

01_einfuehrung

Allgemeines, Ablauf, Statistische Datenanalyse





"Albernes Zauberstabgefuchtel
und kindische Hexereien wird es
hier nicht geben. Daher erwarte
ich von den wenigsten
Begeisterung für die schwierige
Lehre und exakte Kunst der
Statistik."

Warum überhaupt Statistik...

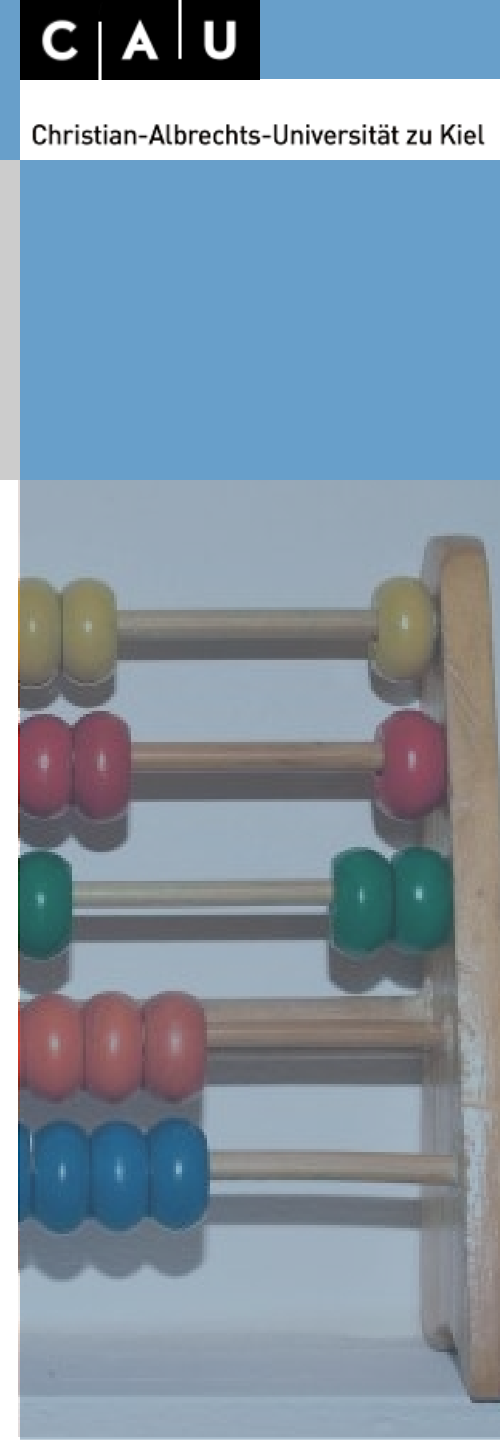
Für Sie:

Es wird gemacht! Wenn Sie es verstehen wollen, müssen Sie sich damit beschäftigen!

Für die Archäologie als Disziplin:

Mit Statistik wird alles einfacher!

- Aussagen werden verständlicher und vor allem nachvollziehbar
- Aussagen sind statistisch richtig oder falsch, unabhängig vom Renomee des Forschers
- Aussagen und Daten werden vergleichbar
- Materialkenntnis für induktives Verstehen von archäolog. Zusammenhängen braucht Jahrzehnte, Erlernen statistischer Verfahren nur mehrere Semester



Figures don't lie, but liars figure.
Samuel Clemens (alias Mark Twain)

Statistik ist nur korrekt, wenn Frage, Ansatz und Methode korrekt sind:

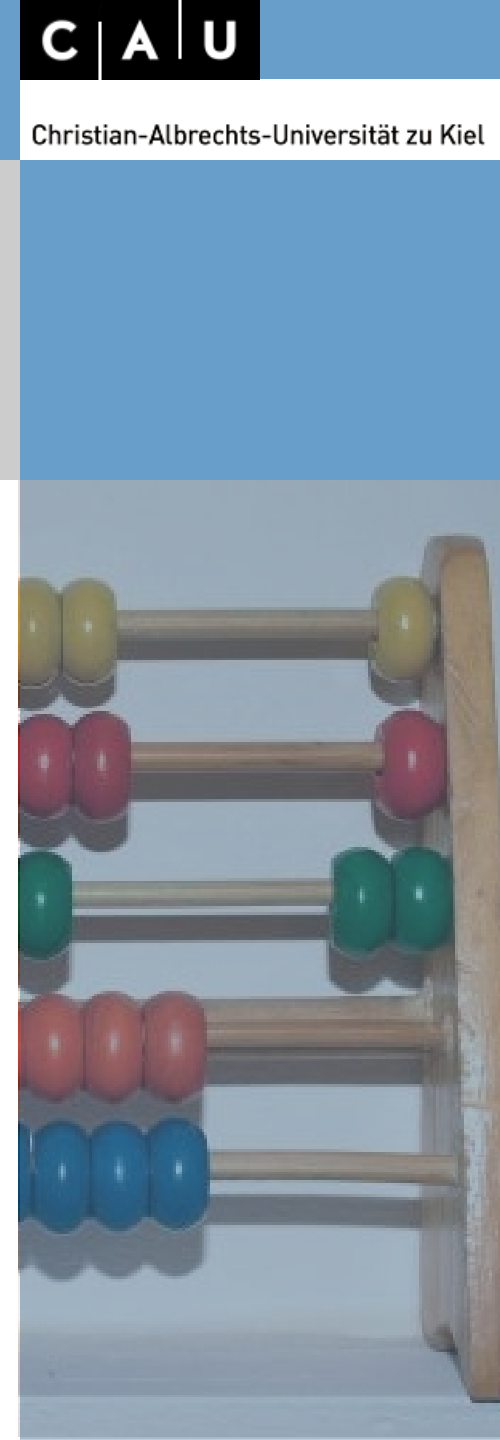
Bsp: Ist soziale Stratifizierung anhand von Metallbeigabe erkennbar?
Oder anhand von Schmuckbeigabe? Was, wenn dies vom (nicht
erkannten) Geschlecht abhängt...

Messen und vor allem Codieren von Messungen erfordern immer subjektive Entscheidungen:

Gründe für die Entscheidungen sind häufig nicht nachzuvollziehen →
Subjektiver Einfluss

Statistik um der Statistik willen?

Ein archäologisch erkennbarer Sinn muß hinter der Untersuchung
stecken. Und die Ergebnisse der Untersuchung muß archäologisch
überprüfbar sein.



Statistikprogramm R: Geschichte (nach Theus)

R ist “Nachfolger von S bzw. S-Plus”

- S Historie:

- 1976-1980: S-Version 1; (Entwicklung bei AT&T Labs) Sammlung von Fortran Routinen
- 1980-1984: S-Version 2 Portierung auf UNIX, Definition der Kommando-Sprache
- 1988-1991: S-Version 3 Portierung auf C, Objekte, Modelle
- 1999-heute:

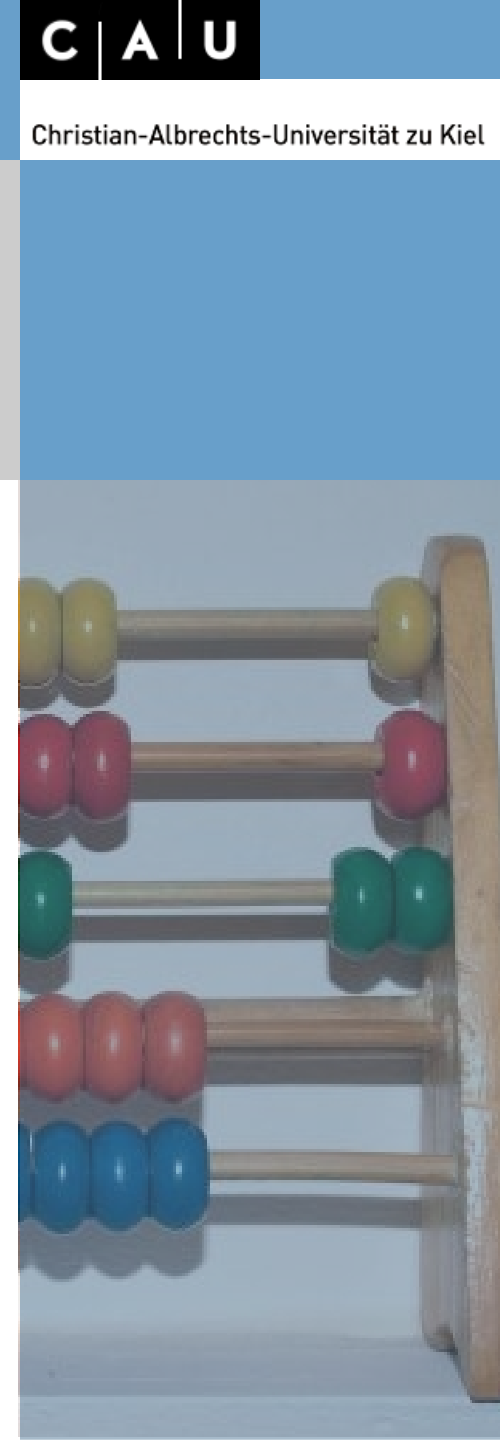
S-Version 4 Verbesserte Objektstruktur

(parallel dazu die kommerzielle Version S-Plus)

- R-Historie

- frühe 90er: Entwicklung in Neuseeland (R. Ihaka, R. Gentleman) Lisp basiert, einzige Plattform war der Mac
- mitte 90er: Erweiterung auf andere Plattformen
- ende 90er: Verteilte Entwicklung durch das R-Core-Team

- R-Core-Team: z.Z. 17 Entwickler aus der ganzen Welt.
- R-”Spezialisten”: z.Z. ca. 50 Contributor
- Entwickler von R-Paketen: hunderte ... täglich mehr



Statistikprogramm R: Warum?

Open Source

Frei zugänglicher Quellcode: Überprüfbarkeit des Programms
Kostenlos: Sie müssen keine horrenden Summen ausgeben oder Raubkopien anfertigen.

Referenz der benutzten Algorithmen

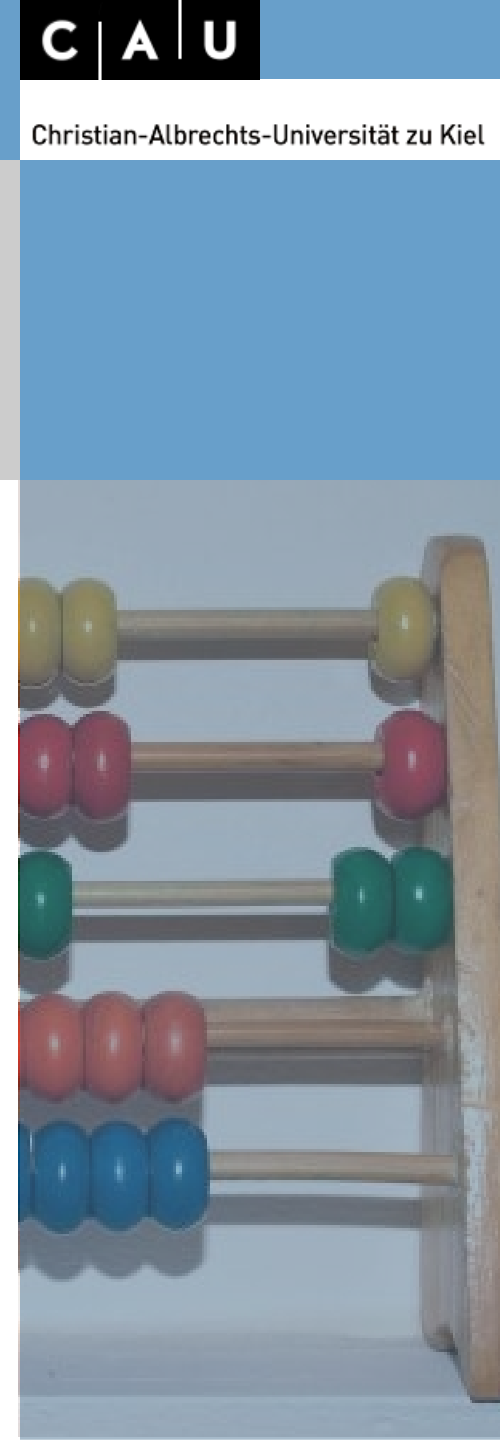
Wissenschaftlich zitierfähig

Mächtigkeit

Das Programm kann alles! Ehrlich!

Verbreitung

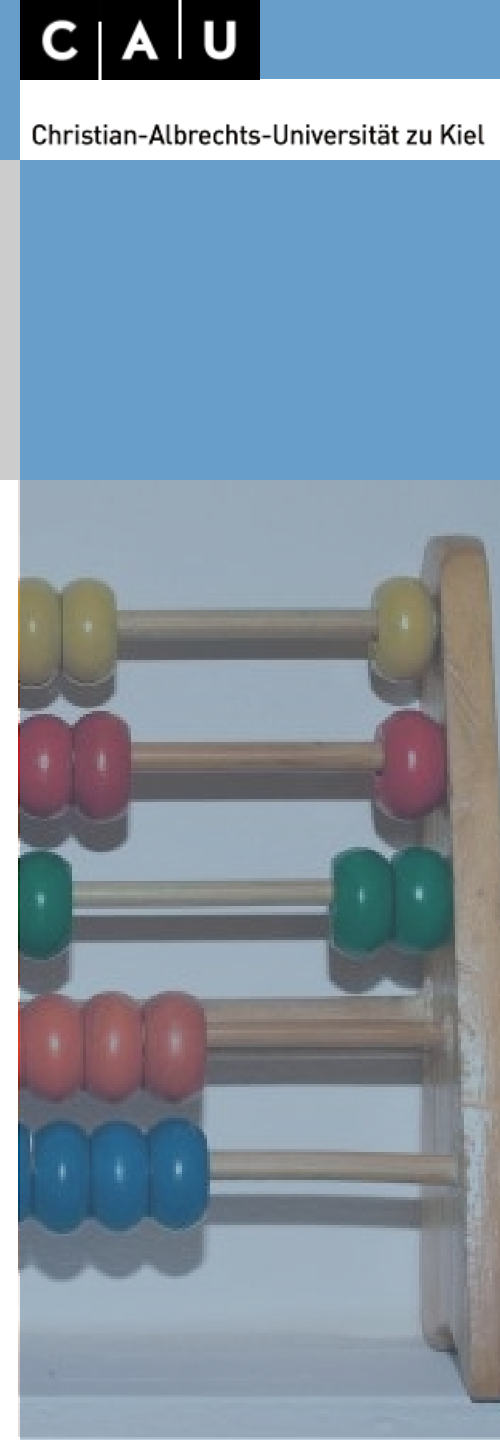
Läuft auf allen Betriebssystemen
Wird im wissenschaftlichen Umfeld (vor allem Naturwissenschaften) häufig genutzt



Statistikprogramm R: Warum?

Nachteile

- Kommandozeile: ungewohnt (neue Art, mit dem Rechner zu arbeiten)
- GUIs sehen unterschiedlich aus
- Englischkenntnisse erforderlich
- Namen von Funktionen und Parametern müssen behalten werden: heißt es `col.names`, `colnames` oder `header`?
- Dokumentation teilweise nicht sehr intuitiv: man sollte wissen, was man sucht

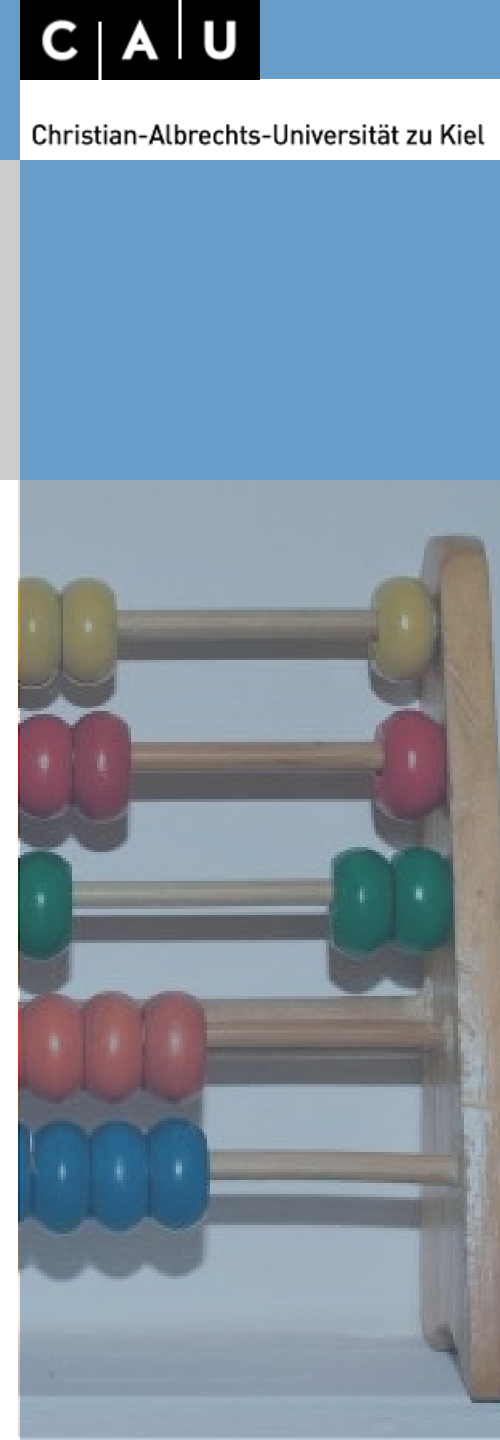


Grundlegende Literatur

Stephan Shennan, Quantifying Archaeology.
Unser Lehrbuch!

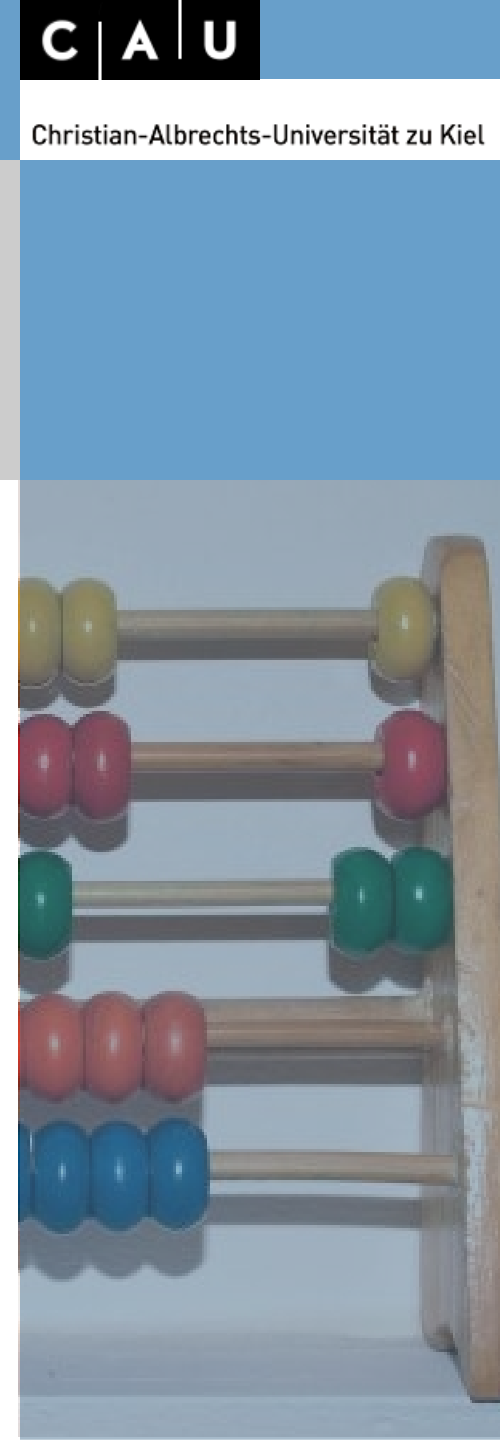
Dubravko Dolić, Statistik mit R.
John Verzani, Using R for Introductory Statistics.
R-spezifische (einführende) Statistikbücher

Lothar Sachs, Angewandte Statistik. Methodensammlung mit R.
<http://www.springerlink.com/content/I32744/>



Weitere Literatur

- M. Fletcher/G. R. Lock, Digging Numbers: Elementary Statistics for Archaeologists. Oxford Univ. Comm. Arch. Monogr. 332 (Oxford 2005).
- M. J. Baxter, Exploratory Multivariate Analysis in Archaeology (Edinburgh 1994).
- M. Baxter, Statistics in Archaeology (London 2003).
- P. Ihm, Statistik in der Archäologie: Probleme der Anwendung, allgemeine Methoden, Seriation und Klassifikation. Archaeo-Physika 9 (Köln 1978).
- J. Bortz, Statistik für Sozialwissenschaftler⁴ (Berlin u. a. 1993).



Datum	Sitzung	Thema	Shennan Kap
28.10.10	1	Allgemeines	1+2
04.11.10	2	Einführung R	
11.11.10	3	Explorative Statistik	3
18.11.10	4	Deskriptive Statistik	4
25.11.10	5	Nicht-parametrische Tests	5
02.12.10	6	Chi-Quadrat und Zusammenhangsmaße	7
09.12.10	7	Stichprobe und Population, Wahrscheinlichkeitstheorie	5, 6, 14
16.12.10	8	Verteilungen	6
23.12.10		frei	
30.12.10		frei	
06.01.11		frei	
13.01.11	9	Parametrische Tests	6
20.01.11	10	Regression und Korrelation	8
27.01.11	11	Clusteranalyse	11
03.02.11	12	Korrespondenzanalyse	13
10.02.11	13	Klausur	

Arten von Statistik

Deskriptive Statistik:

Zusammenfassung und Beschreibung von Daten mittels von Kenngrößen (Mittelwert, Standardabweichung etc.)

(graphische Darstellung):

Darstellung und Zusammenfassung von Daten mittels Diagrammen (Balken-, Kreisdiagrammen etc.)

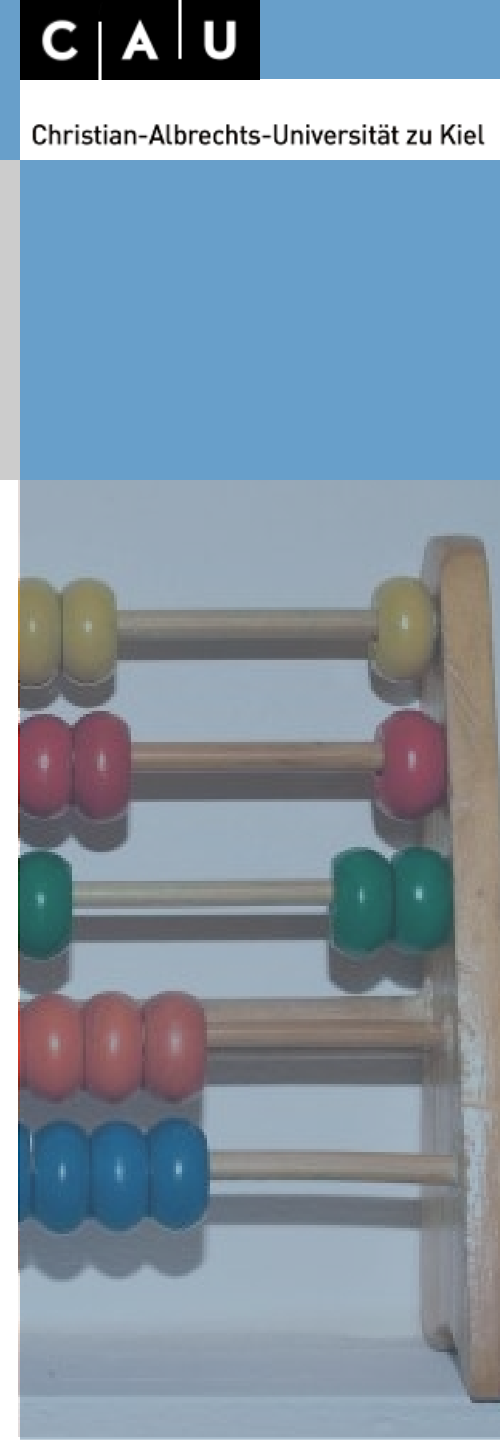
Dient zur Beschreibung wie auch zur Mustererkennung, daher Zwischenstellung

Explorative Statistik:

Darstellung und Zusammenfassung von Daten, um dahinter liegende Muster zu erkennen (z.B. Korrespondenzanalyse)

Induktive Statistik (Inferenzstatistik):

Testen von Hypothesen über die Daten, die dann statistisch signifikant belegt oder verworfen werden (Statistische Tests, z.B. Chi-Quadrat-Test)



Daten, Variablen, Merkmale

Variable oder Merkmal:

Das, was gemessen oder untersucht werden soll

Bsp: Körpergröße

Merkmalsträger

Das, dessen Merkmal gemessen wird.

Bsp: Ich als Besitzer einer Körpergröße, Gräber, Personen...

Variablenausprägungen oder Merkmalswerte (oder einfach Werte):

Tatsächlich gemessene Eigenschaften.

Bsp: Meine Körpergröße beträgt 1.81 m.

Diskrete Variablen:

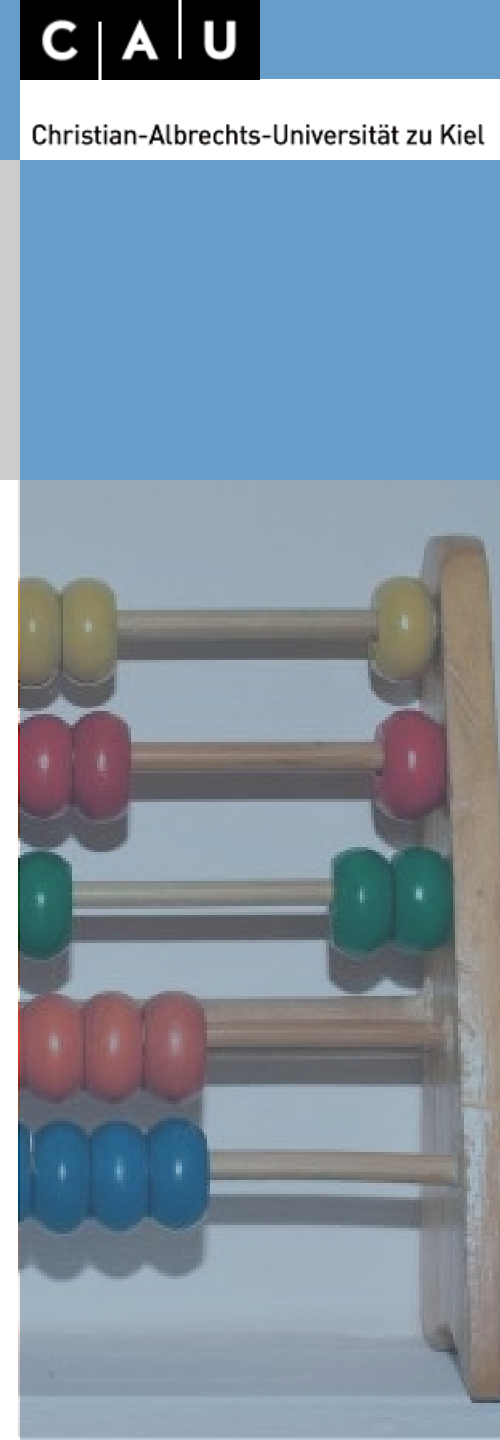
Variable, die nur bestimmte Werte ohne Zwischenwerte annehmen können.

Bsp: Einkommen, Anzahl von Keramikobjekten, Geschlecht (?)

Stetige Variablen:

Variablen, die jeden Wert und jeden Zwischenwert annehmen können.

Bsp: Angaben wie Körpergröße, Temperaturen, Prozentangaben



Arten von Statistik

Univariate Statistik:

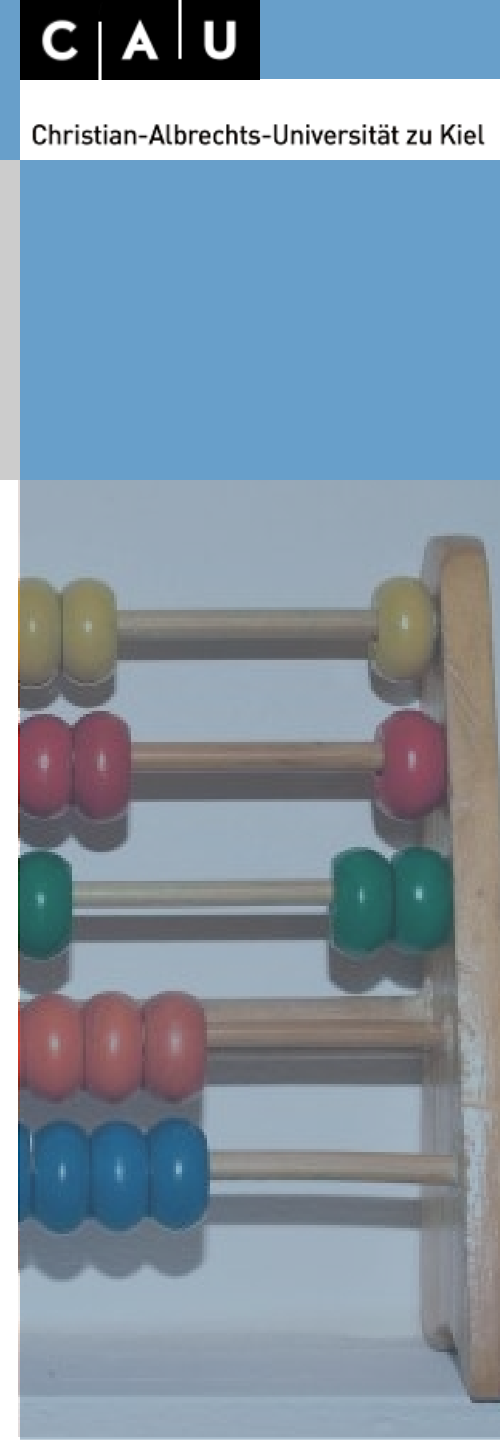
Nur eine Variable ist beteiligt.
z.B. Gewicht von Bronzebeilen

Bivariate Statistik:

Zwei Variablen sind beteiligt, es interessiert ihr Zusammenhang.
z.B. Verhältnis von Länge zu Breite von Bronzebeilen

Multivariate Statistik:

Mehr als zwei Variablen sind beteiligt, es interessiert ihr Zusammenhang.
z.B. Ort des Fundes von Beilen (Grab, Depot, Siedlung) in Abhängigkeit von ihrer chemischen Zusammensetzung (Anteil Kupfer, Zinn, Arsen, Blei etc.)



Unabhängige – Abhängige Variable

Unabhängige Variable:

Die vermutete Ursache eines Zusammenhangs.

Abhängige Variable:

Die vermutete Wirkung der unabhängigen Variable in einem Zusammenhang.

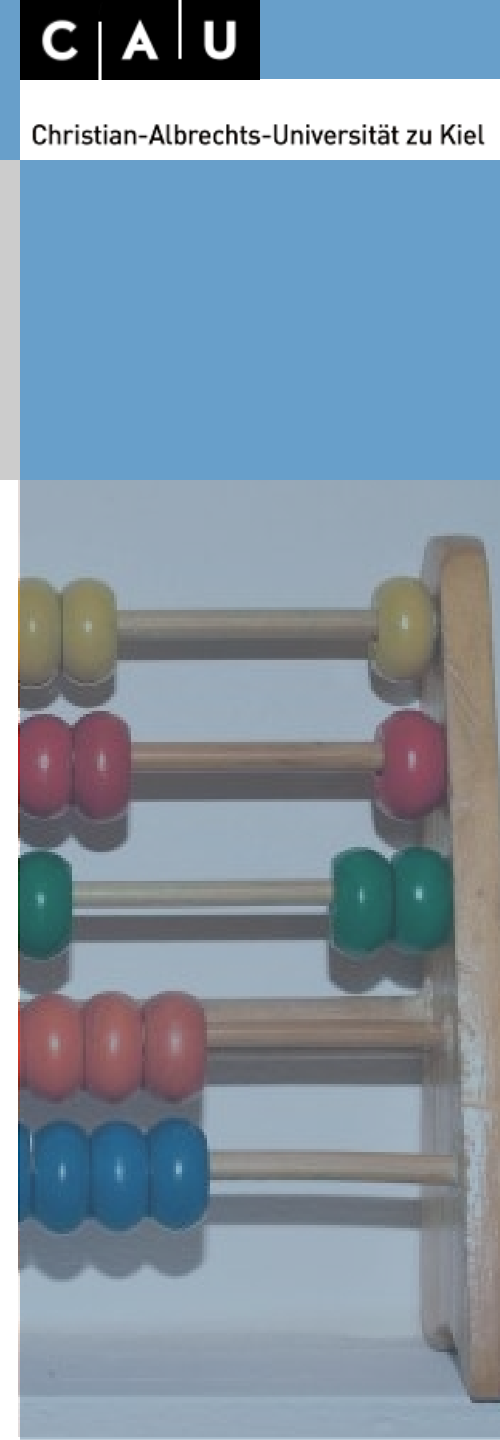
Beispiel:

Anzahl von Perlen in einem Grab	↔	Geschlecht der Bestatteten
Abhängig		Unabhängig

Hypothese: Die Anzahl der Perlen in einem Grab ist abhängig vom Geschlecht der Bestatteten.

Kann (muss) nicht immer festgelegt werden!

Bsp: Volumen und Höhe eines Gefäßes...



Stichprobe und Grundgesamtheit

Grundgesamtheit

Menge aller Merkmalsträger, die für die Untersuchung relevant sind.

Stichprobe

Auswahl von Merkmalsträgern nach bestimmten Kriterien (z.B. Repräsentativität), die an Stelle der Grundgesamtheit untersucht werden

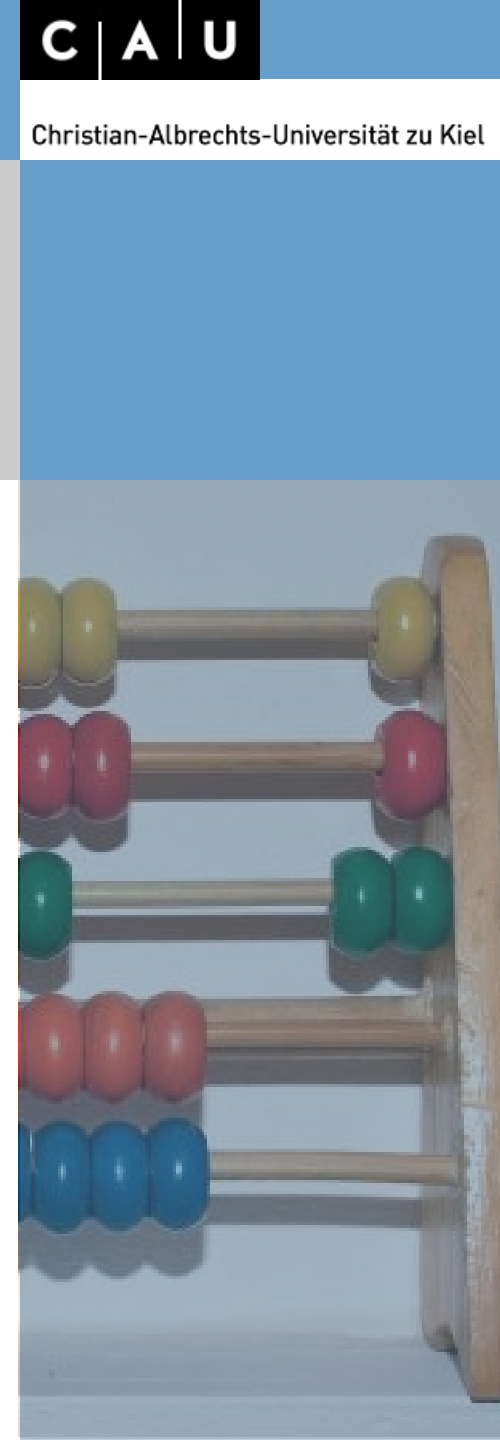
Bsp. Sonntagsfrage:

Grundgesamtheit: Alle Bundesbürger, die eine Politische Meinung haben.

Stichprobe: Diejenigen, die vom Umfrageunternehmen befragt wurden

Totalerhebung ↔ Teilerhebung

In der Archäologie gibt es immer nur Teilerhebungen mittels einer Stichprobe! Die Grundgesamtheit bleibt immer fraglich!



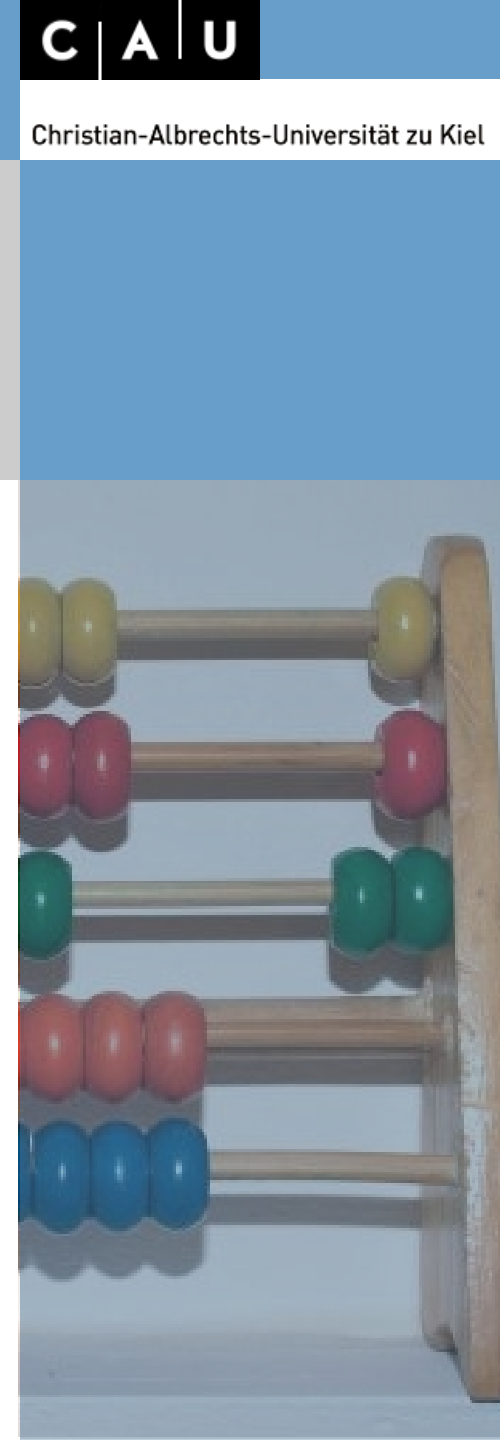
Unabhängige – Abhängige Stichprobe

Abhängige Stichproben:

Das Ergebnis der einen Stichprobe ist teilweise von der anderen abhängig (Untersuchung von Patienten vor/nach Einnahme eines Medikamentes)

Unabhängige Stichproben:

Das Ergebnis der einen Stichprobe ist nicht von der anderen abhängig (Untersuchung von zwei Gräberfeldern)



Gliederung einer statistischen Untersuchung

Datenerhebung

z.B. Grabung, Literaturrecherche

Datenerfassung

z.B. Eingabe der Daten in eine Datenbank

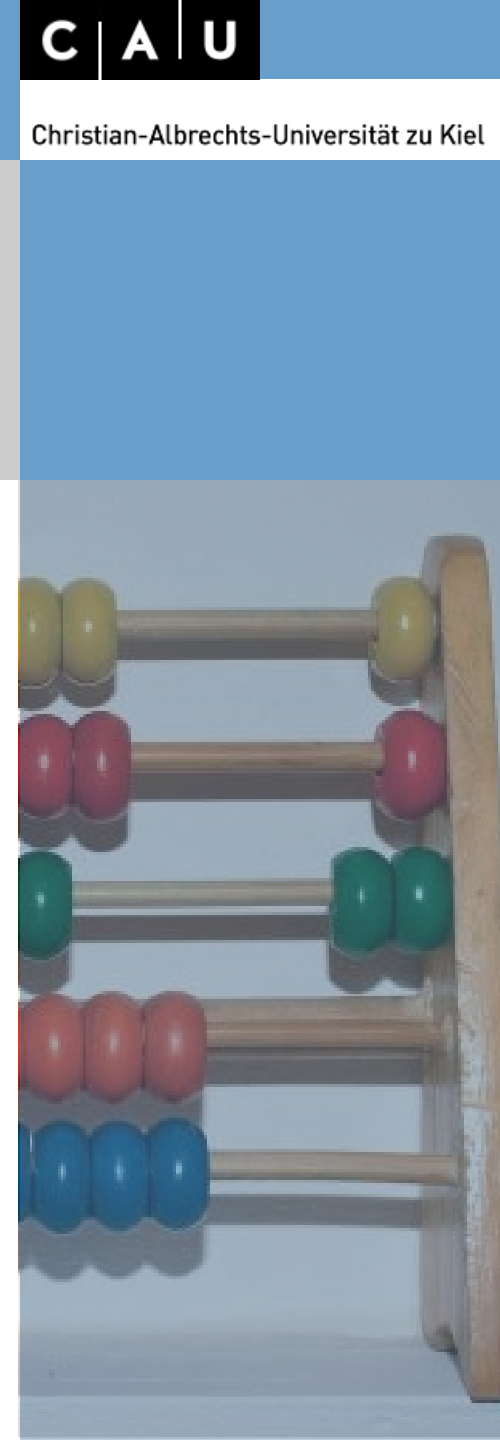
Datenaufbereitung und Datendarstellung

z.B. Eliminieren von Null-Werten

Deskriptive Statistik: Diagramme, um sich Überblick zu verschaffen, Kennwerte erheben (Mittelwert etc.)

Datenanalyse

Prüfen von Hypothesen über die Grundgesamtheit, Inferenzstatistik



Datenskalierungsebenen

nominal:

Variablenkategorien stehen in keinem definierten Verhältnis zueinander; nur Zählen ist erlaubt (Beispiel: Geschlecht)

ordinal:

Variablenkategorien sind vergleichbar, sie unterscheiden sich in Hinsicht auf ihre Ausprägung [Größe/Stärke/Intensität]; ihre Rangfolge ist bestimmbar (Beispiel: Erhaltungsbedingungen)

metrisch:

Variable folgt einem festen Messsystem; alle Rechenoperationen sind möglich. Zu unterscheiden sind:

1. Intervallskala:

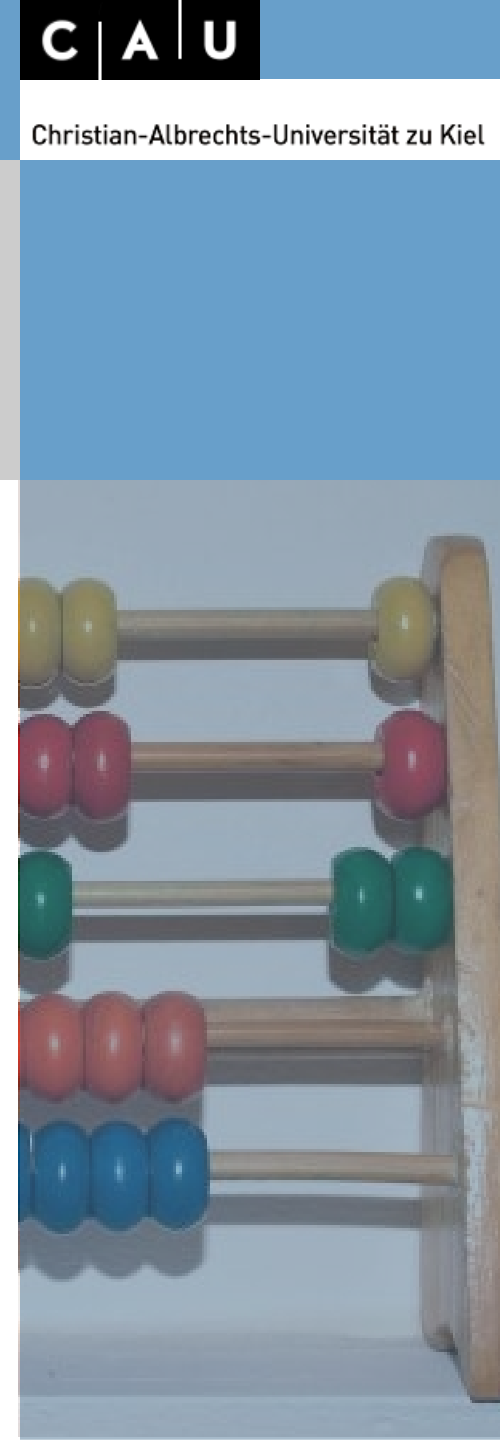
Die Variable besitzt einen willkürlich gewählten Nullpunkt ($^{\circ}\text{C}$)

2. Verhältnisskala (auch Ratioskala):

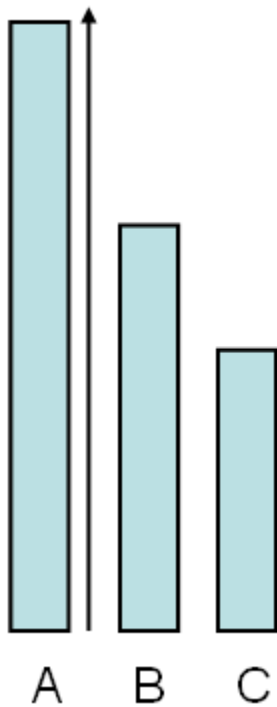
Die Variable besitzt einen absoluten Nullpunkt ($^{\circ}\text{K}$)

manchmal auch benutzt: Absolutskala:

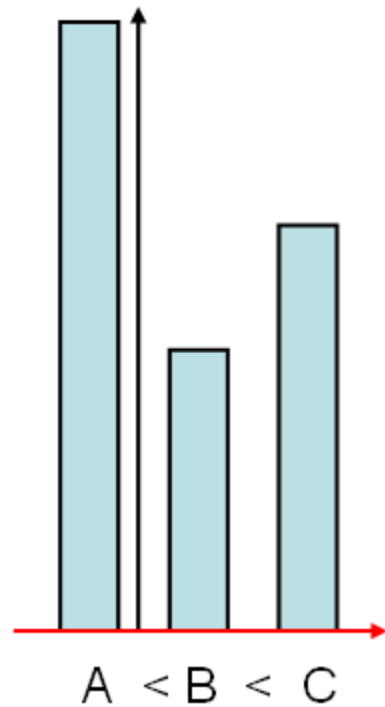
Zählwerte (Einwohneranzahl)



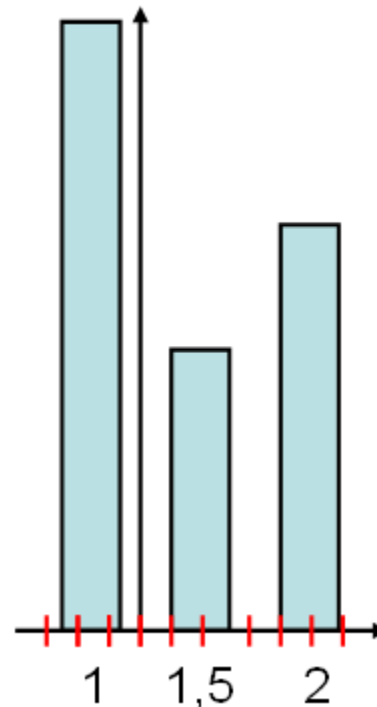
Nominalskala



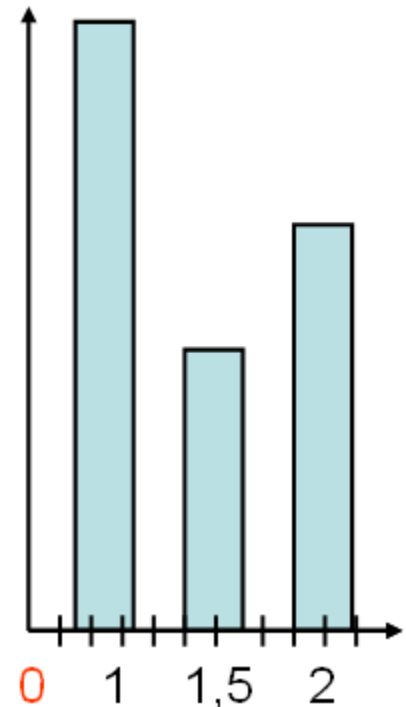
Ordinalskala



Intervallskala

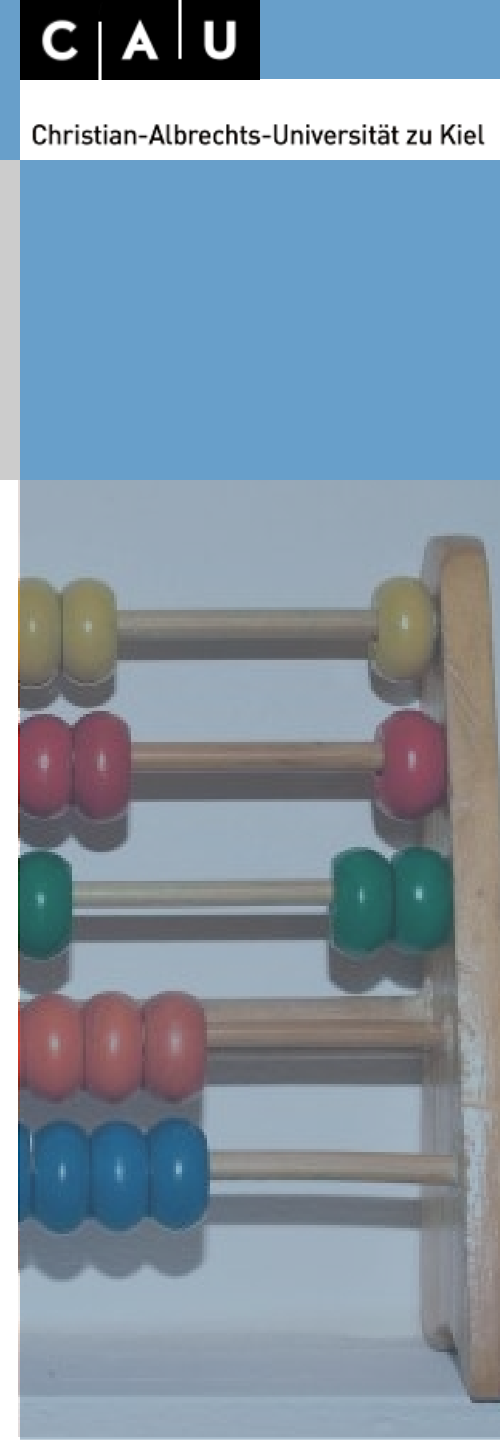


Verhältnisskala



Datenskalierungsebenen

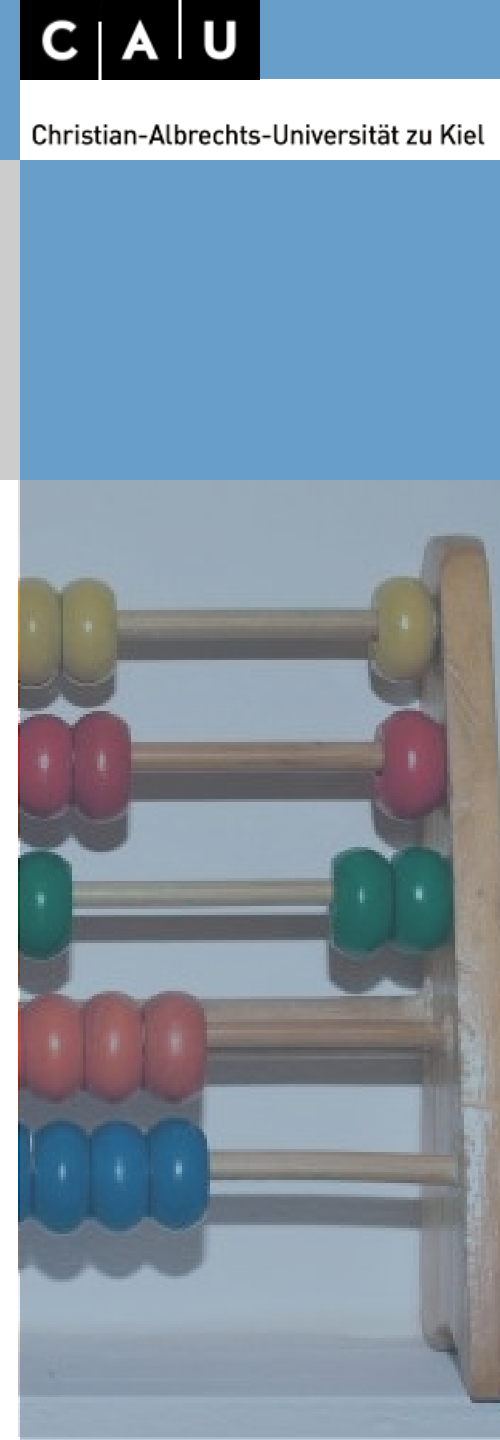
Skalenart	Mögliche Aussagen	Beispiele
Nominalskala	Gleichheit, Verschiedenheit	Telefonnummern, Krankheitsklassifikationen, Keramiktypen
Ordinalskala	Größer-kleiner-Relationen	Windstärken, Akademische Ränge, Reichtumsklassen, Stratigraphie
Intervallskala	Gleichheit von Differenzen	Temperatur (in °C), Kalenderzeit
Verhältnisskala	Gleichheit von Verhältnissen	Längenmessung, Gewichtsmessung, Gefäßhöhe



Datenskalierungsebenen

Skalenart	Sinnvoll interpretierbare Berechnungen			
	Auszählen	Ordnen	Differenz bilden	Quotient bilden
Nominalskala	ja	nein	nein	nein
Ordinalskala	ja	ja	nein	nein
Intervallskala	ja	ja	ja	nein
Verhältnisskala	ja	ja	ja	ja

nach Bortz 2005



Datenskalierungsebenen

Änderungen der Skalierungsebene:

nach unten:

Immer möglich.

Bsp: Klassifizierung von Messergebnissen (klein-mittel-groß)

Aber: Führt aber zu Informationsverlust.

nach oben:

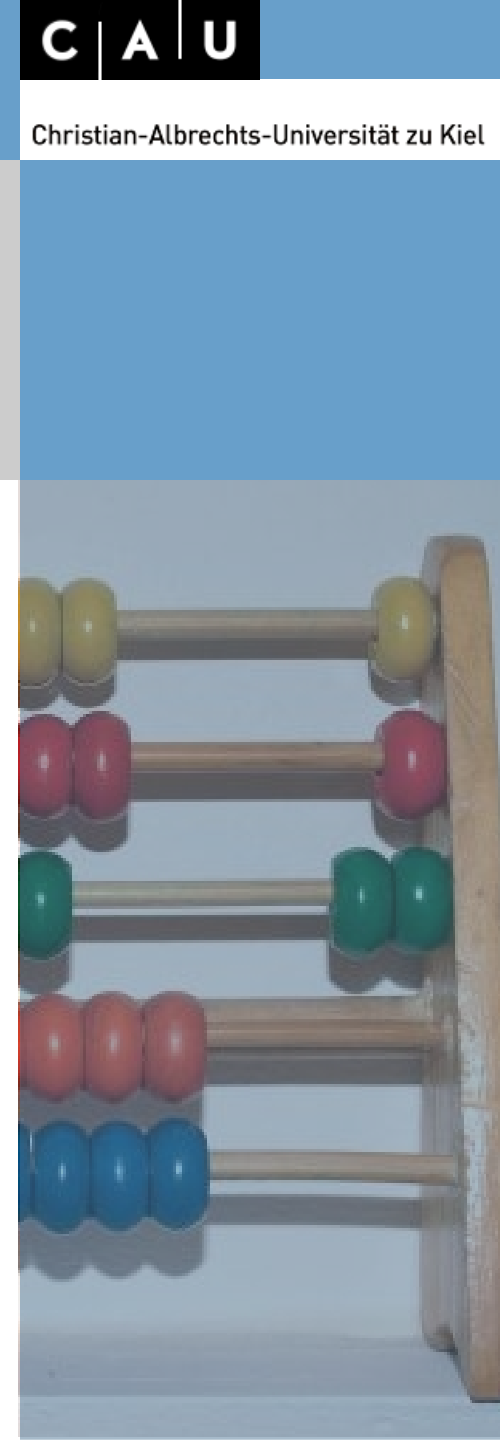
manchmal möglich.

Bsp: Statt Klassifizierung der Keramik in Grob-Feinware Messung der Korngröße

Aber: Führt zu größerem Datenaufkommen und Komplexität der Messung.

Fazit:

Für Analysen ist die am besten geeignete Skalierungsebene zu wählen. Da sich aber Änderungen in den Anforderungen ergeben können, Faustregel: Immer eine Ebene genauer aufnehmen, als man am Ende auswerten will... (wie gesagt, nur eine Faustregel)

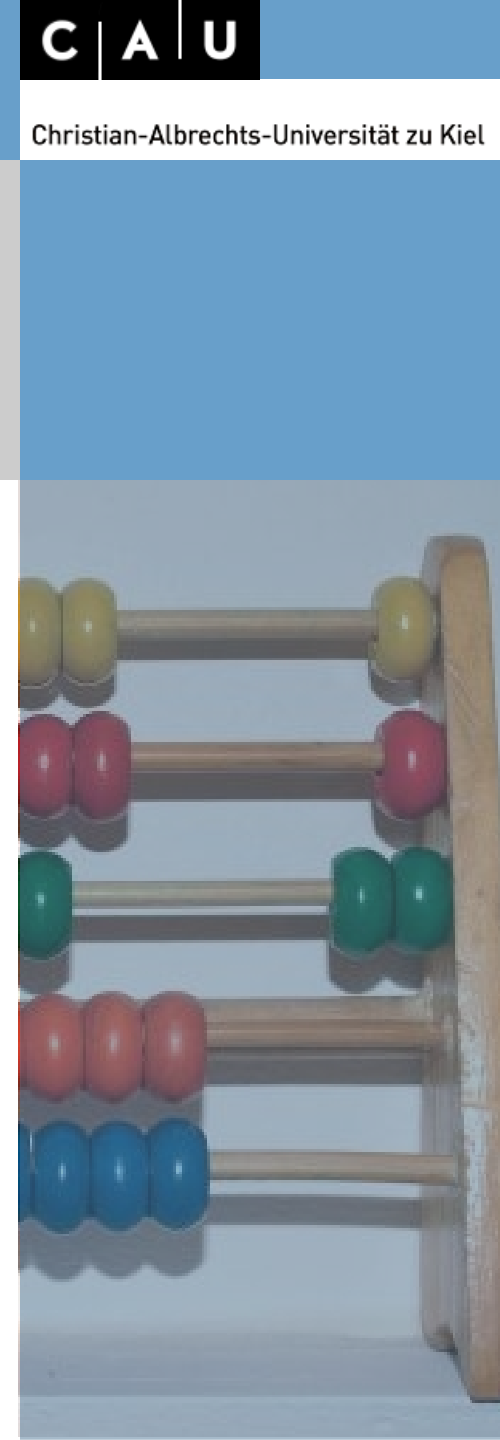


Datenaufnahme: Urliste

Einfache Auflistung von Daten.

Bsp:

154
167
187
165
190
176
167
156
154
165
167
171
154

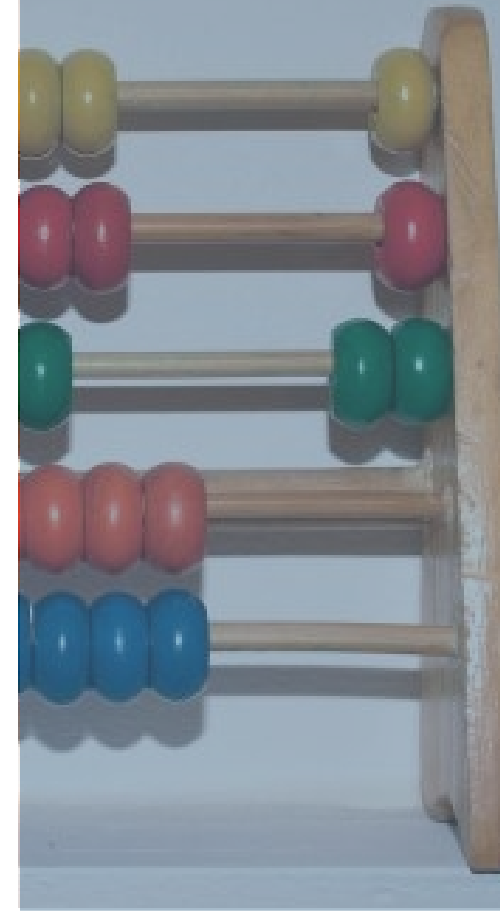


Datenaufbereitung: Datenmatrix

Tabellarische Zusammenfassung mehrerer Variablen je Merkmalsträger

Bsp:

Name	Körpergröße	Geschlecht
Hannah	154	2(weiblich)
Leon	167	1(männlich)
Lukas	187	1
Leonie	165	2
Luka	190	1
Lea	176	2
Lena	167	2
Mia	156	2
Tim	154	1
Fynn	165	1
Anna	167	2
Emily	171	2
Felix	154	1

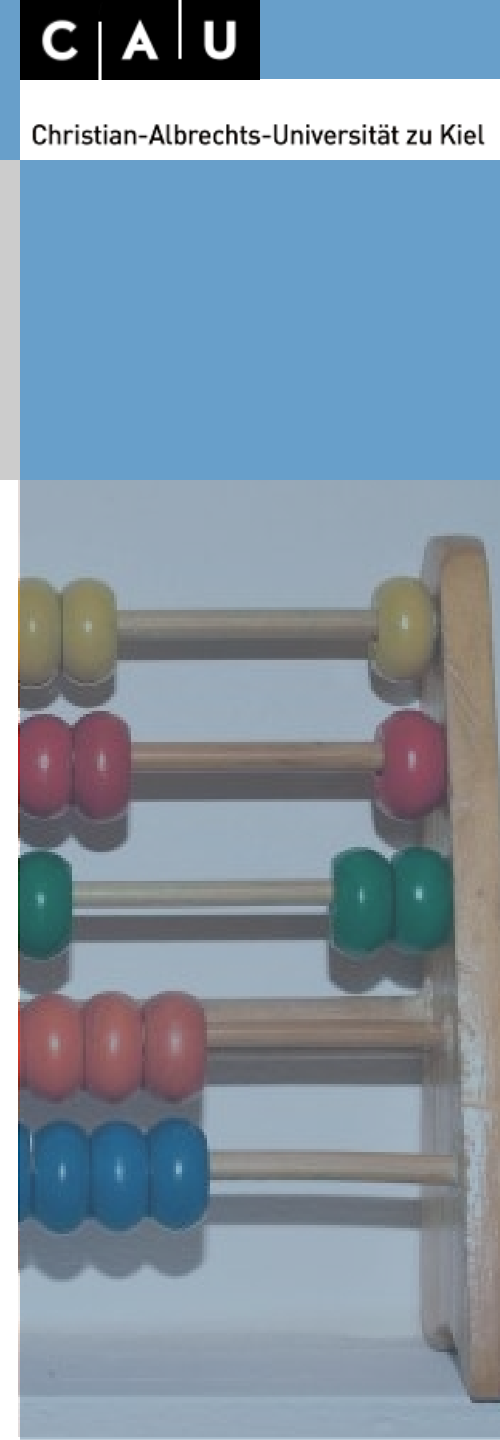


Datenaufbereitung: “Strichliste” / Häufigkeitsverteilung

Tabellarische Zusammenfassung mehrerer Merkmalsträger je Merkmalsausprägung

Bsp:

es interessiert ihr Zusammenhang.		
Körpergröße	„Striche“	Anzahl
154		3
156		1
167		3
165		2
171		1
176		1
187		1
190		1



Datenaufbereitung: Klassifizierung

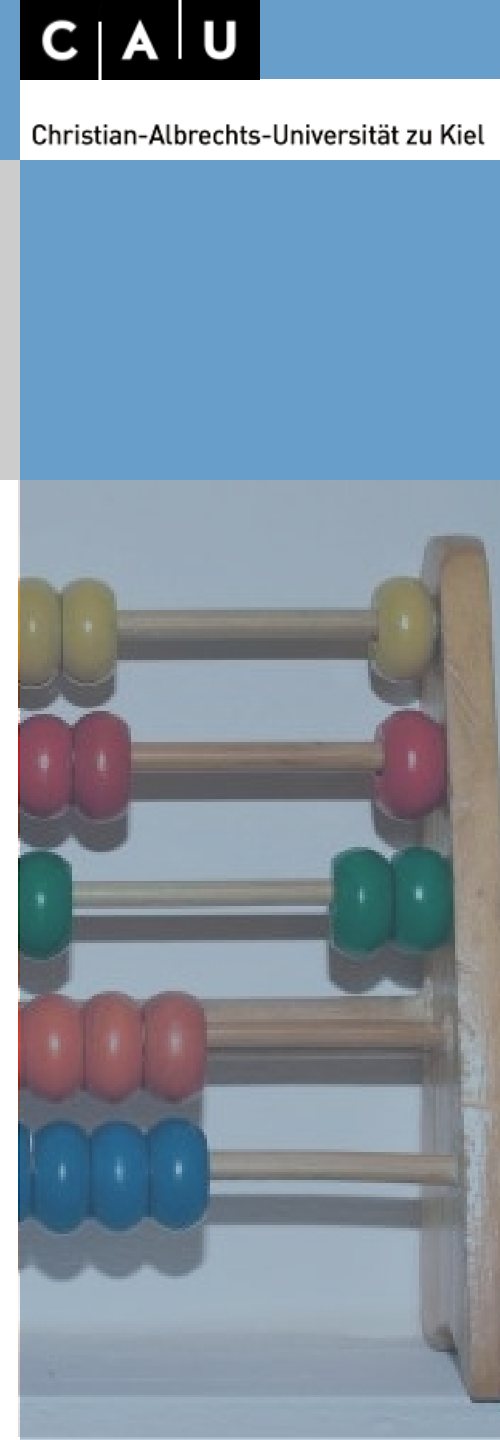
Tabellarische Zusammenfassung mehrerer Merkmalsträger je einer Klasse von Merkmalsausprägung

Bsp:

Körpergröße	„Striche“	Anzahl
<150		0
150-159		4
160-169		5
170-179		2
180-189		1
>190		1

Klassenbreite hier 10 cm

Faustregeln: ca. 8 – 12 Klassen oder
Klassenzahl $k \approx \sqrt{n}$ in diesem Fall also $k \approx \sqrt{13} = 3,605551275 \approx 4$



Formelzeichen

ungefähr

$$a \approx b$$

Anzahl

n

Summe

$$\sum_{i=1}^n x_i$$

Bedeutet

$$x_1=0, x_2=4, x_3=5, x_4=2, x_5=1, x_6=1; n=6$$

Genauso Produkt

$$x_1+x_2+x_3+x_4+x_5+x_6=13=\sum_{i=1}^n x_i$$
$$\prod_{i=1}^n x_i = x_1 * x_2 * x_3 * x_4 * x_5 * x_6 = 0$$

Körpergröße

„Striche“

Anzahl

<150

0

150-159

||||

4

160-169

|||||

5

170-179

||

2

180-189

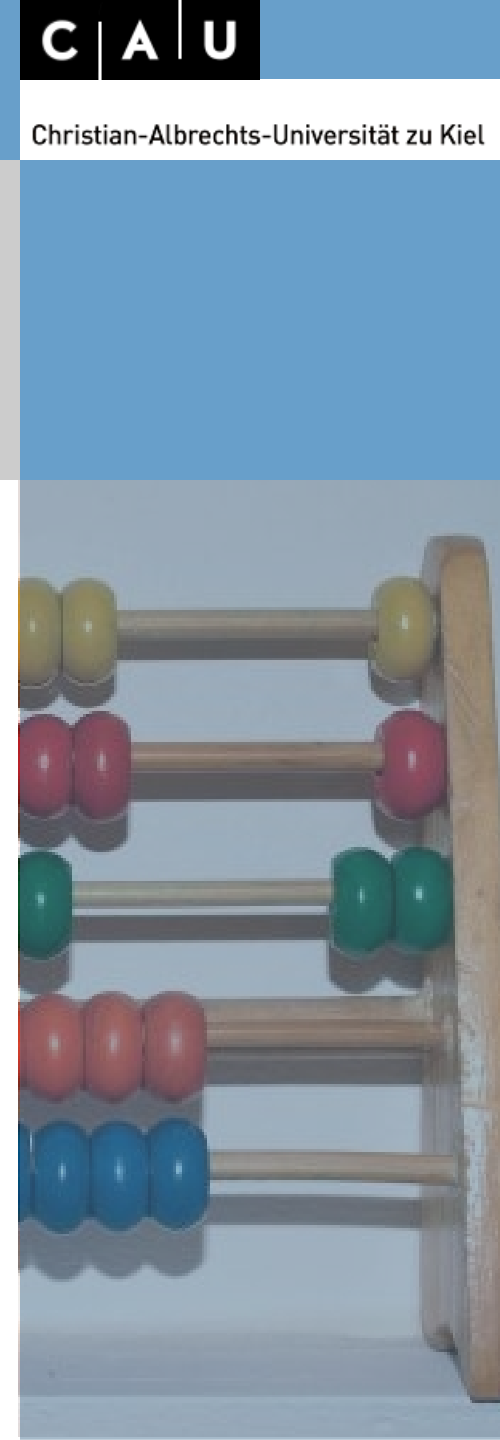
|

1

>190

|

1



Bsp. Arithmetisches Mittel

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

Beobachtungen : $x_i := \{154, 167, 187, 165, 190, 176, 167, 156, 154, 165, 167, 171, 154\}$

Anzahl Beobachtungen : $n = 13$

$$\bar{x} = \frac{154 + 167 + 187 + 165 + 190 + 176 + 167 + 156 + 154 + 165 + 167 + 171 + 154}{13}$$

$$\bar{x} = \frac{2173}{13}$$

$$\bar{x} = 167,153846154$$

Aufgabe: Beschreibung der Kursteilnehmer

Bilden Sie Gruppen je zwei Rechner und führen Sie die Daten nach der Erhebung zusammen:

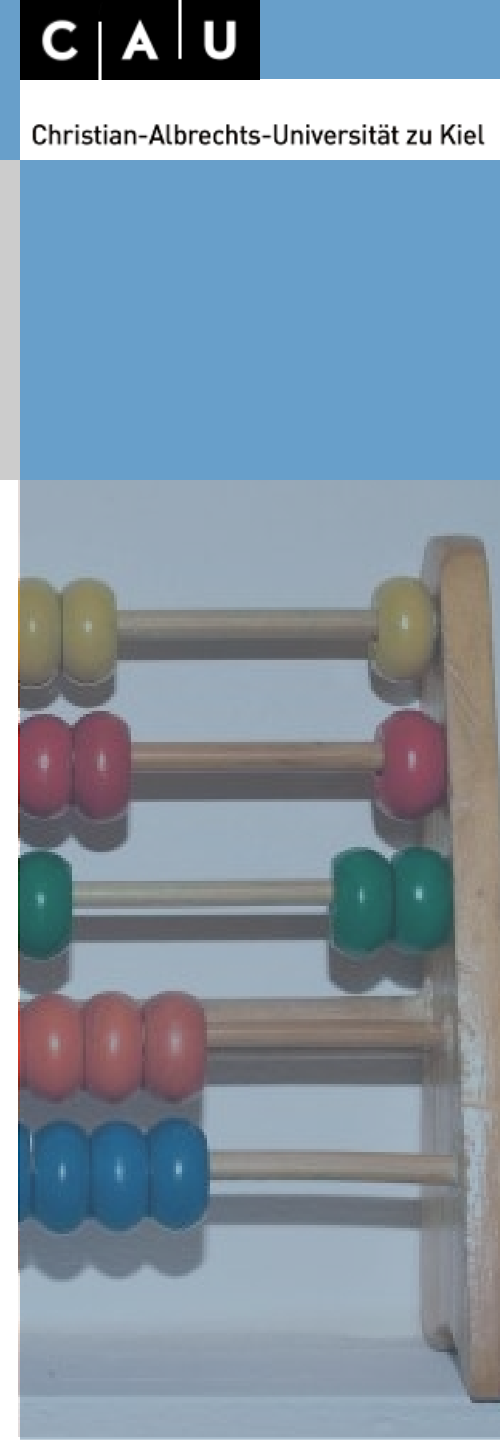
Erhebung der Daten in einzelnen Gruppen

- A) Email, Computername (PC-Lab*).
- B) Geschlecht, Semesterzahl.
- C) Alter, PC ja/nein.
- D) Schuhgröße, Bargeld.

E) angestrebter Abschluss, Körpergröße.

F) Betriebssystem, Geschlecht.

Tragen Sie die Datenmatrix zusammen (Sie dürfen gern ein Tabellenkalkulationsprogramm Ihrer Wahl benutzen), bestimme Sie jeweils das Skalenniveau und präsentieren Sie in 10 Minuten die Ergebnisse.



Aufgabe: Beschreibung der Kursteilnehmer

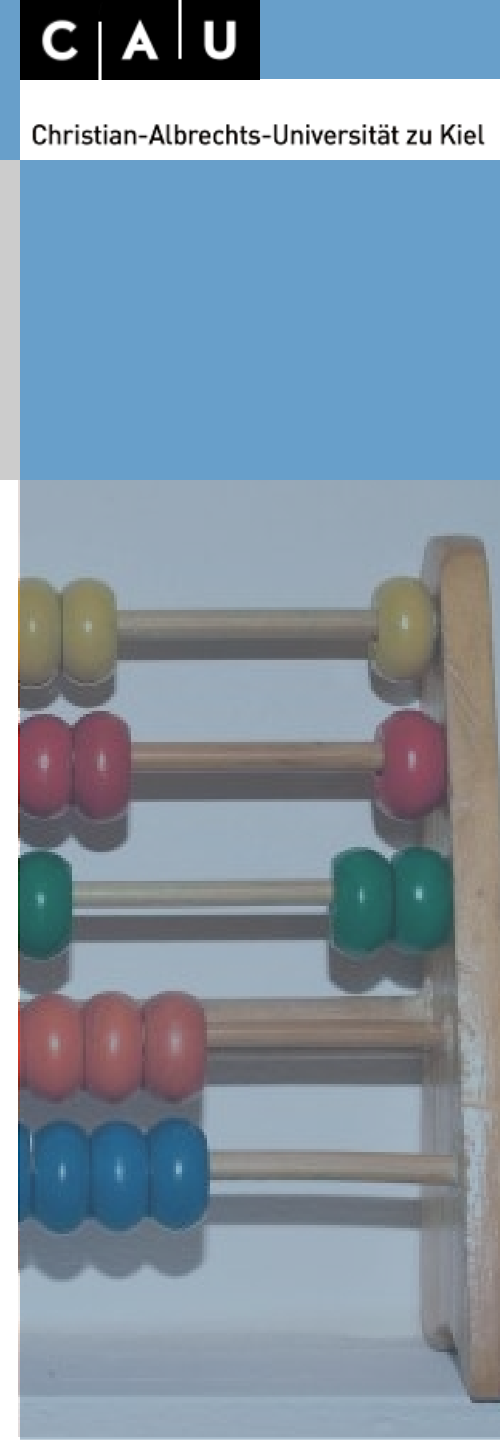
Fazit:


Erhebungen benötigen eine Systematik. Diese wird am besten an einer kleinen Stichprobe entwickelt und ausgewertet ('Pilotstudie'). Anschließend wird die gesamte geplante Erhebung nach einem vorab festgelegten einheitlichem Schema gleichermaßen für alle Fälle durchgeführt.

Es gibt sehr unterschiedliche Arten von Informationen (Skalenniveau).

Jede Art von Informationen geht mit unterschiedlichen Darstellungsmöglichkeiten einher. Es gibt ungeschickte - ungeeignete Arten der Darstellung.

Je nach Aussage gibt es für die gleichen Daten unterschiedliche Arten, sie darzustellen (dazu mehr in der nächsten Sitzung).





Nächste Sitzung 4. November:
Einführung in R

Lesen Sie bitte Shennan Kapitel 1 und 2...