LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Institut für Soziologie
Sabine Düval

Methoden 2

Grundlegende Stata-Bedienung und univariate Statistik

Organisatorisches

Kontakt

Bitte kontaktieren Sie uns bei Fragen ausschließlich über die E-Mailadresse

methoden2@soziologie.uni-muenchen.de

Organisatorisches
Stata-Lizenzen

- Es stehen Stata 14 Lizenzen für Sie bereit, mit denen Sie auch außerhalb des CIP-Pools arbeiten können.
- Näheres auf der Homepage http://www.soziologie.uni-muenchen.de/einrichtungen/it-services/programm-u-geraeteverleih1/programmverleih/index.html

Programm

- Operatoren
- Datenmanagement
- Univariate Statistik mit Stata
- Übungsaufgaben

Grundlegendes zu Stata

Missing values

Missing values werden in Stats mit einem " . " gekennzeichnet. Es gibt z.B. fünf missing values bei der Variablen „rep78".

```
tabulate rep78, missing nolabel
```

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	

Grundlegendes zu Stata

Missing values

Uns interessiert nun, wie viele Autos >= 4 Reperaturaufträge haben.

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	

18 + 11 = 29 Autos

Grundlegendes zu Stata

Missing values

Achtung: Fehlende Werte (sog. Missings) werden in Stata mit einem Wert von + ∞ behandelt

➔ diese Fälle werden bei der Auswahl von Beobachtungen mit > oder >= mit berücksichtigt, z.B.:

```
summarize price if rep78 >= 4
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	34	6073	2315.435	3748	12990

Der Preis für 34 Autos wird ausgegeben, obwohl laut Häufigkeitsauszählung nur 29 Autos mit einem Reparaturauftrag vorhanden sind

- ➔ Grund: Bei den verbleibenden 5 Autos wurde keine Angabe zum Reparaturauftrag gemacht (=missing)
- ➔ + ∞ und damit zulässiger Wert

Grundlegendes zu Stata

Stata-Operatoren: Mathematische Operatoren

Mathematische Operatoren		Befehls-eingabe	Beispiel	
+	Addition (a + b)	a + b	`display 2 + 2`	4
-	Subtraktion (a – b)	a - b	`display 9 – 4`	5
*	Multiplikation (a * b)	a * b	`display 4 * 6`	24
/	Division (a / b)	a / b	`display 32 / 8`	4
^	Potenz (a ^ b)	a ^ b	`display 4 ^ 2`	16

Grundlegendes zu Stata

Stata-Operatoren: Mathematische Operatoren

Mathematische Operatoren	Befehls-eingabe	Beispiel	
Quadratwurzel	sqrt(a)	`display sqrt(9)`	3
Logarithmen (log a)	log(a)	`display log(2)`	.69314…
Exponentialfunktion	exp(a)	`display exp(1.4)`	4.0552
Absolutbetrag (|a|)	abs(a)	`display abs(-8)`	8
Runden	round(a)	`display round (3.2)`	3
Minimum, Maximum	min(a, b,…, z) max(a, b,…,z)	`display min(3, 1, 10, 8, 14 )`	1 / 14
Trigonometr. Funktionen: Sinus, Kosinus, Tangens	sin(a), cos(a), tan(a)	`display sin(120)`	.58061…

Grundlegendes zu Stata

Stata-Operatoren: Relationale und logische Operatoren

Wichtig: Bei einer `if`-Bedingung wird ein doppeltes Gleichheitszeichen verwendet oder andere Zeichen, die eine Bedingung angeben

Relationale und logische Operatoren		Beispiel
<	kleiner als	`list price if weight < 3200`
<=	kleiner als oder gleich wie	`list price if weight <= 3200`
==	gleich wie	`list price if foreign == 1`
>	größer als	`list price if weight > 3200`
>=	größer als oder gleich wie	`list price if weight >= 3200`
!=	ungleich wie	`list price if foreign != 0`
&	und	`list price if weight > 3200 & foreign == 1`
|	oder	`list price if weight < 2000 | weight > 4500`

Erste Schritte mit Stata

Daten-Management

- Die meisten Datensätze sind erst einmal in einer Rohform vorhanden, die nach der Dateneingabe (manuell, elektronisch) entstanden sind
  ➔ Datenprüfung, -bereinigung und -management u.U. notwendig

- Autodatensatz ist bereits bereinigt, aber:
  - Variablen werden ggf. neu erstellt,
  - benannt und
  - verändert

Erste Schritte mit Stata

Benennen von Variablen

- Eine Variable hat immer einen Namen (z.B. rep78, foreign)
- Der Name sollte möglichst kurz sein
- Die Variablen unseres Datensatzes **auto.dta** haben bereits Namen (jedoch Englische)
  ➔ Umbenennen der Variablennamen
- Dafür benutzt man den Befehl `rename`
- Beispiele:
  - `rename weight gewicht`
  - `rename foreign ausland`

Erste Schritte mit Stata

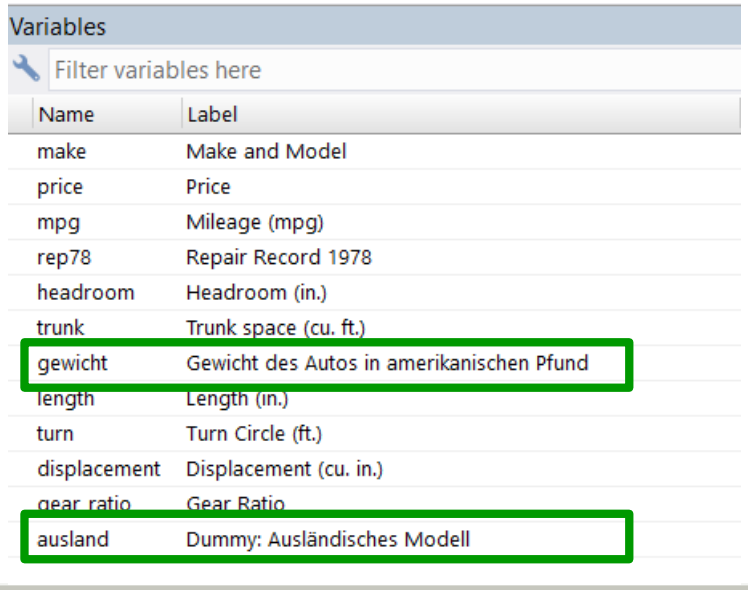
Labeln von Variablen

- Das Label einer Variable sollte länger sein und inhaltliche Informationen zur Variablen liefern
- Dafür benutzt man den Befehl `label variable`

```
label variable gewicht "Gewicht des Autos in ///
amerikanischen Pfund"

label variable ausland "Dummy: Ausländisches Modell"
```

Info an Stata, dass es einen Zeilenumbruch gibt!



Erste Schritte mit Stata
Labeln von Ausprägungen

- Auch Ausprägungen von Variablen sollte man so bezeichnen, dass sie leicht wiedererkennen kann.
- Auch dafür werden eindeutige Labels gebildet, z.B. **labjn** (Variablen mit den Antwortvorgaben Ja/Nein) oder **zustimm** (Variablen mit Zustimmungsskalen, z.B. stimme voll und ganz zu/stimme eher zu/stimme eher nicht zu/stimme überhaupt nicht zu)

- Um Ausprägungen zu labeln, muss man in zwei Schritten vorgehen:
  - Erstellen eines sog. Label-Containers (`label define`)
  - Übertragung des Containers auf die Variable (`label value`)

➔ Vorteil: Einmal erstellte Label-Container können für andere Variablen wiederverwendet werden

Erste Schritte mit Stata

Labeln von Ausprägungen

- Zum Erstellen des Label-Containers benutzt man den Befehl `label define;` Beispiel:
  - `label define labjn 1"ja" 0"nein"`
- Zum Übertragung des Containers auf die Variable (z.B. ausland) benutzt man den Befehl `label value;` Beispiel:
  - `label value ausland labjn`

Hinweis: Damit die Werte **und** die Labels in den Tabellen in Stata sichtbar werden, kann man den Befehl `numlabel, add` verwenden

Dummy: Ausländisches Modell	Freq.	Percent	Cum.
0. nein	52	70.27	70.27
1. ja	22	29.73	100.00
Total	74	100.00	

Befehle für die Tabelle:

```
numlabel, add
tabulate ausland, m
```

Erste Schritte mit Stata
(Neu-)Erstellen von Variablen

- Häufig müssen Variablen neu erstellt werden (z.B. das Alter muss aus dem Geburtsjahr berechnet werden)

- Der Befehl `generate` dient der Erzeugung dieser neuen Variablen mit Hilfe von mathematischen Operationen
  - z.B. wird die Variable Preis logarithmiert:

```
generate preis_log = ln(price)
label variable preis_log "Preis in USD logarithmiert"
```

  - Hypothetisches Beispiel zum Alter (Achtung: Anderer Datensatz notwendig!):

```
generate alter = 2017 - geburt
label variable alter "Alter in Jahren"
```

Erste Schritte mit Stata
(Neu-)Erstellen von Variablen

- Weiteres Beispiel: Berechnen des Preises (der im Datensatz in US$ angegeben ist) in Euro (Wechselkurs am 08.05.2017: 1 US$ = 0,9115 Eur)

```
generate preis_eur = price * 0.9115
label variabel preis_eur "Preis in Euro"
```

Achtung: In Stata müssen Kommas (0,9115) als Punkt eingegeben werden → 0.9115

Erste Schritte mit Stata

Umcodieren von Variablen

- Mit den Befehlen `generate` und `replace` lassen sich z.B. metrische Variablen einfach kategorisieren, z.B. kann man das Gewicht der Autos zusammenfassen, indem man sich an den Quartilen orientiert.
- Zunächst ist es notwendig, sich die Verteilung der Variablen anzusehen; dazu dient uns der Befehl `summarize` mit der Option `detail` (mit dieser Option erhält man neben dem Mittelwert weitere wichtige Verteilungsmaße).
  - `summarize gewicht, detail`

```
. sum weight, d

                        Weight (lbs.)
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1760          1760
 5%         1830          1800
10%         2020          1800       Obs                  74
25%         2240          1830       Sum of Wgt.          74

50%         3190                     Mean           3019.459
                        Largest      Std. Dev.      777.1936
75%         3600          4290
90%         4060          4330       Variance       604029.8
95%         4290          4720       Skewness       .1481164
99%         4840          4840       Kurtosis       2.118403
```

Erste Schritte mit Stata
Umcodieren von Variablen

- Wir erhalten für die Quartile des Gewichts folgende Werte:
  1. Quartil (25%) = 2.240 lbs
  2. Quartil (50%) = 3.190 lbs (=Median)
  3. Quartil (75%) = 3.600 lbs

➔ Folgende Anweisungen für die Kategorisierung der Variablen „gewicht“

```
generate gewicht_kat = .
replace gewicht_kat = 1 if gewicht<=2240
replace gewicht_kat = 2 if gewicht>2240 & gewicht<=3190
replace gewicht_kat = 3 if gewicht>3190 & gewicht<=3600
replace gewicht_kat = 4 if gewicht>3600 & gewicht<=4840

label variable gewicht_kat "Gewicht in lbs: Kategorien"

label define katwei 1"bis 2240 lbs" 2"2241 bis 3190 lbs" ///
3"3191 bis 3600 lbs" 4"über 3600 lbs"

label value gewicht_kat katwei
```

Univariate Statsitik
Häufigkeitsauszählung kategorialer Variablen

- Um sich das Ergebnis der Umcodierung anzeigen zu lassen: `tabulate`
- Mit dem Befehl `tabulate` erzeugt man Tabellen ➔ nützlicher Output, um v.a. die Verteilungen kategorialer Variablen kennenzulernen
  - `numlabel, add` (gibt die Werte der Labels mit aus)
  - `tabulate weight_kateg, missing`
    (miss-Option gibt an, dass auch die fehlenden Werte angezeigt werden sollen)

```
. numlabel, add

. tab weight_kateg, miss
```

Gewicht in lbs: Kategorien	Freq.	Percent	Cum.
1. bis 2240 lbs	19	25.68	25.68
2. 2241 bis 3190 lbs	18	24.32	50.00
3. 3191 bis 3600 lbs	19	25.68	75.68
4. über 3600 lbs	18	24.32	100.00
Total	74	100.00	

- In einfachen Häufigkeitsauszählungen werden neben den **absoluten** Häufigkeiten (Freq.) die **relativen Häufigkeiten** (Percent) sowie die **kumulierten** (aufaddierten) Anteile (Cum.) angegeben

Univariate Statistik
Einführung

- Die Datenanalyse beginnt damit, sich einen Überblick über die Häufigkeitsverteilung der interessierenden Variablen zu verschaffen

- Graphische Beschreibung von Häufigkeitsdaten, zum Beispiel Balken- oder Kreisdiagramm, Histogramm etc.

- Numerische Beschreibung von Variablen
  - Maßzahlen der zentralen Tendenz, bspw. Modus, Median, arithmetischer Mittelwert
  - Maßzahlen der Dispersion, bspw. Spannweite, Quartilsabstand, Varianz, Standardabweichung
  - Maßzahlen der Schiefe und Wölbung, bspw. Skewness- und Kurtosis-Maß

Univariate Statistik
Einführung

- Vor der Auswertung sind folgende Fragen zu beantworten:
  - Welches Skalenniveau liegt vor?
  - Welche Maßzahlen zur Beschreibung der Verteilung sind sinnvoll?

- Dann fordert man diejenigen Outputs an, die man interpretieren möchte

- Anwendungsbezug:
  - Berechnung von beschreibenden Maßzahlen für soziodemografische Variablen im erhobenen Datensatz
  - Vergleich der Maßzahlen mit Kennwerten der amtlichen Statistik
  - Liegt annähernde „Repräsentativität“ vor oder sind bestimmte Gruppen in unsere Stichprobe über/unterrepräsentiert?

Univariate Statistik

Exkurs: Ado-File „fre “

- Mit dem Ado-File `fre` ist es möglich sich Häufigkeitstabellen im SPSS-Format ausgeben zu lassen
- Der Vorteil liegt neben der etwas übersichtlicheren Darstellung darin, dass missings direkt mit ausgegeben werden (die Angabe der Option „missing“ ist also nicht notwendig)

- Wie kommt man an das ado-file:
  - `findit fre`
    Liste dann über Edit – Find mit Suchbegriff „fre“ durchsuchen und die Hompage aufrufen
    Dann „click here to install“
  - Alternativ: `ssc install fre`

Univariate Statistik

Exkurs: Häufigkeitsauszählung mit `frequencies`

- Beispiel einer Tabelle mit dem Befehl `frequencies`:
  - `fre weight_kateg`

```
. fre weight_kateg

weight_kateg — Gewicht in lbs: Kategorien
```

		Freq.	Percent	Valid	Cum.
Valid	1 1. bis 2240 lbs	19	25.68	25.68	25.68
	2 2. 2241 bis 3190 lbs	18	24.32	24.32	50.00
	3 3. 3191 bis 3600 lbs	19	25.68	25.68	75.68
	4 4. über 3600 lbs	18	24.32	24.32	100.00
	Total	74	100.00	100.00	

- Variable mit Missings:
  - `fre rep78`

```
. fre rep78

rep78 — Repair Record 1978
```

		Freq.	Percent	Valid	Cum.
Valid	1	2	2.70	2.90	2.90
	2	8	10.81	11.59	14.49
	3	30	40.54	43.48	57.97
	4	18	24.32	26.09	84.06
	5	11	14.86	15.94	100.00
	Total	69	93.24	100.00	
Missing	.	5	6.76		
Total		74	100.00		

- ✓ Percent: Alle Befragten
- ✓ Gültige Prozente (Valid): Nur Befragte, die eine Angabe gemacht haben

Univariate Statistik

Häufigkeitsauszählung mit `if`-Bedingung

- Manchmal ist es sinnvoll sich für kategoriale Variablen getrennte Häufigkeitsauszählungen ausgeben zu lassen
- Man möchte zum Beispiel die Häufigkeitsverteilung der kategorisierten Variable für das Gewicht der Autos nur für ausländische Marken haben; dazu benötigt man die `if`-Bedingung
  - `tabulate weight_kateg if ausland == 1`

```
. tabulate weight_kateg if foreign == 1
```

Gewicht in lbs: Kategorien	Freq.	Percent	Cum.
1. bis 2240 lbs	13	59.09	59.09
2. 2241 bis 3190 lbs	8	36.36	95.45
3. 3191 bis 3600 lbs	1	4.55	100.00
Total	22	100.00	

Interpretation:

- ✓ Die Mehrheit der ausl. Autos gehört der niedrigsten Gewichtsklasse an (59% wiegen bis 2240 lbs)
- ✓ Unter den ausländischen Autos befindet sich keines der höchsten Gewichtsklasse

Univariate Statistik
Getrennte Häufigkeitsauszählungen

- Es bietet sich auch an, sich für kategoriale Variablen insgesamt getrennte Häufigkeitsauszählungen ausgeben zu lassen
- Dazu benötigt man das Präfix `by`; z.B. möchte man für die einheimischen und ausländischen Autos getrennte Häufigkeits-auszählungen des kategorisierten Gewichts

- `by foreign: tabulate(weight_kateg)`

```
. by foreign: tabulate(weight_kateg)

-------------------------------------------------------------------------------
-> foreign = 0. Domestic

     Gewicht in lbs:
          Kategorien |      Freq.     Percent        Cum.
---------------------+-----------------------------------
      1. bis 2240 lbs |          6       11.54       11.54
2. 2241 bis 3190 lbs |         10       19.23       30.77
3. 3191 bis 3600 lbs |         18       34.62       65.38
     4. über 3600 lbs |         18       34.62      100.00
---------------------+-----------------------------------
                Total |         52      100.00

-------------------------------------------------------------------------------
-> foreign = 1. Foreign

     Gewicht in lbs:
          Kategorien |      Freq.     Percent        Cum.
---------------------+-----------------------------------
      1. bis 2240 lbs |         13       59.09       59.09
2. 2241 bis 3190 lbs |          8       36.36       95.45
3. 3191 bis 3600 lbs |          1        4.55      100.00
---------------------+-----------------------------------
                Total |         22      100.00
```

Univariate Statistik

Beschreibung metrischer Variablen

- Bei metrischen Variablen ist die Darstellung einer Häufigkeitstabelle nicht sinnvoll
- Hier beziehen wir uns auf Maßzahlen zur Beschreibung der Verteilung und ggf. auf graphische Darstellungsmöglichkeiten
- Umfangreiche deskriptive Maßzahlen kann man in Stata durch den Befehl `summarize` mit der Option `detail` anfordern

```
summarize price, detail
```

```
. summarize price, detail

                            Price
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3291            3291
 5%         3748            3299
10%         3895            3667       Obs                  74
25%         4195            3748       Sum of Wgt.          74

50%       5006.5                       Mean           6165.257
                        Largest        Std. Dev.      2949.496
75%         6342           13466
90%        11385           13594       Variance       8699526
95%        13466           14500       Skewness       1.653434
99%        15906           15906       Kurtosis       4.819188
```

Univariate Statistik

Beschreibung metrischer Variablen

```
. summarize price, detail

                            Price
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3291            3291
 5%         3748            3299
10%         3895            3667       Obs                  74
25%         4195            3748       Sum of Wgt.          74

50%       5006.5                       Mean           6165.257
                        Largest        Std. Dev.      2949.496
75%         6342           13466
90%        11385           13594       Variance       8699526
95%        13466           14500       Skewness       1.653434
99%        15906           15906       Kurtosis       4.819188
```

Interpretation (Auswahl):

- **arithmetisches Mittel (=Mittelwert):** 6.165,257 US$
- **Median:** Zentrum der Verteilung bei 5.006,5 US$
- **Standardabweichung/Varianz:** Information über Streuung
- **Schiefe:** >0 → rechtsschiefe Verteilung
- **Kurtosis (Wölbung):** >0 → Verteilung spitzer als Normalverteilung
- **Minimum/Maximum:** Gibt Auskunft über geringsten und höchsten Preis
  → Minimum: 3.291 US$
  → Maximum: 15.906 US$

Univariate Statistik

Beschreibung metrischer Variablen

- Es bietet sich an, sich auch für metrische Variablen insgesamt getrennte Häufigkeitsauszählungen ausgeben zu lassen
- Dazu benötigt man das Präfix `by`; z.B. möchte man für die einheimischen und ausländischen Autos getrennte Häufigkeits-auszählungen des Preises

➔ Interpretation?

- `by foreign: summarice(price), detail`

```
-> foreign = 0. Domestic

                            Price
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3291            3291
 5%         3667            3299
10%         3955            3667       Obs                  52
25%         4184            3799       Sum of Wgt.          52

50%       4782.5                       Mean           6072.423
                        Largest        Std. Dev.      3097.104
75%         6234           13466
90%        11385           13594       Variance       9592055
95%        13594           14500       Skewness       1.777939
99%        15906           15906       Kurtosis       5.090316

-> foreign = 1. Foreign

                            Price
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3748            3748
 5%         3798            3798
10%         3895            3895       Obs                  22
25%         4499            3995       Sum of Wgt.          22

50%         5759                       Mean           6384.682
                        Largest        Std. Dev.      2621.915
75%         7140            9690
90%         9735            9735       Variance       6874439
95%        11995           11995       Skewness       1.215236
99%        12990           12990       Kurtosis       3.555178
```

Univariate Statistik mit Stata

Beschreibung metrischer Variablen

- Graphische Beschreibungen:

Histogramm:

```
histogram price, normal

histogram price, width (500)///
start(0)normal
```

Kerndichteschätzer:

```
kdensity price, bwidth (500)
```

Box-Plot:

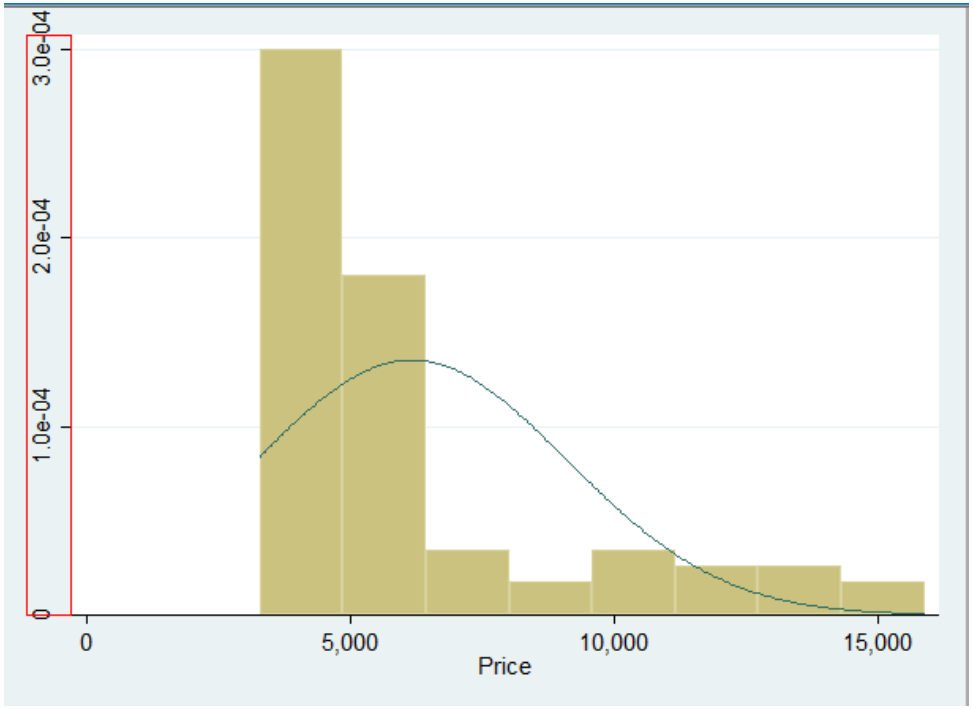
```
graph box price

graph box price, over(foreign)
```

Univariate Statistik mit Stata

Beschreibung metrischer Variablen

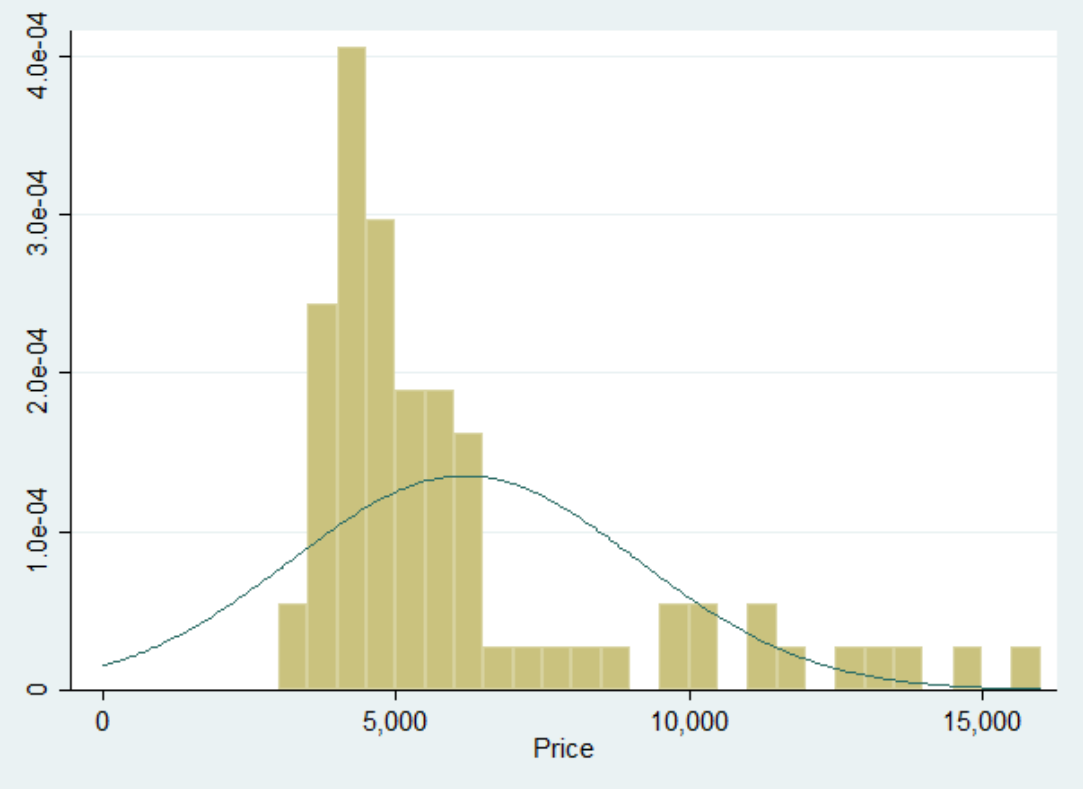
Befehl: `histogram price, normal`



- Histogramm liefert grafischen Eindruck der Verteilung und Vergleich mit Normalverteilung
- Achtung: Histogramm ≠ Balkendiagramm!

Univariate Statistik
Beschreibung metrischer Variablen

Befehl: `histogram price, width (500) start (0) normal`



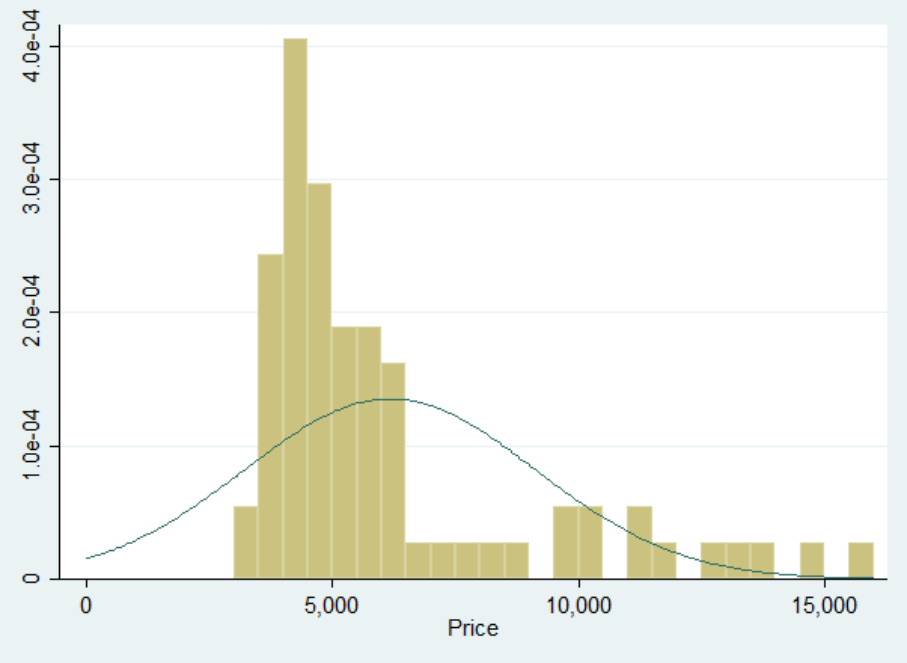
- Mit `width` gibt man die Klassenbreiten an
- Mit `start` den Beginn der Verteilung, die betrachtet werden soll bzw. der Normalverteilungskurve

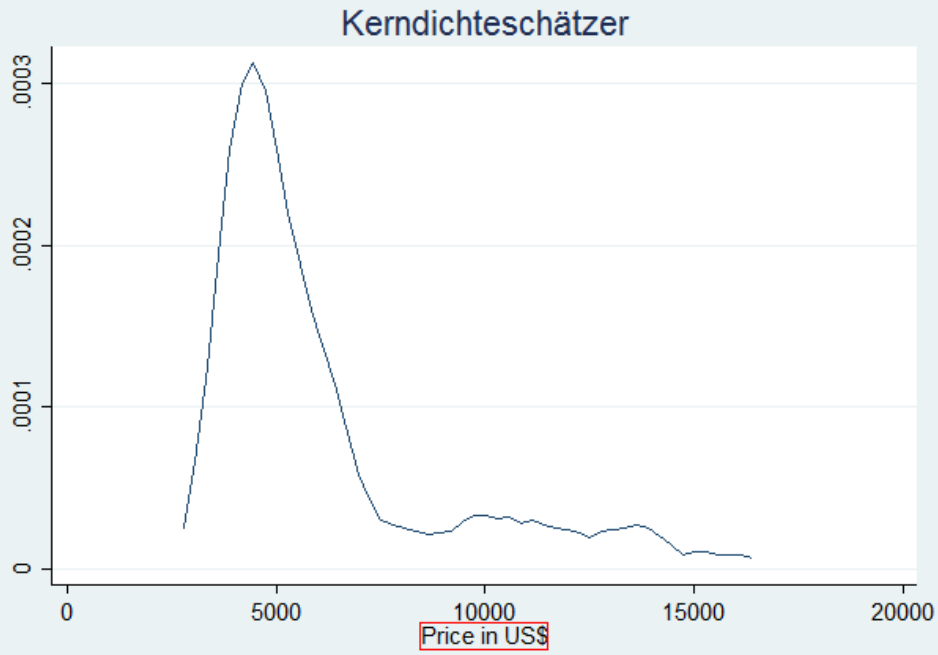
➔ bessere, da detailliertere Darstellung

Univariate Statistik

Beschreibung metrischer Variablen

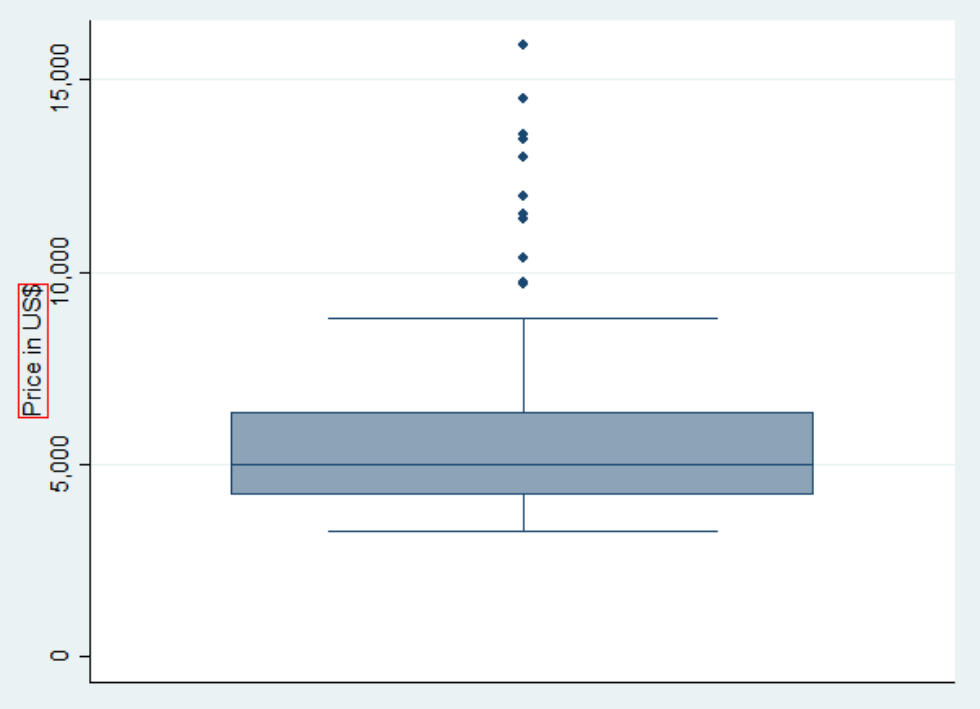
- Der Kerndichteschätzer zeigt die Verteilung einer metrischen Variablen nochmal etwas anschaulicher als ein Histogramm
  - `kdensity price, bwidth (500)`





Univariate Statistik mit Stata

Beschreibung metrischer Variablen

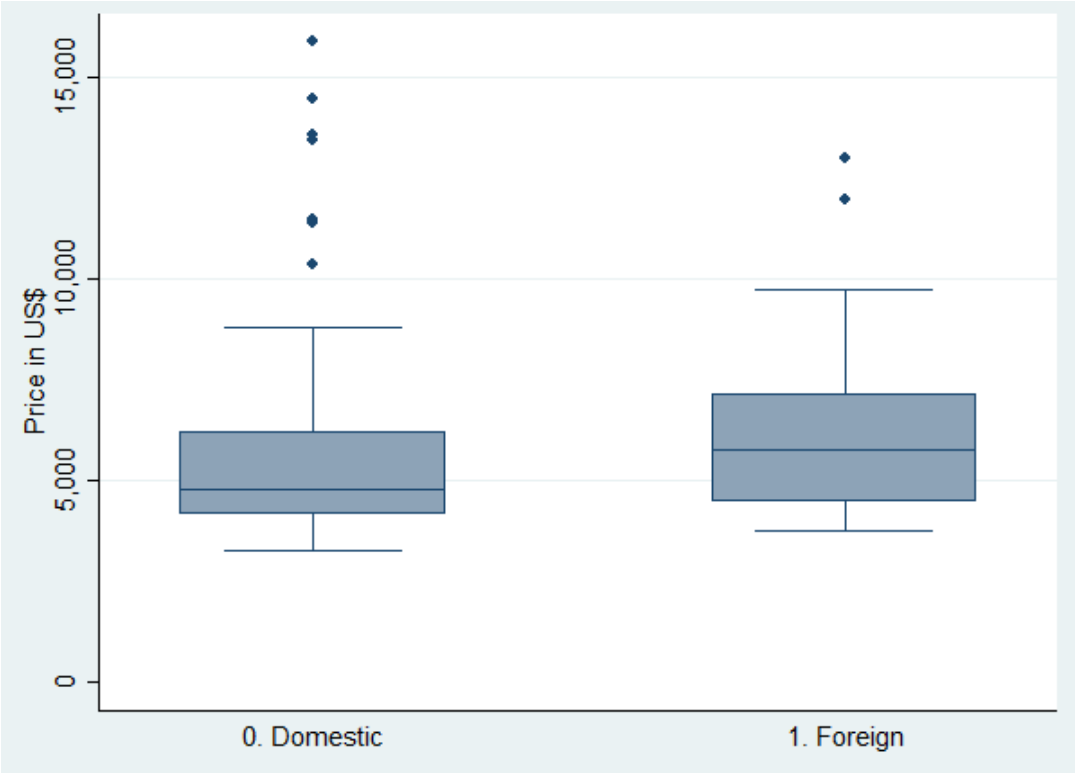


- Befehl: `graph box price`
- Boxplot liefert weitere grafische Darstellung
  → Ausreißer gut erkennbar
- Linie in der Mitte: Median
- Boxgrenzen: 25%- und 75%-Perzentil
- Linien („Whiskers“): 1,5-facher Interquartilsabstand

Univariate Statistik mit Stata

Beschreibung metrischer Variablen

- Auch bei graphischen Beschreibungen → getrennte Darstellung möglich; dazu wird die Option `over` benötigt
  - `graph box price, over(foreign)`



Univariate Statistik mit Stata

Wiederholung

- Heute behandelte Befehle:

```
tabulate                    generate
summarize                   replace
frequencies                 numlabel
display                     findit
rename                      ssc install
label variable              histogram
label define                kdensity
label value                 graph box
```

Übungsaufgaben

1. Öffnen Sie den Datensatz **auto.dta**. **Benennen** Sie die Variablen *length* und *weight* in *laenge* und *gewicht* um.
2. Berechnen sie aus der Variable *gewicht* (im Datensatz in amerikanischen Pfund) eine **neue Variable** *gewicht_kg*, die das Gewicht des Autos in Kilogramm (Umrechnung: 1 Pfund (lb) = 0,45359237 Kilogramm) anzeigt und **labeln** Sie die neue Variable entsprechend.
3. **Bilden** Sie die neue Variable *price_kat*, fassen Sie darin den Preis der Autos anhand der Quartile in vier Kategorien **zusammen**. **Labeln** Sie die Variable und deren Ausprägungen. Lassen Sie sich eine **Häufigkeitstabelle** ausgeben.
4. Lassen Sie sich **getrennte Häufigkeitsauszählungen** der kategorisierten Preisvariablen für einheimische und ausländische Autos ausgeben und interpretieren Sie die Outputs.
5. Führen Sie ein **geeignetes Verfahren zur Beschreibung** der Variablen *gewicht_kg* durch und **interpretieren** Sie die Ergebnisse
6. Lassen Sie sich für das Gewicht der Autos in kg **getrennte Boxplots** für einheimische und ausländische Autos ausgeben.