

Erfassen, Aufbereiten und Darstellen statistischer Daten - Grundwissen



Als **Deskriptive**¹ oder **Beschreibende Statistik**² bezeichnet man das Teilgebiet der Mathematik, das sich damit beschäftigt, wie die in einer **statistischen Erhebung** gewonnenen vielen und daher meist unübersichtlichen **Daten**³ übersichtlich dargestellt, zusammengefasst und ausgewertet werden. Hierdurch soll Typisches oder Wesentliches deutlich werden - Typisches oder Wesentliches für die Vielheit, auf die sich die Daten beziehen, und nicht für die einzelne Einheit.

Als **Grundgesamtheit** bezeichnet man die Menge aller in der Erhebung erfassten Objekte, die 'Vielheit'.

Als **Merkmalsträger** oder **statistische Einheit** bezeichnet man einzelne Elemente dieser Menge, d.h. eines der erfassten Objekte, eine 'Einheit'.

Als **Erhebungsumfang** bezeichnet man die Größe der Grundgesamtheit, d.h. die Anzahl der in der Erhebung erfassten Merkmalsträger. Als Variable für den Erhebungsumfang nimmt man häufig den Buchstaben **n**.

Als **Merkmale** bezeichnet man *Eigenschaften* der erfassten Merkmalsträger, die bei der Erhebung von Interesse sind. Ist bei einer Erhebung nur ein Merkmal von Interesse, so spricht man von einer **univariaten Erhebung**, als Variable für dieses eine Merkmal nimmt man häufig den Buchstaben X. Sind dagegen zwei Merkmale von Interesse, so spricht man von einer **bivariaten Erhebung**, als Variablen für die beiden Merkmale nimmt man häufig dann die beiden Buchstaben X und Y.

Lässt sich ein Merkmal nicht durch Zahlen erfassen, so spricht man von einem **qualitativen Merkmal**. Lässt sich ein Merkmal dagegen numerisch, d.h. durch Zahlenangaben erfassen, spricht man von einem **quantitativen Merkmal**.

Als **Merkmalsausprägungen** (eines Merkmals) bezeichnet man die bei qualitativen Merkmalen die verschiedenen *Kategorien* oder bei quantitativen Merkmalen die verschiedenen *Zahlenwerte*, die einem Merkmalsträger hinsichtlich des Merkmals als Werte zugeordnet werden können.

Alle diese verschiedenen Kategorien oder Zahlenwerte bilden die Menge aller Merkmalsausprägungen des Merkmals X. Steht die Variable **m** für die Anzahl der verschiedenen Merkmalsausprägungen, so ergibt sich eine Menge **{a₁; ...; a_m}**; diese Menge kann aber auch – bei quantitativen Merkmalen – ein Zahlenintervall, z.B. $[a_{\min}; a_{\max}]$, sein, so dass dann die Anzahl der Merkmalsausprägungen unendlich ist.

Beachte:

In der Menge der Merkmalsausprägungen gibt es keine Wiederholungen, jede mögliche Ausprägung wird in der *Menge* der Merkmalsausprägungen natürlich nur genau ein Mal genannt.

¹ desribere: lat. beschreiben, ordnen

² status: lat. Stand, Zustand, Lage, Verfassung

³ dare, datum: lat. geben, gegeben

Als **Messwerte** oder **Merkmalswerte** (eines Merkmalsträgers hinsichtlich eines Merkmals) bezeichnet man die Kategorien bzw. die Zahlenwerte, den man bei einer Erhebung bei den verschiedenen Merkmalsträgern hinsichtlich eines Merkmal tatsächlich beobachtet. Es ergibt sich so eine *Folge* oder *Liste* von Messwerten $(x_1; \dots ; x_n)$, die man als **Urliste** oder **Datenmenge** bezeichnet.

Beachte:

- Da es n verschiedene Merkmalsträger gibt, muss es auch n Messwerte geben.
- In der Urliste gibt es – im Gegensatz zur Menge der Merkmalsausprägungen – normalerweise Wiederholungen, da eine Merkmalsausprägung ja durchaus mehrmals als Messwert bei unterschiedlichen Merkmalsträgern beobachtet werden kann und somit in der *Liste* der Messwerte auch mehrmals aufgeführt werden muss.

Als **Absolute Häufigkeit $H(a_i)$** (einer Merkmalsausprägung a_i , i nummeriert die verschiedenen Merkmalsausprägungen) bezeichnet man die Anzahl der Messwerte bzw. Merkmalsträger mit der entsprechenden Merkmalsausprägung.

Beachte:

- Es gibt immer genau so viele absolute Häufigkeiten wie Merkmalsausprägungen, d.h. die Anzahl der absoluten Häufigkeiten ist immer gleich der Anzahl der Merkmalsausprägungen.
- Die absolute Häufigkeit einer Merkmalsausprägung kann den Wert 0 haben.
- Es gilt selbstverständlich: $H(a_1) + \dots + H(a_m) = n$ (Erhebungsumfang)

Als **Relative Häufigkeit $h(a_i)$** (einer Merkmalsausprägung a_i , i nummeriert die verschiedenen Ausprägungen) bezeichnet man den relativen Anteil der Messwerte bzw. Merkmalsträger mit der entsprechenden Merkmalsausprägung an der Grundgesamtheit, d.h.

$$h(a_i) = \frac{H(a_i)}{n}$$

Die relative Häufigkeit einer Merkmalsausprägung berechnet sich also als Quotient aus der absoluten Häufigkeit der Merkmalsausprägung und dem Erhebungsumfang, d.h.

$$\text{relative Häufigkeit einer Merkmalsausprägung} = \frac{\text{absolute Häufigkeit einer Merkmalsausprägung}}{\text{Erhebungsumfang}}$$

Beachte:

- Relative Häufigkeiten werden oft als Bruch, Dezimalbruch oder in Prozent angegeben.
- Es gibt immer genau so viele relative Häufigkeiten wie absolute Häufigkeiten, d.h. die Anzahl der relativen Häufigkeiten ist ebenfalls immer gleich der Anzahl der Merkmalsausprägungen.
- Es gilt selbstverständlich: $h(a_1) + \dots + h(a_m) = 1$.