

Statistik

CH.2 - Begriff der Häufigkeit

2025 | Prof. Dr. Buchwitz, Sommer, Henke

Wirgeben Impulse

Lernziele

- Kennenlernen der zentralen Grundbegriffe bei statistischen Untersuchungen.
- Definition und Einordnung von Beobachtungsgrößen anhand des Skalenniveaus.
- Diskussion und Definition absoluter und relativer Häufigkeiten.
- Einführung in elementare grafische Darstellungen.
- Schaffen von Verständnis für die empirische Verteilungsfunktion.

Inhaltsübersicht

- 1 Grundbegriffe
- 2 Häufigkeiten

Grundlegende Begriffe

- Statistische Einheit: Träger der Information (Merkmalsträger, Untersuchungsobjekt)
- **Grundgesamtheit:** statistische Masse, Population
- Untersuchungsmerkmal X_i: Eigenschaft der statistischen Einheit
- Merkmalsausprägung: Erscheinungsformen eines Merkmals
- **Beobachtungswert** *x_i*: Die für das *i*-te Untersuchungsobjekt beobachtete Ausprägung des Untersuchungsmerkmals

Beispiel: Grundlegende Begriffe

Im Rahmen einer Vollerhebung sollen verschiedene Eigenschaften von allen Mescheder Studierenden dieser Vorlesung in diesem Raum untersucht werden.

- Grundgesamtheit: Alle Mescheder Studierende (räumliche Identifikation) dieser Vorlesung (sachliche Identifikation), die heute (zeitliche Identifikation) anwesend sind
- Merkmalsträger: Jede/r Studierende
- Merkmale: Alter, Geschlecht, Wohnort, Note in Mathe

Merkmalsausprägungen:

	Alter	Geschlecht	Wohnort	Note
Student:in a	19	m	Meschede	2,3
Student:in b	20	f	Meschede	1,3
Student:in c	22	f	Brilon	4,0
Student:in d	25	m	Winterberg	3,0

Klassifikation von Merkmalen (qualitativ & quantitativ)

- Qualitative Merkmale: variieren artmäßig
 - z.B. Geschlecht, Rechtsform von Unternehmen, Haarfarbe etc.
- Quantitative Merkmale: variieren der Größe nach
 - z.B. Alter, Einkommen, Kinderzahl etc.

Klassifikation von Merkmalen (Skalenniveau)

- Nominalskala: für qualitative Merkmale, für die keine sinnvolle Reihenfolge der Ausprägungen gegeben ist.
 - z.B. Studiengang, Religionszugehörigkeit, Geschlecht
- Ordinalskala: für Merkmale mit einer natürlichen Reihenfolge. Die Abstände zwischen den Ausprägungen sind nicht quantifizierbar.
 - z.B. Schulnoten, Bildungsabschlüsse
- Metrische Skala: für Merkmale, bei denen sowohl die Reihenfolge als auch die Abstände zwischen den Ausprägungen sinnvoll definiert sind.
 - z.B. Umsatz, Alter, Temperatur in °C

Klassifikation von Merkmalen (stetig und diskret)

- Diskrete Merkmale: können nur bestimmte Werte (z.B. nur ganzzahlige) annehmen
 - z.B. Anzahl der Studierenden an einer Hochschule, Anzahl der Einwohner eines Landes
- Stetige Merkmale: können in einem bestimmten Intervall jeden beliebigen Wert annehmen
 - z.B. Längen, Breiten, Gewichte

Inhaltsübersicht

- 1 Grundbegriffe
- 2 Häufigkeiten

Absolute und relative Häufigkeiten

Definition: Absolute Häufigkeit

Beobachtet man an n statistischen Einheiten ein Merkmal X mit k Merkmalsausprägungen, so gilt für die **absolute Häufigkeit** H_j der j-ten Merkmalsausprägung.

$$H_j$$
 = "Anzahl der Beobachtungswerte, die in der
j-ten Ausprägung auftreten" für $j = 1, ..., k$

Definition: Relative Häufigkeit

Die **relative Häufigkeit** h_j gibt den prozentualen Anteil der statistischen Einheiten, die die j-te Merkmalsausprägung tragen, an.

$$h_j = \frac{1}{n} \cdot H_j$$
 für $j = 1, \dots, k$

Beispiel: Absolute und relative Häufigkeit

19 20 21 22 23 ## 0.4 0.2 0.2 0.1 0.1

```
# Daten: Alter befragter Personen
age <- c(19, 21, 20, 21, 19, 23, 22, 19, 20, 19)
# Absolute Häufigkeit
table(age)
## age
## 19 20 21 22 23
## 4 2 2 1 1
# Relative Häufigkeit
proportions(table(age))
## age
```

Beispiel: Graphische Darstellungsmöglichkeiten von Häufigkeiten



Histogramm

- Ein Histogramm ist eine Möglichkeit, die Häufigkeitsverteilung klassierter Daten graphisch darzustellen.
- Die Häufigkeiten werden als aneinander stoßende Rechtecke dargestellt, deren Flächen proportional zur Häufigkeit der Beobachtungswerte der Klassen sind.
- Nicht die H\u00f6he des Rechteckes \u00fcber eine Klasse ist proportional zur Klassenh\u00e4ufigkeit, sondern die Fl\u00e4che selbst.
- Erstellt werden können Histogramme mit dem R-Befehl hist().

Beispiel: Histogramm

In einer Befragung wurden n = 80 Personen nach der Anzahl der verbrachten Urlaubstage pro Jahr befragt. Die Daten wurden in Klassen unterschiedlicher Breiten erfasst und die jeweiligen Häufigkeiten ausgezählt bzw. berechnet.

\$j\$	Klasse \$K_j\$	\$H_j\$	\$h_j\$	Klassenbreite	norm. \$H_j\$	norm. \$h_j\$
1	[0;8)	6	0.075	8	0.750	0.0094
2	[8;18)	16	0.200	10	1.600	0.0200
3	[18;25)	20	0.250	7	2.857	0.0357
4	[25;30)	14	0.175	5	2.800	0.0350
5	[30;35)	12	0.150	5	2.400	0.0300
6	[35;43)	12	0.150	8	1.500	0.0187

Die normierten absoluten und relativen Klassenhäufigkeiten setzen die Klassenhäufigkeiten in Relation zur Klassenbreite, sodass diese Werte einen Vergleich erlauben. Es gilt norm. $H_{\bar{j}} = H_{\bar{j}} / \text{Klassenbreite}$ (norm. $h_{\bar{j}}$ analog).

Beispiel: Histogramm





Die Fläche der zu den Klassen 5 und 6 gehörenden Rechtecke ist identisch, da gilt $H_5 = H_6$ (bzw. $h_5 = h_6$).

Empirische Verteilungsfunktion

Definition: Empirische Verteilungsfunktion

Die empirische Verteilungsfunktion S eines an n Untersuchungseinheiten beobachteten Merkmals X, das k Merkmalsausprägungen a_1, \ldots, a_k besitzt, wird durch die Folge der relativen Summenhäufigkeiten bestimmt.

$$S_j = \frac{1}{n} \sum_{i=1}^{j} H_j = \sum_{i=1}^{j} h_j$$
 für $j = 1, ..., k$

Beispiel: Empirische Verteilungsfunktion

Fortsetzung auf Basis der Befragung zum Urlaubsverhalten von *n* = 80 Personen.

\$j\$	Klasse \$K_j\$	\$H_j\$	\$h_j\$	emp. Verteilungsfkt.
1	[0;8)	6	0.075	0.075
2	[8;18)	16	0.200	0.275
3	[18;25)	20	0.250	0.525
4	[25;30)	14	0.175	0.700
5	[30;35)	12	0.150	0.850
6	[35;43)	12	0.150	1.000

Beispiel: Empirische Verteilungsfunktion





Verständnisfragen

- Kann unsere Hochschule eine Untersuchungseinheit sein?
- Nach welchen Kriterien werden Grundgesamtheiten abgegrenzt?
- Was ist der Unterschied zwischen einem Merkmalsträger und einer Merkmalsausprägung?

Verständnisfragen (Antworten)

- Kann unsere Hochschule eine Untersuchungseinheit sein?
 - Ja, wenn die Grundgesamtheit z.B. aus allen Hochschulen eines Landes besteht.
- Nach welchen Kriterien werden Grundgesamtheiten abgegrenzt?
 - Räumlich, sachlich, zeitlich
- Was ist der Unterschied zwischen einem Merkmalsträger und einer Merkmalsausprägung?
 - Merkmalsträger ist die Person oder das Objekt, das untersucht wird,
 Merkmalsausprägung ist eine der Erscheinungsformen einer Eigenschaft des Merkmalsträgers.

Verständnisfragen

- Wodurch unterscheiden sich absolute von relativen Häufigkeiten?
- Was ist der Unterschied zwischen einem Stabdiagramm und einem Histogramm?
- Was zeigt die empirische Verteilungsfunktion?

Verständnisfragen (Antworten)

- Wodurch unterscheiden sich absolute von relativen Häufigkeiten?
 - ▶ Die absolute Häufigkeit gibt die Anzahl der Beobachtungen wieder, die relative Häufigkeit gibt den Anteil der Beobachtung an der Gesamtanzahl wieder.
- Was ist der Unterschied zwischen einem Stabdiagramm und einem Histogramm?
 - In einem Stabdiagramm gibt die Höhe des Stabs die Häufigkeit wieder, in einem Histogramm ist es die Fläche der Balken, die die Häufigkeit wiedergibt.
- Was zeigt die empirische Verteilungsfunktion?
 - Anhand der empirischen Verteilungsfunktion kann man die relative Häufigkeit ablesen, mit der ein Wert höchstens eine bestimmte Größe hat. Die empirische Verteilungsfunktion gibt so die kumulierten relativen Häufigkeiten bis zu einem bestimmten Wert an.

Anhang

Sys.time()

[9] tibble 3.2.1

ggplot2 3.5.1

```
## [1] "2025-01-02 21:17:19 UTC"
sessionInfo()
## R version 4.4.1 (2024-06-14)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 22.04.5 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so; LAPACK version 3.10.0
##
## locale:
## [1] LC CTYPE=en US.UTF-8
                                  LC NUMERIC=C
## [3] LC TIME=en US.UTF-8
                                 LC COLLATE=en US.UTF-8
## [5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
## [7] LC_PAPER=en_US.UTF-8
                                  LC_NAME=C
## [9] LC ADDRESS=C
                                  LC TELEPHONE=C
## [11] LC MEASUREMENT=en US.UTF-8 LC IDENTIFICATION=C
##
## time zone: Etc/UTC
## tzcode source: system (glibc)
##
## attached base packages:
## [1] grid
                stats
                          graphics grDevices utils datasets methods
## [8] base
##
## other attached packages:
## [1] vcd_1.4-13
                        lubridate_1.9.3 forcats_1.0.0
                                                          stringr_1.5.1
   [5] dplyr_1.1.4
                        purrr 1.0.2
                                         readr 2.1.5
                                                          tidvr 1.3.1
```

tidyverse 2.0.0 texreg 1.39.4