

05_nichtparametrische_tests

Hypothesentests, Kolmogorov-Smirnov, Mann-Whitney-U



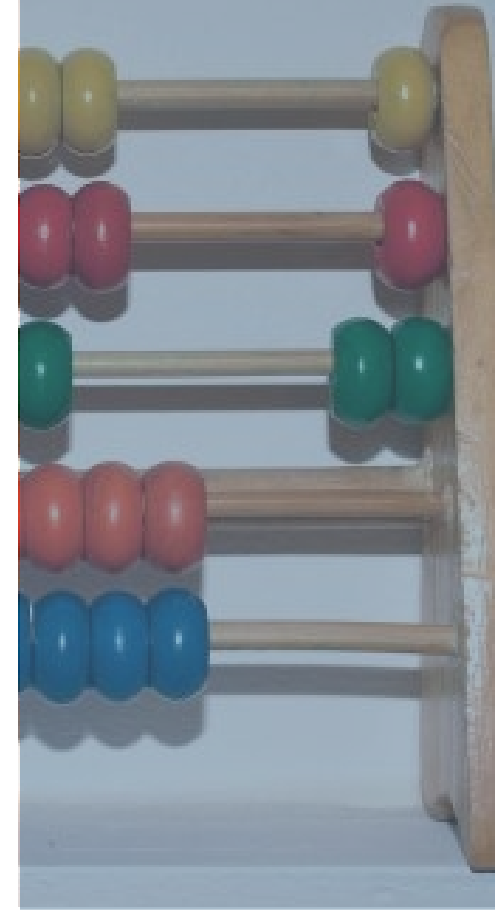
Induktive oder Inferenzstatistik

Dient zur Ableitung von Eigenschaften der Grundgesamtheit aus einer Stichprobe

Die Ergebnisse sind immer statistisch ;-)

d.h., alle Aussagen treffen mit einer bestimmten Wahrscheinlichkeit zu, können aber auch mit einer bestimmten Wahrscheinlichkeit falsch sein.

Grundlage ist die Wahrscheinlichkeitstheorie (Stochastik)



Grundgesamtheit und Stichprobe [1]

Zur Wiederholung:

Grundgesamtheit

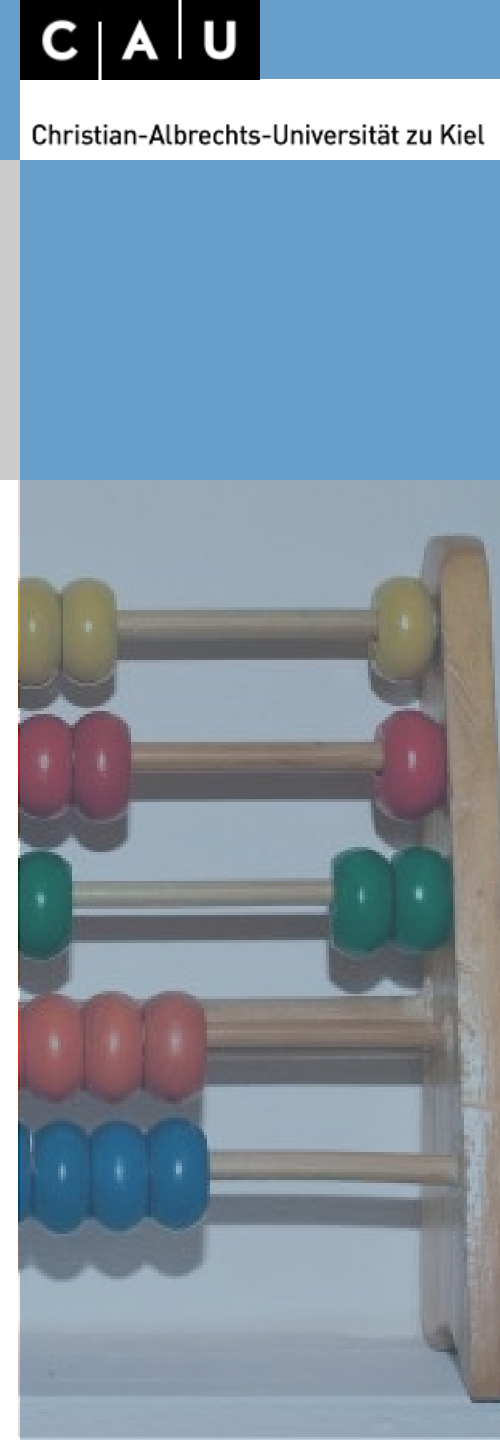
Menge aller Merkmalsträger, die für die Untersuchung relevant sind.

Stichprobe

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

Den Unterschied sollte man sich immer bewußt halten

Archäologen arbeiten (fast) nie mit der Grundgesamtheit



Grundgesamtheit und Stichprobe [2]

Eigenschaften der Grundgesamtheit: Parameter

Parameter gibt es immer, sie sind feste Werte, aber sie sind unbekannt, meist auch nicht überprüfbar

Bsp:

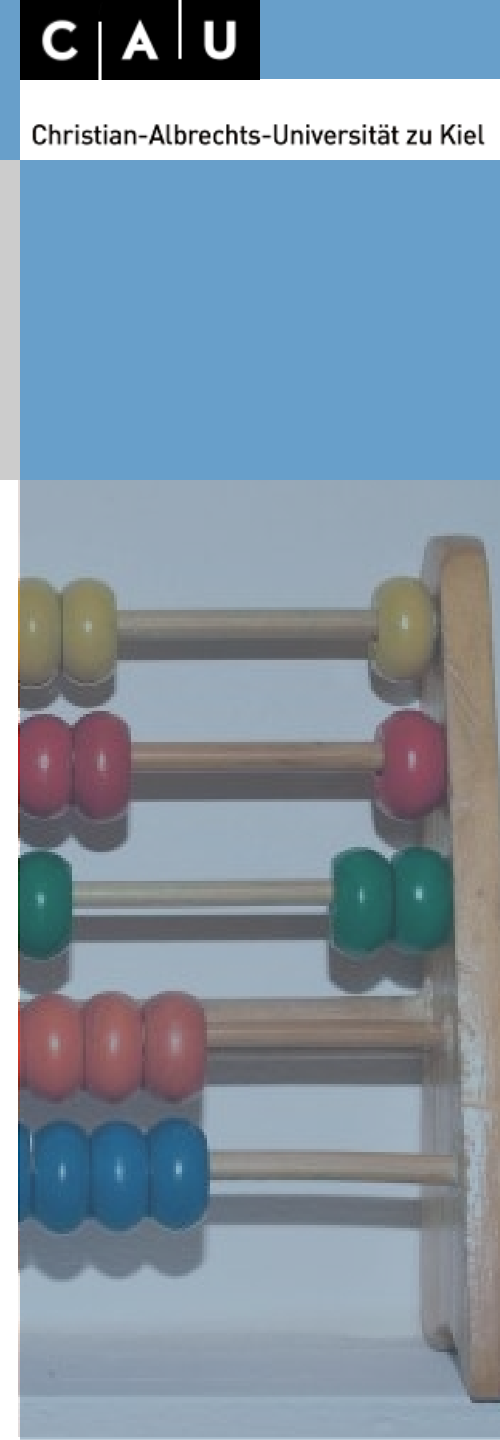
μ : *Arithm. Mittel der Grundgesamtheit*

\bar{x} : *Arithm. Mittel der Stichprobe*

σ : *Standardabweichung der Grundgesamtheit*

s : *Standardabweichung der Stichprobe*

In statistischen Tests können immer nur die Eigenschaften der Stichprobe geprüft werden. Daher hängt die Qualität der Aussage immer von der Wahl der Stichprobe ab (Repräsentativität)



Hypothesen-Test

Überprüfung von Annahmen über die Grundgesamtheit

Es wird eine Annahme (Hypothese) über die Grundgesamtheit aufgestellt und dann anhand der Stichprobe auf ihre Wahrscheinlichkeit getestet.

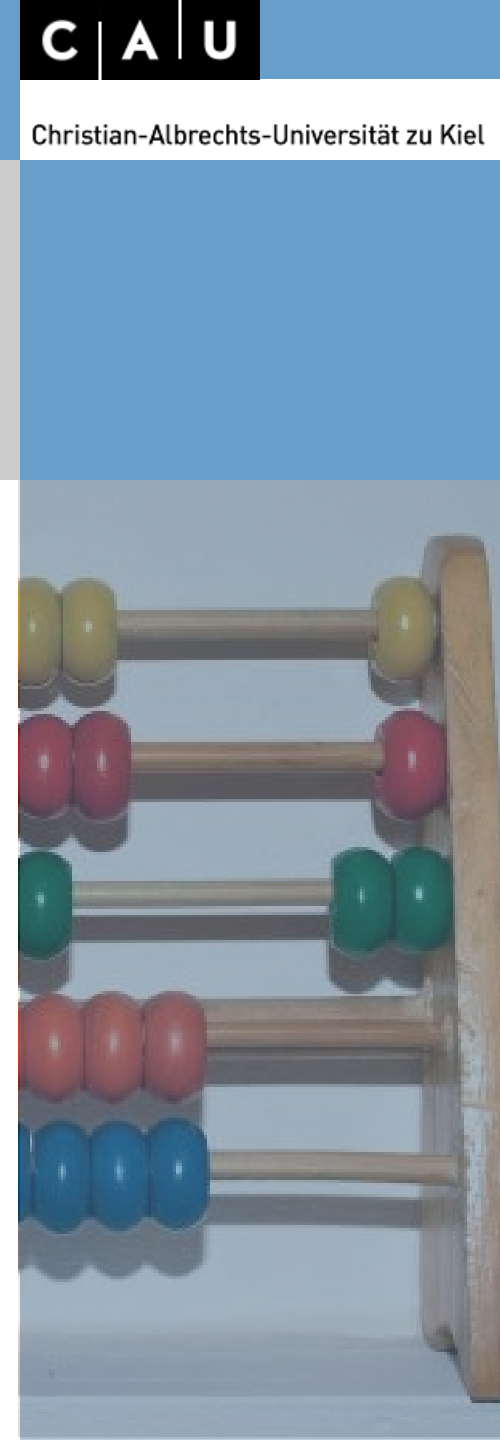
Gängige Fragen:

Wie hoch ist die Wahrscheinlichkeit, dass zwei oder mehr Stichproben von unterschiedlichen Grundgesamtheiten stammen?

(Ist die Ausstattungssitte mit Grabbeigaben zwischen Männern und Frauen so unterschiedlich, dass sich hier zwei unterschiedliche gesellschaftliche Gruppen zeigen?)

Wie hoch ist die Wahrscheinlichkeit, dass eine gegebene Stichprobe von einer Grundgesamtheit mit bestimmten Eigenschaften stammt?

(Ist die Anzahl der Grabbeigaben zufällig oder gibt es ein Muster?)



Null-Hypothese [1]

Validierung durch Falsifikation

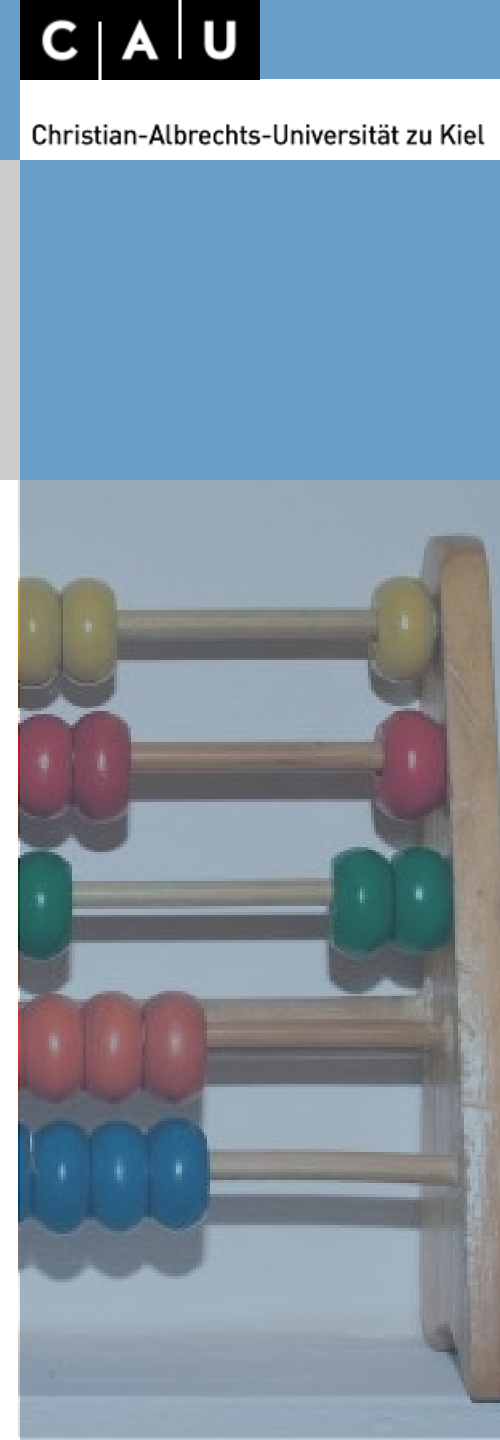
In statistischen Tests prüft man meist nicht die Aussage, die man erwartet, sondern versucht, die Aussage zu widerlegen, die man nicht erwartet: Null-Hypothese. Diese sagt meist aus, das ein Zusammenhang oder ein Unterschied **nicht** besteht und die Verteilung der Beobachtungen lediglich zufällig sind.

Bsp: Wir wollen, ob die Beigabenverteilung zwischen Männern und Frauen unterschiedlich ist.

$$\begin{aligned} H_0 &: \text{Die Beigabenverteilung ist gleich} \\ H_1 &: \text{Die Beigabenverteilung ist unterschiedlich} \end{aligned}$$

Grund:

1. Es ist (technisch) leichter, zu beweisen, das etwas nicht stimmt, als zu beweisen, das etwas stimmt.
2. Eine Nullhypothese ist oft einfacher zu formulieren (Wie genau ist denn die Beigabenverteilung unterschiedlich?). Sie sagt noch nichts über die Natur des Zusammenhangs/Unterschiedes aus.



Null-Hypothese [2]

Ablauf eines Statistischen Testes

Aufstellen einer Alternativhypothese:

Die Beigabenverteilung zwischen Männern und Frauen ist unterschiedlich

Aufstellen der Nullhypothese:

Die Beigabenverteilung zwischen Männern und Frauen ist gleich

Testen der Nullhypothese

Wenn Ergebniss des Testes signifikant:

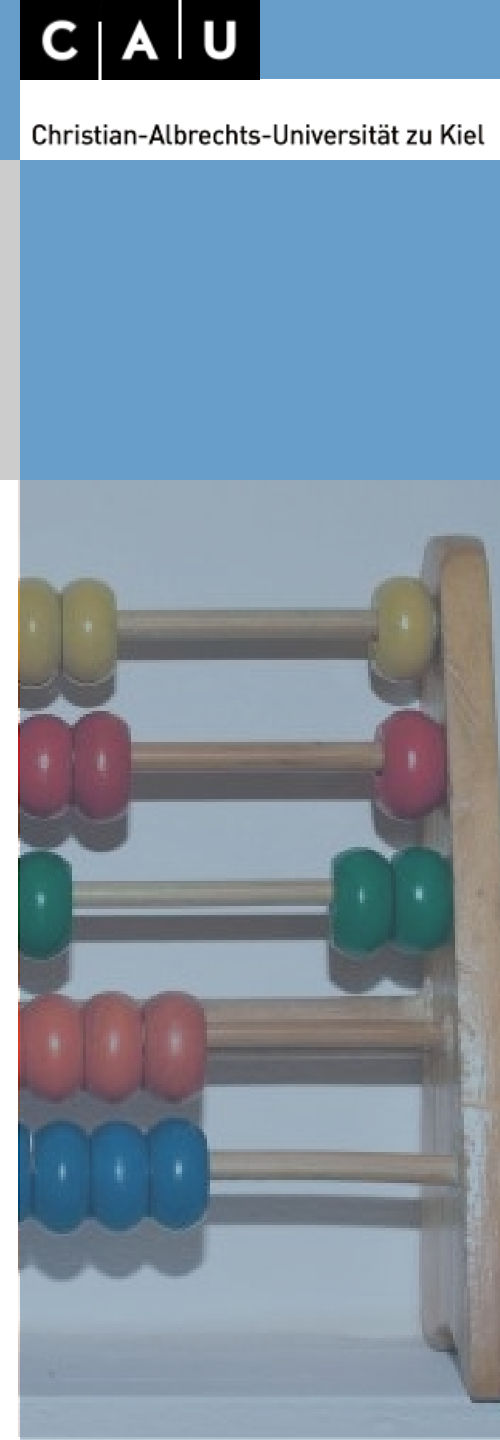
Ablehnen der Nullhypothese, Wahl der Alternativhypothese

Die Beigabenverteilung zwischen Männern und Frauen ist unterschiedlich

Wenn Ergebniss des Testes nicht signifikant:

Die Nullhypothese kann nicht abgelehnt werden

Es kann nicht gesagt werden, ob die Beigabenverteilung zwischen Männern und Frauen unterschiedlich ist



Einseitige/zweiseitige Fragestellung

one-tailed oder two-tailed

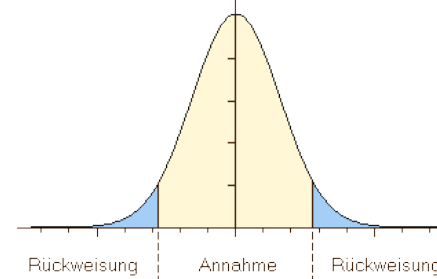
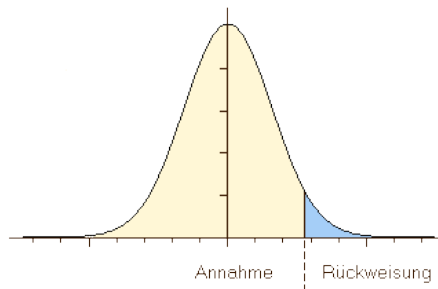
Je nach Frage können sich unterschiedliche Alternativ-Hypothesenanzahlen ergeben

Bsp:

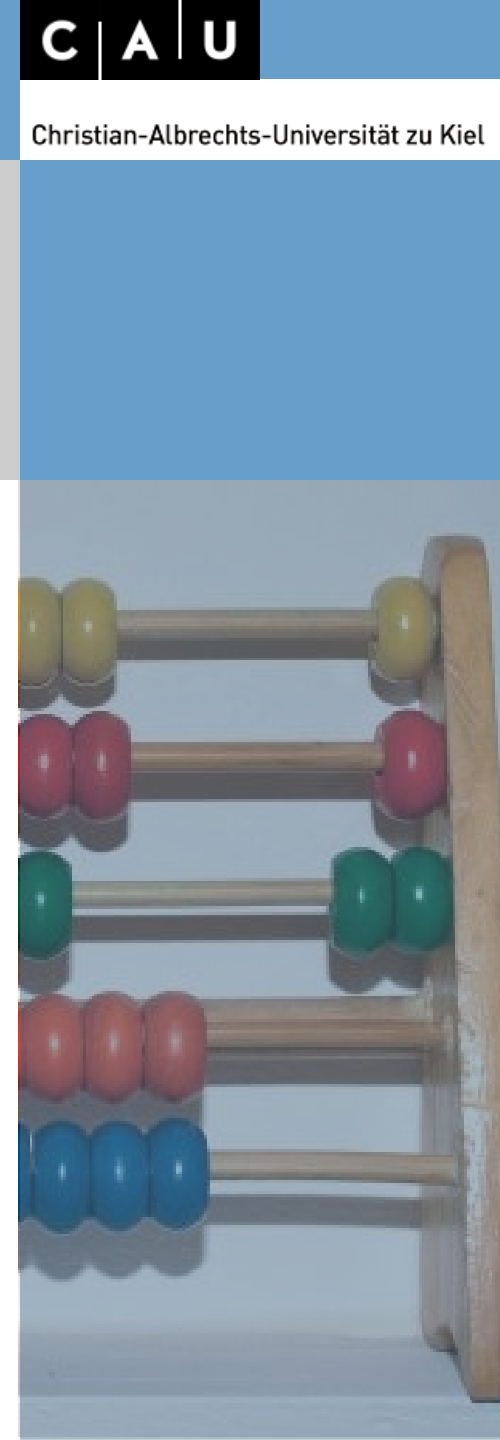
Ist die Beigabenzahl bei Frauen höher als bei Männern?
Einseitige Frage, nur ja oder nein (one-tailed).

Ist die Beigabenzahl bei Frauen anders als bei Männern?
Zweiseitige Frage, kleiner – gleich – größer möglich (two-tailed).

Daher werden bei Stat. Tests oft zwei Signifikanzen angegeben.



Quelle: http://www.statistics4u.info/fundstat_germ/cc_test_one_two_sided.html



Stat. Signifikanz

Wie wahr ist wahr?

Statistische Signifikanz ist im Grunde ein Maß dafür, wie wahrscheinlich ein Irrtum ist.

Auf Basis der Signifikanz wird die Null-Hypothese verworfen und die Alternativ-Hypothese gewählt... oder auch nicht.

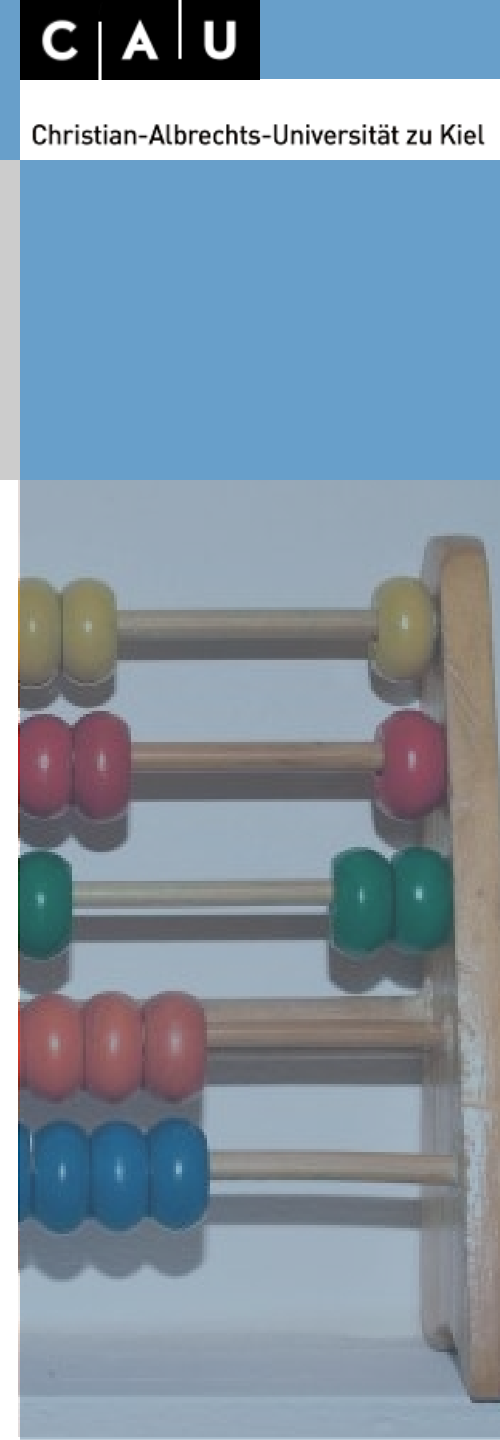
Es gibt klassische Grenzwerte für Signifikanz (Signifikanzniveaus):

0.05: statistisch signifikant, mit 95% Wahrscheinlichkeit ist die Entscheidung korrekt

0.01: sehr signifikant, mit 99% Wahrscheinlichkeit ist die Entscheidung korrekt

0.001: hochsignifikant, mit 99,9% Wahrscheinlichkeit ist die Entscheidung korrekt

Meist mit α oder p-Wert (p-value) bezeichnet



α - und β -Fehler [1]

Wenn die Statistik mal nicht stimmt

Es gibt zwei Arten von möglichen Fehlern:

Die Nullhypothese wird abgelehnt, obwohl sie wahr ist

Fehler 1. Art, falsch positiv oder α -Fehler

Das Ergebnis eines Schwangerschaftstests ist falsch positiv, wenn er eine Schwangerschaft anzeigt, obwohl keine Schwangerschaft vorliegt.

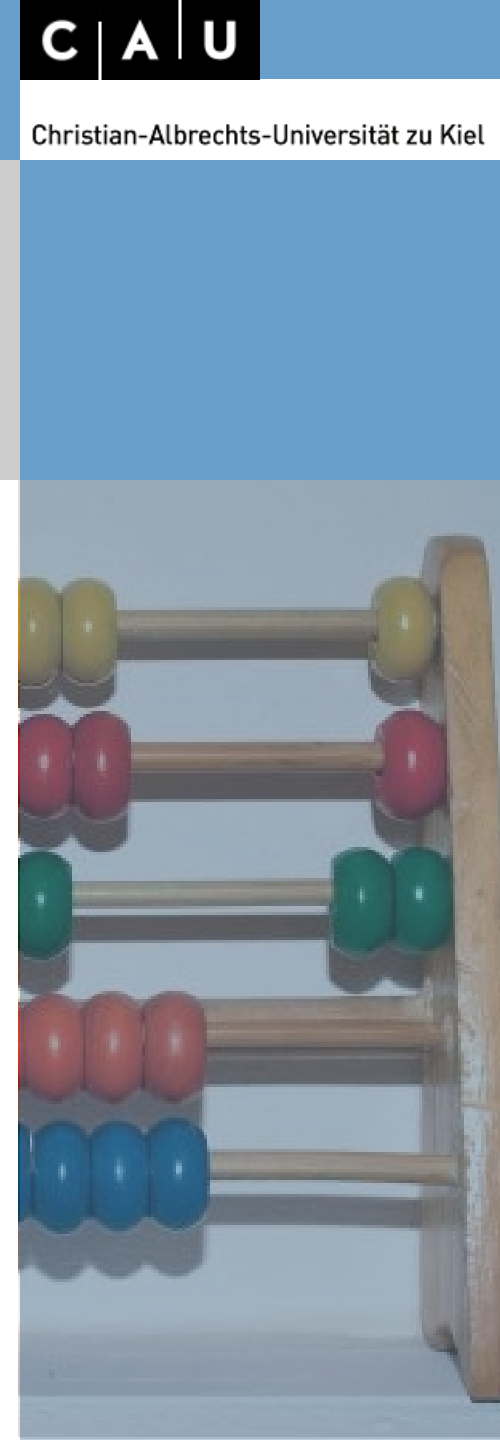
Die Nullhypothese nicht wird abgelehnt, obwohl sie falsch ist

Fehler 2. Art, negativ falsch oder β -Fehler

Das Ergebnis eines Schwangerschaftstests ist falsch negativ, wenn er keine Schwangerschaft anzeigt, obwohl eine Schwangerschaft vorliegt.

	Wahrer Sachverhalt: H0 (Es gibt keinen Unterschied)	Wahrer Sachverhalt: H1 (Es gibt einen Unterschied)
durch einen statistischen Test fällt eine Entscheidung für: H0	richtige Entscheidung	Fehler 2. Art
durch einen statistischen Test fällt eine Entscheidung für: H1	Fehler 1. Art	richtige Entscheidung

Quelle: wikipedia



α - und β -Fehler [1]

Tests und Fehler

Statistische Tests sollten beide Fehlerarten vermeiden

Gradwanderung (zu streng/nicht streng genug)

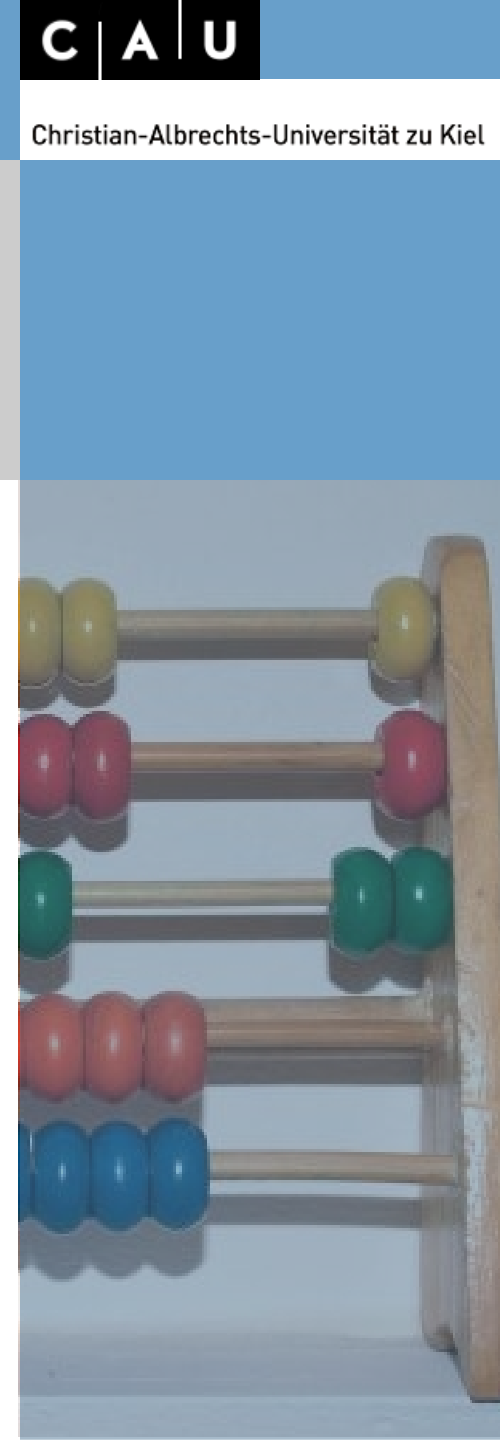
Generell sind Fehler 1. Art meist schwerwiegender als Fehler 2. Art

Dies führt meist zu falschen Annahmen, da bei einem Fehler 2. Art die Null-Hypothese nicht als bewiesen gilt, bei einem Fehler 1. Art die Alternativhypothese hingegen schon.

Teststärke (Power)

Ein Test ist um so stärker (hat mehr Power), umso mehr er Fehler 2. Art vermeidet, ohne Fehler 1. Art zu begehen.

Ein stärkerer Test hilft, Sachverhalte besser zu klären.



Nichtparametrische Tests

parametrisch vs. nicht-parametrisch/parameterfrei:

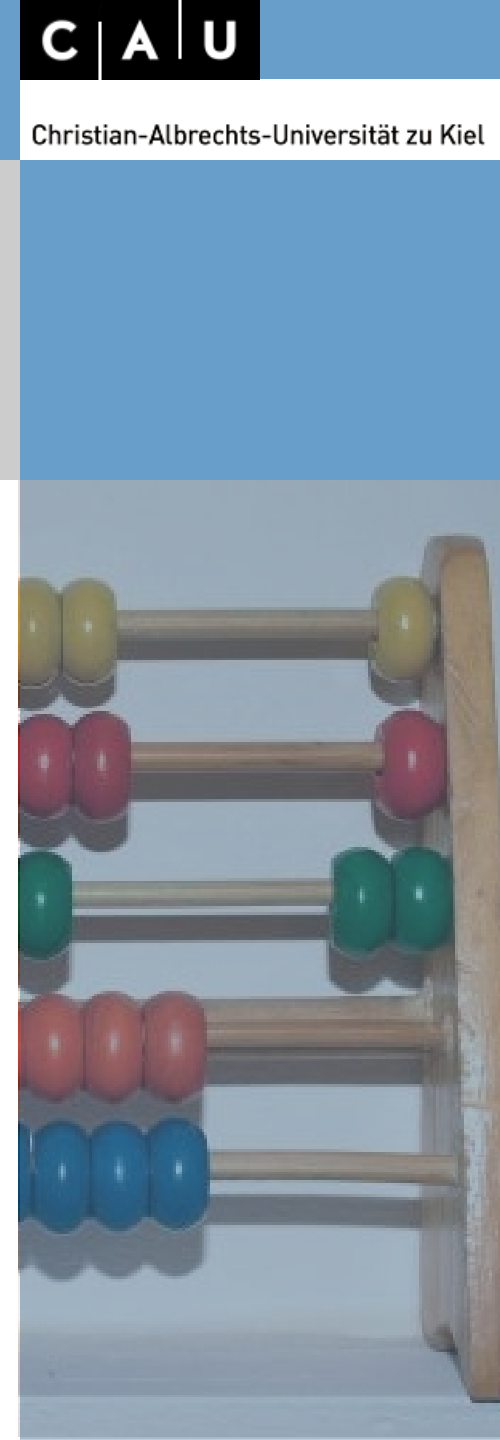
parametrisch: Werte müssen bestimmter Verteilung folgen (z. B. Normalverteilung); Grundannahmen zur Verteilung sind notwendig

nicht-parametrisch: Annahmen zur Werteverteilung entfallen; keine Grundannahmen notwendig

Nicht-parametrische Tests, Vorteile - Nachteile:

Vorteil: Sind auch anwendbar, wenn keine Aussage über die Verteilung möglich ist oder die Verteilung nicht den für parametrische Tests gegebenen Anforderungen entspricht.
Es können auch relativ kleine Stichproben getestet werden.

Nachteil: Haben meist eine geringere Power (Teststärke),



Kolmogorov



+



Test

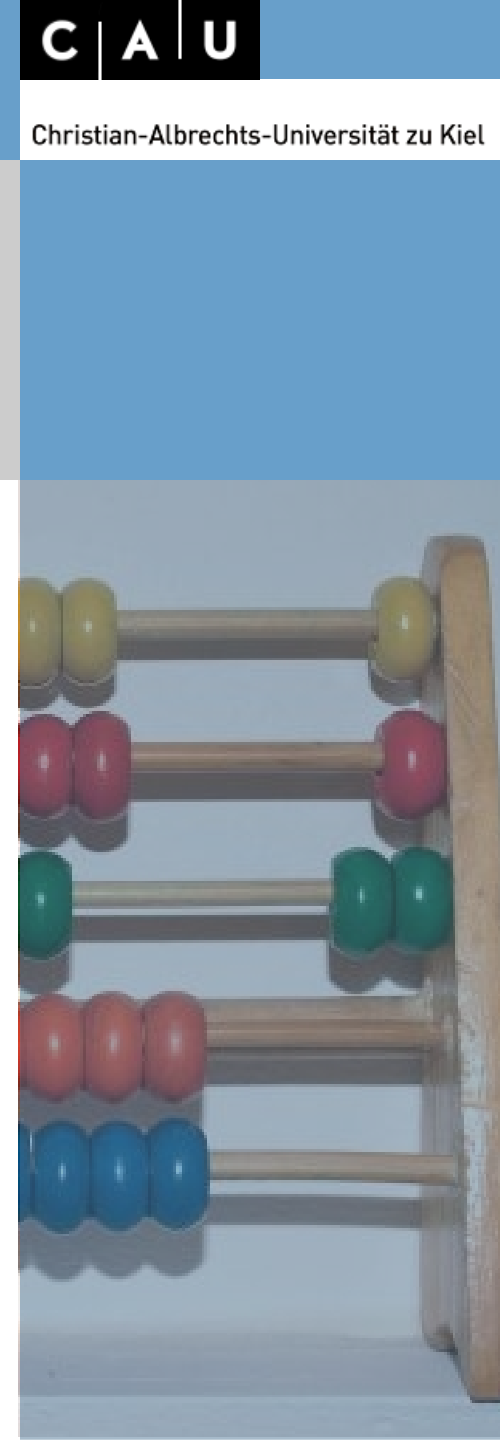
Kolmogorov-Smirnov-Test [1]

Test auf Differenz zweier Verteilungen

Voraussetzung: mindestens 1 ordinalskalierte Variable (bei einer Stichprobe) und 1 nominalskalierte Gruppierungsvariable (bei 2 Stichproben)

Vorgehensweise bei einer Stichprobe: die kumulative prozentuale Häufigkeit der Stichprobe wird mit einer Standardverteilung (meist Normalverteilung) verglichen

Vorgehensweise bei 2 Stichproben: die kumulativen prozentualen Häufigkeiten der Stichproben werden miteinander verglichen (nach Müller-Scheeßel)



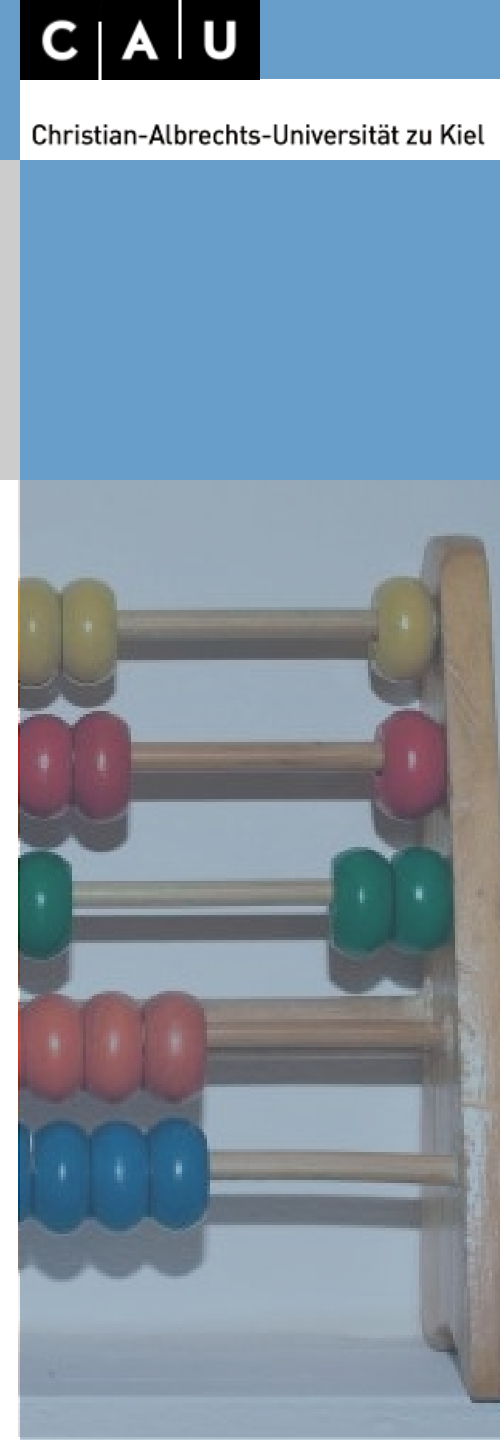
Kolmogorov-Smirnov-Test [2]

Beispiel (nach Shennan)

Weibliche bronzzeitliche Bestattungen auf einem Gräberfeld nach Altersklassen

Alter zum Zeitpunkt des Todes	Reichtumskategorie	
	Reich	Arm
Infans I	6	23
Infans II	8	21
Juvenil	11	25
Adult	29	36
Matur	19	27
Senil	3	4
Gesamt	76	136

Frage: Unterschieden sich die Lebensbedingungen so, daß ein unterschiedliches Alter erreicht wurde?



Kolmogorov-Smirnov-Test [3]

Voraussetzungen

H_0 : Es gibt keinen Unterschied zwischen reichen und armen Gräbern hinsichtlich des Sterbealters

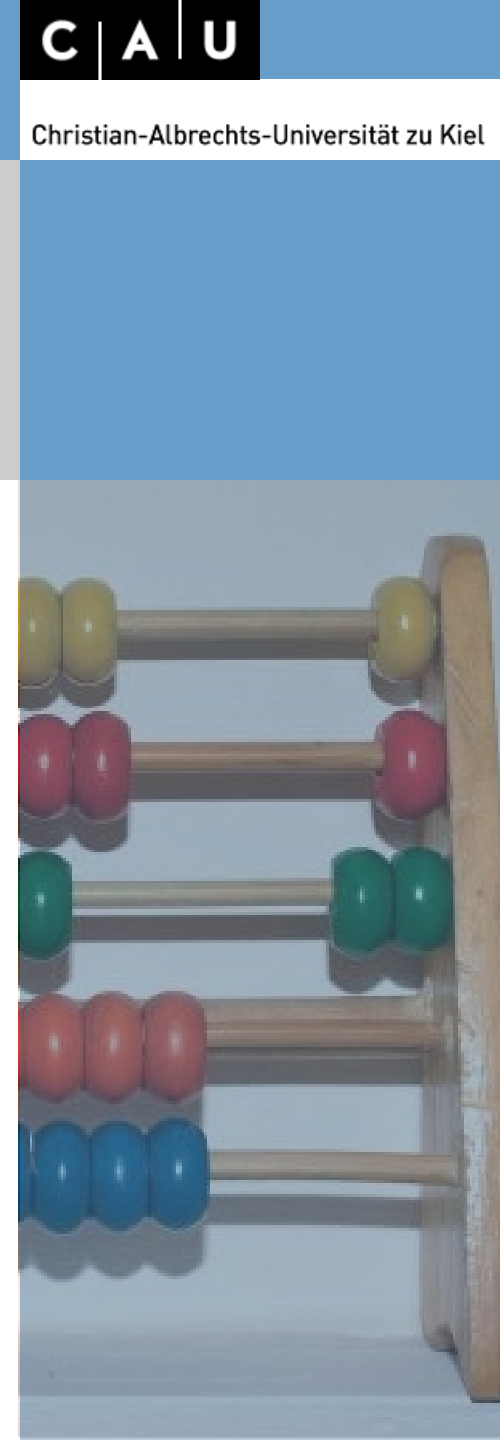
H_1 : Es gibt einen Unterschied zwischen reichen und armen Gräbern hinsichtlich des Sterbealters

Zweiseitiger Test.

Signifikanzniveau: 0.05

Variablen:

1. Ordinal skalierte Altersklassen
2. mindestens nominal (ordinal) skalierte Reichtumsklassen

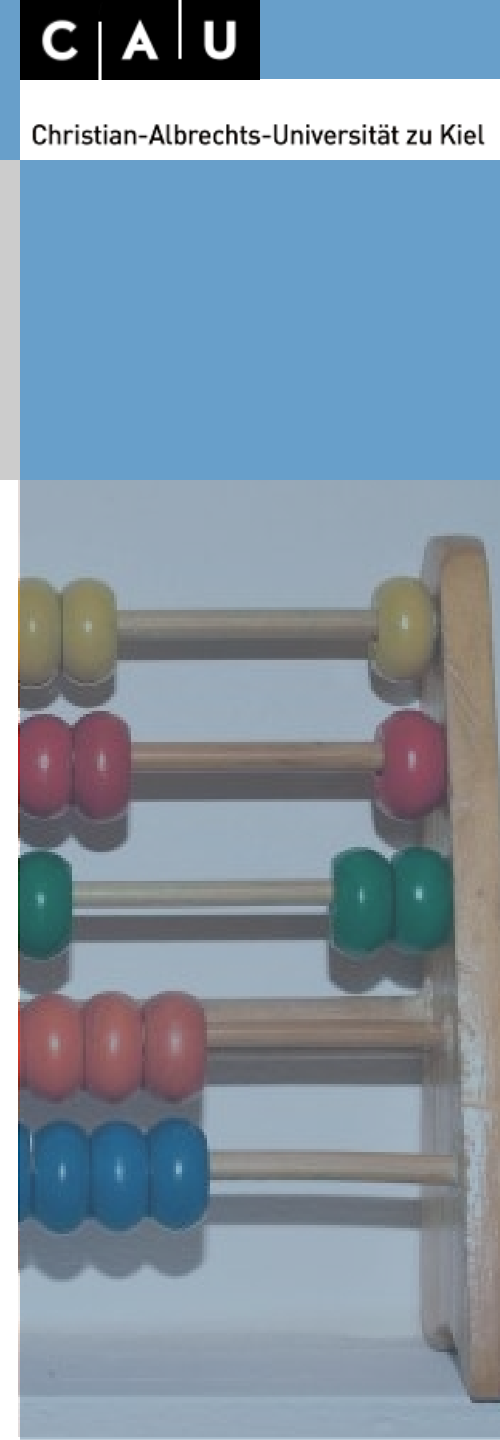


Kolmogorov-Smirnov-Test [4]

Vorgehen: Berechnen der prozentualen Häufigkeit

Teilen jeder Zelle der Spalte durch die Summe der Spalte

Alter zum Zeitpunkt des Todes	Reichtumskategorie			
	Reich		Arm	
Infans I	6	0.079	23	0.169
Infans II	8	0.105	21	0.154
Juvenil	11	0.145	25	0.184
Adult	29	0.382	36	0.265
Matur	19	0.250	27	0.199
Senil	3	0.039	4	0.029
Gesamt	76	1.000	136	1.000

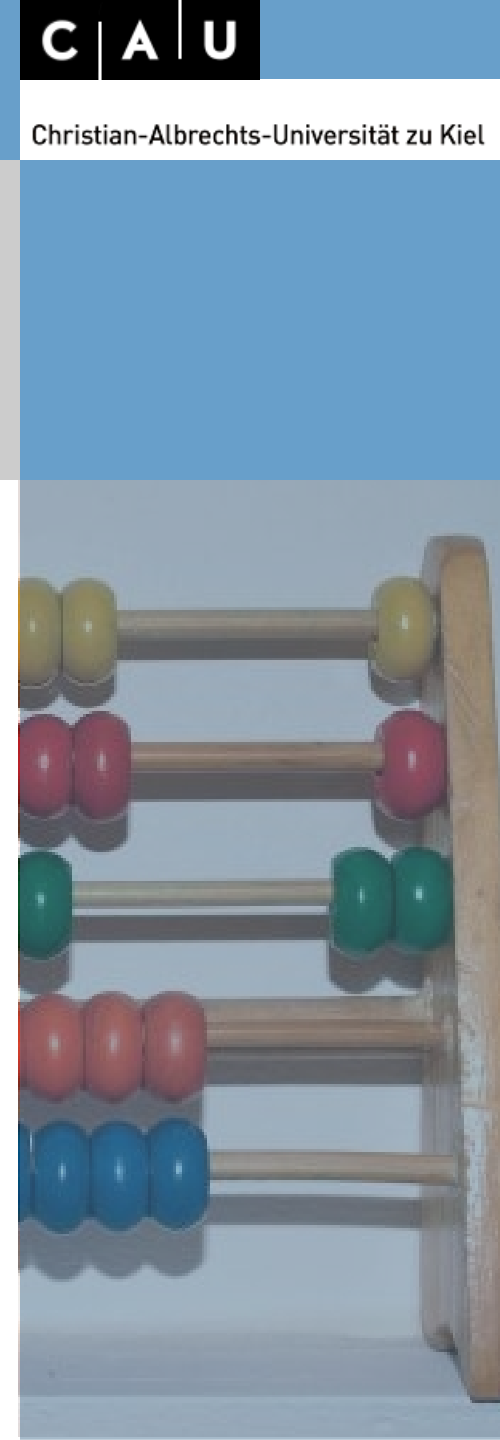


Kolmogorov-Smirnov-Test [5]

Vorgehen: Berechnen der kumulativen prozentualen Häufigkeit

Addieren jeder prozentualen Häufigkeit mit den darunter liegenden Häufigkeiten der Ordinalen Variable

Alter zum Zeitpunkt des Todes	Reichtumskategorie		Arm			
	Reich					
Infans I	6	0.079	0.079	23	0.169	0.169
Infans II	8	0.105	0.184	21	0.154	0.323
Juvenil	11	0.145	0.329	25	0.184	0.507
Adult	29	0.382	0.711	36	0.265	0.772
Matur	19	0.250	0.961	27	0.199	0.971
Senil	3	0.039	1.000	4	0.029	1.000
Gesamt	76	1.000		136	1.000	



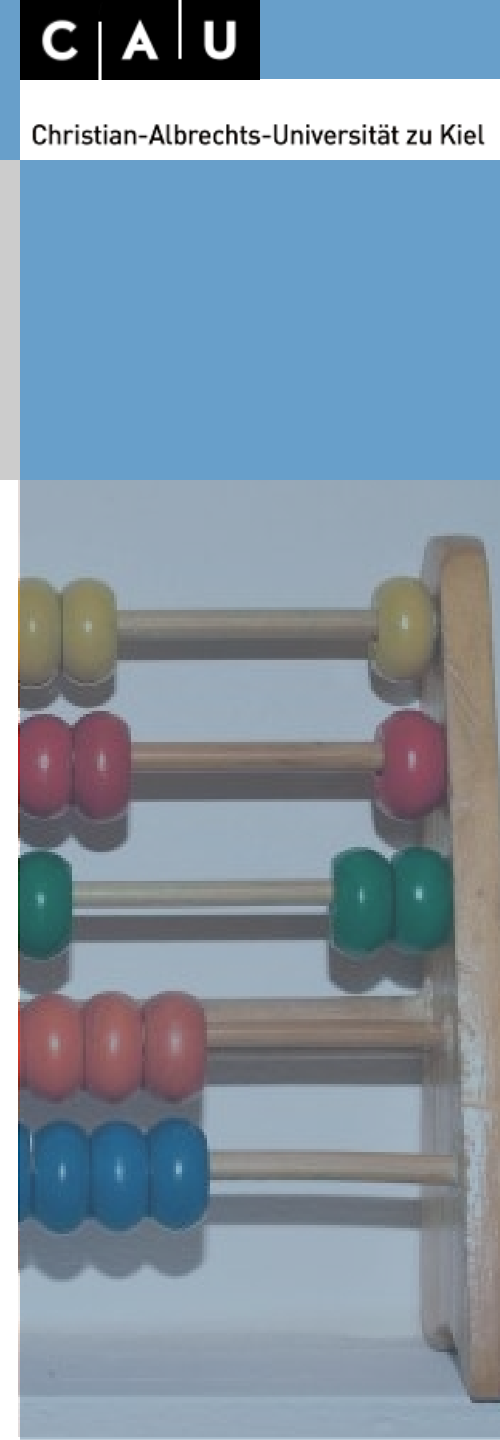
Kolmogorov-Smirnov-Test [6]

Vorgehen: Berechnen der Differenz der kumulativen prozentualen Häufigkeiten

Subtrahieren der kumulativen prozentualen Häufigkeiten voneinander, bilden des Absolut-Wertes

Alter zum Zeitpunkt des Todes	Reichtumskategorie		Differenz
	Reich	Arm	
Infans I	0.079	0.169	0.090
Infans II	0.184	0.323	0.139
Juvenil	0.329	0.507	0.178
Adult	0.711	0.772	0.061
Matur	0.961	0.971	0.010
Senil	1.000	1.000	0.000

Größte Differenz



Kolmogorov-Smirnov-Test [7]

Vergleich der maximalen Differenz mit einem Schwellenwert, der sich aus den Gesamtanzahlen berechnet

Gesamtanzahl Reich: 76
Gesamtanzahl Arm: 136
Differenz max. (D_{\max}): 0.178
Signifikanzniveau: 0.05

Formel:

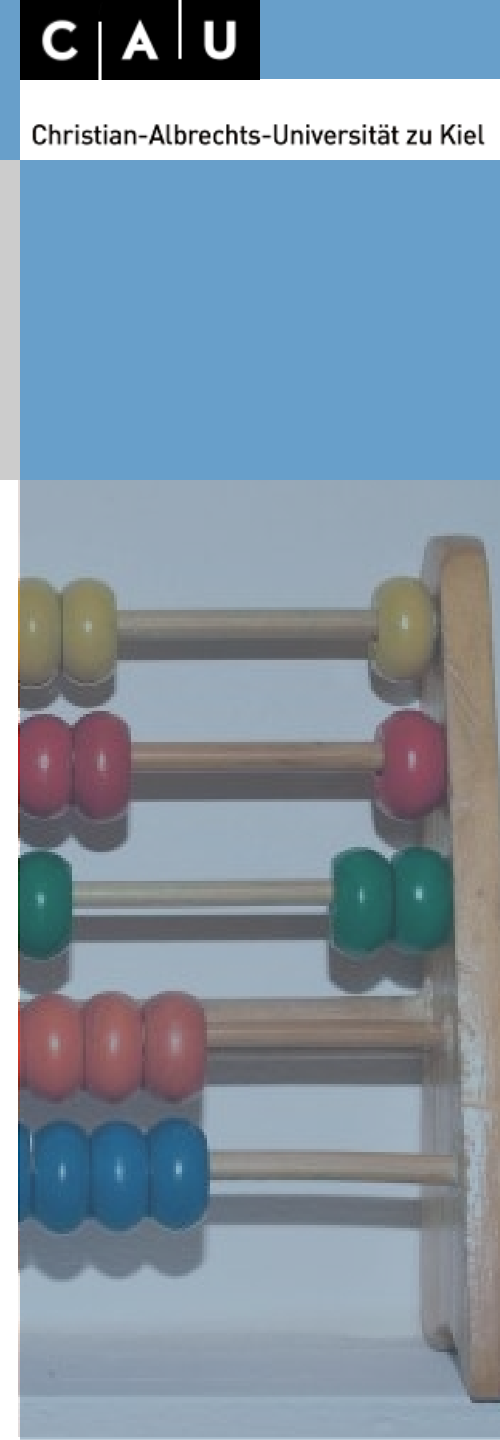
$$\text{Schwellenwert KS-Test} = \text{Faktor } f * \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

Faktor f:
Signifikanzniveau 0.05: 1.36
Signifikanzniveau 0.01: 1.63
Signifikanzniveau 0.001: 1.95

Daher: $\text{Schwellenwert KS-Test} = 1.36 * \sqrt{\frac{76 + 136}{76 * 136}} = 0.195$

$D_{\max} < \text{Schwellenwert}$, kein signifikanter Unterschied feststellbar

Das heißt aber nicht, dass die Verteilungen gleich sind, sondern nur, dass sie sich nicht signifikant unterscheiden.



Kolmogorov-Smirnov-Test [8]

KS-Test in R

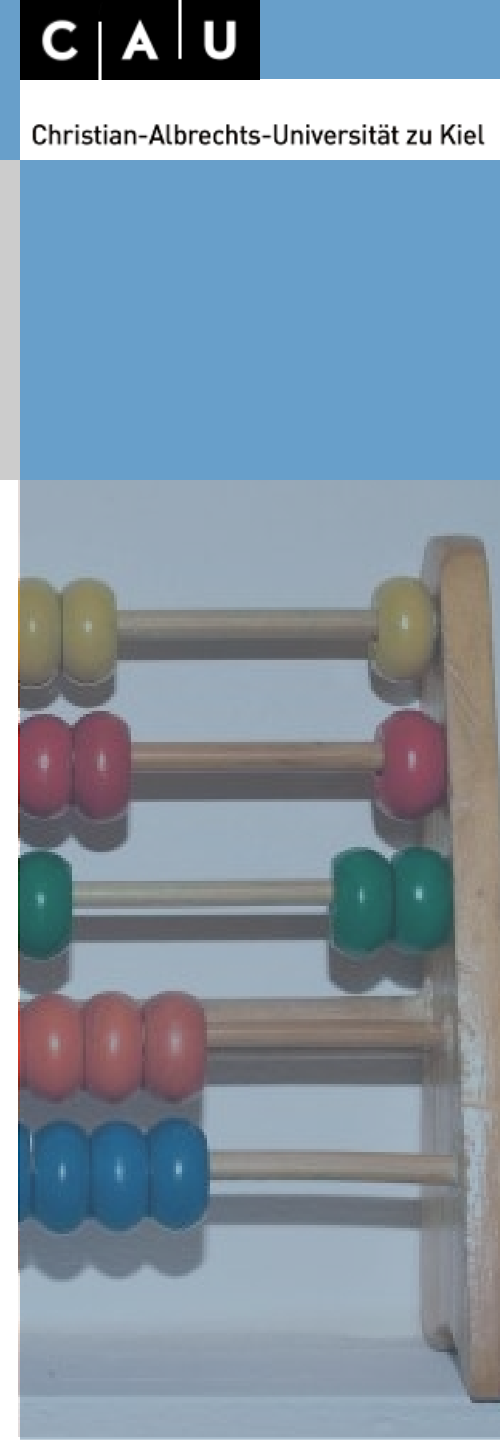
```
> graeberbrz<-read.csv2("graeberbrz.csv",row.names=1)
> table(graeberbrz)
      reichtum
alter arm reich
1      6    23
2      8    21
3     11    25
4     29    36
5     19    27
6      3      4
> alter<-graeberbrz$alter
> reichtum<-graeberbrz$reichtum
> ks.test(alter[reichtum=="arm"],alter[reichtum=="reich"])
```

Two-sample Kolmogorov-Smirnov test

```
data: alter[reichtum == "arm"] and alter[reichtum == "reich"]
D = 0.1784, p-value = 0.08977
alternative hypothesis: two-sided
```

Warning message:

```
In ks.test(alter[reichtum == "arm"], alter[reichtum == "reich"]) :
  kann bei Bindungen nicht die korrekten p-Werte berechnen
```

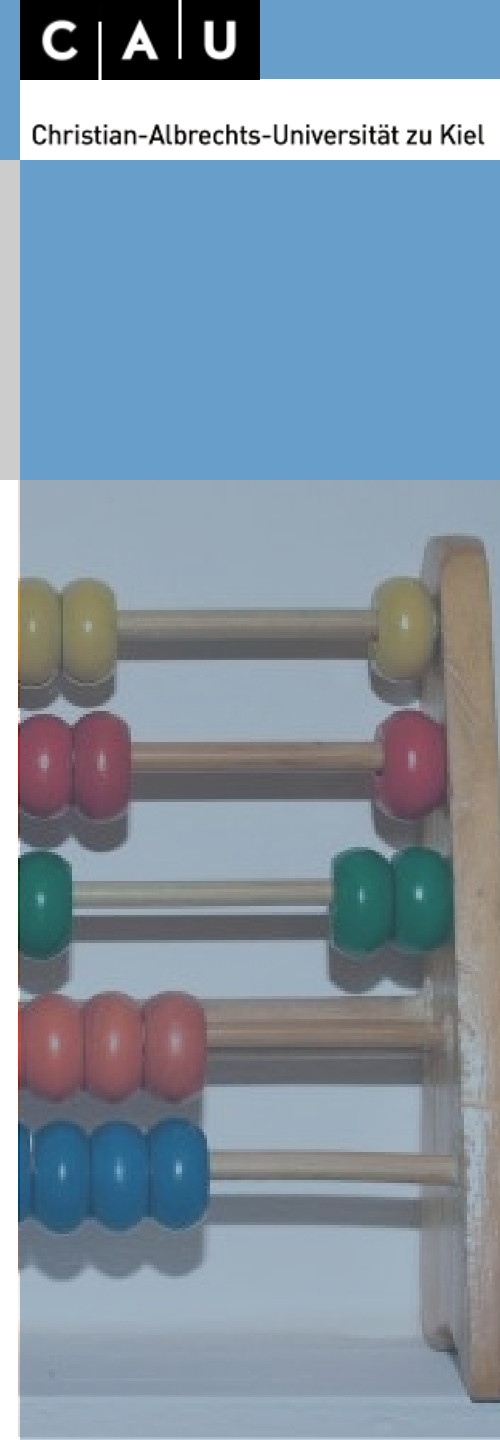


Kolmogorov-Smirnov-Test Aufgabe

Tassen aus relativ geschlossenen spätneolithischen Inventaren (Müller 2001)

Stellen Sie mittels des Kolmogorov-Smirnov-Tests fest, ob sich die Höhen der ein- und zweigliedrigen Tassen signifikant auf einem 0.05-Niveau unterscheiden.

Datei: mueller2001.csv



Kolmogorov-Smirnov-Test Lösung

Tassen aus relativ geschlossenen spätneolithischen Inventaren (Müller 2001)

Stellen Sie mittels des Kolmogorov-Smirnov-Tests fest, ob sich die Höhen der ein- und zweigliedrigen Tassen signifikant auf einem 0.05-Niveau unterscheiden.

Datei: mueller2001.csv

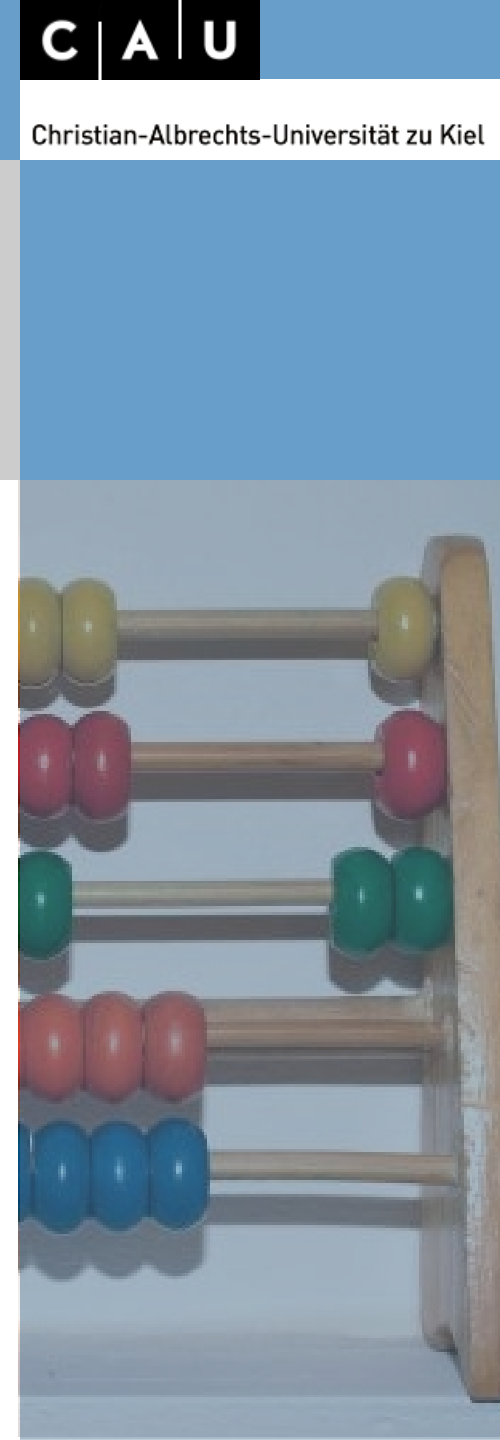
```
> mueller<-read.csv2("mueller2001.csv")
> tassentyp<-mueller$tassentyp
> hoehe<-mueller$hoehe
> ks.test(hoehe[tassentyp=="eingliedrig"],hoehe[tassentyp=="zweigliedrig"])
```

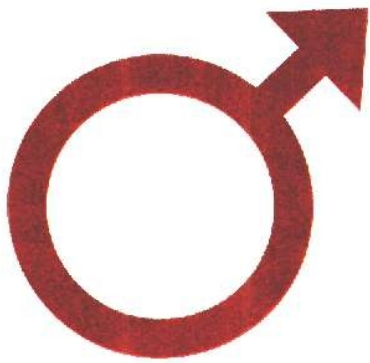
Two-sample Kolmogorov-Smirnov test

data: hoehe[tassentyp == "eingliedrig"] and hoehe[tassentyp == "zweigliedrig"]
D = 0.2519, p-value = 0.1020
alternative hypothesis: two-sided

Warning message:

In ks.test(hoehe[tassentyp == "eingliedrig"], hoehe[tassentyp == "zweigliedrig"]) :
kann bei Bindungen nicht die korrekten p-Werte berechnen





+



+



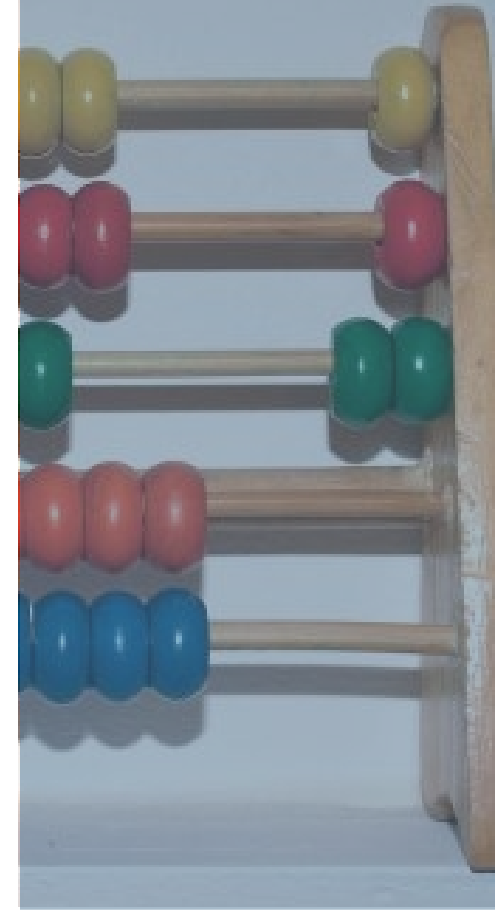
Test

Mann-Whitney-U-Test [1] (=Wilcoxon-Rangsummentest)

Test auf Differenz zweier Verteilungen

Voraussetzung: mindestens 1 intervall- oder ordinalskalierte Variable und 1 nominalskalierte Gruppierungsvariable

Vorgehensweise: die Daten werden in eine Rangfolge gebracht und für jede Gruppe werden die Rangplätze verglichen



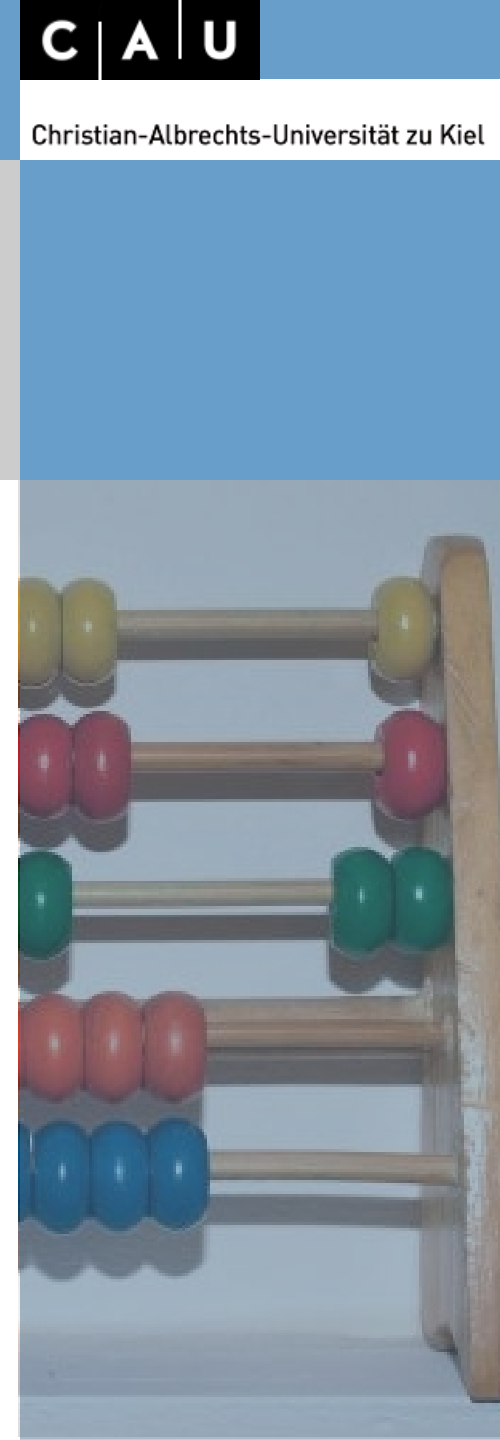
Mann-Whitney-U-Test [2]

Beispiel (nach Müller-Scheeßel)

Kammergrößen von eisenzeitlichen Kammergrößen mit Angabe des Geschlechtes

Kammergröße	Geschlecht
11,7	m
4,4	w
35,9	m
8,0	w
23,0	m
5,1	w
9,2	m
15,8	w
26,1	m
7,3	w

Frage: Unterscheidet sich die Größe der Gräber abhängig vom Geschlecht?

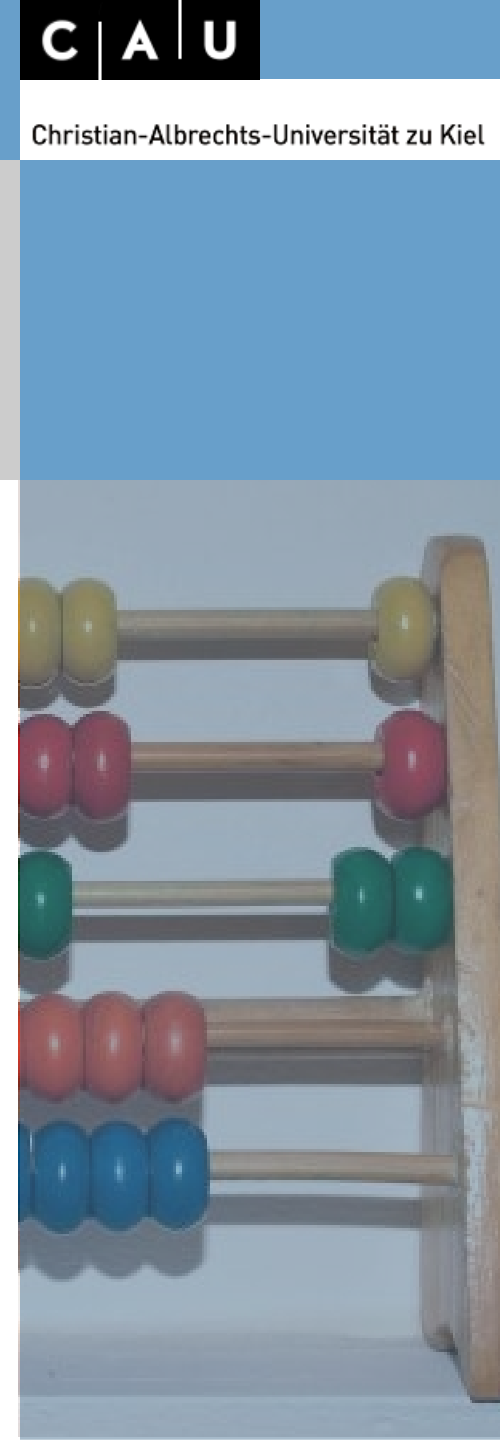


Mann-Whitney-U-Test [3]

Vorgehen

Bestimmen des Rangplatzes der Gräber nach der Größe

Kammergröße	Geschlecht	Rang
11,7	m	5
4,4	w	10
35,9	m	1
8,0	w	7
23,0	m	3
5,1	w	9
9,2	m	6
15,8	w	4
26,1	m	2
7,3	w	8

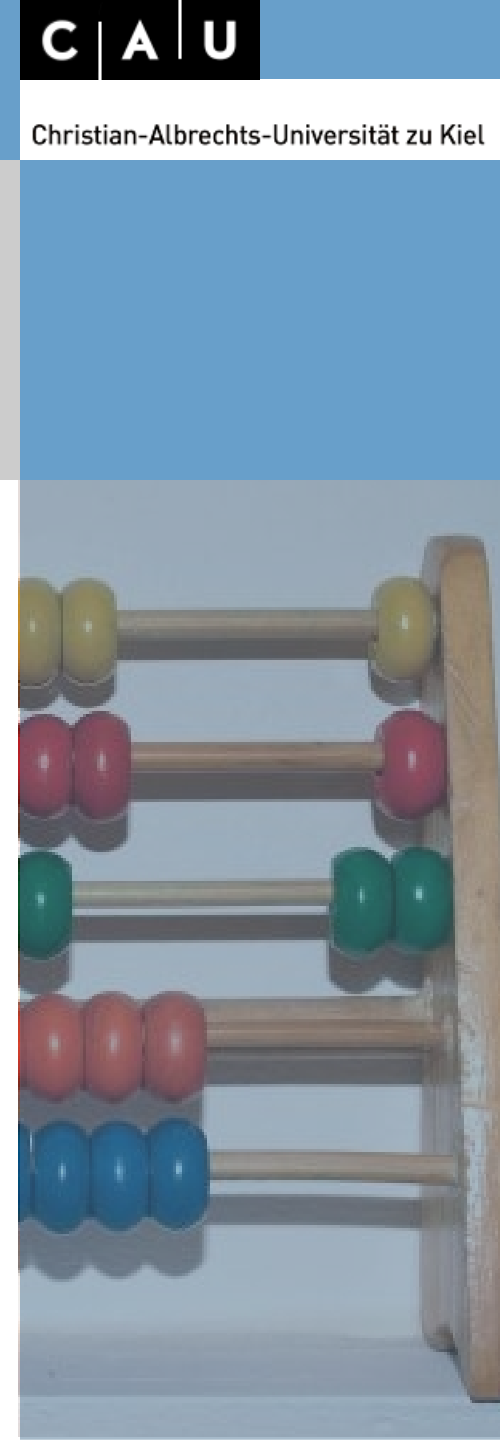


Mann-Whitney-U-Test [4]

Vorgehen

Sortieren nach Rängen

Kammergröße	Geschlecht	Rang
35,9	m	1
26,1	m	2
23,0	m	3
15,8	w	4
11,7	m	5
9,2	m	6
8,0	w	7
7,3	w	8
5,1	w	9
4,4	w	10

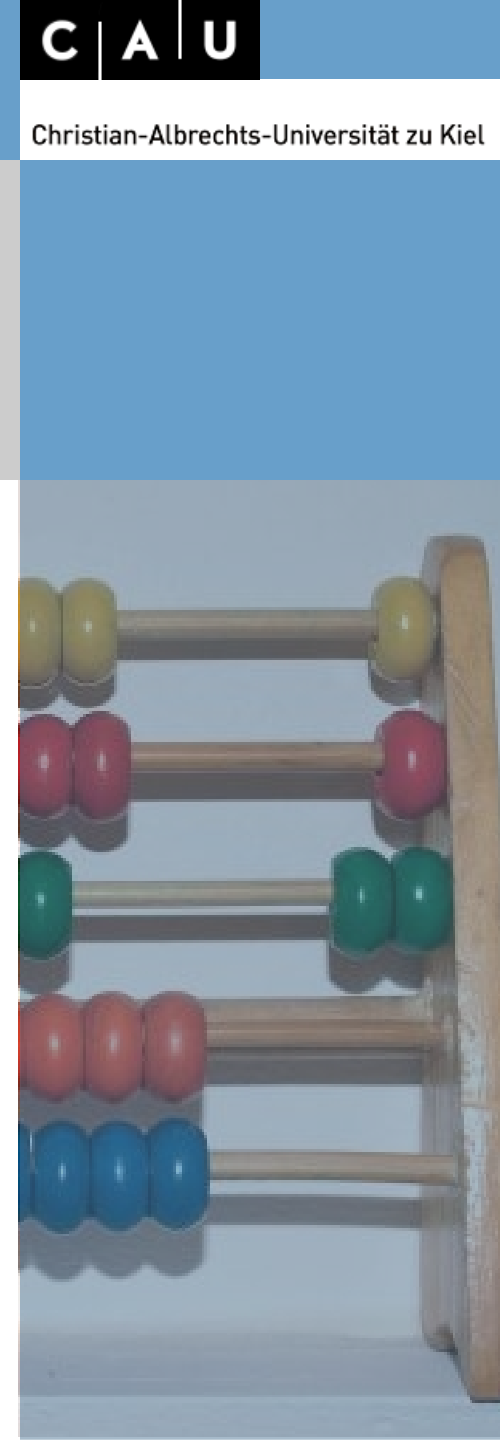


Mann-Whitney-U-Test [5]

Vorgehen

Zählen, wie viele Werte der jeweils anderen Kategorie unterhalb der einzelnen Werte liegen

Kammergröße	Geschlecht	Rang	M unterhalb	F unterhalb
35,9	m	1		5
26,1	m	2		5
23,0	m	3		5
15,8	w	4	2	
11,7	m	5		4
9,2	m	6		4
8,0	w	7		
7,3	w	8		
5,1	w	9		
4,4	w	10		
Summe			2	23



Mann-Whitney-U-Test [6]

Vorgehen

Zahl der männlichen Bestattungen: 5

Zahl der weiblichen Bestattungen: 5

Summe der Ränge männliche Bestattungen: 23

Summe der Ränge weibliche Bestattungen: 2

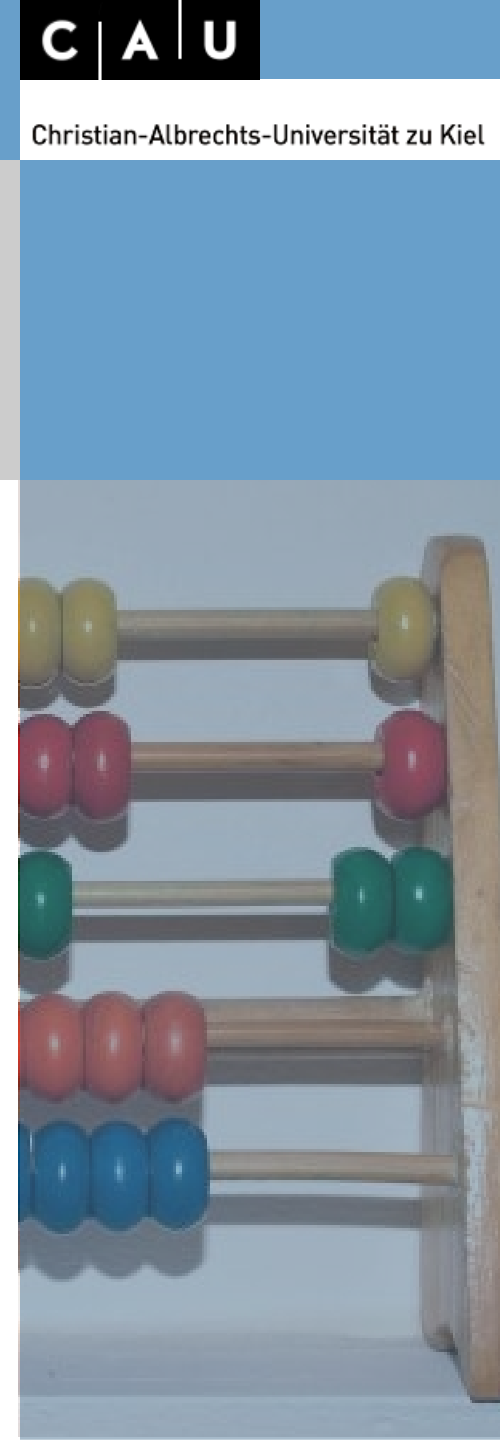
$$5 \cdot 5 = 25 = 23 + 2$$

Der kleinere Wert wird ausgewertet: 2

Nachschlagen in Tabelle (z.B. Shennan 1997, Tabelle B):

Schwellenwert für Signifikanz 0.05 bei $n_1=5$ und $n_2=5$: 2

Die Kammergrößen unterscheiden sich signifikant voneinander.



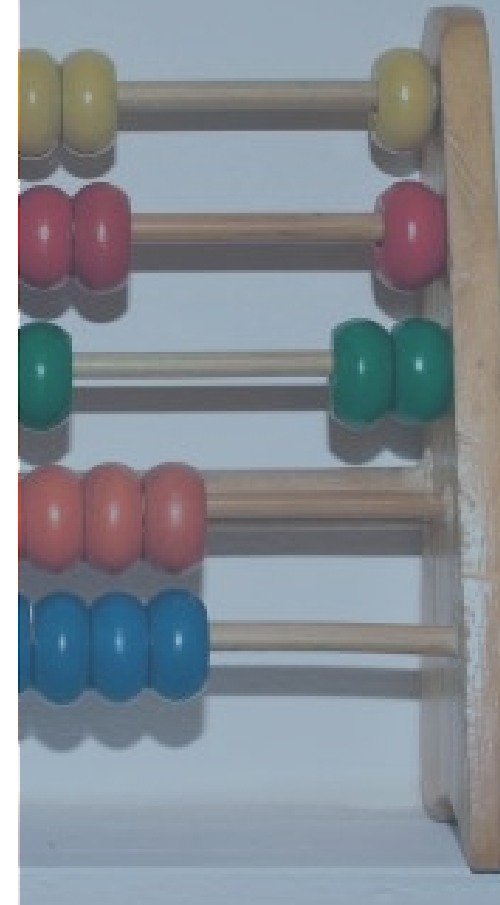
Mann-Whitney-U-Test [7]

Mann-Whitney-U-Test in R

```
> kammergroesse<-read.csv2("kammergroesse_mueller-scheessel.csv")
> kammergroesse
  kammergroesse geschlecht
1          35.9          m
2          26.1          m
3          23.0          m
4          15.8          w
5          11.7          m
6           9.2          m
7           8.0          w
8           7.3          w
9           5.1          w
10          4.4          w
> wilcox.test(kammergroesse$kammergroesse ~
kammergroesse$geschlecht)
```

Wilcoxon rank sum test

```
data: kammergroesse$kammergroesse by kammergroesse$geschlecht
W = 23, p-value = 0.03175
alternative hypothesis: true location shift is not equal to 0
```

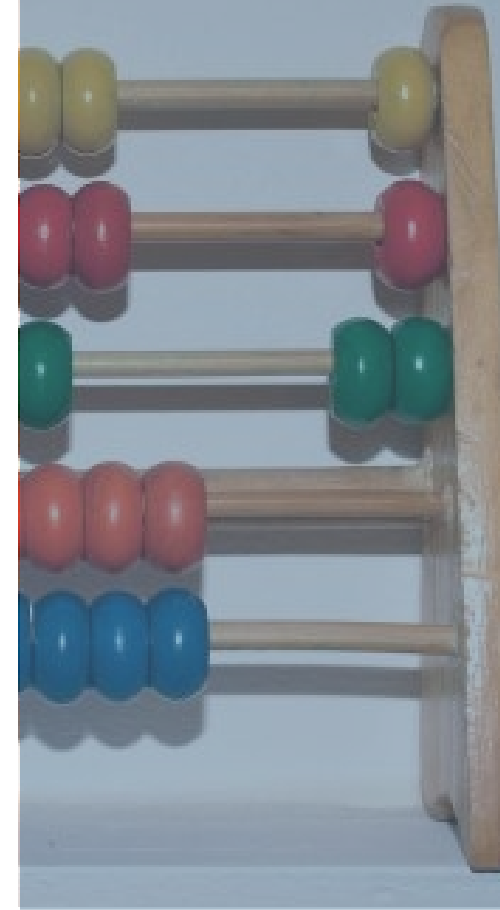


Mann-Whitney-U-Test Aufgabe

Längen von Randleistenbeilen der Typen Bikun und Cegun (Cullberg 1968)

Stellen Sie mittels des Mann-Whitney-U-Test fest, ob sich die Längen der Randleistenbeile vom Typ Bikun und Cegun signifikant auf einem 0.05-Niveau unterscheiden.

Datei: cullberg1968.csv



Mann-Whitney-U-Test Lösung

Längen von Randleistenbeilen der Typen Bikun und Cegun (Cullberg 1968)

Stellen Sie mittels des Mann-Whitney-U-Test fest, ob sich die Längen der Randleistenbeile vom Typ Bikun und Cegun signifikant auf einem 0.05-Niveau unterscheiden.

Datei: cullberg1968.csv

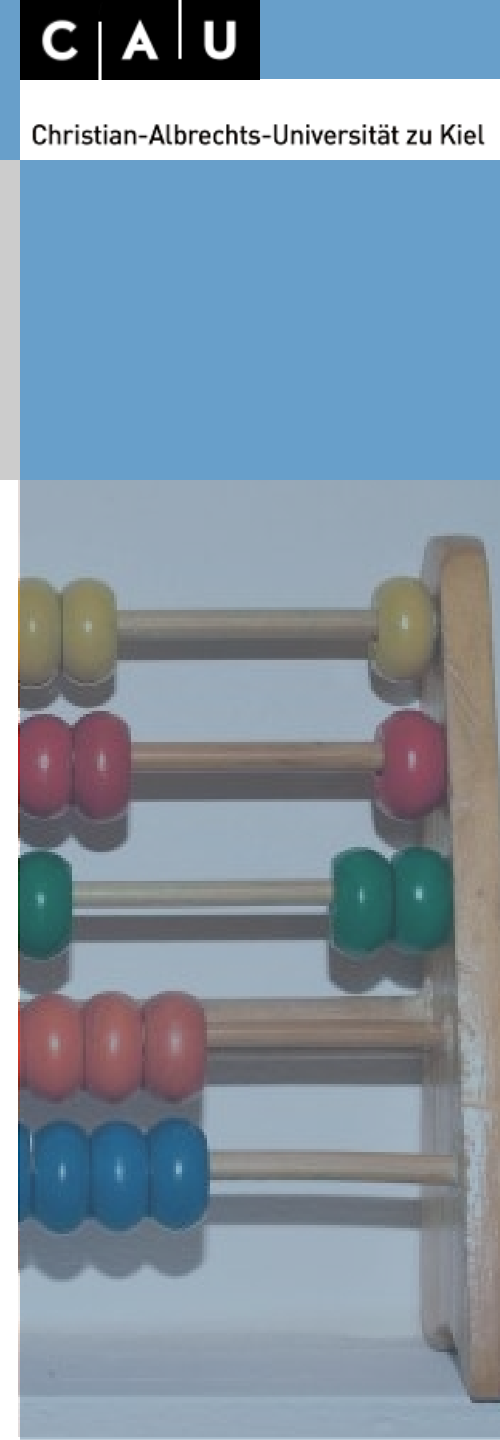
```
> cullberg<-read.csv2("cullberg1968.csv")
> laenge<-cullberg$laenge
> typ<-cullberg$typ
> wilcox.test(laenge[typ=="Bikun"],laenge[typ=="Cegun"])
```

Wilcoxon rank sum test with continuity correction

```
data: laenge[typ == "Bikun"] and laenge[typ == "Cegun"]
W = 17.5, p-value = 0.02673
alternative hypothesis: true location shift is not equal to 0
```

Warning message:

```
In wilcox.test.default(laenge[typ == "Bikun"], laenge[typ ==
"Cegun"]) :
  kann bei Bindungen keinen exakten p-Wert Berechnen
```



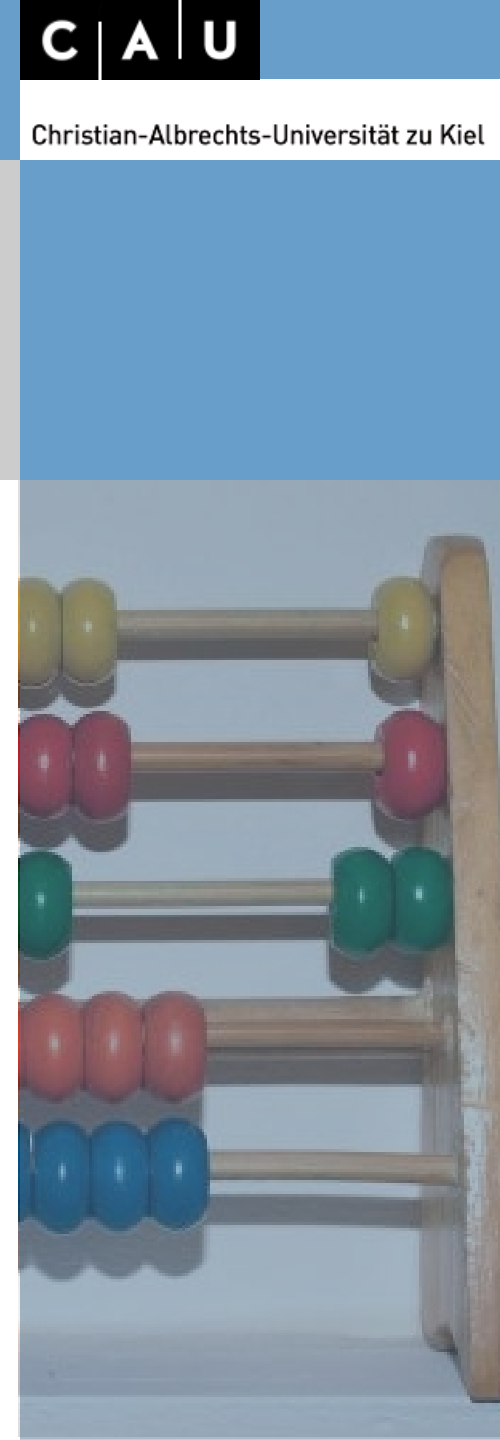
Interpretation von Signifikanztests

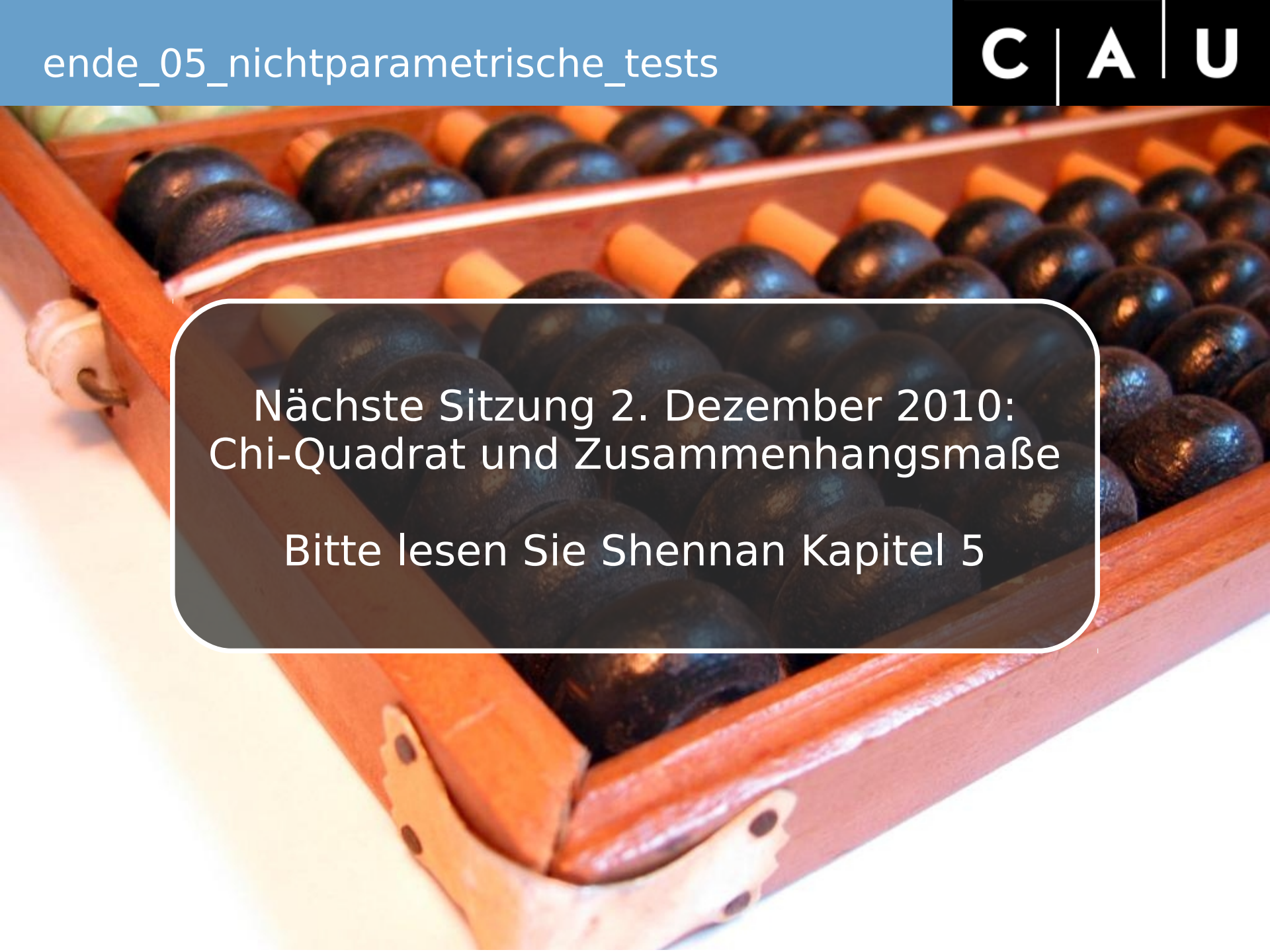
Vorsicht, auch wenn die Statistik klar scheint

Nach dem Test wie vor dem Test: Die Interpretation entscheidet über das Ergebnis!

Statistische Signifikant \neq Archäologischer Signifikanz!

Statistische Ergebnisse bleiben statistisch: Signifikanz ist Wahrscheinlichkeit, dass das Ergebniss stimmt, aber es bleibt immer eine Restmöglichkeit, dass Zufall im Spiel ist!





Nächste Sitzung 2. Dezember 2010:
Chi-Quadrat und Zusammenhangsmaße

Bitte lesen Sie Shennan Kapitel 5