Versuchsreihen. Relative Häufigkeit

Versuchsreihen

Ein Versuch zu einer Zvn. X werde s-mal durchgeführt (Versuchsreihe). Wir notieren jedes auftretende Merkmal in einer Liste S (Urliste, Protokoll, **Stichprobe**):

$$S = (x_1, x_2, \ldots, x_s)$$

Die natürliche Zahl s heißt **Länge** der Versuchsreihe (auch: **Umfang** der Stichprobe S); sie kann stark variieren.

N.B.: In einer Stichprobe S müssen keinesfalls alle Elemente des zugehörigen Merkmalraums Ω vorkommen. Andererseits können Merkmale in S mehrfach auftreten.

Relative Häufigkeit

■ Relative Häufigkeit von Merkmalen

Sei $x \in \Omega$ irgendein Merkmal, S die Urliste einer Versuchsreihe. Wir durchmustern die in S auftretenden Merkmale der Reihe nach und prüfen, ob sie mit x übereinstimmen. Gibt es a Übereinstimmungen, so heißt a **absolute Häufigkeit** und der Quotient $\frac{a}{s}$ **relative Häufigkeit des Merkmals** x **in** S (bzw. in der Versuchsreihe). Die relative Häufigkeit eines Merkmals gibt den Anteil seines Vorkommens innerhalb einer Stichprobe an.

$$\text{Relative Häufigkeit von Merkmal } x = \frac{\text{Anzahl der Vorkommen von } x \text{ in der Stichprobe}}{\text{Umfang der Stichprobe}}$$

Um diesen Ausdruck formal zu präzisieren, setzen wir $\delta(x, y) = 1$, falls $x = y$, und sonst: $\delta(x, y) = 0$. Damit lässt sich die relative Häufigkeit $H_S(x)$ von Merkmal x in S wie folgt schreiben:

$$H_S(x) = \frac{1}{s} \sum_{i=1}^{s} \delta(x, x_i)$$

Die Bezugnahme auf S (insbesondere die Indizierung von H) kann fortgelassen werden, wenn sie aus dem Zusammenhang hervorgeht.

Offenkundig sind folgende Eigenschaften von H_S:

1. $0 \leq H_S(x) \leq 1$ für alle $x \in \Omega$
2. $\sum_{x \in \Omega} H_S(x) = 1$

Ein Merkmal x kommt in der Stichprobe S genau dann nicht vor, wenn seine relative Häufigkeit 0 beträgt. Relative Häufigkeit 1 bedeutet, dass x das einzige in S auftretende Merkmal ist. In der Summe (unter 2.) addieren sich die Anteile aller tatsächlich in S vorkommenden Merkmale zu 1 auf.

■ Relative Häufigkeit von Ereignissen

Es ist naheliegend, den Begriff der relativen Häufigkeit auf Ereignisse auszudehnen. Dies lässt sich auf natürliche Weise durchführen:

Die relative Häufigkeit eines Elementarereignisses $X = x$ (dessen Merkmalmenge aus x allein besteht: $\{x\}$) identifizieren wir direkt mit $H_S(x)$. Setzt sich das Ereignis aus mehreren Merkmalen $x_1, \ldots, x_m$ zusammen, d.h. besitzt es die Merkmalmenge bzw. Extension $A = \{x_1, \ldots, x_m\}$, so wird definiert:

$$H_S(A) = \sum_{i=1}^{m} H_S(x_i)$$

Das hier verwendete Definitionsverfahren heißt **Summenregel**. Es garantiert, dass die relative Häufigkeit eines Ereignisses in einer Stichprobe S gerade der Anteil ist, den sämtliche Merkmale bilden, deren Auftreten das betreffende Ereignis nach sich zieht.

Wo notationstechnisch bequem, wollen wir Ereignisse durch ihre Merkmalmengen $A, B, C, \ldots$ ($\subseteq \Omega$) wiedergeben.

Es folgt eine Zusammenstellung einfacher Grundtatsachen (wobei eine feste Stichprobe S zu Grunde liegt, auf die nicht jedesmal Bezug genommen wird):

1. $H(\emptyset) = 0$, $H(\Omega) = 1$
2. $0 \leq H(A) \leq 1$ für alle Ereignisse $A \subseteq \Omega$
3. $A \subseteq B \Longrightarrow H(A) \leq H(B)$
4. $H(A \cup B) + H(A \cap B) = H(A) + H(B)$

Beweis:

1.) und 2.) sind sofort ersichtlich, 3.) eine unmittelbare Konsequenz der Summenregel. 4.) ergibt sich daraus, dass die Merkmale im Durchschnitt von A und B links und rechts vom Gleichheitszeichen jeweils doppelt gezählt werden. ■

Die Eigenschaft 3 heißt Monotonie. Für disjunkte (unverträgliche) Ereignisse A, B geht Eigenschaft 4 über in: $H(A \cup B) = H(A) + H(B)$ (Additivität).

■ Erweiterung einer Stichprobe

In der Praxis kommt es häufig vor, dass eine vorliegende Stichprobe $S = (x_1, x_2 \ldots, x_s)$ durch eine zusätzliche Anzahl k von Versuchen erweitert wird: $S^* = (x_1, \ldots, x_s, x_{s+1}, \ldots x_{s+k})$.

Wir wollen untersuchen, wie sich die relative Häufigkeit eines Merkmals bei einer solchen Erweiterung ändert.

Offenbar gilt für die absolute Häufigkeit a von x in der erweiterten Stichprobe S^*:

$$a = s \cdot H_S(x) + \delta(x, x_{s+1}) + \ldots + \delta(x, x_{s+k})$$

Die relative Häufigkeit von x ergibt sich nach Division der absoluten Häufigkeit durch den Umfang $(s + k)$ der erweiterten Stichprobe S^*:

$$H_{S^*}(x) = \frac{a}{s+k} = \frac{s}{s+k} H_S(x) + \frac{1}{s+k} \sum_{i=1}^{k} \delta(x, x_{s+i})$$

Hieraus folgt für die Häufigkeitendifferenz in den Stichproben S und S^*:

$$H_{S^*}(x) - H_S(x) = \frac{k}{s+k} \left(-H_S(x) + \frac{1}{k} \sum_{i=1}^{k} \delta(x, x_{s+i}) \right)$$

Der zweite Summand in der Klammer ist die relative Häufigkeit von x in der Reststichprobe $(x_{s+1}, \ldots, x_{s+k})$. Da sie ≥ 0 und ≤ 1 ist, lässt sich der Klammerausdruck dem Betrage nach mit dem größeren der beiden Werte $H_S(x)$ und $1 - H_S(x)$ abschätzen.

Damit ergibt sich die Ungleichung:

$$\left| H_{S^*}(x) - H_S(x) \right| \leq \frac{k}{s+k} \max \{H_S(x), 1 - H_S(x)\}$$

Folgerung: Der Absolutbetrag der Änderung, welche die relative Häufigkeit eines Merkmals erleidet, wenn man die zu Grunde liegende Stichprobe vom Umfang s um k Elemente erweitert, ist durch den Bruch $\frac{k}{s+k}$ beschränkt. Bei festem k strebt dieser Betrag daher gegen Null für $s \longrightarrow \infty$, d.h. in Abschnitten konstanter Größe werden die Schwankungen der relativen Häufigkeit beliebig klein.

Warnung: Dies ist *nicht* gleichbedeutend damit, dass die Häufigkeitenfolge konvergiert!

Die folgende Grafik zeigt (bei einer größeren Stichprobe) deutlich den Verlauf der Schranke (rechte Seite der obigen Ungleichung) oberhalb der Punkte, welche die Differenzbeträge der relativen Häufigkeiten (linke Seite der Ungleichung) darstellen: