BERUFSAKADEMIE SACHSEN
STAATLICHE STUDIENAKADEMIE
RIESA
UNIVERSITY OF COOPERATIVE EDUCATION

Chi²-Unabhängigkeitstest

Ziel des Chi²-Unabhängigkeitstest

Daten: Sie haben z.B. zwei Gruppen (Frauen / Männer) und für alle Befragten die Zustimmung / Ablehnung zu einem Thema.

Anzahl Variablen:

Skalenniveau: r

Problemstellung: Sie wollen nun wissen. Stimmen Frauen eher oder weniger zu im Vergleich zu den Männer.

Sinn des Tests: H_0 =Die Häufigkeiten in den untersuchten Gruppen sind gleich.

Chi²-Unabhängigkeitstest: χ^2-Wert

$$\chi^2 = \sum_{i=1}^{q} \sum_{j=1}^{r} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

$\hat{n} = erwartete\ Häufigkeit$

$n = beobachtete\ Häufigkeit$

Sie vergleichen den ermittelten χ^2 nun in der χ^2-Tabelle

Sie brauchen noch die Freiheitsgrade (F): F = n-1

Die Formel erklärt (formal)

Wir brauchen die Tabelle der beobachteten Häufigkeiten a

$$
\begin{array}{cc|cccc|c}
 & & & Y & & & \\
 & & 1 & 2 & \dots & q & i \\
\hline
 & 1 & n_{11} & n_{12} & \dots & n_{1q} & n_{1.} \\
X & 2 & n_{21} & n_{22} & \dots & n_{2q} & n_{2.} \\
 & \dots & \dots & \dots & \dots & \dots & \dots \\
 & r & n_{r1} & n_{r2} & \dots & n_{rq} & n_{r.} \\
\hline
 & j & n_{.1} & n_{.2} & \dots & n_{.q} & n
\end{array}
$$

Erwartete Häufigkeiten: $\hat{n}_{ij} = \frac{n_i * n_j}{n}$

Tabelle für die Chi²-Verteilung

f	1−α 0,900	0,950	0,975	0,990	0,995	0,999
1	2,71	3,84	5,02	6,63	7,88	10,83
2	4,61	5,99	7,38	9,21	10,60	13,82
3	6,25	7,81	9,35	11,34	12,84	16,27
4	7,78	9,49	11,14	13,28	14,86	18,47
5	9,24	11,07	12,83	15,09	16,75	20,52
6	10,64	12,59	14,45	16,81	18,55	22,46
7	12,02	14,07	16,01	18,48	20,28	24,32
8	13,36	15,51	17,53	20,09	21,95	26,12
9	14,68	16,92	19,02	21,67	23,59	27,88
10	15,99	18,31	20,48	23,21	25,19	29,59
11	17,28	19,68	21,92	24,72	26,76	31,26
12	18,55	21,03	23,34	26,22	28,30	32,91
13	19,81	22,36	24,74	27,69	29,82	34,53
14	21,06	23,68	26,12	29,14	31,32	36,12
15	22,31	25,00	27,49	30,58	32,80	37,70
16	23,54	26,30	28,85	32,00	34,27	39,25
17	24,77	27,59	30,19	33,41	35,72	40,79
18	25,99	28,87	31,53	34,81	37,16	42,31
19	27,20	30,14	32,85	36,19	38,58	43,82
20	28,41	31,41	34,17	37,57	40,00	45,31
21	29,62	32,67	35,48	38,93	41,40	46,80
22	30,81	33,92	36,78	40,29	42,80	48,27
23	32,01	35,17	38,08	41,64	44,18	49,73
24	33,20	36,42	39,36	42,98	45,56	51,18
25	34,38	37,65	40,65	44,31	46,93	52,62
26	35,56	38,89	41,92	45,64	48,29	54,05
27	36,74	40,11	43,19	46,96	49,64	55,48
28	37,92	41,34	44,46	48,28	50,99	56,89
29	39,09	42,56	45,72	49,59	52,34	58,30
30	40,26	43,77	46,98	50,89	53,67	59,70
40	51,81	55,76	59,34	63,69	66,77	73,40
50	63,17	67,50	71,42	76,15	79,49	86,66
60	74,40	79,08	83,30	88,38	91,95	99,61
70	85,53	90,53	95,02	100,43	104,21	112,32

Vorgehen

1. Chi-Quadrat-Wert ausrechnen
2. Freiheitsgrade bestimmen und in die entsprechende Zeile gehen
3. In der Zeile die nächstgelegene kleinere Zahl suchen.
4. Oben in der Spalte schauen, zu welcher Wahrscheinlichkeit der Chi-Quadrat-Wert gehört: dies ist der gesuchte p-Wert.
5. Prüfen, liegt der p-Wert über oder unter dem festgelegtem Signifikanzniveau von 0,05.

Quelle: Chi-Quadrat-Test (2020)
Wikipedia:
https://de.wikipedia.org/wiki/Chi-Quadrat-Test
Datum: 18.06.2020

Chi²-Unabhängigkeitstest: Die Frage

Beobachte Häufigkeit			
	Männer	Frauen	*Summe*
Ja	20	29	40
Nein	10	30	40
Summe	30	50	80

Geschlecht	Antwort
M	Ja
M	Nein
W	Ja
M	Ja
…	…

Aus den Rohdaten, die meist in Tabellenform vorliegen bilden Sie eine Kreuztabelle.

Chi²-Unabhängigkeitstest: Erwartete Häufigkeiten

Beobachte Häufigkeit			
	Männer	Frauen	*Summe*
Ja	20	29	**40**
Nein	10	30	40
Summe	**30**	50	**80**

Erwartete Häufigkeit

	Männer	Frauen
Ja	15	25
Nein	15	25

Erwartete Häufigkeiten
= Spaltensumme *
Zeilensumme /n
z.B.: 716 = 30 * 40/**80**

Bedeutung der erwarteten Häufigkeiten: Sie geben an, welche Häufigkeiten bei Gültigkeit von H_0 auftreten würden.

Differenzen

Beobachte Häufigkeit			
	Männer	Frauen	*Summe*
Ja	20	29	40
Nein	10	30	40
Summe	30	50	80

Erwartete Häufigkeit

	Männer	Frauen
Ja	15	25
Nein	15	25

$$1{,}6 = (20 - 15)^2/15$$

Zwischenrechnung (Differenzen zwischen beobachteten und erwarteten Häufigkeiten)

	Männer	Frauen
Ja	1,6	1,6
Nein	1,6	1,6

= (Beobachtete Häufigkeit - Erwartete Häufigkeit)^2/ Erwartete Häufigkeit

χ^2 - Wert

Summe der „Zwischenrechnung“ = Chi²-Wert

Zwischenrechnung (Differenzen zwischen beobachteten und erwarteten Häufigkeiten)

	Männer	Frauen
Ja	1,6	1,6
Nein	1,6	1,6

1,6+ 1,6 + 1,6 +1,6 = $\boxed{6{,}66 = \chi^2}$

Freiheitsgrade (df/degres of freedom)

df = (Anzahl der Spalten - 1) * (Anzahl der Zeilen - 1)

Beobachte Häufigkeit			
	Männer	Frauen	*Summe*
Ja	20	29	40
Nein	10	30	40
Summe	30	50	80

(2-1) * (2-1) = 1 = Freiheitsgrade

Nutzung der Tabelle für Chi²-Verteilung

Tabelle für
Verteilungsform
der Chi²-Verteilung

Das ist das relevante
Fläche.

Fläche df	0,975	0,990	0,995	0,999
1	5,023	6,634	7,879	10,828
2	7,377	9,210	10,596	13,816
3	…	11,3449	12,838	16,266
4	…	…	…	…

1 (df)

6,66 (χ^2)

Frage?
Was ist die nächst kleiner
Zahl zum meinem
gefundenem Chi² - in der
Zeile 1 df?

χ^2: Signifikanzniveau

0,999

Umrechnung: 1 - 0,999 = 0,001

Das ist das relevante Signifikanzniveau

Mögliche Aussage:
“Die Nullhypothese, dass beide Geschlechte auf die Frage gleich mit Ja/Nein Antworten wird zurückgewiesen zugungsten der Alternativehypothese. Es kann davon ausgegangen werden dass es einen signifikanten Unterschied zwischen dem Antwortverhalten bei Frauen und Männern gibt.”

Voraussetzung beim Chi²-Test

Die erwarteten Häufigkeiten sollten alle größer als 5 sein.

Voraussetzung nicht erfüllt

Erwartete Häufigkeit

	Männer	Frauen
Berufsschulen	20	0
Berufliche Gymnasien	12	14

Was macht man wenn das nicht erfüllt ist?

Falls möglich Kategorien zusammenfassen.

Auf einen Test verzichten und darauf hinweisen, dass die Gruppen zu klein sind für „weitere Tests“.

Chi-Quadrat-Test: Was man noch wissen sollte.

χ^2-Wert macht **keine Aussage, in welcher Gruppenkombination** es zu signifikant erhöhten Häufigkeiten kommt.

Abhilfe: Beobachtung der Differenzen zwischen erwartete und beobachtete Häufigkeiten. Es können zur Interpretation auch gern die Prozente betrachtet werden.

Die **Stärke des Zusammenhangs** wird mittels des p-Werts nicht angegeben. Dafür ist der Kontingenzkoeffizient notwendig (Siehe Kronthaler, F. (2016): Statistik angewandt: Datenanalyse ist (k)eine Kunst Excel Edition Springer Spektrum.)

Eine **dritte Variable** kann berücksichtigt werden, wenn zwei Chi-Quadrat-Test miteinander verglichen werden (Schichtung in SPSS)

Übungsaufgabe

$$\chi^2 = \sum_{i=1}^{q} \sum_{j=1}^{r} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Kleine Erhebung im Seminarraum

Beobachte Häufigkeit			
	Linke Seite	Rechte Seite	*Summe*
ÖPNV: Bus/Bahn			
Privat: Auto/zu Fuß			
Summe			

Wie lautet die Nullhypothese/Alternativhypothese.

Berechnen Sie den Chi-Quadrat-Wert. Wie interpretieren Sie ihn?

Chi-Quadrat-Test: Signifikanzniveau in Excel

Sie benötigen eine Tabelle mit beobachteten und erwarteten Messwerten, so wie es ein paar Folien weiter oben aufgezeigt wurde.

=CHIQU.TEST(Beobachtete_Messwerte;Erwarteter_Häufigkeiten)

z.B. =CHIQU.TEST(B3:C4;B9:C10)

Ergebnis z.B. 0,002

Das Ergebnis ist der p-Wert des Chi-Quadrat-Rests (Wir fragen uns nur: liegt der Wert über oder unter dem Signifikanzniveau von 0,05)

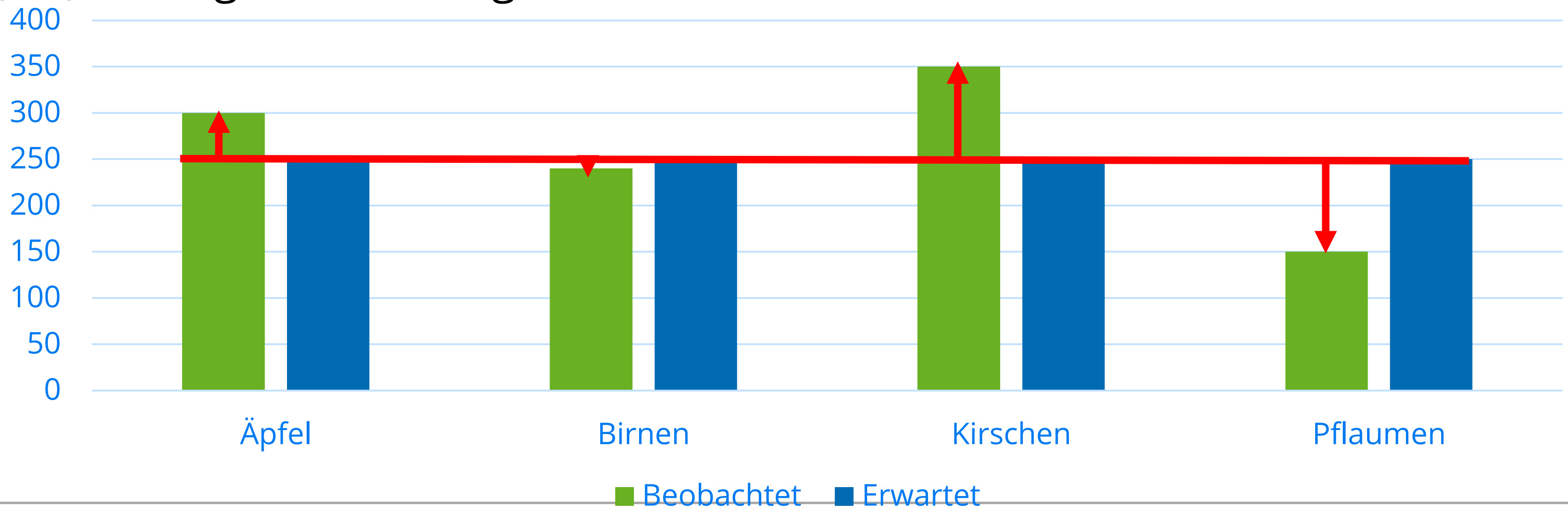
Mögliche Aussage hier: „Die Unterschiede in den Häufigkeiten zwischen den Gruppen sind signifikant.“

Chi^2-Anpassungstest

Ziel des Chi²-Anpassungstest

Sie haben Daten erfasst. Ihre Frage ist, ist die Verteilung der Daten zum Beispiel normalverteilt oder gleichverteilt?

Zum Beispiel könnte ein Obstbauer behaupten, er erntet von allen Früchten ungefähr gleich viel Kilogramm im Jahr (250 Kg). In der Grafik sind grün die beobachteten Werte den zu erwartenden gegenüber gestellt. Man sieht Abweichungen. Der Chi²-Anpassungstest schaut nun, ob die Abweichungen (rot) zufällig sind oder signifikant.



Chi²-Anpassungstest

	Beobachtet	Erwartet
Äpfel	300	250
Birnen	240	250
Kirschen	350	250
Pflaumen	150	250

Sie erhalten wieder einen χ^2 den Sie in der Tabelle für χ^2 nachschlagen können bei den Freiheitsgraden F: m-1 = 3.

$$\chi^2 = \sum_{j=1}^{m} \frac{(n_j - n)^2}{n}$$

$$n = \ erwartete\ Häufigkeit$$

$$n_j = beobachtete\ Häufigkeit$$

Chi²-Anpassungstest: kompliziertere Verteilungen - Nachdenken

Der Chi²-Anpassungstest kann theoretisch für jede Verteilungsform genutzt werden. Wie würden Sie vorgehen, wenn die Hypothese lautet, ihre Beobachtete Verteilung entspricht einer Normalverteilung?

Analysis of Variance (ANOVA)

ANOVA: Varianzanalysen im Überblick

Einfaktorielle Varianzanlayse		
Gruppe I	Gruppe II	Gruppe III
3	3	2
2	2	3
1	4	4
	5	5
	2	

Zweifaktorielle Varianzanlyse ohne Messwertwiederholung

Person	Mathe	Deutsch	Sport
A	3	3	2
B	2	2	3
C	1	4	4
D	3	5	5

Zweifaktorielle Varianzanalyse mit Messwertwiederholung

Person	Gruppe	Mathe	Deutsch	Sport
A	I	3	3	2
B	I	2	2	3
C	I	1	4	4
D	I	3	5	5
E	II	3	1	4
F	II	4	2	3
G	II	2	3	2
H	II	4	5	1

ANOVA: Einfaktorielle Varianzanlayse

Problem: Sind die Mittelwerte in drei (oder mehr) Gruppen unterschiedlich groß?

Nullhypothese =

Skalenniveau:

abhängige Variable: metrisch;

unabhängige Variable: nominal/ordinal mit zwei Ausprägungen (z. B. Geschlecht; Gruppe A/B)

Vorname	Leistungstest	**Gruppe**
Ilsa	22	A
Peter	31	A
Lisa	12	B
Klaus	46	B
Max	34	C
Mara	23	C

Einfaktoriele Varianzanalyse: Formel

$$F = \frac{MQA}{MQR} = \frac{\sum_{i=1}^{k} n_1 (\bar{x}_{i.} - \bar{x}_{..})^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{i.})^2 \,/ (N-k)}$$

Berechnen

A.

	Gruppe A	Gruppe B	Gruppe C	Alle
	1	1	1	
	2	2	2	
	3	3	3	

B.

Mittelwerte	2	2	2	2
Faktorstufen	1. Stufe	2. Stufe	3. Stufe	3
Anzahl Werte	3	3	3	9

C. Quadrierten Restabweichungen vom Mittelwert innerhalb der Faktorstufen.
Formel: $\text{Für Gruppe A} = (x_{Ai} - \bar{x}_A)^2$

	1	1	1	
	0	0	0	
	1	1	1	
Summe (SQR)				6
Freiheitsgrade (FG1). $= Anzahl\ Werte\ Alle\ - Anzahl\ Faktorstufen$				6
MQR: Mittlere quadrierten Restabweichungen innerhalb der Faktorstufe $= SQR/FG1$				1

D. Quadratsumme der Abweichung der einzelnen Mittelwerte vom Gesamtmittelwert x Anzahl Messwerte in der Faktorstufe: $\text{Für Gruppe } A = (\bar{x}_A - \bar{x})^2 * Anzahl\ Werte$

	0,00	0,00	0,00	
Summe (SQA)				0
Freiheitsgrade (FG2) $= Anzahl\ Faktorstufen\ - 1$				2
MQA: Mittlere quadrierte Abweichung zwischen den Faktorstufen. $= SQA/FG2$				0

E.

F-Wert: $F = MQA/MQR$				0

BA UNIVERSITY

Interpretation des F-Werts

$F = MQA/MQR$

Signifikanzniveau ermitteln: F-Wert wieder in einer Tabelle anschauen. Wie müssen die Freiheitsgrade des Zählers (in Spalten) und Nenners (in Zeilen) dafür nutzen. Abgelesen wird dort der kritische Wert. Liegt der ermittelte F-Wert über dem kritischem Wert aus der Tabelle, wird H0 zurückgewiesen:

https://de.wikibooks.org/wiki/Statistik:_Tabelle_der_F-Verteilung

0 = kein Zusammenhang (siehe vorherige Folie, H0 ablehnen)

Größere Werte stehen für größere Zusammenhänge

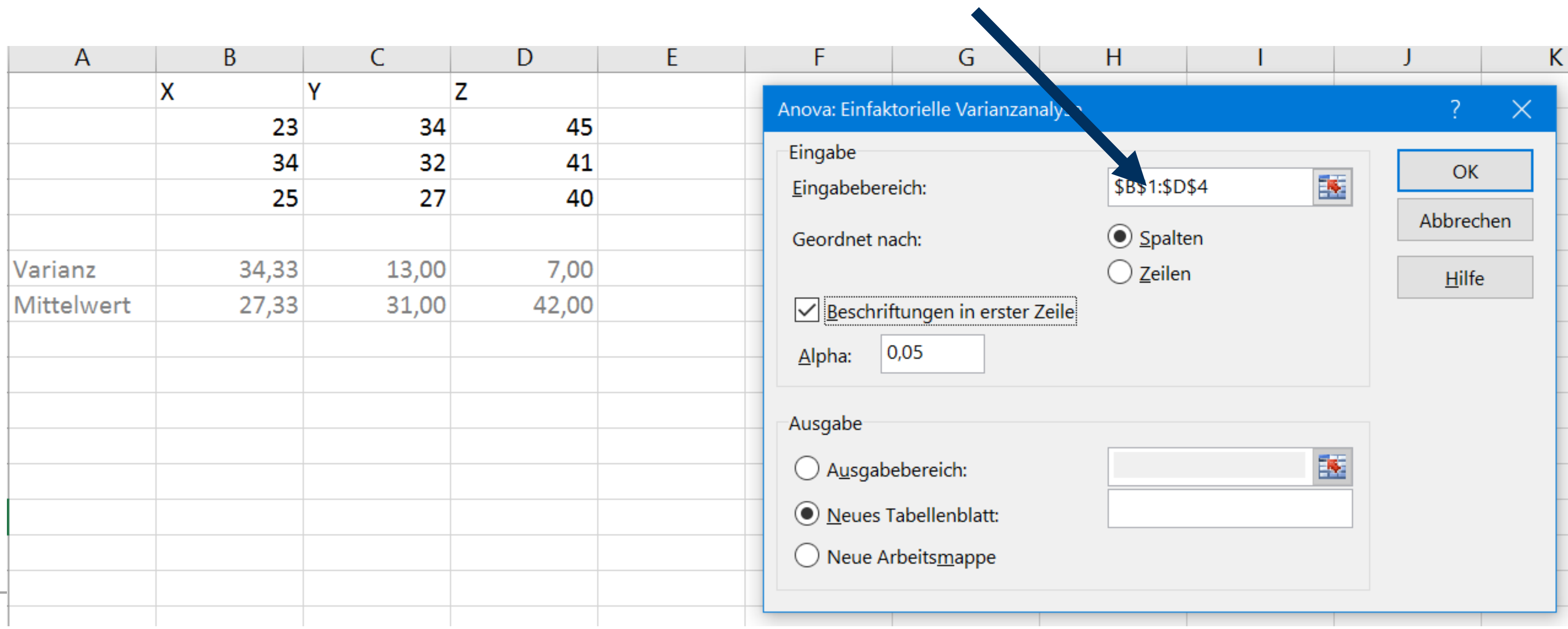
Mögliche Kombinationen von MQA und MQR und der resultierende F-Wert:

F-Wert wird:		MQA	
		Klein	Groß
MQR	Klein	-	Groß
	Groß	Klein	-

ANOVA: in Excel

Das Excel-Add-In Datenanalyse muss aktiviert sein

Daten → Datenanalyse → Einfaktorielle Varianzanalyse
Hier festlegen wie Ihre Daten geordnet sind



ANOVA: Ausgabe und Interpretation

Mittelwerte der drei Gruppen

p-Wert
Hier liegt er unter 0,05
→ Es konnte ein signifikanter Unterschied gefunden werden

Anova: Einfaktorielle Varianzanalyse

ZUSAMMENFASSUNG

Gruppen	*Anzahl*	*Summe*	*Mittelwert*	*Varianz*
X	3	82	27,33	34,33
Y	3	93	31	13
Z	3	126	42	7

ANOVA

Streuungsursache	*Quadratsummen (SS)*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*P-Wert*	*kritischer F-Wert*
Unterschiede zwischen den Gruppen	349,556	2	174,778	9,650	0,013	5,143
Innerhalb der Gruppen	108,667	6	18,111			
Gesamt	458,222	8				

Interpretation:
Die Nullhypothese, dass die Mittelwerter aller drei Gruppen gleich groß sind kann nicht beibehalten werden, sie wird zugunsten der Alternativhypothese – die Mittelwerte der Gruppen unterscheiden sich voneinander – zurückgewiesen.

Zweifaktorielle Varianzanalyse mit Messwertwiederholung: Einführung

	Früh	Mittags	Faktor 2
Stichprobe 1	1	4	
	2	5	
Stichprobe 2	4	1	
	5	2	
Faktor 1			

Zweifaktorielle steht für die Analyse von zwei Faktoren: Faktor 1 z.B.: zwei/drei/… Stichproben (z.B. Gruppe) Faktor 2 z.B.: und zwei/drei/… Tageszeiten/Schulfächer

Die Nullhypothese der zweifaktoriellen Varianzanalyse lautet: Es gibt keine Wechselwirkung (Interaktion) zwischen den beiden Faktoren

Zweifaktorielle Varianzanalyse mit Messwertwiederholung: Überlegung

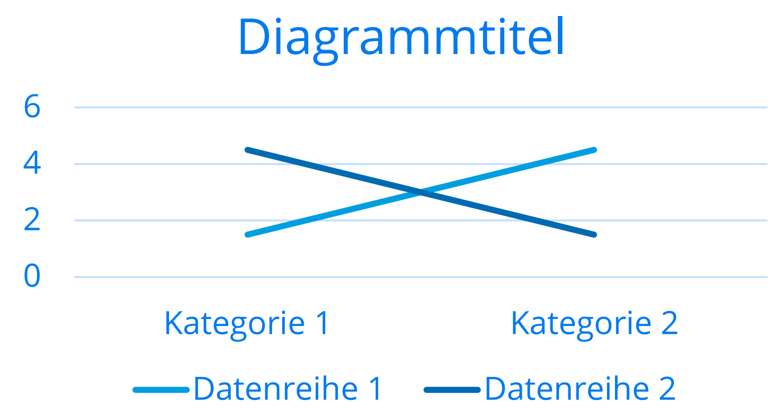
	Früh	Mittags
Stichprobe 1	1	4
	2	5
Stichprobe 2	4	1
	5	2

Welche Mittelwerte haben Stichprobe 1 und Stichprobe 2 jeweils?

Welche Mittelwerte haben „Früh“ und „Mittags“ jeweils?

Zweifaktorielle Varianzanalyse mit Messwertwiederholung: Interaktion

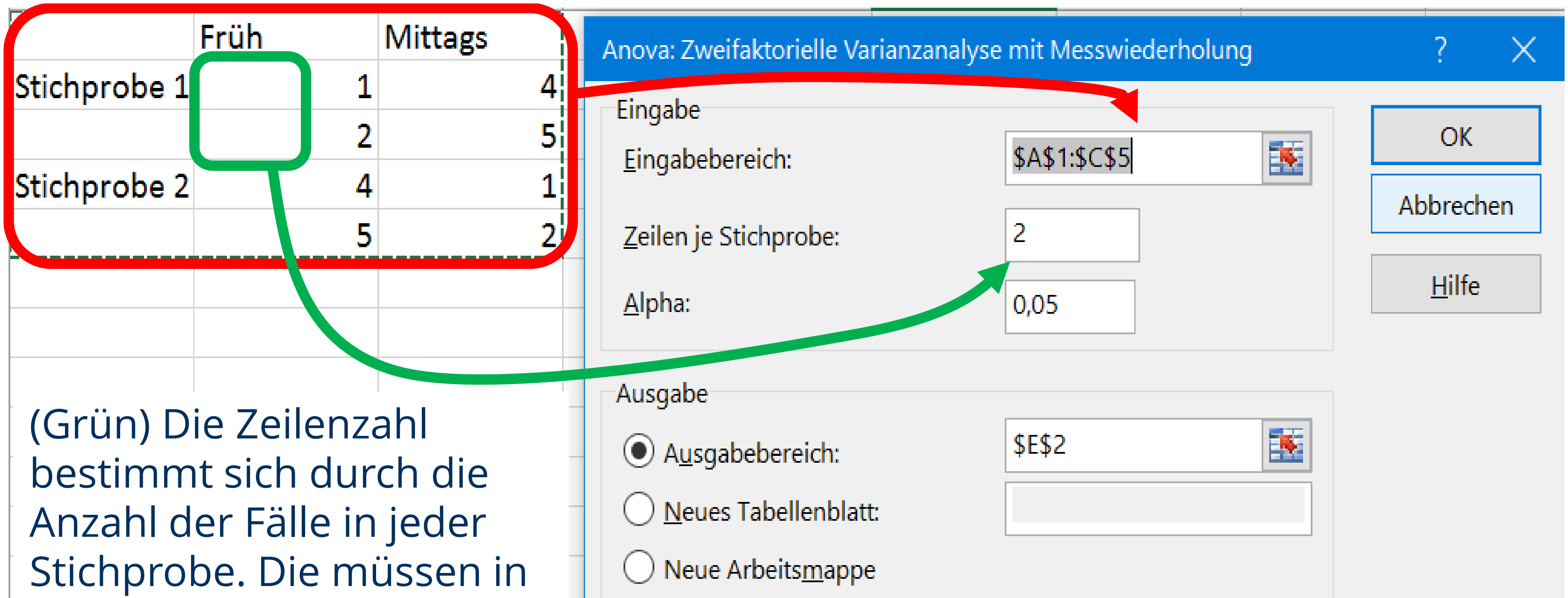
	Früh	Mittags
Stichprobe 1	1	4
	2	5
Stichprobe 2	4	1
	5	2



Die zweifaktorielle Varianzanalyse fragt, ob es eine Interaktion zwischen beiden Faktoren gibt.
In unserem Beispiel sehen wir, dass in Stichprobe 1 früh die Werte geringer sind als Mittags, wobei in Stichprobe 2 das genau umgedreht ist, das ist die Interaktion.

Zweifaktorielle Varianzanalyse mit Messwertwiederholung in Excel

(Rot) Die gesamte Tabelle mit Beschriftung eingeben. Die Daten müssen in dieser Form vor liegen.



(Grün) Die Zeilenzahl bestimmt sich durch die Anzahl der Fälle in jeder Stichprobe. Die müssen in beiden Stichproben gleich sein!

Zweifaktorielle Varianzanalyse mit Messwertwiederholung interpretieren

Der p-Wert für den Vergleich von Stichprobe 1 und Stichprobe 2 ist 1 also größer 0,05 also insignifkant.

Der p-Wert für den Vergleich von früh/mitags ist auch 1 also größer 0,05 also insignifikant.

ANOVA						
Streuungsursache	*adratsummen (*	*Freiheitsgrade (df)*	*Mittlere Quadratsumme (*		*t*	*kritischer F-Wert*
Stichprobe	0	1	0	0	1	7,708647422
Spalten	0	1	0	0	1	7,708647422
Wechselwirkung	18	1	18	36	0.003882537	7,708647422
Fehler	2	4	0,5			
Gesamt	20	7				

Das ist der p-Wert für die zweifaktorielle Varianzanalyse. Er sagt aus, ob es eine Wechselwirkung gibt. Liegt der Wert unter 0,05 so wird die Nullhypothese abgelehnt und wird entscheiden uns für die Alternativehypothese.

Mögliche Aussage:

Die zweifaktorielle Varianzanalyse erbrachte ein signifkantes Ergebnis, dies bedeutet, dass eine Wechselwirkung zwischen den Stichproben und der Tageszeit (früh, mittags) vorliegt.

Zusätzliche Infos: Einen signifikanten Unterschied zwischen den Stichproben alleine gibt es hingegen nicht. Ebenso konnte kein signifikanter Unterschied allein zwischen den Tageszeiten festgestellt werden.

Zweifaktorielle Varianzanalyse: 2. Beispiel (Anzahl der Messwerte und Gruppen erhöhen)

	Früh	Mittags	Abends
Gruppe 1	1	4	5
	2	5	4
Gruppe 2	4	1	2
	5	2	1
Gruppe 3	1	1	4
	2	2	5

ANOVA: Such nach den Unterschieden (Bonferoni Korrektur)

Die ANOVA sagt nicht, in welchen Gruppen ein Unterschied besteht.

Wegen der Gefahr von zu vielen Signifikanztests können wir nicht einfach zwischen allen Gruppen einen t-Test machen.

Aber wir können das machen, wenn wir das Signifikanzniveau anpassen. Teilung des „Standard"-Signifikanzniveaus (0,05) durch die Anzahl der Tests (Bonferoni-Post-Hoc-Test).

Beispiel:

Aus einer ANOVA mit drei Gruppen mache ich drei t-Tests zwischen Gruppe 1 und Gruppe 2, zwischen 1 und 3 und 2/3.

Als Signifikanzniveau wähle ich 0,05 geteilt durch 3 (weil ich drei Tests mache). Das neue Signifikanzniveau ist 0,01667.

Ich vergleiche nun die p-Werte meiner drei t-Test jeweils mit den neuem Signifikanzniveau (nach Bonferoni-Korrektur) 0,01667.

Liegt ein p-Wert unter 0,01667 so unterscheiden sich diese beiden Gruppen signifikant voneinander.

Wiederholung

Quiz Teil I

Frage	Richtig	Falsch
Bei der einfachen linearen Regression beeinflusst ein oder mehrere Variablen die abhängige Variable.		
Das konstante Glied der einfachen linearen Regressionsanalyse entspricht dem Wert des Y-Achsenschnittpunktes der linearen Regressionsgeraden.		
Der 1. Schritt in jedem stat. Testverfahren besteht in der Entscheidung, ob die Nullhypothese oder die Alternativhypothese getestet werden soll.		
Der Korrelationskoeffizient kann nur Wert zwischen 0 und 1 annehmen		
Der Median wird von Ausreißern beeinflusst.		
Der Mittelwert ist gegenüber Ausreißern robust.		

Quiz Teil II

Frage	Richtig	Falsch
Der Modal wert ist der Wert, der genau in der Mitte der geordneten Verteilung liegt.		
Der Regressionskoeffizient entspricht der Steigung der linearen Regression		
Die Spannweite wird nie von Ausreißern beeinflusst.		
Die Standardabweichung berechnet sich als positive Wurzel aus der Varianz.		
Die Wahrscheinlichkeiten aller möglichen Elementarereignisse eines Zufallsvorgang ergeben zusammenaddiert den Wert 2.		
Diskrete Variablen mit sehr vielen Ausprägungen gelten auch als quasi stetig.		

Quiz Teil III

Frage	Richtig	Falsch
Ein korrekt durchgeführter stat. Test gestattet eine definitive Aussage über die Korrektheit von Null- und Alternativhypothese.		
Ein Korrelationskoeffizient von -0,85 deutet auf eine starke lineare Korrelation hin?		
Ein Zufallsexperiment ist die beliebig häufige Wiederholung eines Zufallsvorgang unter gleichen Rahmenbedingungen.		
Eine zufällig gezogene Stichprobe mit hoher Rücklaufquote ist unabhängig von ihrem Umfang stets repräsentativ.		
Erwartungswert, Median und Modus einer normalverteilten Variablen alle ungleich groß.		
Es gibt keine Zufallsvariablen die diskret sind.		

Quiz Teil IV

Frage	Richtig	Falsch
Je mehr Hypothesen man an einem Datensatz testet, desto höher wird die Wahrscheinlichkeit, dass eine davon fehlerhaft als zutreffend angenommen wird.		
Kreisdiagramme eignen sich eher für stetige als für diskrete Daten.		
Nominalskalierte Daten können in eine natürliche Reihenfolge gebracht werden.		
Ordinalskalierte Daten können in eine natürliche Reihenfolge gebracht werden.		
Stetige Daten sollten vor der Erstellung von Säulendiagrammen klassiert werden.		
Streudiagramm zeigen die Verteilung von zwei Variablen.		

Chi²-Streuungstest

Ziel des Chi²-Streuungstest

Sie vergleichen die Varianz Ihrer Daten mit der Varianz in der Grundgesamtheit. Die Frage ist, weicht Ihre Varianz signifikant davon ab.

Vorrausetzung: Normalverteilung (zur didaktischen Vereinfachung wird im Beispiel aber nicht normalverteilte Daten verwendet).

Vergleich von einer Varianz mit der Varianz der Grundgesamtheit

Stichprobe	Grundgesamtheit
1	1
2	3
3	5
4	7
Varianz: 1,6	**Varianz: 6,6**

$$\chi^2 = \frac{(n-1) * s^2}{\sigma^2}$$

$$s^2 = Varianz\ der\ Stichprobe$$

$$\sigma^2 = Varianz\ der\ Grundgesamtheit$$

Gaustest

Aufgabe Gaußtest / z-Test
Wissen selbst erarbeiten

Wie lautet die Nullhypothese und die Alternativhypothese?

Erfinden Sie ein Anwendungsbeispiel bei den es um den Mittelwert von Streichhölzern geht.

Welche Informationen über die Stichprobe und über die Grundgesamtheit braucht der Test.

Welche Voraussetzungen in den Daten müssen für den Test erfüllt sein?

Wie lautete die Formel für den Gaußtest.

Was bedeutet ein signifikantes und ein nicht signifikantes Ereignis, was würde dies jeweils beim Streichholzbeispiel bedeuten.

Konstruieren Sie ein einfaches Rechenbeispiel aufbauend auf dem Streichholzbeispiel.

Suchen Sie sich am Ende eine Kommilitonin oder Kommilitonen (oder eine Gruppe) mit der Sie Ihr Wissen abgleichen.

Verteilungsformen

Verschiedene Verteilungen

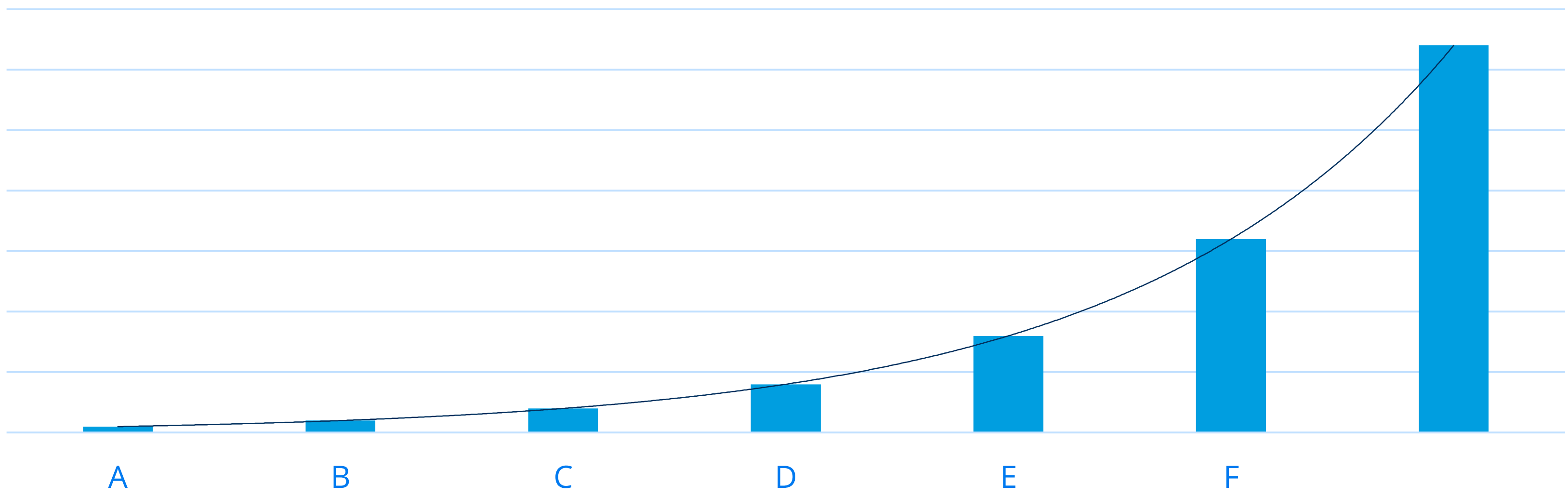
Bereits behandelt:

- **Zentraler Grenzwertsatz:** Normalverteilung
- **Häufigkeitsverteilung:** Gleichverteilung, Zweigipflige Verteilung
- **Wahrscheinlichkeitsrechnung:** Hypergeometrische-, Binominal- und Geometrische Verteilung
- **Testverteilungen:** t-Verteilung, Chi²-Verteilung, F-Verteilung, Normalverteilung

Weitere Verteilungen:

- Exponentielle Verteilung

Exponentielle Verteilung, Poissonverteilung

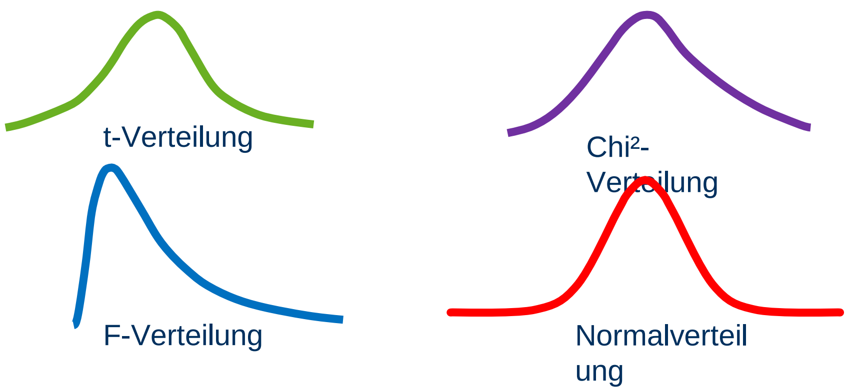


Exponentialverteilung

Eigenschaften von Verteilungen

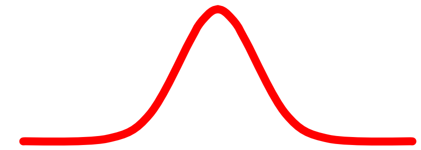
Jede dieser (idealen) Verteilung kann mittels Varianz, Erwartungswert, Verteilungsfunktion und Dichtefunktion beschreiben werden:

https://de.wikipedia.org/wiki/Liste_univariater_Wahrscheinlichkeitsverteilungen



- Für empirische Verteilungen in den Sozialwissenschaften ähneln solchen theoretischen Verteilungen übrigens ... nie. Oder sagen wir so gut wie nie.

Nachtrag zur Normalverteilung: Schiefe und Wölbung



Die Normalverteilung kann abweichen von der Idealform. Dafür gibt es zwei Messwerte: Schiefe und Wölbung. Was heißt das, wie lautet die Formel zur Berechnung? Erstellen Sie eine Beispielaufgabe.

Schiefe

https://www.statistik-nachhilfe.de/ratgeber/statistik/deskriptive-statistik/masszahlen/parameter-der-form/schiefe

Wölbung

https://www.statistik-nachhilfe.de/ratgeber/statistik/deskriptive-statistik/masszahlen/parameter-der-form/woelbung-exzess-kurtosis

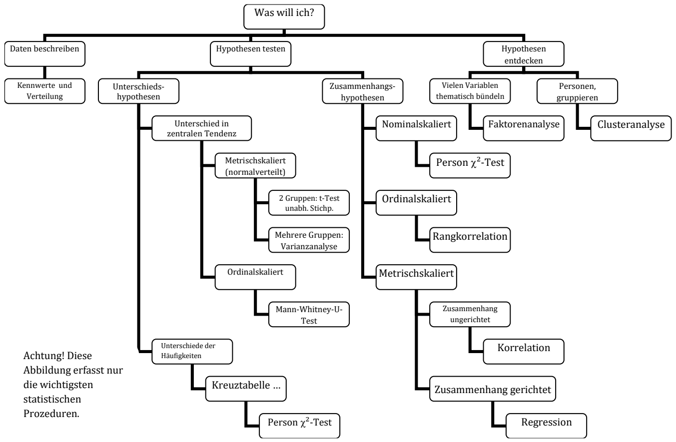
Entscheidungsbaum

Entscheidungsbaum für die Statistik und quantitative Methoden Veranstaltung

Aufgabe: Gesucht ist ein Entscheidungsbaum, der alle behandelten Verfahren der Veranstaltung beinhaltet.

Hinweise: Bedenken Sie, dass es mehrere Möglichkeiten gibt für den endgültigen Entscheidungsbaum.

Beispiel: https://www.methodenberatung.uzh.ch/de/datenanalyse_spss.html



Zwischenschritte/Teilfragen:

- Welche statistischen Verfahren wurden behandelt?
- Was definieren Sie als Verfahren
- Nach welchen Kriterien möchten Sie den Baum einteilen

Ihr Entscheidungsbaum

Ergebnisse veröffentlichen und Reflektieren

Ethik und Statistik

„Glauben Sie nur der Statistik, die Sie selbst gefälscht haben."

Was bedeutet es wenn Sie mit Statistik „lügen" für

1. Interne unternehmerische Ziele
2. Für Marketingzwecke
3. Für medizinische Zwecke

Veröffentlichung einer empirischen und statistischen Arbeit: Übersicht

- Einleitung
- Forschungsstand und Hypothesen
- Methoden und Daten
- Ergebnisse
- Ausblick/Diskussion
- Zusammenfassung
- Anhang

Im Folgenden nur die wesentlichen Stichpunkte. Suchen Sie nach „Aufbau einer empirischen Arbeit“ dann finden Sie ausführlichere Tipps. z.B.
http://www.uni-bielefeld.de/ikg/daten/Gliederung_einer_empirischen_Arbeit.pdf
https://bildungswissenschaften.psychologie.sowi.uni-mainz.de/files/2015/06/emp_arbeit.pdf

Veröffentlichung: Einleitung

Verortung, Bedeutung des Themas

Evtl. Historie des Themas

Was macht der Artikel (wie und wo)

„In dieser Arbeit wurde untersucht wie A und B auch ohne C auskommt. Dafür habe ich ABC und XYZ in einer Gruppe von Menschen gemessen. Im Absatz Methoden wird der Fragebogen vorgestellt … die Arbeit schließt mit einer Zusammenfassung.“

Abgrenzung: Was macht die Arbeit nicht (was man eigentlich Erwarten könnte)

Veröffentlichung: Forschungsstand und Hypothesen

(Hinweise: Hier nur zusätzliche Informationen für Arbeiten in denen methodisch und statistisch gearbeitet wurde, Weiteres entnehmen Sie vermutlich auch anderen Vorlesungen)

Zentrale Begriffe oder Zusammenhänge erläutern

(Der Theorieteil ist in empirischen Arbeiten kurz und es reicht, wenn Sie auf andere Literatur verweisen.)

Relevante Ergebnisse anderer kurz vorstellen.

Auf Forschungslücken und Widersprüche verweisen.

Hinüberleiten zu eigenen Forschungshypothesen (die sie testen werden)

Evtl. theoretische Begründung Ihres vermuteten Zusammenhangs erläutern.

Evtl. Ihre neuen Begriffe einführen.

Veröffentlichung: Methoden und Daten I

Begründung der Erhebungsmethoden (Fragebogen), welche Alternativen gäbe es.

Beschreibung der Erhebungsmethode. Welche Pretests, Woher kommen Fragen, Wie wurden Fragen entwickelt.

Operationalisierung der Hypothesen: Welche Frage misst welches theoretischen Konstrukt aus Ihren Hypothesen.

Beschreibung der Stichprobe (hier evtl. wichtige deskriptive Statistiken). Wer und wann und wo?

Erwähnung von Besonderheiten während der Erhebung.

… Fortsetzung nächste Folie

Veröffentlichung: Methoden und Daten II

Verwendete Variablen beschreiben: Was bedeuten die Werte. Welche Datentransformationen wurden vorgenommen. Wie berechnen sich neue Variablen (Summen, Indizes)

Verwendetes statistisches Verfahren nennen mit Bezug auf die Hypothese:

- „Um die Hypothese XY zu testen habe ich eine Regression mit der abhängigen Variable Y verwendet und der unabhängigen Variablen X1 und X2. Die Nullhypothese der Regression besagt, dass die unabh. Var. Keinen Einfluss haben. Eine Betätigung meiner Hypothese XY ist dann gegeben, wenn die Koeffizienten von X1 und X2 signifikant auf dem Niveau von 0,05 sind.“

Nur spezielle Statistiken erklären

Veröffentlichung: Ergebnisse I

Mit wichtigen einfachen Ergebnissen beginnen. Auch spannende Grafiken. Jede Tabelle, jede Grafik im Text beschreiben.

Große Tabellen mit nebensächlicher Information nur in Anhang.

Evtl. Tabellen mit wichtigen Ergebnissen für statistische Tests, falls mehr als 3 Zahlen gezeigt werden sollen.

Veröffentlichung: Ergebnisse II

Signifikante Ergebnisse: Darauf eingehen, ob eine Hypothese abgelehnt wird oder nicht. Auch nicht signifikante Ergebnisse sind erwähnenswert!

„Die Ergebnisse der Zweifaktoriellen Varianzanalyse mit Messwertwiederholung finden sich in Abbildung X. Die Berechnung der unterschiede in den Stichproben und in den Spalten (Früh, Mittags) ist nicht signifikant. Das bedeutet, die Nullhypothesen – es gibt keine Unterschiede zwischen den Stichproben und den Spalten - wird beibehalten. Hingegen ist der p-Wert für die Wechselwirkung der Faktoren (Gruppen/Tageszeit) signifikant. Das bedeutet es gibt eine Wechelwirkung zwischen Tageszeit und Gruppen. Ein Blick auf das Häufigkeitsdiagramm zeigt, … "

Evtl. zusätzliche Analysen vorstellen und begründen warum sie notwendig sind

Veröffentlichung: Ausblick/Diskussion

Inhaltliche Interpretation der Ergebnisse (eigene Interpretation und mit Hilfe der Literatur); Bedeutung für Fragestellung.

Eventuell neue Hypothesen formulieren für zukünftige Forschung

Andeuten was man in zukünftigen Untersuchungen anders machen könnte

Schlussfolgerungen für die Praxis ziehen

Grenzen der Untersuchung klar machen (es ist nur eine Stichprobe, Besonderheit der untersuchten Personen).

Selbstkritik an den verwendeten Prozeduren/Skalen

Veröffentlichung: Zusammenfassung

Die Ziele des Artikels wiederholen, von den Ergebnissen nur die Bedeutendsten erwähnen. Nichts Neus hier!

In wissenschaftlichen Artikeln gibt es meist noch eine zweite Zusammenfassung die aus weniger als 10 Sätzen besteht. Anhand dieses „Abstracts“ (engl.) können Leser schnell einschätzen, ob der Artikel für sie relevant ist oder nicht. (Für den Autor ist er auch gut, weil er so seine Kerngedanken und Ergebnisse noch mal hervorbringt.)

Da Leser wenig Lust haben viele Aufsätze vollständig zu lesen, sollte die Zusammenfassung tatsächlich alles zusammenfassen.

Veröffentlichung: Anhang

Jede Tabelle und Grafik sollte aus sich heraus so gut es geht allein verständlich sein.

Nur Tabellen hier die zu groß sind für den Text und dort auch nur am Rande erwähnt werden.

Oft werden deskriptive Statistiken (Häufigkeiten, Prozentangaben) hier veröffentlicht.

Im Anhang ist auch Platz für verwendete Fragebögen

Komplizierte Berechnungen (neue Variablen, unbekannte statistische Verfahren) können hier auch gezeigt werden.

Weitere Tipps für empirische Arbeit

- Protokoll führen (so lange bis man es nicht mehr brauch)
- Sicherheitskopie vom Originaldatensatz
- Zwischenziele aufschreiben
- Notieren, welche Fälle/Variable ausgewählt worden
- Erwähnenswerte (Zwischen-)ergebnisse notieren oder aus der Auswertung in eine extra (Excel-) Datei kopieren
- Konfuse Ergebnisse löschen
- Ergebnisse die veröffentlichungswert erscheinen, in Sprache umformen. Dann erkennt man was vielleicht falsch gemacht wurde bzw. welche weiteren Schritte man unternehmen könnte.
- Bei komplizierten Vorgängen genau aufschreiben was man vor hat
- Eigene Ideen/Kritik an der Methode festhalten

Tipps für Diagramme

Diagramme sollten meist bearbeitet werden (so wenig wie möglich Informationen aber alle nötigen Informationen: Skalentitel, Legende, Einheiten).

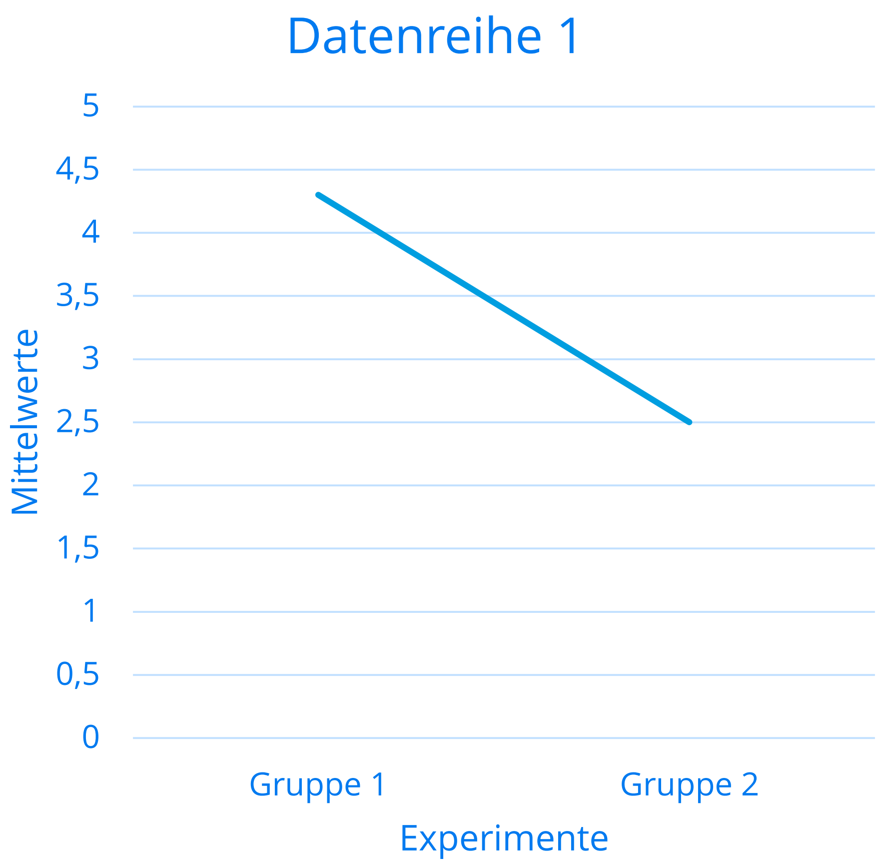
Das Anklicken einzelner Elemente erlaubt detaillierte Bearbeitung.

Es geht einfacher ein kompliziertes Diagramm zu erstellen, wenn man vorher auf Papier skizziert wie es im Endeffekt aussehen soll.

Excel macht im Vergleich zu andere Statistikprogrammen sehr gute Diagramme.

Kleine Fehler finden

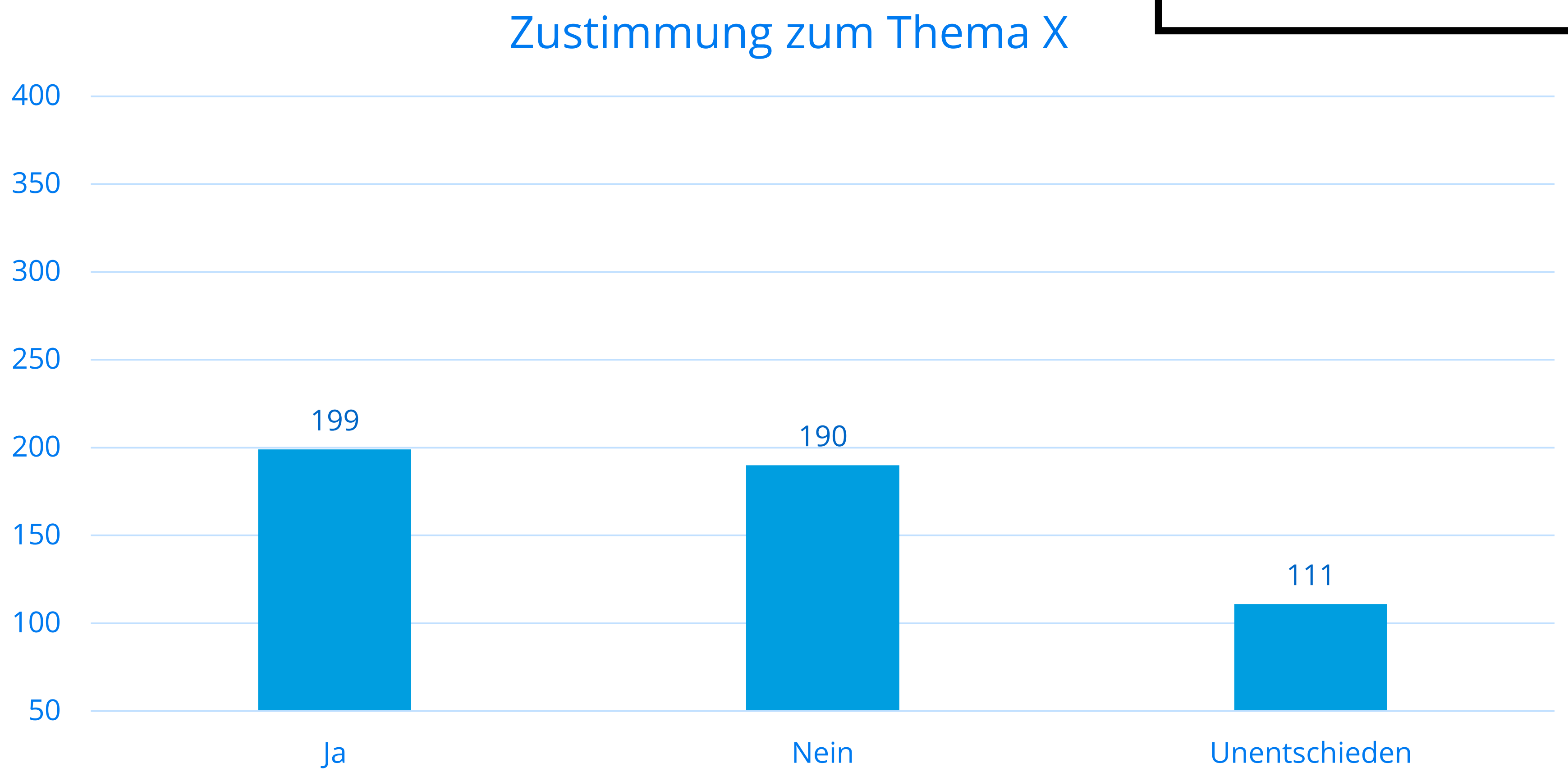
Schlechtes Beispiel: Grafik



Was ist hier das Problem?

Schlechtes Beispiel: Säulendiagramm

Zwei Fehle gibt es in dieser Grafik. Was ist an dieser Grafik nicht gut gelungen?



Schlechte Beispiele: Tabelle

	Unternehmen A	Unternehmen B	N
Frauen	700	500	1200
	31,111%	22,222%	53,333%
Männer	400	650	1050
	17,777%	28,888%	46,666%
n	1100	1150	2250

Welche Kritikpunkte sehen Sie?:

-
-
-
-
-
-

Schlechte Beispiele: Manipulation des Mittelwerts

Mittelwert Notendurchschnitt von zwei Klassen. Welche Schlussfolgerung könnten Sie ziehen? Ist diese sinnvoll?

Klasse 1	
1	
1	
1	
1	
1	
5	
Mittelwert:	
Median:	

Klasse 2	
1	
1	
1	
2	
2	
3	
Mittelwert:	
Median:	

Schlechte Beispiele: Prozente

In einer Analyse eines Fragebogens steht die Aussage: „Der Anteil der Männer (40 %) war in der Stichprobe geringer als der von Frauen (60 %).“ Wie ensteht die Aussage bei einer Gesamtzahl von 5 Befragten? Warum ist es nicht sinnvoll, diese Aussage zu machen?

Rohdaten:

Nummer	Geschlecht
1	
2	
3	
4	
5	

Stichproben wählen

Befragung zum Thema Lärmbelästigung von Anwohnern.

Stichprobe: Alle Anwohner, die Mitglieder des Vereins „Grüne Stadt“ sind.

Was ist nicht gut an dieser Stichprobe?

Bezugspunkte ändern

Ein Sportverein hat die Gruppen Junioren und Senioren. Leute mittleren Altern dürfen frei wählen. Was passiert, wenn der 35-jährige wechselt?

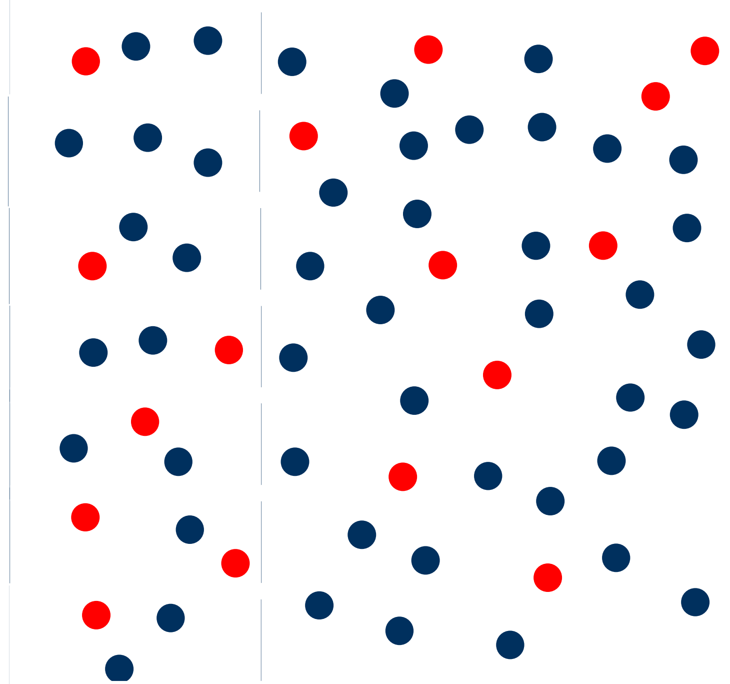
	Junioren	Senioren
	5	
	10	40
	15	45
	20	50
	25	55
	30	65
	35	70
Mittelwert **vor** Wechsel:	20	57,1

	Junioren	Senioren
Mittelwert **nach** Wechsel:		

Stichprobengröße: Ein kleines Dorf mit sehr alten Leuten

- Es gibt kleine Dörfer in dem sehr viele Menschen sehr alt sind.
- Liegt das am gesunden Lebensstil oder auch an der Statistik?
-

Räumliche Verteilung von Menschen nach Altersgruppen



Fehlschluss: Große Stichproben sind immer gut

Was passiert hier durch die größere Stichprobe?

Gruppe 1	Gruppe 2
1	2
2	3

Gruppe 1	Gruppe 2
1	2
2	3
1	2
2	3

p-Wert des t-Tests:
0,293

p-Wert des t-Tests:
0,050

Trends zu gewagt

Körpergröße von 18 jährigen Männern in Metern. Welcher Trend lässt sich anhand der Daten formulieren. Wäre das sinnvoll?

1970: 1,80

1990: 1,85

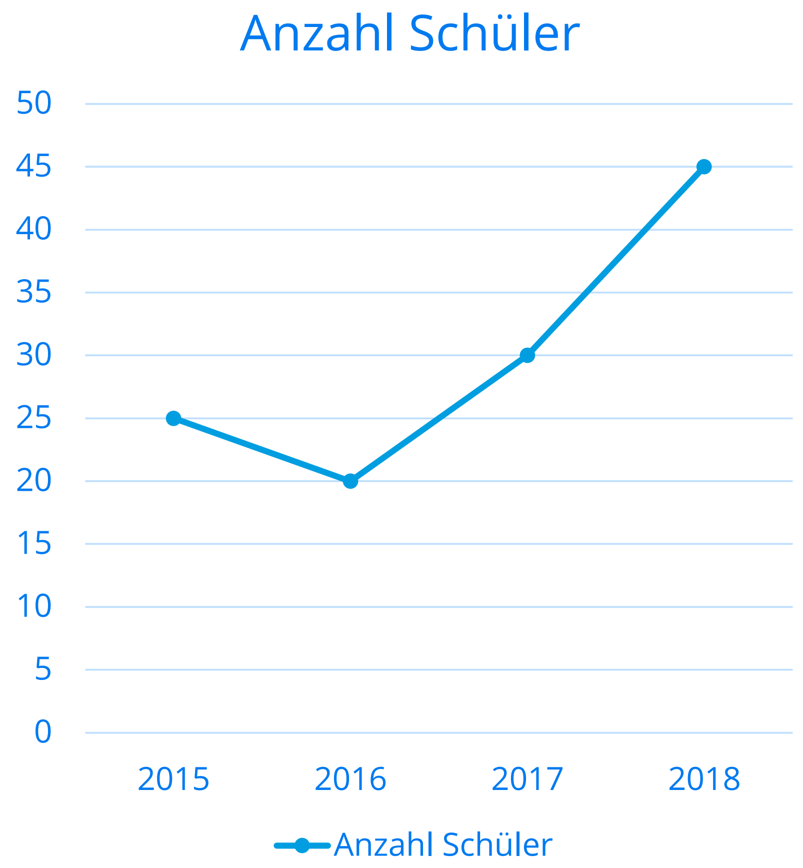
2010: 1,90

2030:

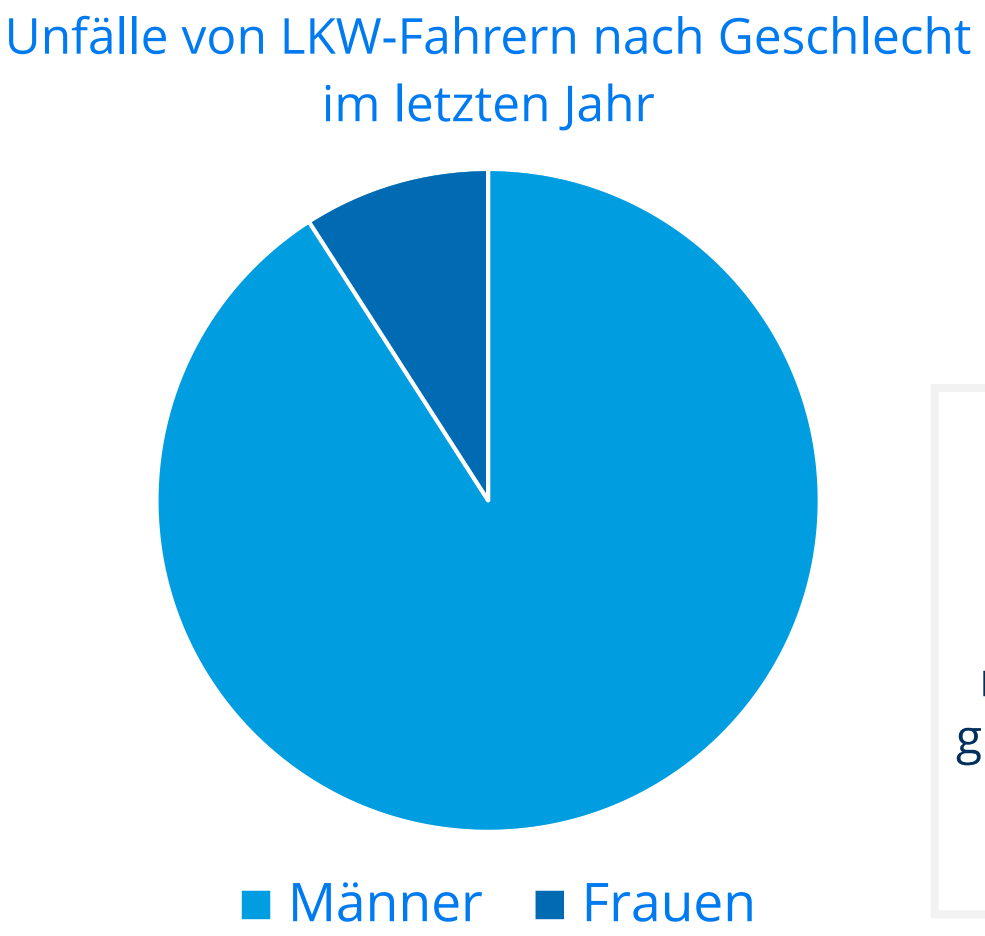
Quelle: Statista. Lügen mit Statistiken https://de.statista.com/statistik/lexikon/definition/8/luegen_mit_statistiken/
Abgerufen: 10.12.2018

Trends zurechtbiegen

Was könnte die nebenstehende Grafik verheimlichen?



Bezugsrahmen verheimlichen



Ihr Unternehmen analysiert die Unfallstatistik bei den angestellten LKW-Fahrenden nach Geschlecht. Jemand zeiht den Schluss: es müssen mehr Frauen eingestellt werden, weil die weniger Unfälle machen.

Was fehlt hier? Erfinden Sie Daten, die genau die gegenteilige Aussage nahelegen würden, aber zum gleichen Kreisdiagramm führen könnten.

Prozente können schön klingen

Partei X 2010:
Partei X 2015:

Finden Sie Datenbeispiele die die nebenstehenden Aussagen zulassen.

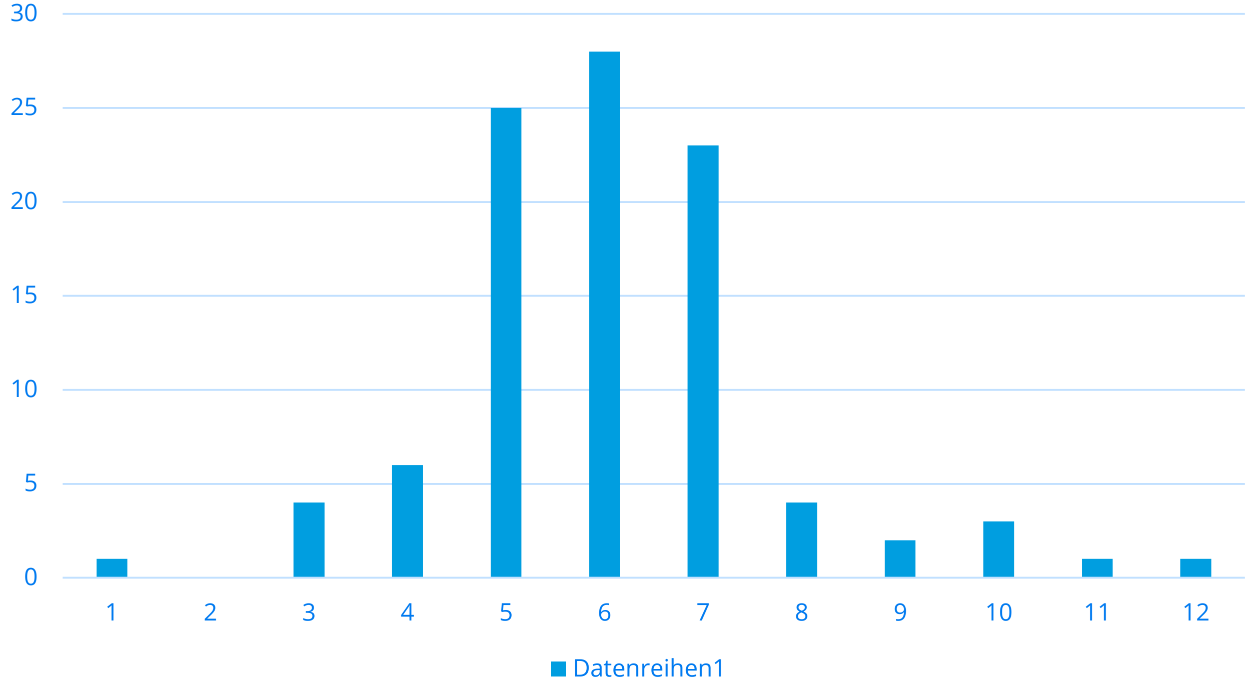
Verdopplung des Frauenanteils in einer Partei X

Anstieg um nur 20 Prozent in Partei Y

Partei Y 2010:
Partei Y 2015:

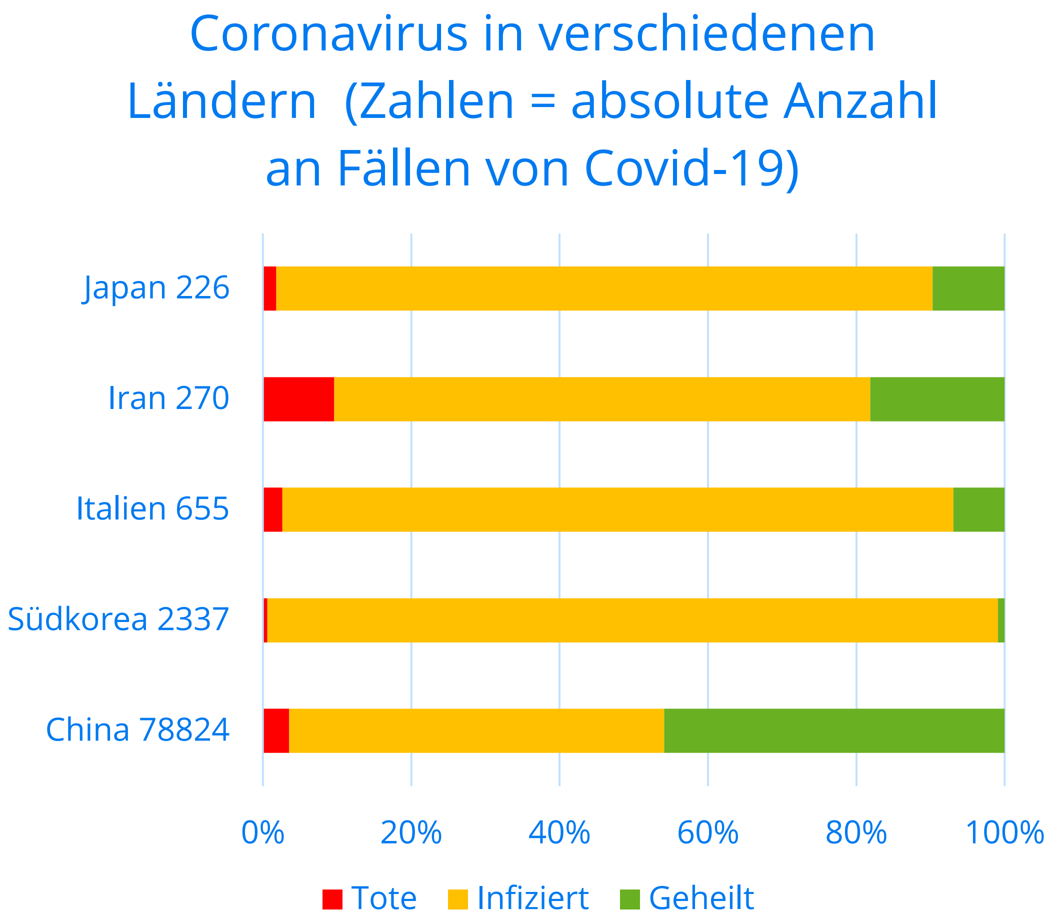
„Normalerweise passiert das nicht“

Abgebildet sind die Lieferzeiten Ihres Möbelwarengeschäfts. Ein Kunde beschwert sich, er hat eine Lieferzeit von 12 Tagen und meint, das ist doppelt so lang wie der Durchschnitt. Das ist doch nicht normal.



Was antworten Sie dem Kunden?

Zeitpunkt der Statistik



Datenquelle de.Statista.com
28.02.2020

Probleme mit dieser Statistik

-
-
-

Datenbanken

Woher nehme ich Daten, wenn ich oder meine Organisation sie nicht selbst erhoben haben?

(Daten-)Datenbanken

Wer stellt sie bereit?

- Staatliche Institutionen
- Forschungseinrichtungen und nichtstaatliche Organisationen (Themengebunden)

Warum werden sie bereitgestellt

- Berichtspflichten von staatlichen Einrichtungen
- Interesse von Organisationen

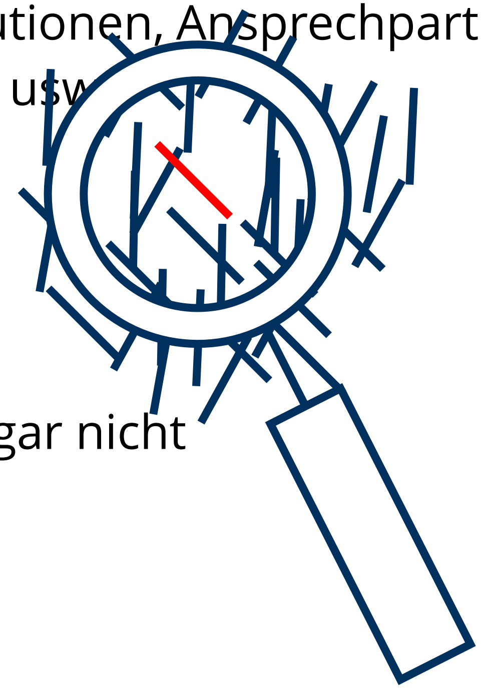
Was wird dort angeboten?

- Inhalt
- Variablen
- Fälle
- Anleitung zum Lesen von Daten
- Evtl. verwendete Fragebögen

Problem bei Suche nach Datenbanken

Unter dem Begriff „Datenbanken“ versteckt sich sehr viel:

- Datenbanken für Zeitschriften
- Datenbanken für Listen mit Organisationen, Institutionen, Ansprechpartner
- Datenbanken für Informationsseiten, News-Artikel usw.
- Datenbanken für Projekte, Geldgeber
- Datenbanken für …
- Datenbanken für Daten



Daten zu finden ist nicht leicht.

Daten zu spezifischen Fragestellungen gibt es evtl. gar nicht

Wie findet man DATEN-Datenbanken: Metasuchen

Registry of Research Data Repositories

https://www.re3data.org

Suche einschränken Hummanitiers and Social Sciences und Empirical Social Research

→ Browse

→ Browse by subject

Weitere Einschränkung z.B. „country“ Germany und „subjects“ auf der linken Seite

Fachdatenbanken in Bibliotheken

http://rzblx10.uni-regensburg.de/dbinfo/fachliste.php?bib_id=slub

Datenbank-Infosystem (DBIS)
Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden

SLUB Wir führen Wissen.

SLUB-Katalog | Digitale Sammlungen | Beratung | Literaturverwaltung | Open Access / Bibliometrie | Veranstaltungen

Schnelle Suche

Go!

Erweiterte Suche

Aktuelles
Fachübersicht
Alphabetische Liste
Sammlungen
Hinweise zur Benutzung
Ansprechpartner
Bibliotheksauswahl / Einstellungen
Über DBIS

Gefördert durch:

Fachübersicht

Fachgebiete	Anzahl
Allgemein / Fachübergreifend	1382
Allgemeine und vergleichende Sprach- und Literaturwissenschaft	332
Anglistik, Amerikanistik	187
Archäologie	178
Architektur, Bauingenieur- und Vermessungswesen	229
Biologie	344
Chemie	172
Elektrotechnik, Mess- und Regelungstechnik	62
Energie, Umweltschutz, Kerntechnik	161
Ethnologie (Volks- und Völkerkunde)	154
Geographie	236
Geowissenschaften	146
Germanistik, Niederländische Philologie, Skandinavistik	478

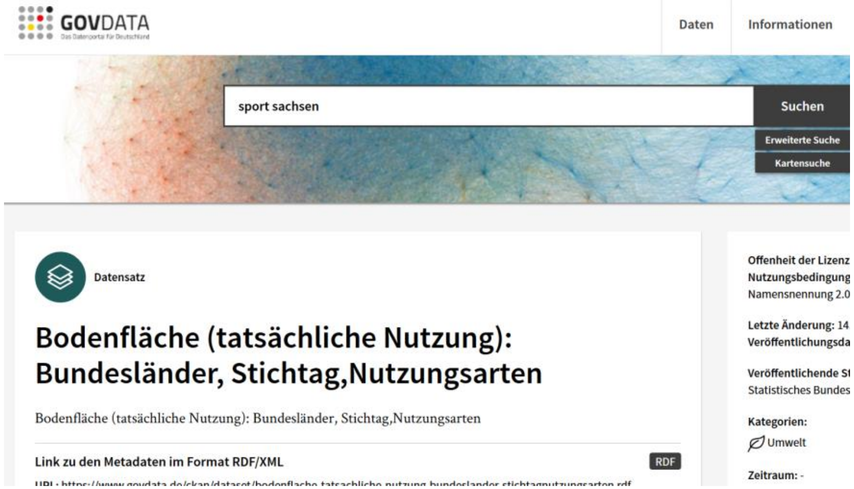
Beispiele für eine Datenbanken, die Sie über die

Amadeus (Finanzdaten von Unternehmen)

https://amadeus.bvdinfo.com/version-2019919/Search.QuickSearch.serv?_CID=1&context=36H0I3VCVYADEY6

GOVDATA: Beispiel Bodenfläche Nutzung. Gefunden über die Suchbegriffe „Sport Sachsen“

https://www.govdata.de/web/guest/suchen/-/details/bodenflache-tatsachliche-nutzung-bundeslander-stichtagnutzungsarten



Format von Datensätzen

Datensätze werden in verschiedenen Formaten angeboten.

.pdf: meist Beschreibung von Metadaten, also Informationen zum Datensatz

.csv: „Comma seperatet Value“ kann einfach mit Excel oder anderen Programmen zur Datenverarbeitung geöffnet werden.

.xlsx: Exceldatei

.sav: Datensätze für die Statistikprogramme SPSS oder PSPP

.dta: Datensätze für das Statistikprogramm STATA

.xml: für automatisches Auslesen mit einem Programm