

Sandra Hansen-Morath
Sascha Wolfer

STATISTIK MIT R

Skalenniveaus

MESSEN UND SKALEN

- Datenerhebung ist Variablenmessung.
- Die einzelnen Messungen müssen notiert werden.
- Dies geschieht immer auf einer bestimmten Skala.




GRUNDLAGEN

- **Variable**
 - Untersuchungsmerkmal, Bezeichnung für eine interessierende Eigenschaft, die in unterschiedlichen Varianten (Ausprägungen) auftritt
 - Sie kann in mindestens zwei Abstufungen vorkommen
 - Bsp.: *Tempus; Geschlecht*
- **Merkmalsausprägungen**
 - Werte, die eine Variable annehmen kann
 - Bsp.: *Präsens, Perfekt; männlich, weiblich*

SKALEN- / MESSNIVEAUS

- Die Art und Genauigkeit der Messung beeinflusst das Skalenniveau, auf dem eine Variable notiert wird.
- Je höher das Skalenniveau, desto mehr statistische Verfahren können angewendet werden.
- Ziel: Das höchstmögliche Skalenniveau! Ansonsten wird Information verschenkt.
- Sind Messungen nicht hinreichend durchdacht, können die Daten u.U. nur mit großem Aufwand gerettet werden (wenn überhaupt).

SKALENNIVEAUS

- Je höher das Skalenniveau, desto mehr, präzisere und stärkere statistische Verfahren können angewendet werden.
 - Mögliche Skalenniveaus sind
 - Nominalskala
 - Ordinalskala / Rangskala
 - Intervallskala
 - Verhältnisskala / Rationalskala
- Metrische Skala /
Kardinalskala
- 

NOMINALSKALA

- Notation von Variablen, die keine inhärente Ordnung ausweisen.
- Lediglich die Gleich- bzw. Verschiedenheit kann aus der Skala abgelesen werden.
- Beispiele:
 - Biologisches Geschlecht
 - Muttersprache
 - Richtig / Falsch
- Mögliche Operationen:
 - Auszählen von Häufigkeiten
 - Tests: bspw. χ^2 -Test



ORDINALSKALA

- Notation von Variablen, die eine inhärente Ordnung aufweisen, aber keine Aussage über die Größe der Abstände möglich ist.
- **Zusätzlich** kann die Geordnetheit / Rangfolge aus einer Skala abgelesen werden.
- Beispiele:
 - Altersklassen
 - Härte von Mineralien (Mohssche Härteskala)
- Zusätzliche Operationen:
 - Aufstellen von Rangfolgen (größer/kleiner)
 - Tests: bspw. Mann-Whitney-U-Test



INTERVALLSKALA

- Notation von Variablen, bei denen Aussagen über die Größe von Abständen möglich sind. Der Nullpunkt der Skala ist nicht absolut.
- **Zusätzlich** können Aussagen über die Differenzen (die Intervalle) zwischen Werten getroffen werden (und die Verhältnisse von Differenzen).
- Beispiele:
 - Reaktionszeiten
 - Temperatur in °C oder °F
 - Datum
- Zusätzliche Operationen:
 - Errechnen des Mittelwerts, der Varianz/Standardabweichung
 - Vergleich von Differenzen
 - Tests: bspw. t-Test und ANOVA



RATIONALSKALA

- Notation von Variablen deren Nullpunkt absolut ist.
- Zusätzlich können Aussagen über die Verhältnisse zwischen Messwerten gemacht werden.
- Beispiele:
 - Token pro Text
 - Gehalt
 - Gewicht
 - Temperatur in K
- Zusätzliche Operation:
 - Aussagen der Form „A ist doppelt so schwer wie B.“
 - keine zusätzlichen Tests
- Der Unterschied zwischen Intervall- und Rationalskala ist für uns meist irrelevant.



ÜBERBLICK

	Nominalskala	Ordinalskala	Kardinalskala
Eigenschaften des numerischen Relativs	Identität	Identität und Geordnetheit	Identität, Geordnetheit, Definiertheit der Abstände
Ableitbare Interpretationen	Gleich- / Verschiedenheit von Elementen	+ Größer-/Kleiner-Relationen	+ Relationen und Gleich- / Verschiedenheit von Intervallen
Beispiele für zulässige statistische Kennwerte	Modus	+ Median	+ Mittelwert, Varianz
Variablentyp	diskret	diskret	kontinuierlich

NOMINAL? ORDINAL? INTERVALL?

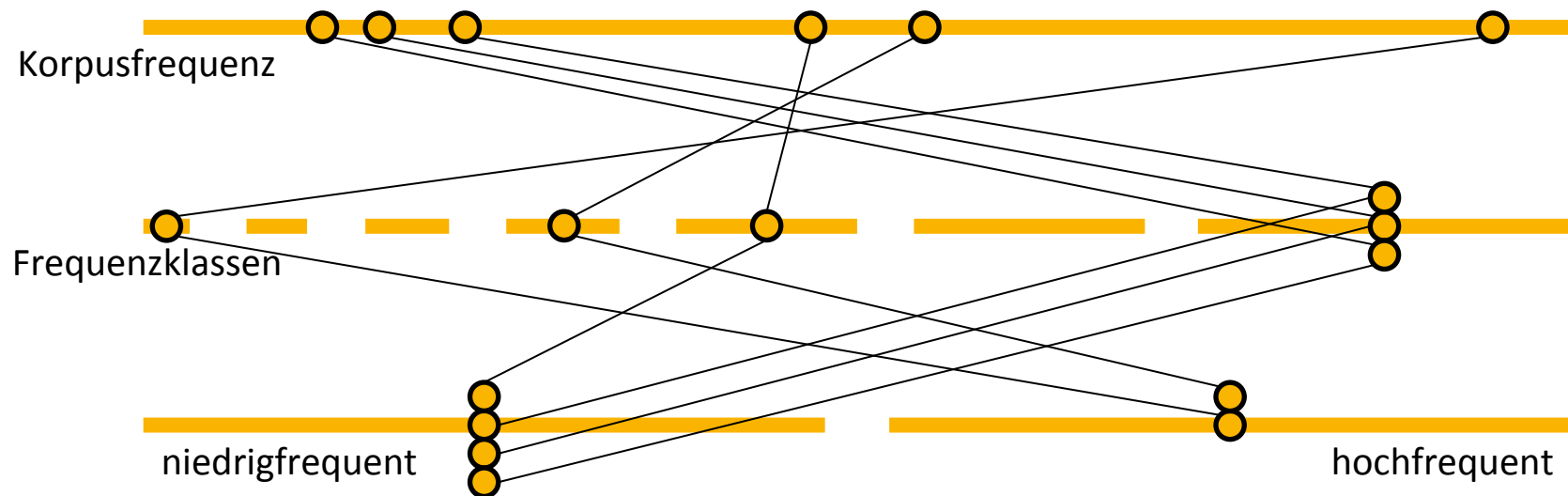
- Tokenfrequenz
- Akzeptabilitätsurteile auf Likert-Skala
- Schulabschluss
- Schwierigkeit von Vokabeln (einfach/mittel/schwer)
- IQ
- Spanisch: „ser“ oder „aver“ als Hilfsverb?
- Lesezeit auf einem Wort
- Einbettungstiefe einer Nominalphrase
- Herkunft eines Korpus
- Korpusgröße
- Formalitätsgrad

BEGRIFFLICHES

- Hier wurden die Begriffe Rangskala und Ordinalskala synonym verwendet.
- Die enge Definition einer Rangskala ist, dass jeder Rang nur einmal vergeben werden darf.
 - Man spricht in diesem Fall davon, dass keine **Bindungen** vorliegen.
- In der Praxis ist dieser Unterschied nur selten wichtig. Statistische Tests korrigieren i.A. für gebundene Ränge.

ÜBERFÜHRBARKEIT

- Merkmale, die auf einer Skala notiert werden, können immer auch auf **niedrigeren** Skalen notiert werden.
- Dabei geht aber Information verloren!
- Beispiel: Wortfrequenz



ÜBERFÜHRBARKEIT

- Auf der Intervallskala (rohe Frequenz) sind alle Wörter und auch die Intervalle zwischen ihnen noch unterscheidbar.
- Auf der Rangskala (Frequenzklassen) fallen manche Wörter bereits in eine Klasse. Ein Unterschied zwischen diesen Wörtern ist damit nicht mehr detektierbar.
- Auf der Nominalskala (niedrig- vs. hochfrequent) wird die ursprüngliche Skala einfach in der Mitte getrennt. Das Problem verschärft sich also noch.

DATENERHEBUNG

- Ziel einer jeden Datenerhebung ist es, auf dem höchstmöglichen Skalenniveau zu messen.
- Außerdem gilt: Soweit möglich sollte jeder Einzelfall dokumentiert werden!
- Beispielstudie: Wird ein spanisches Wort mit dem Auxiliar „ser“ (sein) oder „aver“ (haben) gebildet? Als potentielle Einflussfaktoren werden erhoben
 - Jahrhundert
 - Zustandsveränderung ja/nein
- Datenbasis: 1000 mit Auxiliar gebildete Verbformen pro Jahrhundert (17. und 18. Jahrhundert).

DATENERHEBUNG

17. Jahrhundert		Zustandsveränderung	
		nein	ja
Hilfsverb	ser	150	250
	aver	320	280

18. Jahrhundert		Zustandsveränderung	
		nein	ja
Hilfsverb	ser	50	150
	aver	550	250

Verb	Zust.v.	Jahr	Hilfsv.
1	ja	1654	ser
2	nein	1688	ser
3	nein	1672	aver
4	ja	1753	aver
5	ja	1792	aver
6	nein	1702	aver
7	ja	1634	ser
...
2000	nein	1777	ser

Die Kreuztabellenlösung (links) ist zwar kompakter und übersichtlicher, die Datentabellenlösung (rechts) ist aber viel **flexibler** und **mächtiger**! Außerdem kann die linke Lösung sehr leicht daraus erstellt werden.

SKALENNIVEAUS IN R

- In R gilt jeder numerische Vektor als kardinalskaliert.
 - Das kann aber umkodiert werden (`?factor`).
- Character-Vektoren (Vektoren mit Zeichenketten) sind grundsätzlich nominalskaliert.
- Faktorenvektoren können als „ordered“ (geordnet) behandelt werden, sind aber per default nominalskaliert.
- Liest man eine Datentabelle ein, werden (in der Defaulteinstellung) alle Spalten, in denen Zeichenketten stehen, als Faktoren behandelt.



BEGRIFFE

- Variable
- Ausprägung
- Skalenniveau / Messniveau
- Nominalskala
- Ordinal- / Rangskala
- Intervallskala
- Rational- / Verhältnisskala
- Kardinalskala