die Dichte $f(x)$ von X, skizzieren Sie $f(x)$ und interpretieren Sie $f(x)$ anschaulich!

Skizze zu $F(x)$:



Bei der Modellbildung geht man auch häufig umgekehrt vor: Gegeben ist eine Dichte, die die Verteilungsfunktion eindeutig bestimmt.

Man erhält die Verteilungsfunktion durch

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

und das Wahrscheinlichkeitsmaß P über

$$P(a < X \leq b) = \int_{a}^{b} f(x)dx.$$

Bsp. 1.57.

Gegeben sei die Funktion

$$f_c(x) \quad = \quad \begin{cases} c \cdot x & x \in [0,1] \\ 0 & \text{sonst,} \end{cases}$$

a) Wie ist c zu wählen, dass f_c eine Dichte ist?

b) Berechnen Sie die Verteilungsfunktion und $P(X \in [\frac{1}{4}, \frac{3}{4}])$!