Institut für Soziologie
Sabine Düval

Methoden 2

Kontingenztabellen

Chi-Quadrat-Unabhängigkeitstest

Aufbau der Sitzung

- Wiederholung
- Exkurs zur Inferenzstatistik
- Was sind Kontingenztabellen?
  - Recodieren, Kategorisieren von Variablen
  - Wofür werden Kontingenztabellen verwendet?
  - Aufbau und Interpretation der Tabelle
  - Chi-Quadrat-(χ^2)-Unabhängigkeitstest
- Beispiel mit Stata:
  - Recodieren, Kategorisieren von Variablen
  - Erstellen von Kontingenztabellen
  - Durchführen von Chi-Quadrat-Tests

Aufbau der Sitzung

- **Wiederholung**
- Exkurs zur Inferenzstatistik
- Was sind Kontingenztabellen?
  - Recodieren, Kategorisieren von Variablen
  - Wofür werden Kontingenztabellen verwendet?
  - Aufbau und Interpretation der Tabelle
  - Chi-Quadrat-(χ^2)-Unabhängigkeitstest
- Beispiel mit Stata:
  - Recodieren, Kategorisieren von Variablen
  - Erstellen von Kontingenztabellen
  - Durchführen von Chi-Quadrat-Tests

Wiederholung

https://pingo.upb.de/

Passwort: **409475**

Aufbau der Sitzung

- Wiederholung
- **Exkurs zur Inferenzstatistik**
- Was sind Kontingenztabellen?
  - Recodieren, Kategorisieren von Variablen
  - Wofür werden Kontingenztabellen verwendet?
  - Aufbau und Interpretation der Tabelle
  - Chi-Quadrat-(χ^2)-Unabhängigkeitstest
- Beispiel mit Stata:
  - Recodieren, Kategorisieren von Variablen
  - Erstellen von Kontingenztabellen
  - Durchführen von Chi-Quadrat-Tests

Einschub: Inferenzstatistik

- In der **Inferenzstatistik** wird versucht, anhand der empirischen Daten aus der Stichprobe auf die Grundgesamtheit zu schließen

- Ablauf eines **Signifikanztests**:

  1. Festlegung von H_0 und H_1
  2. Wahl der Irrtumswahrscheinlichkeit (α) für H_0
  3. Festlegung einer Prüfgröße bzw. Teststatistik unter Berücksichtigung der Modellvoraussetzungen
  4. Berechnung der Prüfgröße und Entscheidung über die Verwerfung von H_0

Einschub: Inferenzstatistik

1. Festlegung von H_0 und H_1
   - Hypothesentest erfolgt immer für H_0 („Nullhypothese“)
   - H_0: z. B. Mittelwerte in der Grundgesamtheit (GG) sind gleich
   - H_1: z. B. Mittelwerte in der GG sind <u>nicht</u> gleich

2. Wahl der Irrtumswahrscheinlichkeit (α) für H_0
   - Entspricht der maximal zugelassenen Wahrscheinlichkeit dafür, dass H_0 abgelehnt wird, obwohl H_0 wahr ist
   - **Fehler 1. Art**:

     H_0 wird verworfen, obwohl H_0 wahr ist (**false rejection, α-Fehler**)
   - **Fehler 2. Art**:

     H_0 wird beibehalten, obwohl H_1 wahr ist (**false acceptance, β-Fehler**)

Einschub: Inferenzstatistik

3. Festlegung einer **Prüfgröße** (z.B. *t*-Test) und damit gleichzeitig deren Verteilung (z.B. Normalverteilung, *t*-Verteilung, etc.)

   Entscheidung für eine bestimmte Prüfgröße, je nach:

   - Skalenniveau der abhängigen Variablen
   - Art der Stichprobe (abhängig/unabhängig)
   - Anzahl der Ausprägungen der unabhängigen Variable

4. Berechnung der Prüfgröße

   - Dann: Vergleich mit dem kritischen Wert des Tests (d.h. liegt die berechnete Prüfgröße im Ablehnungsbereich?)
   - Anschließend: Entscheidung über die Verwerfung der H_0.

Einschub: Inferenzstatistik

- Erklärung anhand des **Binomialtests**

  Weicht die Wahrscheinlichkeit eines bestimmten Ereignisses von einem Testwert (gegebene Wahrscheinlichkeit) ab?

  - H_0: keine Abweichung
  - H_1: Abweichung

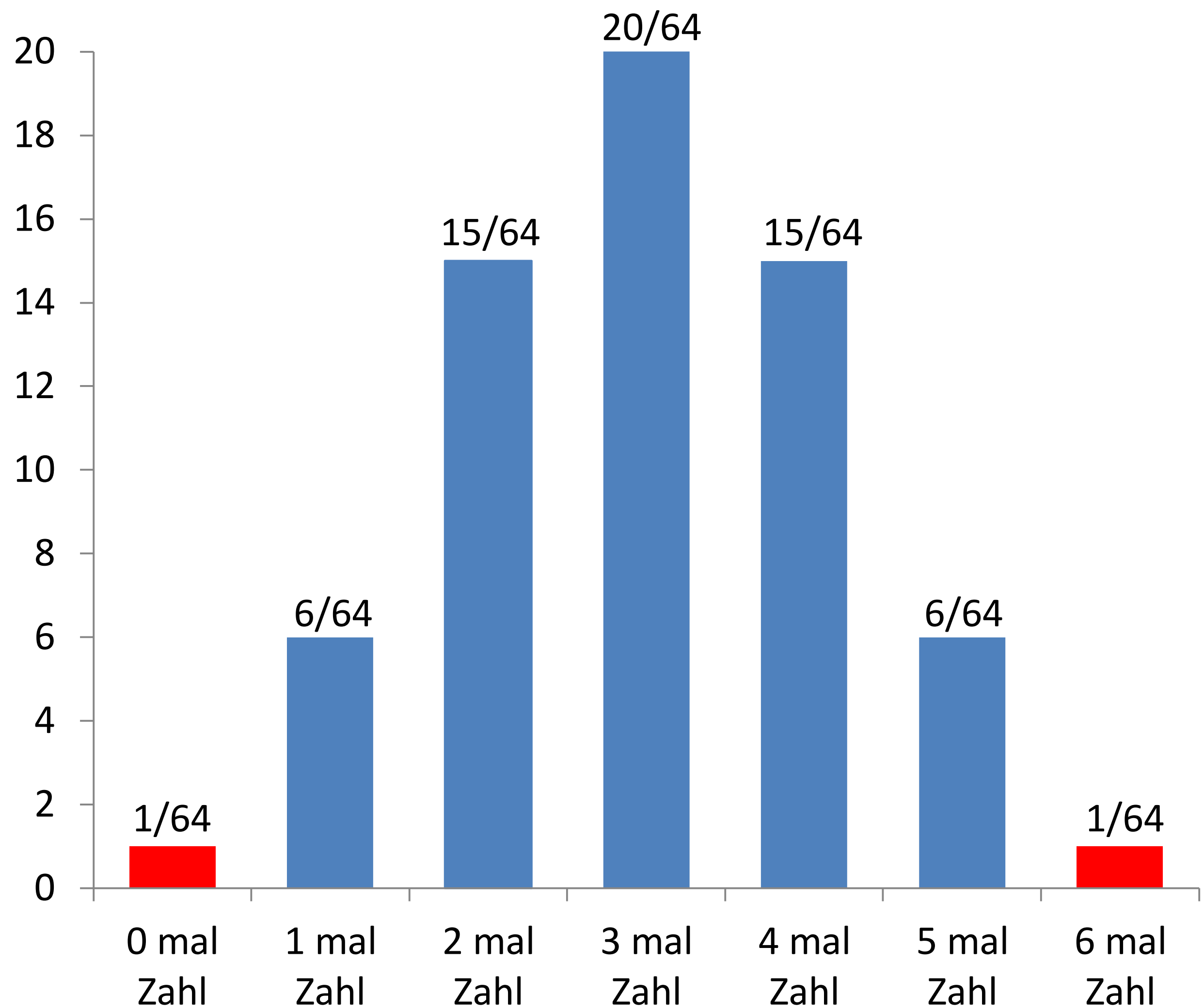
  Beispiel: **Münzwurf**

  Ist die Münze fair, d.h. beträgt die Wahrscheinlichkeit für Kopf und Zahl jeweils 0,5?

  Wie lauten H_0 und H_1?

Einschub: Inferenzstatistik

1	K K K K K K
2	Z K K K K K
3	K Z K K K K
4	K K Z K K K
5	K K K Z K K
6	K K K K Z K
7	K K K K K Z
8	Z Z K K K K
9	Z K Z K K K
10	Z K K Z K K
11	Z K K K Z K
12	Z K K K K Z
13	K Z Z K K K
14	K Z K Z K K
	…
63	K Z Z Z Z Z
64	Z Z Z Z Z Z



Einschub: Inferenzstatistik

Teststatistik: Diese ergibt sich aus der Häufigkeit mit der Zahl bei allen n Münzwürfen beobachtet wurde.

$$n_{ZAHL} = \sum_{i=1}^{n} y_i \qquad \text{für} \qquad y_i = \begin{cases} 0 \; wenn \; Kopf \; fällt \\ 1 \; wenn \; Zahl \; fällt \end{cases}$$

Ablehnbereich: Bei einem zweiseitigen Test besteht der Ablehnbereich aus den Ereignissen, bei welchen Zahl zu selten oder zu häufig auftritt, um die H_0 beizubehalten.

- Der untere kritische Wert ist dasjenige Ereignis, bei dem die Wahrscheinlichkeit, dass Zahl seltener oder genauso oft beobachtet wird, kleiner als das Signifikanzniveau ist.
- Der obere kritische Wert ist dasjenige Ereignis, bei dem die Wahrscheinlichkeit, dass Zahl häufiger oder genauso oft beobachtet wird, kleiner als das Signifikanzniveau ist.

Wird also ein Ereignis beobachtet das kleiner als der untere kritische Wert (hier 0 mal Zahl) oder größer als der obere kritische Wert (hier 6 mal Zahl) ist, so wird die Nullhypothese abgelehnt.

Einschub: Inferenzstatistik

- Was Signifikanztests NICHT leisten können:
  - Aussagen über die **Stärke** des Zusammenhangs
  - Aussagen über die **theoretische Wichtigkeit** eines Effekts
  - **Beweise** über die **tatsächliche Existenz** eines Effekts
    (man kann lediglich mit einer bestimmten Irrtumswahrscheinlichkeit α auf die Grundgesamtheit schließen).

Aufbau der Sitzung

- Wiederholung

- Exkurs zur Inferenzstatistik

- **Was sind Kontingenztabellen?**
  - Recodieren, Kategorisieren von Variablen
  - Wofür werden Kontingenztabellen verwendet?
  - Aufbau und Interpretation der Tabelle
  - Chi-Quadrat-(χ^2)-Unabhängigkeitstest

- Beispiel mit Stata:
  - Recodieren, Kategorisieren von Variablen
  - Erstellen von Kontingenztabellen
  - Durchführen von Chi-Quadrat-Tests

Kontingenztabellen

Was sind Kontingenztabellen?

- Gemeinsame Häufigkeitsverteilung zweier kategorialer Variablen
  - zwei „gekreuzte“ Häufigkeitstabellen (= Kreuztabelle)
  - Untersuchung des gemeinsamen Auftretens zweier Merkmale (mit relativ wenigen Ausprägungen)
- Skalenniveau: **Nominalskala** oder **Ordinalskala**
- Bei höherem Skalenniveau (metrisch) kann prinzipiell kategorisiert werden.
  - Dies ist jedoch nur in Ausnahmefällen anzuraten, da durch das Kategorisieren viel Informationsgehalt verloren geht!
  - Nutzen Sie stattdessen das metrische Skalenniveau aus und verwenden ein aussagekräftigeres Verfahren (z.B. Mittelwertvergleich, Varianzanalyse)!

Kontingenztabellen
Wozu verwende ich Kontingenztabellen?

- Testen von Zusammenhangshypothesen

  → Besteht ein Zusammenhang zwischen X (unabhängige Variable) und Y (abhängige Variable)?

  → Sind X und Y voneinander abhängig?

- Signifikanztest: Chi-Quadrat-Unabhängigkeitstest

  → Kann von einem Zusammenhang zwischen X und Y auch in der Grundgesamtheit ausgegangen werden?

- Zusammenhangsmaße in der nächsten Sitzung

  → Wie stark ist der Zusammenhang zwischen X und Y?

Kontingenztabellen

Aufbau der Tabelle (1)

Konvention:

- Unabhängige Variable (X) in den Tabellenspalten
- Abhängige Variable (Y) in den Tabellenzeilen

Zellen können enthalten:

- absolute Häufigkeiten (Anzahl)
- relative Häufigkeiten (%)
- erwartete Häufigkeiten
  (theoretische Häufigkeiten, wenn kein Zusammenhang besteht (=Nullhypothese))

Kontingenztabellen
Aufbau der Tabelle (2)

Unabhängige Variable (UV, X)

Absolute Häufigkeiten

Abhängige Variable (AV, Y)

	x_1	x_2	
y_1	a	b	$a+b$
y_2	c	d	$c+d$
	$a+c$	$b+d$	$N=a+b+c+d$

Zeilensumme

Randhäufigkeiten

Fallzahl

Spaltensumme

Kontingenztabellen
Interpretation der Prozentangaben

Fragestellung: Gibt es bei Autos einen Zusammenhang zwischen dem Herkunftsland und der Preisstufe?

Abhängige Variable

Unabhängige Variable

Absolute Häufigkeiten

Relative Häufigkeiten

Preisstufe		**Herkunftsland** **US-Autos**	**Ausl. Autos**	
	Billige Autos	*29* *55,77*	*8* *36,36*	*37*
	Teure Autos	*23* *44,23*	*14* *64,64*	*37*
		52	*22*	*N=74*

Kontingenztabellen
χ^2-Unabhängigkeitstest (Schritt 1)

Berechnung erwarteter Häufigkeiten (Indifferenztabelle)

→ für jede Zelle dazugehörige Randhäufigkeiten miteinander multiplizieren und durch Fallzahl teilen!
= theoretische Häufigkeiten, die sich ergäben, wenn kein Zusammenhang zwischen X und Y bestünde

		X			
		1	…	*m*	
Y	*1*	$\frac{h_{1\cdot}\, h_{\cdot 1}}{N}$	…	$\frac{h_{1\cdot}\, h_{\cdot m}}{N}$	$h_{1\cdot}$
	2	$\frac{h_{2\cdot}\, h_{\cdot 1}}{N}$	…	$\frac{h_{2\cdot}\, h_{\cdot m}}{N}$	$h_{2\cdot}$
	⋮	⋮		⋮	⋮
	k	$\frac{h_{k\cdot}\, h_{\cdot 1}}{N}$	…	$\frac{h_{k\cdot}\, h_{\cdot m}}{N}$	$h_{k\cdot}$
		$h_{\cdot 1}$	…	$h_{\cdot m}$	N

Erwartete Häufigkeiten (Indifferenztabelle)

Erwartete Häufigkeit

		Herkunftsland		
		US-Autos	Ausl. Autos	
Preisstufe	Billige Autos	*29* *26,0*	*8* *11,0*	*37*
	Teure Autos	*23* *26,0*	*14* *11,0*	*37*
		52	*22*	*N=74*

$$\frac{37 * 52}{74} = 26$$

Kontingenztabellen
χ^2-Unabhängigkeitstest (Schritt 2)

- **Ziel:** Prüfung, ob in der Stichprobe festgestellter Zusammenhang zwischen UV und AV auch für die Grundgesamtheit gilt $\rightarrow$ Signifikanztest
- Getestet wird dabei die **Nullhypothese (H_0):** Es besteht kein Zusammenhang zwischen UV und AV
  gegen die **Forschungshypothese (H_1):** Es besteht ein Zusammenhang zwischen UV und AV
- Allgemeine Logik von Signifikanztests: Bestimmung von statistischen Prüfgrößen (die einer bestimmten Verteilung folgen – hier **χ^2 -Verteilung**), welche die Berechnung der Irrtumswahrscheinlichkeit erlauben
- Prüfgröße für Kontingenztabellen: **χ^2-Koeffizient**
  Idee: Vergleich der tatsächlich beobachteten gemeinsamen Häufigkeitsverteilung mit der **Indifferenztabelle** (erwartete gemeinsame Häufigkeitsverteilung)

Kontingenztabellen
χ^2-Unabhängigkeitstest (Schritt 2)

Berechnung des χ^2-Koeffizienten:

- Summe aller quadrierten Abweichungen der beobachteten Werte von den erwarteten Werten

$$\chi^2 = \sum_k \sum_m \frac{(O_{km} - E_{km})^2}{E_{km}}$$

O_{km}: beobachtete Häufigkeiten
E_{km} : erwartete Häufigkeiten

- Intuitive Testlogik aus der Berechnungsformel:
  Je größer der Wert der Testgröße, desto eher kann die Nullhypothese verworfen werden.

Kontingenztabellen
χ^2-Unabhängigkeitstest (Schritt 3)

- Stata berechnet nicht nur die Prüfgröße, sondern führt auch den Signifikanztest durch und gibt direkt die Irrtumswahrscheinlichkeit (p-Wert) an

- **Sozialwissenschaftliche Konvention**: Wenn $p \leq 0,05$, dann wird H_0 abgelehnt.

Irrtumswahrscheinlichkeit	Bedeutung (Symbol)
$p > 0,05$	nicht signifikant (ns)
$p \leq 0,05$	signifikant (*)
$p \leq 0,01$	hoch signifikant (**)
$p \leq 0,001$	höchst signifikant (***)

- **Sprachgebrauch**: Effekte, Zusammenhänge, Assoziationen etc. sind signifikant bzw. statistisch überzufällig. Bitte nicht davon sprechen, dass ein Koeffizient, eine Variable oder eine Hypothese signifikant ist!

Kontingenztabellen

χ^2-Unabhängigkeitstest (Schritt 3)

- χ^2-Wert = 2,3287 bei einem Freiheitsgrad
- p-Wert: 0,127

Interpretation:

- χ^2-Wert unterschreitet mit 2,3287 den kritischen χ^2-Wert von 3,8415 für das 5%-Niveau (kann man Übersichtstabellen entnehmen; z.B. Jann 2005, S. 195). **Der Zusammenhang ist nicht signifikant (p = 0,127).**
- Forschungshypothese muss verworfen werden. Die Nullhypothese wird (vorläufig) angenommen! Es besteht **kein signifikanter Zusammenhang** zwischen X und Y!
- Je geringer die Irrtumswahrscheinlichkeit, desto größer ist umgekehrt die Wahrscheinlichkeit, dass in der GG ein Zusammenhang zwischen den Variablen vorliegt.

Aufbau der Sitzung

- Wiederholung
- Exkurs zur Inferenzstatistik
- Was sind Kontingenztabellen?
  - Recodieren, Kategorisieren von Variablen
  - Wofür werden Kontingenztabellen verwendet?
  - Aufbau und Interpretation der Tabelle
  - Chi-Quadrat-(χ^2)-Unabhängigkeitstest
- **Beispiel mit Stata**:
  - Recodieren, Kategorisieren von Variablen
  - Erstellen von Kontingenztabellen
  - Durchführen von Chi-Quadrat-Tests

Kontingenztabellen mit Stata
1. Vorbereiten der Variablen

Hypothese: Amerikanische und ausländische Autos unterscheiden sich in ihrer Reparaturanfälligkeit.

- Welche ist die unabhängige und welche die abhängige Variable?
  UV: Herkunftsland der Autos (*foreign*)
  AV: Reparaturaufträge seit 1978 (*rep78*)

```
. numlabel, add

. tab foreign
```

Car type	Freq.	Percent	Cum.
0. Domestic	52	70.27	70.27
1. Foreign	22	29.73	100.00
Total	74	100.00	

```
. tab rep78, m
```

Repair Record 1978	Freq.	Percent	Cum.
1	2	2.70	2.70
2	8	10.81	13.51
3	30	40.54	54.05
4	18	24.32	78.38
5	11	14.86	93.24
.	5	6.76	100.00
Total	74	100.00	

Syntax für einfache Kreuztabelle (absolute Häufigkeiten) - allgemein

```
tabulate AV UV
```

Repair Record 1978	Car type: 0. Domest	Car type: 1. Foreig	Total
1	2	0	2
2	8	0	8
3	27	3	30
4	9	9	18
5	2	9	11
Total	48	21	69

Befehl: `tabulate rep78 foreign`

Berechnung bedingter Häufigkeiten (Spaltenprozente)

Repair Record 1978	Car type: 0. Domest	Car type: 1. Foreig	Total
1	2 4.17	0 0.00	2 2.90
2	8 16.67	0 0.00	8 11.59
3	27 56.25	3 14.29	30 43.48
4	9 18.75	9 42.86	18 26.09
5	2 4.17	9 42.86	11 15.94
Total	48 100.00	21 100.00	69 100.00

Befehl: `tabulate rep78 foreign, column`

Berechnung erwarteter Häufigkeiten (Indifferenztabelle)

Repair Record 1978	Car type 0. Domest	1. Foreig	Total
1	2 1.4	0 0.6	2 2.0
2	8 5.6	0 2.4	8 8.0
3	27 20.9	3 9.1	30 30.0
4	9 12.5	9 5.5	18 18.0
5	2 7.7	9 3.3	11 11.0
Total	48 48.0	21 21.0	69 69.0

Befehl: `tabulate rep78 foreign, expected`

3. Durchführung eines χ^2-Tests

```
Pearson chi2(4) =  27.2640   Pr = 0.000
```

Befehl: `tabulate rep78 foreign, chi2`

Eigenschaften der χ^2-Teststatistik:

- χ^2-Wert ist immer positiv, daher ist mit χ^2-Wert allein keine Interpretation über die Richtung des Zusammenhangs möglich
- Konkreter χ^2-Wert hängt auch von Tabellen- und Stichprobengröße ab. Also: primäre Orientierung am p-Wert!

3. Durchführung eines χ^2-Tests

```
Pearson chi2(4) =  27.2640   Pr = 0.000
```

Interpretation:

- χ^2-Wert = 27,260 bei 4 Freiheitsgraden
- p-Wert: 0,000

  - χ^2-Wert überschreitet mit 27,2640 deutlich den kritischen χ^2-Wert von 9.4877 für das 5%-Niveau (kann man Übersichtstabellen entnehmen; z.B. Jann 2005, S. 195). Der Zusammenhang ist signifikant (p = 0,000)
  - Die Nullhypothese muss verworfen werden. Die Forschungshypothese wird (vorläufig) angenommen! Es besteht **ein signifikanter Zusammenhang** zwischen X und Y!
  - Je geringer die Irrtumswahrscheinlichkeit, desto größer ist umgekehrt die Wahrscheinlichkeit, dass in der GG ein Zusammenhang zwischen den Variablen vorliegt.

Exkurs 1
`tabulate` und `tabulate1`

Vorsicht beim Befehl `tabulate`:

- Der Befehl `tabulate` wird in Stata sowohl für einfache Häufigkeitsauszählungen verwendet als auch für Kreuztabellen

- Möchte man z.B. jeweils eine einfache Häufigkeitsauszählung für das kategorisierte Gewicht und das Herkunftsland des Autos, kann man also nicht den Befehl

  ```
  tabulate rep78 foreign
  ```

  eingeben, da man hier die Kreuztabelle erhält (siehe Folie #27)
- Um zwei getrennte Häufigkeitsauszählungen für das Gewicht und das Herkunftsland zu erhalten, ist der Befehl `tab1` zu verwenden

  ```
  tab1 rep78 foreign
  ```

Codierung von Dummies mit `tabulate`:

- Der Befehl `tabulate` kann auch dazu genutzt werden, um eine kategoriale Variable in k Dummies zu codieren, z.B.:

  `tabulate weight_kat, generate(weight_dum)`

- Man erhält 4 neue Varaible (weight_dum1, … , weight_dum4)

```
. tab weight_dum1

weight_kat= |
   =1. bis  |
  2000 lbs  |      Freq.     Percent        Cum.
------------+-----------------------------------
         0  |         67       90.54       90.54
         1  |          7        9.46      100.00
------------+-----------------------------------
     Total  |         74      100.00
```

Codierung von Dummies mit tabulate

- Diese vier neuen Variablen kann man nun noch entsprechend umbenennen und labeln

```
rename weight_dum1 leicht

rename weight_dum2 mittel1

rename weight_dum3 mittel2

rename weight_dum4 schwer
```

```
. tab leicht
```

weight_kat==1. bis 2000 lbs	Freq.	Percent	Cum.
0. nein	67	90.54	90.54
1. ja	7	9.46	100.00
Total	74	100.00	

```
label define labjn 1"ja" 0"nein"

label value leicht mittel1 mittel2 schwer labjn
```

Befehlsübersicht zur heutigen Sitzung

```
tabulate AV UV
tabulate … , column
tabulate … , expected
tabulate … , chi2
tab1
tabulate … , generate(var_dum)
rename
```

Literatur

Acock (2012): A Gentle Introduction to Stata. College Station: Stata Press. S. 119-128.

Jann (2005): Einführung in die Statistik. München: Oldenbourg. S. 59-66, 121f.

Pollock (2015): A Stata Companion to Political Analysis. London: Sage. S. 53-55.

Übungsaufgaben

1. *(Wiederholungsaufgabe):* **Berechnen** Sie die Länge des Autos (Variable „length"), die im Datensatz in inch angegeben ist in Meter (Umrechnung: 1 inch = 2,54 cm), **vergeben** Sie für die neue Variable den Namen „laenge" und **labeln** Sie die neue Variable entsprechend.
2. *(Wiederholungsaufgabe):* Führen Sie ein geeignetes Verfahren zur **Beschreibung der Variablen** „Laenge in Metern" durch, **interpretieren Sie die Ergebnisse** und lassen Sie sich für die neue Variable einen **Boxplot** ausgeben.
3. *(Wiederholungsaufgabe):* Fassen Sie die Länge der Autos in Metern (Variable „laenge") in drei **Kategorien** zusammen („bis 4 Meter", „4,01 bis 5 Meter" und „über 5 Meter"). L**abeln** Sie Variable und Ausprägungen, **zählen** Sie die neue Variable mit den zugewiesenen Werten **aus** und überprüfen Sie, ob ihre Recodierung korrekt durchgeführt wurde mit einer **Kreuztabelle**.
4. Erstellen Sie für die Variable der kategorisierten Länge in Metern sowie die Variable „foreign" **Häufigkeitsauszählungen mit einem Befehl**.
5. **Prüfen Sie folgende Hypothese**: Amerikanische Autos sind länger als ausländische Autos. Erstellen Sie für diese Hypothese eine **Kreuztabelle** mit absoluten, relativen und erwarteten Häufigkeiten. Führen Sie einen **χ^2-Test** durch und interpretieren Sie Ihren Befund.
6. Erzeugen Sie aus der Variablen der kategorisierten Länge in Metern **drei Dummies**, **benennen** Sie diese neuen Variablen um, **labeln** Sie die neuen Variablen sowie deren Werte und lassen sich eine der Dummy-Variablen **auszählen.**