Karlsruher Institut für Technologie

Institut für Volkswirtschaftslehre (ECON)

Lehrstuhl für Ökonometrie und Statistik

Kapitel IV - Häufigkeitsverteilungen

Deskriptive Statistik

Prof. Dr. W.-D. Heller
Hartwig Senska
Carlo Siebenschuh

Agenda

1. **Beispiel**
2. Darstellung
3. Absolute und relative Häufigkeiten
4. Klassierung
5. Summenhäufigkeiten
6. Empirische Verteilungsfunktion

Beispiel 4.1 - Personalstatistik des Unternehmens usw-it.com der Jungunternehmerin Ulrike Schmidt-Wohlfahrt

	Zugehörigkeit	Alter	Geschlecht	Ausbildung	Funktion	Religion	Gehalt
K.A.	43	29	w	U	S,K,E	ev	3700
U.D.	60	34	w	U	S,K,V	rk	4250
S.F.	13	23	m	B	W,H	rk	2100
R.G.	27	28	m	FH	P,K	rk	2950
A.H.	36	28	w	BA	P,K	ev	2850
F.L.	40	31	m	U	S,F	sonst.	4450
P.S.	18	25	m	FH	P,W,K	sonst.	2800
K.W.	21	22	w	B	B,SK	ev	1950

Agenda

1. Beispiel
2. **Darstellung**
3. Absolute und relative Häufigkeiten
4. Klassierung
5. Summenhäufigkeiten
6. Empirische Verteilungsfunktion

Darstellung

Erhebungsergebnis in allgemeiner Form:

Statistische Einheit	s_1	s_2	s_3	$\cdots$	s_n
Merkmalswert	x_1	x_2	x_3	$\cdots$	x_n

Anmerkung: Für jede statistische Einheit s_i kann der zugehörige Merkmalswert x_i abgelesen werden. Es wird zunächst nur ein Merkmal untersucht. Die Anzahl der statistischen Einheiten sei mit n bezeichnet.

Darstellung

Erhebungsergebnis in allgemeiner Form:

Statistische Einheit	s_1	s_2	s_3	$\cdots$	s_n
Merkmalswert	x_1	x_2	x_3	$\cdots$	x_n

Anmerkung: Für jede statistische Einheit s_i kann der zugehörige Merkmalswert x_i abgelesen werden. Es wird zunächst nur ein Merkmal untersucht. Die Anzahl der statistischen Einheiten sei mit n bezeichnet.

Darstellung

Urliste (stat. Reihe, Liste der Merkmalswerte):

$$x_1, x_2, x_3, \ldots, x_n,$$

Anmerkung: Grundlage für weitere Auswertung
Für die folgenden Überlegungen ist es wichtig, die Liste der Merkmalswerte (Urliste) und die Menge M der möglichen Merkmalsausprägungen genau zu unterscheiden. Aus diesem Grund verwenden wir bei der Urliste die Buchstaben x und y und die Bezeichnung **Merkmalswert** und bei der Menge M die Buchstaben a und b und die Bezeichnung **Merkmalsausprägung**.

Darstellung

Urliste (stat. Reihe, Liste der Merkmalswerte):

$$x_1, x_2, x_3, \ldots, x_n,$$

Anmerkung: Grundlage für weitere Auswertung
Für die folgenden Überlegungen ist es wichtig, die Liste der Merkmalswerte (Urliste) und die Menge M der möglichen Merkmalsausprägungen genau zu unterscheiden. Aus diesem Grund verwenden wir bei der Urliste die Buchstaben x und y und die Bezeichnung **Merkmalswert** und bei der Menge M die Buchstaben a und b und die Bezeichnung **Merkmalsausprägung**.

Darstellung - Urliste

Beispiel 4.2
Im Leistungskurs Mathematik eines Jahrgangs eines Gymnasiums wurden folgende Ergebnisse erzielt:

Schüler	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}
Punktzahl	14	9	13	8	12	9	11	12	9	12	14

Schüler	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
Punktzahl	8	10	12	9	7	11	12	13	9

Zugehörige Urliste lautet also:

$$14, 9, 13, 8, 12, 9, 11, 12, 9, 12, 14, 8, 10, 12, 9, 7, 11, 12, 13, 9.$$

Darstellung - Urliste

Beispiel 4.2
Im Leistungskurs Mathematik eines Jahrgangs eines Gymnasiums wurden folgende Ergebnisse erzielt:

Schüler	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}
Punktzahl	14	9	13	8	12	9	11	12	9	12	14

Schüler	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
Punktzahl	8	10	12	9	7	11	12	13	9

Zugehörige Urliste lautet also:

$$14, 9, 13, 8, 12, 9, 11, 12, 9, 12, 14, 8, 10, 12, 9, 7, 11, 12, 13, 9.$$

Darstellung

Geordnete Urliste:

$$x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}.$$

Anmerkung: Zur übersichtlicheren Gestaltung können bei einem quantitativen Merkmal die Werte der Urliste entsprechend der Ordnung der reellen Zahlen umsortiert bzw. "geordnet" werden. Man erhält bei aufsteigender Reihenfolge die geordnete Urliste. Diese besteht aus derselben Anzahl von Merkmalswerten, wobei gegenüber der ursprünglichen Liste nur eine Umsortierung vorgenommen wurde.
Auch bei einem Rangmerkmal kann die Urliste geordnet werden, wobei man dann entsprechend der natürlichen Reihenfolge der Merkmalsausprägungen vorgeht.

Die geordnete Urliste zum vorangegangenen **Beispiel** lautet also:

$$7, 8, 8, 9, 9, 9, 9, 9, 10, 11, 11, 12, 12, 12, 12, 12, 13, 13, 14, 14$$

Darstellung

Geordnete Urliste:

$$x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}.$$

Anmerkung: Zur übersichtlicheren Gestaltung können bei einem quantitativen Merkmal die Werte der Urliste entsprechend der Ordnung der reellen Zahlen umsortiert bzw. "geordnet" werden. Man erhält bei aufsteigender Reihenfolge die geordnete Urliste. Diese besteht aus derselben Anzahl von Merkmalswerten, wobei gegenüber der ursprünglichen Liste nur eine Umsortierung vorgenommen wurde.
Auch bei einem Rangmerkmal kann die Urliste geordnet werden, wobei man dann entsprechend der natürlichen Reihenfolge der Merkmalsausprägungen vorgeht.

Die geordnete Urliste zum vorangegangenen **Beispiel** lautet also:

$$7, 8, 8, 9, 9, 9, 9, 9, 10, 11, 11, 12, 12, 12, 12, 12, 13, 13, 14, 14$$

Darstellung

Geordnete Urliste:

$$x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}.$$

Anmerkung: Zur übersichtlicheren Gestaltung können bei einem quantitativen Merkmal die Werte der Urliste entsprechend der Ordnung der reellen Zahlen umsortiert bzw. "geordnet" werden. Man erhält bei aufsteigender Reihenfolge die geordnete Urliste. Diese besteht aus derselben Anzahl von Merkmalswerten, wobei gegenüber der ursprünglichen Liste nur eine Umsortierung vorgenommen wurde.
Auch bei einem Rangmerkmal kann die Urliste geordnet werden, wobei man dann entsprechend der natürlichen Reihenfolge der Merkmalsausprägungen vorgeht.

Die geordnete Urliste zum vorangegangenen **Beispiel** lautet also:

$$7, 8, 8, 9, 9, 9, 9, 9, 10, 11, 11, 12, 12, 12, 12, 12, 13, 13, 14, 14$$

Darstellung

Stiel- und Blatt-Darstellung:
Bei einer größeren Anzahl von Daten als hier in diesem Beispiel wird eine reine Auflistung der Zahlen schnell unübersichtlich.

Die geordnete Urliste zum vorangegangenen **Beispiel** lautet also:

$$7, 8, 8, 9, 9, 9, 9, 9, 10, 11, 11, 12, 12, 12, 12, 12, 13, 13, 14, 14$$

Beispiel 4.3
Bei den Altersangaben von 40 Beschäftigten einer Behörde

37	58	63	17	28	46	57	26	39	47
16	62	44	39	48	27	35	59	19	26
55	36	37	48	28	46	18	62	25	37
38	45	59	61	29	36	28	42	29	37

...

Darstellung

... erhalten wir als geordnete Urliste

16	17	18	19	25	26	26	27	28	28
28	29	29	35	36	36	37	37	37	37
38	39	39	42	44	45	46	46	47	48
48	55	57	58	59	59	61	62	62	63

Diese Liste wird offensichtlich wesentlich übersichtlicher, wenn wir sie nach der Zehnerziffer zeilenweise ordnen:

1	6	7	8	9						
2	5	6	6	7	8	8	8	9	9	
3	5	6	6	7	7	7	7	8	9	9
4	2	4	5	6	6	7	8	8		
5	5	7	8	9	9					
6	1	2	2	3						
7										

Stiel- und Blatt-Dartstellung

Links der senkrechten Linie sind die Zehnerziffern, rechts davon die zugehörigen Einerziffern der Merkmalswerte der Urliste aufgeführt.

Darstellung

... erhalten wir als geordnete Urliste

16	17	18	19	25	26	26	27	28	28
28	29	29	35	36	36	37	37	37	37
38	39	39	42	44	45	46	46	47	48
48	55	57	58	59	59	61	62	62	63

Diese Liste wird offensichtlich wesentlich übersichtlicher, wenn wir sie nach der Zehnerziffer zeilenweise ordnen:

1	6	7	8	9						
2	5	6	6	7	8	8	8	9	9	
3	5	6	6	7	7	7	7	8	9	9
4	2	4	5	6	6	7	8	8		
5	5	7	8	9	9					
6	1	2	2	3						
7										

Stiel- und Blatt-Dartstellung

Links der senkrechten Linie sind die Zehnerziffern, rechts davon die zugehörigen Einerziffern der Merkmalswerte der Urliste aufgeführt.

Darstellung

... erhalten wir als geordnete Urliste

16	17	18	19	25	26	26	27	28	28
28	29	29	35	36	36	37	37	37	37
38	39	39	42	44	45	46	46	47	48
48	55	57	58	59	59	61	62	62	63

Diese Liste wird offensichtlich wesentlich übersichtlicher, wenn wir sie nach der Zehnerziffer zeilenweise ordnen:

1	6	7	8	9						
2	5	6	6	7	8	8	8	9	9	
3	5	6	6	7	7	7	7	8	9	9
4	2	4	5	6	6	7	8	8		
5	5	7	8	9	9					
6	1	2	2	3						
7										

Stiel- und Blatt-Dartstellung

Links der senkrechten Linie sind die Zehnerziffern, rechts davon die zugehörigen Einerziffern der Merkmalswerte der Urliste aufgeführt.

Darstellung

Anmerkung: Bei **qualitativen Merkmalen** ist eine Ordnung der Urliste in dieser Art nicht möglich. Hier kann nur so umsortiert werden, dass identische Merkmalswerte benachbart sind.

Beispiel 4.4 Die Spieler eines Bundesligavereins (17 Spieler) wurden nach ihrem letzten Schul- bzw. Hochschulabschluss befragt. Dabei bedeuten: H Hauptschule, B Berufsschule, R Realschule, A Abitur, FO Fachoberschule, FH Fachhochschule, U Universität:

Spieler	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
Abschluss	A	H	H	H	U	R	B	H	B	B

Spieler	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}
Abschluss	R	R	FO	H	A	A	R

Darstellung

Beispiel 4.4
Zugehörige Urliste lautet:

$$A, H, H, H, U, R, B, H, B, B, R, R, FO, H, A, A, R.$$

Umsortierung:

$$A, A, A, H, H, H, H, H, U, R, R, R, R, B, B, B, FO$$

oder kürzer:

$$A(3), H(5), U(1), R(4), B(3), FO(1), (FH(O))$$

Darstellung

Beispiel 4.4
Zugehörige Urliste lautet:

$$A, H, H, H, U, R, B, H, B, B, R, R, FO, H, A, A, R.$$

Umsortierung:

$$A, A, A, H, H, H, H, H, U, R, R, R, R, B, B, B, FO$$

oder kürzer:

$$A(3), H(5), U(1), R(4), B(3), FO(1), (FH(O))$$

Darstellung

Beispiel 4.4
Zugehörige Urliste lautet:

$$A, H, H, H, U, R, B, H, B, B, R, R, FO, H, A, A, R.$$

Umsortierung:

$$A, A, A, H, H, H, H, H, U, R, R, R, R, B, B, B, FO$$

oder kürzer:

$$A(3), H(5), U(1), R(4), B(3), FO(1), (FH(O))$$

Agenda

1. Beispiel
2. Darstellung
3. **Absolute und relative Häufigkeiten**
4. Klassierung
5. Summenhäufigkeiten
6. Empirische Verteilungsfunktion

Absolute und relative Häufigkeiten

Absolute Häufigkeit: Die absolute Häufigkeit einer Merkmalsausprägung a ist die Anzahl der Merkmalswerte der Urliste, die mit a übereinstimmen; sie wird mit $h(a)$ bezeichnet:

$$h(a) = \#\{i \mid i = 1, 2, 3, ..., n : x_i = a\}.$$

Absolute Häufigkeitsverteilung: Die Zusammenstellung der absoluten Häufigkeiten $h(a)$ für alle $a \in M$ nennt man absolute Häufigkeitsverteilung. Mathematisch formuliert handelt es sich also um eine Abbildung $h : M \to \mathbb{N} \cup \{0\}$, die im allgemeinen in Tabellenform wiedergegeben wird.

Absolute und relative Häufigkeiten

Absolute Häufigkeit: Die absolute Häufigkeit einer Merkmalsausprägung a ist die Anzahl der Merkmalswerte der Urliste, die mit a übereinstimmen; sie wird mit $h(a)$ bezeichnet:

$$h(a) = \#\{i \mid i = 1, 2, 3, ..., n : x_i = a\}.$$

Absolute Häufigkeitsverteilung: Die Zusammenstellung der absoluten Häufigkeiten $h(a)$ für alle $a \in M$ nennt man absolute Häufigkeitsverteilung. Mathematisch formuliert handelt es sich also um eine Abbildung $h : M \to \mathbb{N} \cup \{0\}$, die im allgemeinen in Tabellenform wiedergegeben wird.

Absolute Häufigkeitsverteilung

Häufigkeitsverteilung zum Beispiel 4.2

Punktzahl	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Absolute H.	0	0	0	0	0	0	0	1	2	5	1	2	5	2

Punktzahl	14	15
Absolute H.	2	0

Und zu Beispiel 4.4

Abschluss	A	H	U	R	B	FO	FH
Absolute H.	3	5	1	4	3	1	0

Absolute Häufigkeitsverteilung

Häufigkeitsverteilung zum Beispiel 4.2

Punktzahl	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Absolute H.	0	0	0	0	0	0	0	1	2	5	1	2	5	2

Punktzahl	14	15
Absolute H.	2	0

Und zu Beispiel 4.4

Abschluss	A	H	U	R	B	FO	FH
Absolute H.	3	5	1	4	3	1	0

Absolute und relative Häufigkeiten

Relative Häufigkeit: Um auch ohne Kenntnis der Gesamtzahl der Merkmalswerte eine Aussage machen zu können, werden die absoluten Häufigkeiten mit der Gesamtzahl der Merkmalswerte der statistischen Reihe, also der Anzahl der statistischen Einheiten, in Beziehung gesetzt, indem man den Anteil der absoluten Häufigkeit einer Merkmalsausprägung an der Gesamtzahl der aufgeführten Merkmalswerte angibt.

Sei n die Anzahl der statistischen Einheiten und $h(a)$ die absolute Häufigkeit der Merkmalsausprägung a, so wird das Verhältnis

$$\frac{h(a)}{n} \quad \text{als} \quad \textbf{relative Häufigkeit}$$

...

Absolute und relative Häufigkeiten

Relative Häufigkeit: Um auch ohne Kenntnis der Gesamtzahl der Merkmalswerte eine Aussage machen zu können, werden die absoluten Häufigkeiten mit der Gesamtzahl der Merkmalswerte der statistischen Reihe, also der Anzahl der statistischen Einheiten, in Beziehung gesetzt, indem man den Anteil der absoluten Häufigkeit einer Merkmalsausprägung an der Gesamtzahl der aufgeführten Merkmalswerte angibt.

Sei n die Anzahl der statistischen Einheiten und $h(a)$ die absolute Häufigkeit der Merkmalsausprägung a, so wird das Verhältnis

$$\frac{h(a)}{n} \quad \text{als} \quad \textbf{relative Häufigkeit}$$

...

Absolute und relative Häufigkeiten

... und

$$\frac{h(a)}{n} \cdot 100\% \quad \text{als} \quad \textbf{prozentuale relative Häufigkeit}$$

bezeichnet.

Relative Häufigkeitsverteilung: Die Zusammenstellung aller relativen Häufigkeiten $p(a), a \in M$ nennt man analog rel. Häufigkeitsverteilung (mathematisch formuliert die Abbildung $p : M \mapsto [0,1]$ bzw. $[0,100]$).

Absolute und relative Häufigkeiten

... und

$$\frac{h(a)}{n} \cdot 100\% \quad \text{als} \quad \textbf{prozentuale relative Häufigkeit}$$

bezeichnet.

Relative Häufigkeitsverteilung: Die Zusammenstellung aller relativen Häufigkeiten $p(a), a \in M$ nennt man analog rel. Häufigkeitsverteilung (mathematisch formuliert die Abbildung $p : M \mapsto [0,1]$ bzw. $[0,100]$).

Relative Häufigkeitsverteilung

Zum Beispiel 4.2

Punktzahl	0	1	2	3	4	5	6	7	8	9	10	11	12
Relative H. in %	0	0	0	0	0	0	0	5	10	25	5	10	25

Punktzahl	13	14	15
Relative H. in %	10	10	0

Agenda

1. Beispiel
2. Darstellung
3. Absolute und relative Häufigkeiten
4. **Klassierung**
5. Summenhäufigkeiten
6. Empirische Verteilungsfunktion

Klassierung von Merkmalsausprägungen

Eine *Klassierung von Merkmalsausprägungen* ist eine Zusammenfassung von Merkmalsausprägungen zu *Klassen*. Dabei muss jede Merkmalsausprägung genau einer Klasse angehören.

Klassen: Bei quantitativen Merkmalen Intervalle z.B. über $\mathbb{R}$. Die Klasseneinteilung erfolgt durch Angabe von *unteren* und *oberen Klassengrenzen*, wobei jeweils anzugeben ist, ob die Klassengrenze zur Klasse gehört ([ ]) oder nicht (( )). Obere Klassengrenze einer Klasse ist gleichzeitig untere Klassengrenze der nächsten Klasse.

Offene Randklassen: Haben nur eine Klassengrenze und zwar eine obere Klassengrenze bei den *nach links offenen* Randklassen und eine untere bei den *nach rechts offenen* Randklassen. Dabei werden alle Merkmalsausprägungen zusammengefasst, die links bzw. rechts von der Klassengrenze liegen. Die Klassengrenze kann dabei zur Klasse gehören oder auch nicht. Offene Randklassen werden immer dann verwendet, wenn die Mehrzahl der Werte in einem begrenzten Bereich und einige wenige Werte weit verstreut liegen.

Klassierung von Merkmalsausprägungen

Eine *Klassierung von Merkmalsausprägungen* ist eine Zusammenfassung von Merkmalsausprägungen zu *Klassen*. Dabei muss jede Merkmalsausprägung genau einer Klasse angehören.

Klassen: Bei quantitativen Merkmalen Intervalle z.B. über $\mathbb{R}$. Die Klasseneinteilung erfolgt durch Angabe von *unteren* und *oberen Klassengrenzen*, wobei jeweils anzugeben ist, ob die Klassengrenze zur Klasse gehört ([ ]) oder nicht (( )). Obere Klassengrenze einer Klasse ist gleichzeitig untere Klassengrenze der nächsten Klasse.

Offene Randklassen: Haben nur eine Klassengrenze und zwar eine obere Klassengrenze bei den *nach links offenen* Randklassen und eine untere bei den *nach rechts offenen* Randklassen. Dabei werden alle Merkmalsausprägungen zusammengefasst, die links bzw. rechts von der Klassengrenze liegen. Die Klassengrenze kann dabei zur Klasse gehören oder auch nicht. Offene Randklassen werden immer dann verwendet, wenn die Mehrzahl der Werte in einem begrenzten Bereich und einige wenige Werte weit verstreut liegen.

Klassierung von Merkmalsausprägungen

Eine *Klassierung von Merkmalsausprägungen* ist eine Zusammenfassung von Merkmalsausprägungen zu *Klassen*. Dabei muss jede Merkmalsausprägung genau einer Klasse angehören.

Klassen: Bei quantitativen Merkmalen Intervalle z.B. über $\mathbb{R}$. Die Klasseneinteilung erfolgt durch Angabe von *unteren* und *oberen Klassengrenzen*, wobei jeweils anzugeben ist, ob die Klassengrenze zur Klasse gehört ([ ]) oder nicht (( )). Obere Klassengrenze einer Klasse ist gleichzeitig untere Klassengrenze der nächsten Klasse.

Offene Randklassen: Haben nur eine Klassengrenze und zwar eine obere Klassengrenze bei den *nach links offenen* Randklassen und eine untere bei den *nach rechts offenen* Randklassen. Dabei werden alle Merkmalsausprägungen zusammengefasst, die links bzw. rechts von der Klassengrenze liegen. Die Klassengrenze kann dabei zur Klasse gehören oder auch nicht. Offene Randklassen werden immer dann verwendet, wenn die Mehrzahl der Werte in einem begrenzten Bereich und einige wenige Werte weit verstreut liegen.

Klassierung von Merkmalsausprägungen

Beispiel 4.5 - Die Messung der Körpergröße von 20 Personen ergab folgende Urliste:

$$1.49, 1.87, 1.91, 1.53, 1.68, 1.75, 1.66, 1.82, 1.76, 1.80,$$

$$1.92, 1.71, 1.77, 1.69, 1.57, 1.83, 1.84, 1.47, 1.79, 1.81.$$

Damit ist die absolute (relative) Häufigkeit für jeden Wert der Urliste 1 (0.05) und 0 (0) für alle übrigen Merkmalsausprägungen.

Klassierung von Merkmalsausprägungen

Beispiel 4.5 - Die Messung der Körpergröße von 20 Personen ergab folgende Urliste:

$$1.49, 1.87, 1.91, 1.53, 1.68, 1.75, 1.66, 1.82, 1.76, 1.80,$$

$$1.92, 1.71, 1.77, 1.69, 1.57, 1.83, 1.84, 1.47, 1.79, 1.81.$$

Damit ist die absolute (relative) Häufigkeit für jeden Wert der Urliste 1 (0.05) und 0 (0) für alle übrigen Merkmalsausprägungen.

Klassierung von Merkmalsausprägungen

Durch eine geeignete Klassierung erhält man folgende Tabelle:

Klasse	1 1.40 b.u. 1.50	2 1.50 b.u. 1.60	3 1.60 b.u. 1.70
Absolute H.	2	2	3
Relative H.	0.1	0.1	0.15

Klasse	4 1.70 b.u. 1.80	5 1.80 b.u. 1.90	6 1.90 b.u. 2.00	$\sum$
Absolute H.	5	6	2	20
Relative H.	0.25	0.3	0.1	1

Kenngrößen der Klassierung

Klassenbreite $\Delta_I = \beta_I - \alpha_I$: Differenz aus oberer und unterer Klassengrenze. Die Klassenbreite sollte in der Regel (natürlich bis auf offene Randklassen) einheitlich sein.

Klassenmitte $z_I = 0.5(\alpha_I + \beta_I)$: Arithmetisches Mittel aus unterer und oberer Klassengrenze. Repräsentant der Klasse, Ersatzwert für die unbekannten Merkmalswerte in der Klasse. Die Klassengrenzen sind so zu wählen, dass die Klassenmitte ein sinnvoller Repräsentant ist.

Kenngrößen der Klassierung

Klassenbreite $\Delta_I = \beta_I - \alpha_I$: Differenz aus oberer und unterer Klassengrenze. Die Klassenbreite sollte in der Regel (natürlich bis auf offene Randklassen) einheitlich sein.

Klassenmitte $z_I = 0.5(\alpha_I + \beta_I)$: Arithmetisches Mittel aus unterer und oberer Klassengrenze. Repräsentant der Klasse, Ersatzwert für die unbekannten Merkmalswerte in der Klasse. Die Klassengrenzen sind so zu wählen, dass die Klassenmitte ein sinnvoller Repräsentant ist.

Kenngrößen der Klassierung

Anzahl der Klassen: Je nach Anzahl der Beobachtungswerte festzulegen. Je größer die Klassen sind, desto mehr Information geht verloren. Es ist also die Relation zwischen der Anzahl von Beobachtungswerten und der Anzahl von Klassen zu berücksichtigen.

DIN-55302:

Anzahl der Beobachtungswerte		Anzahl der Klassen	
bis	100	mindestens	10
etwa	1.000	mindestens	13
etwa	10.000	mindestens	16
etwa	100.000	mindestens	20

Kenngrößen der Klassierung

Anzahl der Klassen: Je nach Anzahl der Beobachtungswerte festzulegen. Je größer die Klassen sind, desto mehr Information geht verloren. Es ist also die Relation zwischen der Anzahl von Beobachtungswerten und der Anzahl von Klassen zu berücksichtigen.

DIN-55302:

Anzahl der Beobachtungswerte		Anzahl der Klassen	
bis	100	mindestens	10
etwa	1.000	mindestens	13
etwa	10.000	mindestens	16
etwa	100.000	mindestens	20

Klassierung von Merkmalsausprägungen

Anmerkung: Bei stetigen Merkmalen werden die Merkmalswerte mit einer (in der Regel vorher festgelegten) *Messgenauigkeit* erhoben. Dies bedeutet, dass nicht der exakte Wert ermittelt wird, sondern nur ein *Näherungswert.*

Beispielsweise wird durch die Angabe "Körpergröße: 1.72m" abkürzend wiedergegeben, dass der exakte Wert mindestens 1.71500...m und weniger als 1.72500...m ist, dass also der exakte Wert in das Intervall $[1.715, 1.725)$ fällt.

Die Messgenauigkeit entspricht also einer "versteckten" Klassierung.

Klassierung von Merkmalsausprägungen

Die Tabelle zur Klassierung von Beispiel 4.4 müsste korrekt lauten:

Klasse	1 1.395 b.u. 1.495	2 1.495 b.u. 1.595	3 1.595 b.u. 1.695
Absolute H.	2	2	3
Relative H.	0.1	0.1	0.15

Klasse	4 1.695 b.u. 1.795	5 1.795 b.u. 1.895	6 1.895 b.u. 1.995	$\sum$
Absolute H.	5	6	2	20
Relative H.	0.25	0.3	0.1	1

Klassenmitten sind also nicht 1.45, 1.55,..., sondern korrekterweise 1.445, 1.545,...

Der Unterschied wirkt sich jedoch i.a. nicht sehr aus und wird daher oft nicht beachtet.

Klassierung von Merkmalsausprägungen

Die Tabelle zur Klassierung von Beispiel 4.4 müsste korrekt lauten:

Klasse	1 1.395 b.u. 1.495	2 1.495 b.u. 1.595	3 1.595 b.u. 1.695
Absolute H.	2	2	3
Relative H.	0.1	0.1	0.15

Klasse	4 1.695 b.u. 1.795	5 1.795 b.u. 1.895	6 1.895 b.u. 1.995	$\sum$
Absolute H.	5	6	2	20
Relative H.	0.25	0.3	0.1	1

Klassenmitten sind also nicht 1.45, 1.55,..., sondern korrekterweise 1.445, 1.545,...

Der Unterschied wirkt sich jedoch i.a. nicht sehr aus und wird daher oft nicht beachtet.

Agenda

1. Beispiel
2. Darstellung
3. Absolute und relative Häufigkeiten
4. Klassierung
5. **Summenhäufigkeiten**
6. Empirische Verteilungsfunktion

Summenhäufigkeiten

Hauptarten des Datenmaterials für weitere Auswertung:

1. **Urliste** (geordnet oder ungeordnet)
2. **Häufigkeitsverteilung** (absolut oder relativ)
3. **Häufigkeitsverteilung der Klassen** (absolut oder relativ)

Summenhäufigkeiten

Hauptarten des Datenmaterials für weitere Auswertung:

1. **Urliste** (geordnet oder ungeordnet)
2. **Häufigkeitsverteilung** (absolut oder relativ)
3. **Häufigkeitsverteilung der Klassen** (absolut oder relativ)

Summenhäufigkeiten

Hauptarten des Datenmaterials für weitere Auswertung:

1. **Urliste** (geordnet oder ungeordnet)
2. **Häufigkeitsverteilung** (absolut oder relativ)
3. **Häufigkeitsverteilung der Klassen** (absolut oder relativ)

Summenhäufigkeiten

Weitere Darstellungsform bei Rangmerkmalen und quantitativen Merkmalen:

- **Rangmerkmal** - Liste der Merkmalsausprägungen in ihrer natürlichen Reihenfolge
- **Quantitatives Merkmal** - $\preceq$ - Beziehung

Summenhäufigkeit
Sei a die Merkmalsausprägung. Wieviele statistische Einheiten haben einen Merkmalswert unterhalb a oder übereinstimmend mit a? (Erfordert Reihenfolge bei den Merkmalsausprägungen!)

Summenhäufigkeiten

Weitere Darstellungsform bei Rangmerkmalen und quantitativen Merkmalen:

- **Rangmerkmal** - Liste der Merkmalsausprägungen in ihrer natürlichen Reihenfolge
- **Quantitatives Merkmal** - $\preceq$ - Beziehung

Summenhäufigkeit
Sei a die Merkmalsausprägung. Wieviele statistische Einheiten haben einen Merkmalswert unterhalb a oder übereinstimmend mit a? (Erfordert Reihenfolge bei den Merkmalsausprägungen!)

Summenhäufigkeiten

Weitere Darstellungsform bei Rangmerkmalen und quantitativen Merkmalen:

- **Rangmerkmal** - Liste der Merkmalsausprägungen in ihrer natürlichen Reihenfolge
- **Quantitatives Merkmal** - $\preceq$ - Beziehung

Summenhäufigkeit
Sei a die Merkmalsausprägung. Wieviele statistische Einheiten haben einen Merkmalswert unterhalb a oder übereinstimmend mit a? (Erfordert Reihenfolge bei den Merkmalsausprägungen!)

Summenhäufigkeiten

Weitere Darstellungsform bei Rangmerkmalen und quantitativen Merkmalen:

- **Rangmerkmal** - Liste der Merkmalsausprägungen in ihrer natürlichen Reihenfolge
- **Quantitatives Merkmal** - $\preceq$ - Beziehung

Summenhäufigkeit
Sei a die Merkmalsausprägung. Wieviele statistische Einheiten haben einen Merkmalswert unterhalb a oder übereinstimmend mit a? (Erfordert Reihenfolge bei den Merkmalsausprägungen!)

Summenhäufigkeiten

Absolute Summenhäufigkeit der Merkmalsausprägung a:

$$H(a) = \sum_{a' \text{ kommt vor } a \text{ oder } a'=a} h(a') \qquad \text{(Rangmerkmale)}$$

$$H(a) = \sum_{a' \leq a} h(a') \qquad \text{(Quant. Merkmale)}$$

Relative Summenhäufigkeit der Merkmalsausprägung a:

$$F(a) = \sum_{a' \text{ kommt vor } a \text{ oder } a'=a} p(a') \qquad \text{(Rangmerkmale)}$$

$$F(a) = \sum_{a' \leq a} p(a') \qquad \text{(Quant. Merkmale)}$$

Summenhäufigkeiten

Absolute Summenhäufigkeit der Merkmalsausprägung a:

$$H(a) = \sum_{a' \text{ kommt vor } a \text{ oder } a'=a} h(a') \qquad \text{(Rangmerkmale)}$$

$$H(a) = \sum_{a' \leq a} h(a') \qquad \text{(Quant. Merkmale)}$$

Relative Summenhäufigkeit der Merkmalsausprägung a:

$$F(a) = \sum_{a' \text{ kommt vor } a \text{ oder } a'=a} p(a') \qquad \text{(Rangmerkmale)}$$

$$F(a) = \sum_{a' \leq a} p(a') \qquad \text{(Quant. Merkmale)}$$

Summenhäufigkeiten

Beispiel 4.6 - Häufigkeitsverteilung der Gesamtnote bei Studienabgängern eines Jahres:

nicht bestanden	ausreichend	befriedigend	gut	sehr gut
7	28	27	33	5

Man erhält damit folgende Summenhäufigkeitsverteilung:

Abschluss	Anzahl
nicht bestanden	7
ausreichend oder schlechter	35
befriedigend oder schlechter	62
gut oder schlechter	95
sehr gut oder schlechter	100

Summenhäufigkeiten

Beispiel 4.6 - Häufigkeitsverteilung der Gesamtnote bei Studienabgängern eines Jahres:

nicht bestanden	ausreichend	befriedigend	gut	sehr gut
7	28	27	33	5

Man erhält damit folgende Summenhäufigkeitsverteilung:

Abschluss	Anzahl
nicht bestanden	7
ausreichend oder schlechter	35
befriedigend oder schlechter	62
gut oder schlechter	95
sehr gut oder schlechter	100

Summenhäufigkeiten

Zu Beispiel 4.2 - Summenhäufigkeitsverteilung (Punkteverteilung LK Mathematik)

Punkte	Anzahl
bis einschließlich 6	0
bis einschließlich 7	1
bis einschließlich 8	3
bis einschließlich 9	8
bis einschließlich 10	9
bis einschließlich 11	11
bis einschließlich 12	16
bis einschließlich 13	18
bis einschließlich 14	20
bis einschließlich 15	20

Summenhäufigkeiten

Summenhäufigkeit bei klassierten Daten:

$$H(a) = \sum_{I:\beta_I{}^1 \leq a^2} h(I) \quad \text{bzw.} \quad F(a) = \sum_{I:\beta_I \leq a} p(I).$$

Anmerkung: Bei klassierten Daten sind Summenhäufigkeiten exakt nur für Klassengrenzen bestimmbar. Bei der Interpretation ist dabei noch darauf zu achten, ob die Klassengrenze zur Klasse gehört oder nicht.

[1] obere Klassengrenze der Klasse I

[2] Klassengrenze

Summenhäufigkeiten

Summenhäufigkeit bei klassierten Daten:

$$H(a) = \sum_{I:\beta_I{}^1 \leq a^2} h(I) \quad \text{bzw.} \quad F(a) = \sum_{I:\beta_I \leq a} p(I).$$

Anmerkung: Bei klassierten Daten sind Summenhäufigkeiten exakt nur für Klassengrenzen bestimmbar. Bei der Interpretation ist dabei noch darauf zu achten, ob die Klassengrenze zur Klasse gehört oder nicht.

[1] obere Klassengrenze der Klasse I
[2] Klassengrenze

Summenhäufigkeiten

Summenhäufigkeitsverteilung zu Beispiel 4.4:

Körpergröße in cm	140	150	160	170	180	190	200
abs. Summenhäufigkeit	0	2	4	7	12	18	20
rel. Summenhäufigkeit	0	0.1	0.2	0.35	0.6	0.9	1

Agenda

1. Beispiel
2. Darstellung
3. Absolute und relative Häufigkeiten
4. Klassierung
5. Summenhäufigkeiten
6. **Empirische Verteilungsfunktion**

Empirische Verteilungsfunktion

Bei einem quantitativen Merkmal ist $M \subseteq \mathbb{R}$ und dementsprechend kann für jede reelle Zahl x der Anteil aller Merkmalswerte festgestellt werden, die diese Merkmalsausprägung nicht überschreiten. Man erhält auf diese Weise eine Funktion, die jeder reellen Zahl einen Wert zwischen 0 und 1 einschließlich zuordnet, ähnlich der relativen Summenhäufigkeit.

Sei $x_1, x_2, ..., x_n$ die Urliste. So heißt die Funktion $F : \mathbb{R} \to [0, 1]$ **empirische Verteilungsfunktion:**

$$F(x) = \frac{1}{n} \#\{i \mid i = 1, ..., n : x_i \leq x\}$$

Empirische Verteilungsfunktion

Bei einem quantitativen Merkmal ist $M \subseteq \mathbb{R}$ und dementsprechend kann für jede reelle Zahl x der Anteil aller Merkmalswerte festgestellt werden, die diese Merkmalsausprägung nicht überschreiten. Man erhält auf diese Weise eine Funktion, die jeder reellen Zahl einen Wert zwischen 0 und 1 einschließlich zuordnet, ähnlich der relativen Summenhäufigkeit.

Sei $x_1, x_2, ..., x_n$ die Urliste. So heißt die Funktion $F : \mathbb{R} \to [0, 1]$ **empirische Verteilungsfunktion:**

$$F(x) = \frac{1}{n} \#\{i \mid i = 1, ..., n : x_i \leq x\}$$

Empirische Verteilungsfunktion der unklassierten Daten des Beispieles 4.5:

$$F(x) = \begin{cases} 0 & \text{für} \quad x < 1.47 \\ 0.05 & \text{für} \quad 1.47 \leq x < 1.49 \\ 0.1 & \text{für} \quad 1.49 \leq x < 1.53 \\ 0.15 & \text{für} \quad 1.53 \leq x < 1.57 \\ 0.2 & \text{für} \quad 1.57 \leq x < 1.66 \\ 0.25 & \text{für} \quad 1.66 \leq x < 1.68 \\ 0.3 & \text{für} \quad 1.68 \leq x < 1.69 \\ 0.35 & \text{für} \quad 1.69 \leq x < 1.71 \\ 0.4 & \text{für} \quad 1.71 \leq x < 1.75 \\ 0.45 & \text{für} \quad 1.75 \leq x < 1.76 \\ 0.5 & \text{für} \quad 1.76 \leq x < 1.77 \\ 0.55 & \text{für} \quad 1.77 \leq x < 1.79 \\ 0.6 & \text{für} \quad 1.79 \leq x < 1.80 \\ 0.65 & \text{für} \quad 1.80 \leq x < 1.81 \\ 0.7 & \text{für} \quad 1.81 \leq x < 1.82 \\ 0.75 & \text{für} \quad 1.82 \leq x < 1.83 \\ 0.8 & \text{für} \quad 1.83 \leq x < 1.84 \\ 0.85 & \text{für} \quad 1.84 \leq x < 1.87 \\ 0.9 & \text{für} \quad 1.87 \leq x < 1.91 \\ 0.95 & \text{für} \quad 1.91 \leq x < 1.92 \\ 1 & \text{für} \quad 1.92 \leq x \end{cases}$$

Empirische Verteilungsfunktion

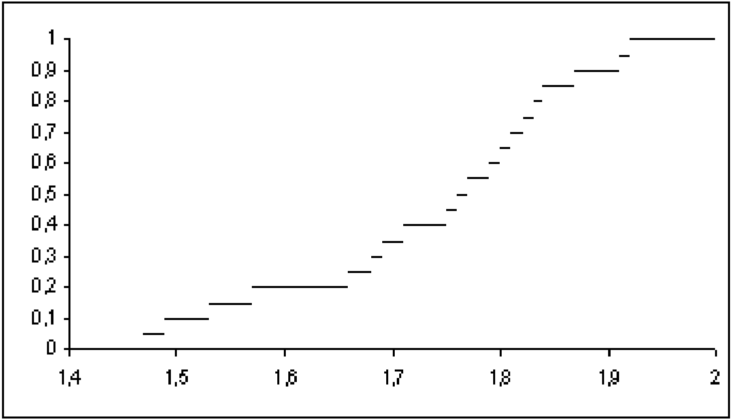

Abbildung: Empirische Verteilungsfunktion zu Beispiel 4.5

Typische Eigenschaften empirischer Verteilungsfunktionen

1. Die Funktion ist von links nach rechts ansteigend (isoton).
2. Da zwischen zwei benachbarten Werten $x_{(i)}$ und $x_{(i+1)}$ der geordneten Urliste keine Merkmalswerte beobachtet wurden, ist die Funktion in diesem Bereich konstant. Man erhält eine sogenannte "Treppenfunktion", wobei die "Treppenstufen" unterschiedliche Länge aufweisen können.
3. Die Höhe der Treppenstufen ist in der Regel (zumindest bei stetigen Merkmalen) $\frac{1}{n}$. Nur bei aufeinanderfallenden Merkmalswerten ist sie ein entsprechendes Vielfaches von $\frac{1}{n}$.
4. An der Sprungstelle nimmt die Funktion den höheren Wert an (Die Funktion ist rechtsseitig stetig.).
5. Minimalwert der Funktion ist 0, Maximalwert ist 1.

Typische Eigenschaften empirischer Verteilungsfunktionen

1. Die Funktion ist von links nach rechts ansteigend (isoton).
2. Da zwischen zwei benachbarten Werten $x_{(i)}$ und $x_{(i+1)}$ der geordneten Urliste keine Merkmalswerte beobachtet wurden, ist die Funktion in diesem Bereich konstant. Man erhält eine sogenannte "Treppenfunktion", wobei die "Treppenstufen" unterschiedliche Länge aufweisen können.
3. Die Höhe der Treppenstufen ist in der Regel (zumindest bei stetigen Merkmalen) $\frac{1}{n}$. Nur bei aufeinanderfallenden Merkmalswerten ist sie ein entsprechendes Vielfaches von $\frac{1}{n}$.
4. An der Sprungstelle nimmt die Funktion den höheren Wert an (Die Funktion ist rechtsseitig stetig.).
5. Minimalwert der Funktion ist 0, Maximalwert ist 1.

Typische Eigenschaften empirischer Verteilungsfunktionen

1. Die Funktion ist von links nach rechts ansteigend (isoton).
2. Da zwischen zwei benachbarten Werten $x_{(i)}$ und $x_{(i+1)}$ der geordneten Urliste keine Merkmalswerte beobachtet wurden, ist die Funktion in diesem Bereich konstant. Man erhält eine sogenannte "Treppenfunktion", wobei die "Treppenstufen" unterschiedliche Länge aufweisen können.
3. Die Höhe der Treppenstufen ist in der Regel (zumindest bei stetigen Merkmalen) $\frac{1}{n}$. Nur bei aufeinanderfallenden Merkmalswerten ist sie ein entsprechendes Vielfaches von $\frac{1}{n}$.
4. An der Sprungstelle nimmt die Funktion den höheren Wert an (Die Funktion ist rechtsseitig stetig.).
5. Minimalwert der Funktion ist 0, Maximalwert ist 1.

Typische Eigenschaften empirischer Verteilungsfunktionen

1. Die Funktion ist von links nach rechts ansteigend (isoton).
2. Da zwischen zwei benachbarten Werten $x_{(i)}$ und $x_{(i+1)}$ der geordneten Urliste keine Merkmalswerte beobachtet wurden, ist die Funktion in diesem Bereich konstant. Man erhält eine sogenannte "Treppenfunktion", wobei die "Treppenstufen" unterschiedliche Länge aufweisen können.
3. Die Höhe der Treppenstufen ist in der Regel (zumindest bei stetigen Merkmalen) $\frac{1}{n}$. Nur bei aufeinanderfallenden Merkmalswerten ist sie ein entsprechendes Vielfaches von $\frac{1}{n}$.
4. An der Sprungstelle nimmt die Funktion den höheren Wert an (Die Funktion ist rechtsseitig stetig.).
5. Minimalwert der Funktion ist 0, Maximalwert ist 1.

Typische Eigenschaften empirischer Verteilungsfunktionen

1. Die Funktion ist von links nach rechts ansteigend (isoton).
2. Da zwischen zwei benachbarten Werten $x_{(i)}$ und $x_{(i+1)}$ der geordneten Urliste keine Merkmalswerte beobachtet wurden, ist die Funktion in diesem Bereich konstant. Man erhält eine sogenannte "Treppenfunktion", wobei die "Treppenstufen" unterschiedliche Länge aufweisen können.
3. Die Höhe der Treppenstufen ist in der Regel (zumindest bei stetigen Merkmalen) $\frac{1}{n}$. Nur bei aufeinanderfallenden Merkmalswerten ist sie ein entsprechendes Vielfaches von $\frac{1}{n}$.
4. An der Sprungstelle nimmt die Funktion den höheren Wert an (Die Funktion ist rechtsseitig stetig.).
5. Minimalwert der Funktion ist 0, Maximalwert ist 1.

Abschließende Bemerkung: Man beachte, dass durch die Bildung der Summenhäufigkeitsverteilung bzw. der empirischen Verteilungsfunktion kein Informationsverlust entsteht, d.h. es ist möglich, aus diesen die Häufigkeitsverteilung bzw. die Urliste zu ermitteln, sofern die Anzahl der Beobachtungen bekannt ist.