JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

3. Einheit

Plan für heute

- **Univariate Datenanalyse: Lagemaße (auch: Maße der zentralen Tendenz)**

Lernziele

- Kenntnis von Lagemaßen der univariaten Statistik
- Bestimmung und Berechnung von Lagemaßen der univariaten Statistik: Modus, Median, arithmetisches Mittel, Quantile

Einführung

- Tabellen und grafische Darstellungen verschaffen einen globalen Eindruck von der Verteilung eines interessierenden Merkmals
- In Häufigkeitstabellen wird abgetragen, welche Merkmalsausprägung mit welcher (absoluten, prozentualen und/oder relativen kumulierten) Häufigkeit beobachtet wurde
- Bei kontinuierlichen Variablen bietet sich die Zusammenfassung der Merkmalsausprägungen in Gruppen an
- Häufigkeitsverteilungen können grafisch z.B. als Säulendiagramm oder Histogramm dargestellt werden.
- Grafiken nehmen aber viel Platz ein, und der Vergleich von Verteilungsmerkmalen via Grafiken ist schwierig.

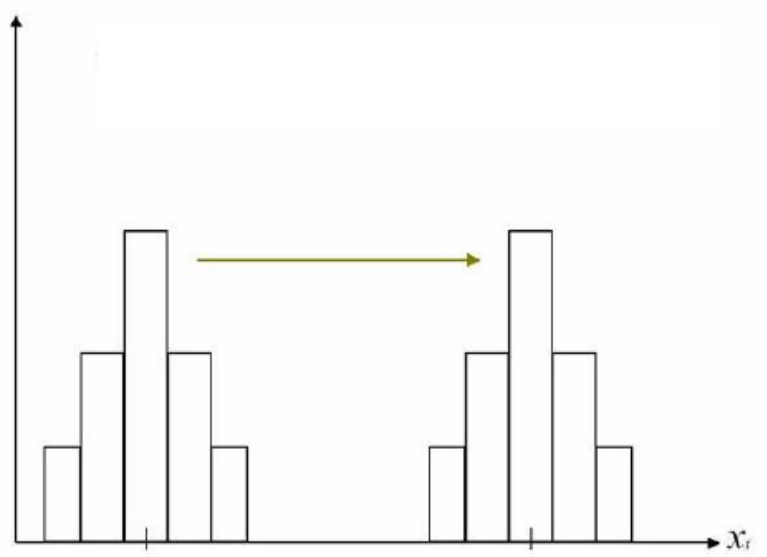
→ Informationsverdichtung nötig → **statistische Kennwerte**

Statistische Kennwerte

- beschreiben spezifische Eigenschaften einer Merkmalsverteilung: **Unterscheidung Lage- und Streu(ungs)maße**
- **Lagemaße** (auch Maße der zentralen Tendenz):
  - Geben Auskunft über das Zentrum bzw. typische Werte einer Verteilung
  - Durch welchen Wert wird die Merkmalsverteilung am besten repräsentiert?
  - Wichtigste Lagemaße: **Modus, Median und arithmetisches Mittel**

Lagemaße

- Lagemaße als „Mittelwerte“
- Wo auf der x-Achse befinden sich die Merkmalswerte einer Verteilung „im Mittel“ ?
- Beispiel: Identische Verteilungen, Unterschied lediglich in der „Lage“



Lagemaße

Durch welchen Wert wird die Merkmalsverteilung am besten repräsentiert?

→ 3 Möglichkeiten

1. Durch den Wert, der in der Verteilung am häufigsten vorkommt → **Modus (auch Modalwert)**
2. Durch den Wert, der die Menge der Beobachtungseinheiten in zwei gleich große Teile teilt → **Median** (auch Zentralwert)
3. Durch den Durchschnitt aller Werte → **Arithmetisches Mittel**

Lagemaße

Welches Lagemaß verwendet wird hängt auch vom Skalenniveau des interessierenden Merkmals ab!

	Nominalskala	Ordinalskala	ab Intervallskala
Modus	Ja	Ja	Ja
Median	Nein	Ja	Ja
Arithmetisches Mittel	Nein	Nein	Ja

Quelle: Eigene Darstellung

Modus

- Definition: die Merkmalsausprägung, die in der Verteilung am häufigsten vorkommt
- Notation: x_{Mo}
- vor allem dann sinnvoll für die Charakterisierung einer Verteilung, wenn er sich deutlich von den anderen Werten abhebt

Beispiel Modus

- x_{Mo}=1

Studiengang x_k	Absolute Häufigkeit fx_k bzw. Hx_k
BA Social Sciences (1)	80
Jura (2)	10
Medizin (3)	40
BA Psycho (4)	30
Summe	160

Im Beispiel liegt der Modus für das Merkmal „Studiengang“ bei der Merkmalsausprägung „1“ (SoWi)

Übung: Modus

Bei welchem Wert liegt in diesem Beispiel der Modus?

Schulnote	Absolute Häufigkeiten
1 („sehr gut“)	150
2 („gut“)	230
3 („befriedigend“)	400
4 („ausreichend“)	190
5 („mangelhaft“)	25
6 („ungenügend“)	5
Gesamt	1000

Lösung: Modus

- Wo liegt in diesem Beispiel der Modus?
- x_{Mo}= 3

Schulnote	Absolute Häufigkeiten
1 („sehr gut“)	150
2 („gut“)	230
3 („befriedigend“)	400
4 („ausreichend“)	190
5 („mangelhaft“)	25
6 („ungenügend“)	5
Gesamt	1000

Lagemaße

„Durch welchen Wert wird die Merkmalsverteilung am besten repräsentiert?

→ 3 Möglichkeiten

1. Durch den Wert, der in der Verteilung am häufigsten vorkommt → **Modalwert (auch Modus)**
2. Durch den Wert, der die Menge der Beobachtungseinheiten in zwei gleich große Teile teilt → **Median** (auch Zentralwert)
3. Durch den Durchschnitt aller Werte → **Arithmetisches Mittel**

Median

- Definition: diejenige Ausprägung, die in der Mitte einer geordneten Verteilung (nach ihrer Größe sortierten Messwerte) steht.
- Notation: $\tilde{x}$
- Der Median teilt die Verteilung einer Variablen exakt in zwei Hälften, das bedeutet unter- und oberhalb des Medians liegen jeweils 50% der Untersuchungseinheiten
- Voraussetzung für die Bestimmung des Medians ist ordinales Skalenniveau

Median

- Berechnung unterscheidet sich, je nachdem ob eine gerade oder ungerade Anzahl von Messwerten vorliegt (für eine gerade Anzahl von Messwerten gibt es keine „Mitte“)
- Die Position des Medians bei einer Verteilung mit ungerader Fallzahl entspricht dem Wert des Elements auf dem Rangplatz $\frac{n+1}{2}$

Beispiel Median, ungerades n

- Werte der Tabelle bereits aufsteigend geordnet
- Formel für Median bei ungerader Anzahl von Messwerten: $\tilde{x} = x_{\frac{n+1}{2}}$
- $\tilde{x} = x_{\frac{9+1}{2}}$

	Median								
Rang-platz	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
Wert	28	32	38	41	42	42	55	59	78

$\tilde{x}$=42

Median, gerades n

- Für eine Verteilung mit gerader Fallzahl kann dem Median keine ganze Zahl zugeordnet werden
- bestimmt sich dann als Mittelwert der beiden zentral gelegenen Realisationen
- Genauer: entspricht dem Mittelwert der Realisationen der beiden mittleren Rangplätze

$$\frac{n}{2} \text{ und } \frac{n}{2} + 1$$

→ Bei gerader Anzahl von Merkmalsträgern:

$$\tilde{x} = \frac{1}{2} * (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$$

Beispiel Median, gerades n

- Werte der Tabelle bereits aufsteigend geordnet
- Formel für Median bei gerader Anzahl von Messwerten:

$$\tilde{x} = \frac{1}{2} * (x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$$

- $\tilde{x}$=0.5x(41+42) =41.5

	Median							
Rang-platz	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$
Wert	28	32	38	41	42	42	55	59

Median – kumulierte relative Häufigkeiten

- Sind die Werte einer Häufigkeitstabelle nach der Größe geordnet, so entspricht der Median dem Wert, bei dem die kumulierten Anteilswerte mindestens 50% betragen

Median – kumulierte relative Häufigkeiten

- Sind die Werte einer Häufigkeitstabelle nach der Größe geordnet, so entspricht der Median dem Wert, bei dem die kumulierten Anteilswerte mindestens 50% betragen

Schulnote	Absolute Häufigkeiten	Relativer Anteil (%)	Kum. relative Häufigkeit (%)
„sehr gut“	150	15	15
„gut“	230	23	38
„befriedigend“	400	40	78
„ausreichend“	190	19	97
„mangelhaft“	25	2.5	99.5
„ungenügend“	5	0.5	100
Gesamt	1000		

Lagemaße

„Durch welchen Wert wird die Merkmalsverteilung am besten repräsentiert?

→ 3 Möglichkeiten

1. Durch den Wert, der in der Verteilung am häufigsten vorkommt → **Modalwert (auch Modus)**
2. Durch den Wert, der die Menge der Beobachtungseinheiten in zwei gleich große Teile teilt → **Median** (auch Zentralwert)
3. Durch den Durchschnitt aller Werte → **Arithmetisches Mittel**

Arithmetisches Mittel

- Definition: Summe aller Messwerte geteilt durch ihre Anzahl
- Notation: $\bar{x}$
- Umgangssprachlich: „der Mittelwert“ oder „Durchschnitt“
- Berechnung:

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i}{n}$$

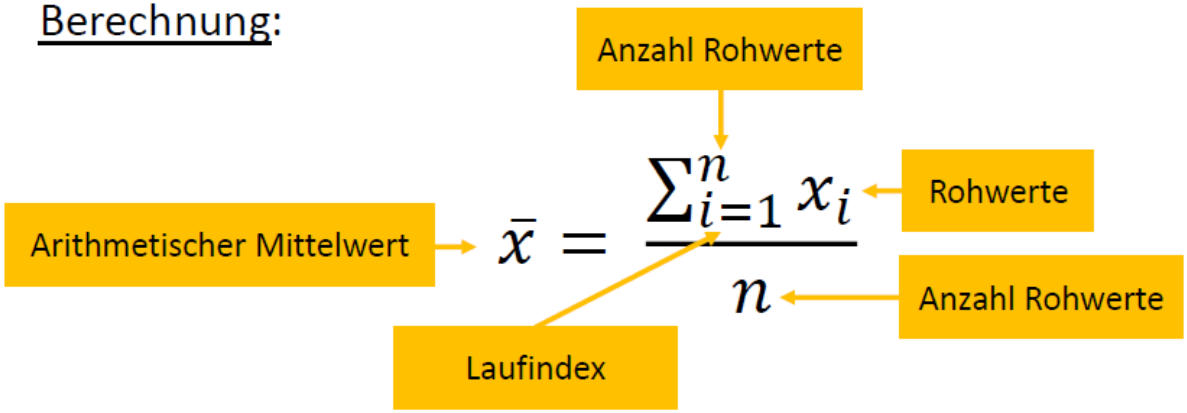
i = Laufindex Merkmalsträger $(i = 1, \ldots, n)$

x_i = Merkmalsausprägung x des i-ten Merkmalsträgers

n = Anzahl der Merkmalsträger

Berechnung: Arithmetisches Mittel

Berechnung:



i = Laufindex Merkmalsträger ($i = 1, \ldots, n$)

x_i = Merkmalsausprägung x des i-ten Merkmalsträgers

n = Anzahl der Merkmalsträger

Beispiel: Arithmetisches Mittel

- Gegeben ist folgende Urliste des Merkmals X: 2, 4, 5, 4, 3, 1, 3, 4, 2, 5 (n=10); wie lautet das arithmetische Mittel?

- Berechnung:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{1}{10}(2+4+5+4+3+1+3+4+2+5) = 3.3$$

Arithmetisches Mittel

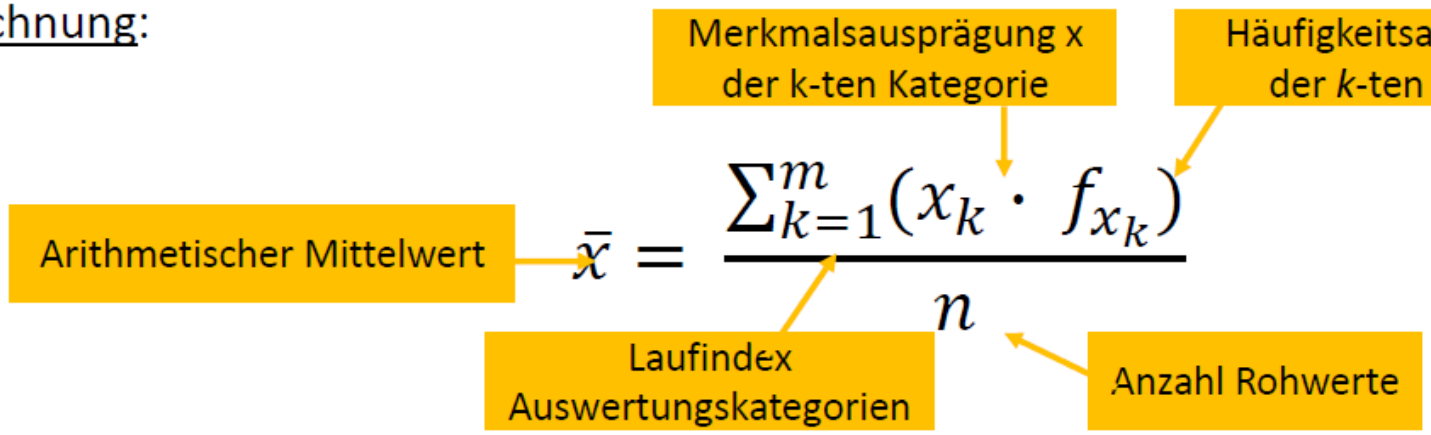
- Berechnung bei kleinem n problemlos und intuitiv „händisch“ möglich
- In der sozialwissenschaftlichen Praxis häufig große Fallzahlen (ALLBUS Umfrage, n=ca. 3000)
- Arithmetisches Mittel kann auch auf Basis der Häufigkeitstabelle berechnet werden
- Beispiel:

Schulnote	Absolute Häufigkeiten
1 („sehr gut“)	150
2 („gut“)	230
3 („befriedigend“)	400
4 („ausreichend“)	190
5 („mangelhaft“)	25
6 („ungenügend“)	5
Gesamt	1000

$$\bar{x} = (1*150)+(2*230)+(3*400)+(4*190)+(5*25)+(6*5)/1000=2725/1000= 2.725$$

Arithmetisches Mittel

Berechnung:



k = Laufindex über die Kategorien $(k = 1, \ldots, m)$

m = Anzahl der Kategorien

x_k = Merkmalsausprägung x der k-ten Kategorie

f_{x_k} = Häufigkeitsausprägung x der k-ten Kategorie

n = Anzahl der Merkmalsträger

Arithmetisches Mittel

- enthält am Meisten Information über eine Verteilung (alle Messwerte werden berücksichtigt)
- Voraussetzung: (Pseudo-) metrisches Skalenniveau (z.B. auch Notendurchschnitt)
- Teilweise ohne sinnvolle empirische Entsprechung (z.B. „Deutsche Frauen gebären im Durchschnitt 1,58 Kinder“)

Arithmetisches Mittel

- Vorteil: alle verfügbaren Informationen werden bei der Berechnung ausgeschöpft (alle Messwerte werden berücksichtigt)
- Nachteil: sensibel für Extremwerte („Ausreißer“) (daher v.a. geeignet, wenn eine symmetrische, unimodale Verteilung zugrunde liegt)

Beispiel: Arithmetisches Mittel

- Vorteil: alle verfügbaren Informationen werden bei der Berechnung ausgeschöpft (alle Messwerte werden berücksichtigt)
- Nachteil: sensibel für Extremwerte („Ausreißer“) (daher v.a. geeignet, wenn eine symmetrische, unimodale Verteilung zugrunde liegt)

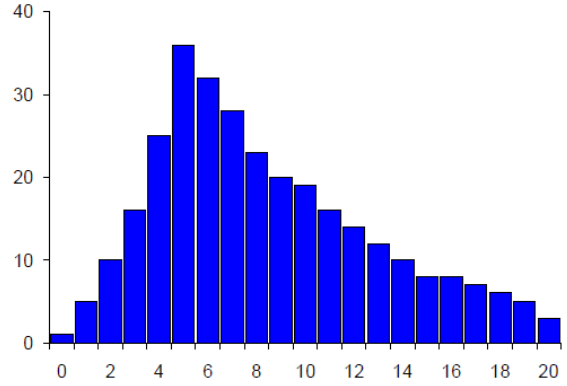
Beispiel (Alter)	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	$\bar{x}$
Gruppe 1	20	25	25	25	25	25	30	30	**35**	26,67
Gruppe 2	20	25	25	25	25	25	30	30	**70**	30,56

Lagemaße und Verteilungsformen

- Das Verhältnis von Modus, Median und arithmetischem Mittel erlaubt Rückschlüsse auf die Form der Verteilung

Linkssteile/rechtsschiefe Verteilung:

- Häufigste Ausprägung weiter links; Modus ist der kleinste der drei Mittelwerte
- Median liegt in der Mitte, daher größer als der Modus
- Arithmetisches Mittel wird stärker durch Ausreißer weit rechts beeinflusst, deshalb noch größer als der Median
- *Modus< Median< arithmetisches Mittel*

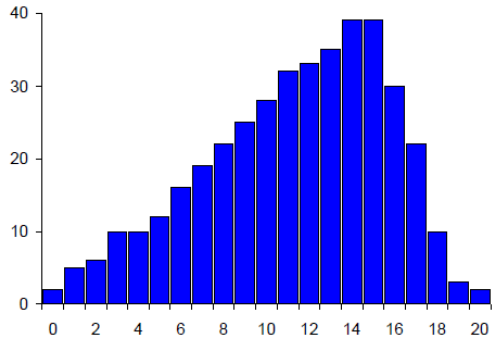


Lagemaße und Verteilungsformen

- Das Verhältnis von Modus, Median und arithmetischem Mittel erlaubt Rückschlüsse auf die Form der Verteilung

Rechtssteile/linksschiefe Verteilung:

- Häufigste Ausprägung weiter rechts; Modus ist der größte der drei Mittelwerte
- Median liegt in der Mitte, daher kleiner als der Modus
- Arithmetisches Mittel wird stärker durch Ausreißer links beeinflusst, deshalb noch kleiner als der Median
- *Modus> Median> arithmetisches Mittel*

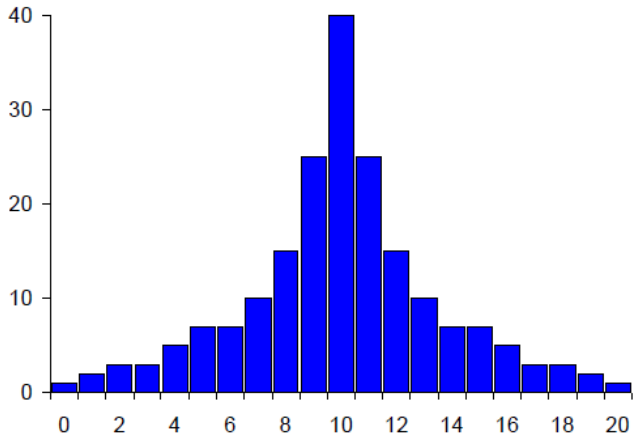


Lagemaße und Verteilungsformen

- Das Verhältnis von Modus, Median und arithmetischem Mittel erlaubt Rückschlüsse auf die Form der Verteilung

Symmetrische unimodale Verteilung:

- Modus, Median und arithmetisches Mittel nehmen sehr ähnliche Werte an

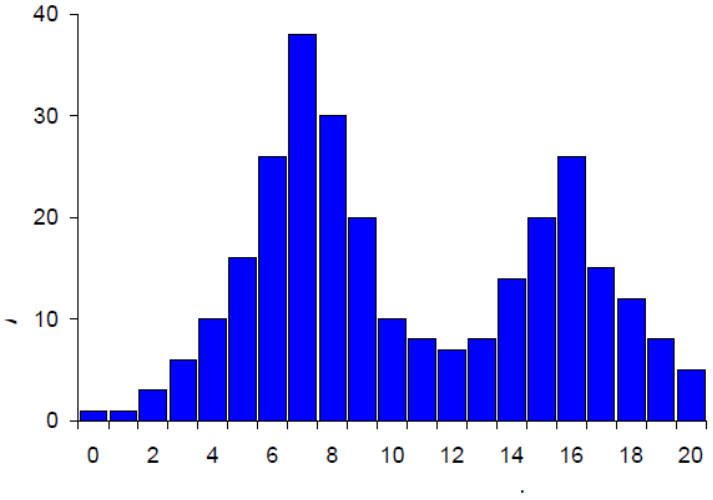


Lagemaße und Verteilungsformen

- Das Verhältnis von Modus, Median und arithmetischem Mittel erlaubt Rückschlüsse auf die Form der Verteilung

Bimodale Verteilung:

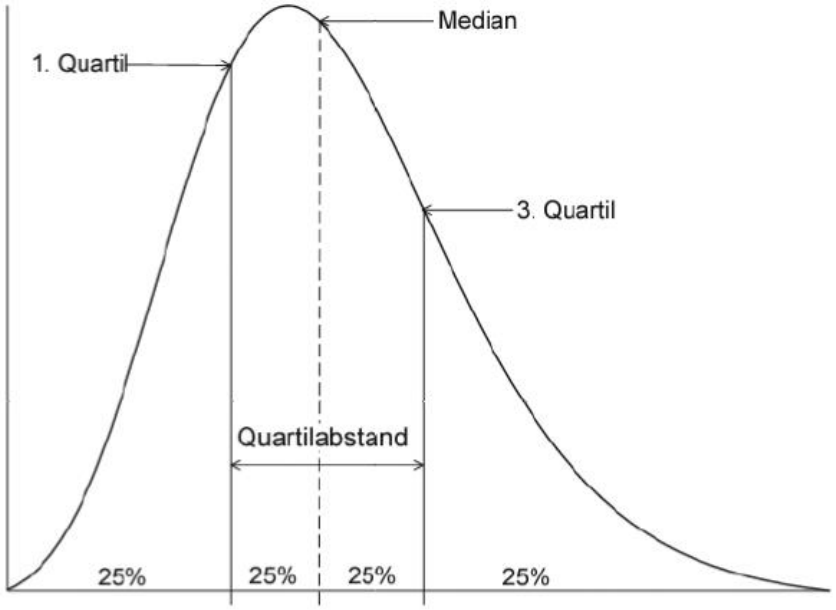
- Median und arithmetisches Mittel nehmen sehr ähnliche Werte an
- Modus teilweise nicht klar interpretierbar (ggfs. 2 Modalwerte angeben)



Quantile (auch Perzentile)

- Median entspricht dem Wert, bei dem die kumulierten relativen Anteile 50% erreichen oder übertreffen
- Analog hierzu lassen sich beliebige Quantile bilden (= Anteilswerte)
- Der Vergleich mehrerer Quantile erleichtert die Charakterisierung einer Verteilung. Zu den besonders gebräuchlichen Quantilen zählen die Quartile.
- **Quartile** unterteilen geordnete Daten in vier gleichgroße Gruppen (jeweils 25% der Datenwerte)

Quantile (auch Perzentile)



- 25% der Werte sind kleiner oder gleich und 75% der Werte sind größer oder gleich dem 1. Quartil ($Q_{0.25}$)
- Das zweite Quartil ist der Median. Der Median ($Q_{0.5}$) zerteilt eine Verteilung in zwei gleich große Hälften.
- 75% der Werte sind kleiner oder gleich und 25% der Werte größer oder gleich dem 3. Quartil ($Q_{0.75}$).

Quantile (auch Perzentile)

- Quantilwerte können aus der Häufigkeitstabelle abgelesen werden (analog zur Bestimmung des Medians)
- Der Quantilwert ist die Ausprägung, bei der in der Spalte mit den kumulierten Anteilen bzw. kumulierten Prozentwerten erstmals der Quantilanteil **erreicht oder überschritten** wird

Beispiel Quantile

- Wo liegt im Beispiel das 1. Quartil? → bei 11 Semestern
- Wo liegt das 3. Quartil?

Semesterzahl	Absolute Häufigkeit	%	Kumulierte %
10	1	9.1	9.1
11	2	18.2	27.3
12	3	27.3	54.6
13	2	18.2	72.8
14	1	9.1	81.9
15	1	9.1	91
20	1	9.1	100
∑	*11*	*100*	*100*

Beispiel Quantile

- Wo liegt im Beispiel das 1. Quartil? → bei 11 Semestern
- Wo liegt das 3. Quartil?

Semesterzahl	Absolute Häufigkeit	%	Kumulierte %
10	1	9.1	9.1
11	2	18.2	27.3
12	3	27.3	54.6
13	2	18.2	72.8
14	1	9.1	81.9
15	1	9.1	91
20	1	9.1	100
∑	*11*	*100*	*100*

Lernziele

- Kenntnis von Lagemaßen der univariaten Statistik
- Bestimmung und Berechnung von Lagemaßen der univariaten Statistik: Modus, Median, arithmetisches Mittel, Quantile