SE Geographie und Ökonomie

Einheit 2:
Univariate Datenanalyse: Deskriptive Statistik

Dieter Pennerstorfer
dieter.pennerstorfer@jku.at

Institut für Volkswirtschaftslehre
Johannes Kepler Universität Linz Linz

Lernziele der Einheit 2

Sie können ein Merkmals einer Stichprobe oder einer Grundgesamtheit auf folgende Arten **beschreiben bzw. darstellen**:

- Darstellung der Verteilung eines Merkmals als **Häufigkeitsverteilung (tabellarisch)**
- Darstellung der Verteilung als **Stabdiagramm oder Histogramm (grafisch)**
- (Räumliche Daten können auch als **Karten** dargestellt werden.)
- Beschreibung einer Variable durch **Lage- und Streuungsmaße**

Sinn und Zweck:

- Eine "Urliste" eines Merkmals ist unübersichtlich. Die Daten müssen daher verdichtet und in **übersichtlicher Form dargestellt** werden, um Dinge erkennen zu können (obwohl dabei Informationen verloren gehen).

Häufigkeitsverteilung

Bei **diskreten** (insbesondere nominal oder ordinal skalierten) **Merkmalen**.

Bezeichnungen:

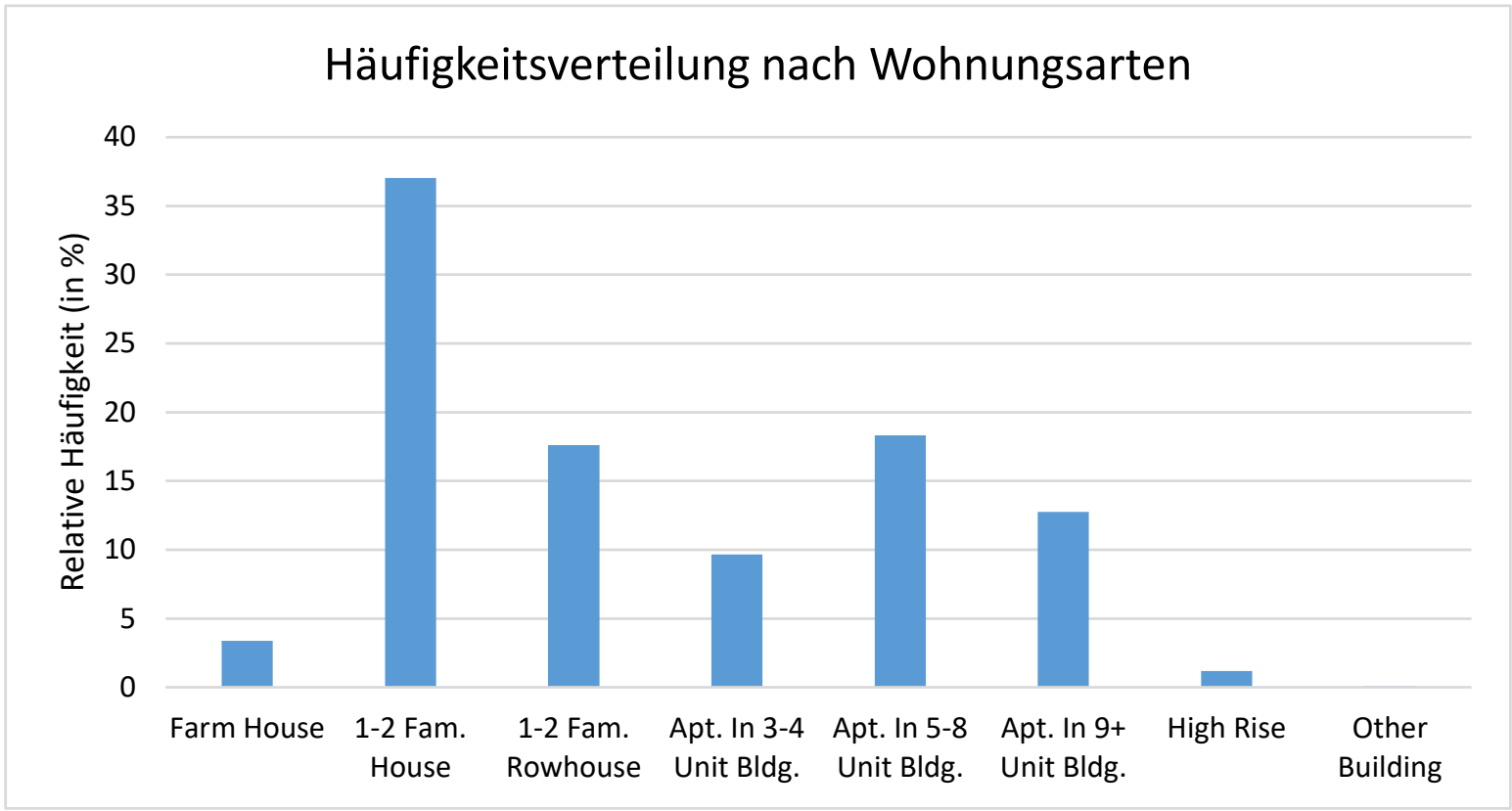
N	Untersuchungsumfang
n	Stichprobenumfang
r	Anzahl an verschiedenen Ausprägungen
x_m	Ausprägung, $m = 1, ..., r$
h_m	absolute Häufigkeit der Ausprägung x_m
p_m	$= h_m/n$ relative Häufigkeit der Ausprägung x_m
P_m	$= 100 \cdot p_m$ relative Häufigkeit der Ausprägung x_m in Prozent

Häufigkeitsverteilung: Tabelle (Wohnungsart)

Klasse	Ausprägung	Absolute Häufigkeiten	Relative Häufigkeiten	Relative Häufigkeiten (in Prozent)
m	x_m	h_m	p_m	P_m (in %)
1	Farm House	179	0.034	3.4
2	1-2 Fam. House	1,959	0.370	37.0
3	1-2 Fam. Rowhouse	931	0.176	17.6
4	Apt. In 3-4 Unit Bldg.	510	0.096	9.6
5	Apt. In 5-8 Unit Bldg.	969	0.183	18.3
6	Apt. In 9+ Unit Bldg.	674	0.127	12.7
7	High Rise	63	0.012	1.2
$(r =)$ 8	Other Building	5	0.001	0.1
Summe	$(n =)$	5,290	1	100

Anmerkung: Der Datensatz beinhaltet eigentlich 5,411 Erhebungseinheiten, aber 121 Personen haben keine Angaben zur Wohnungsart gemacht. Diese **fehlenden Werte** sollten in EXCEL mit leeren Zellen kodiert sein (und nicht mit ".", kA, 9999, ...) und werden bei sämtlichen Berechnungen ausgelassen (**Fallausschluss**). Das ist dann zulässig, wenn die Werte "zufällig" fehlen. Generell sind fehlende Werte oft ein Problem.

Häufigkeitsverteilung: Stabdiagramm (Wohnungsart)



Häufigkeitsverteilung: Tabelle (Wohnungsgröße)

Bei **stetigen Merkmalen** ist es für die Erstellung einer Häufigkeitstabelle zielführend, den gesamten Wertebereich in **Intervalle** zu gliedern.

Änderungen zu diskreten Variablen:

- e_{m-1} ist die Unter- und e_m die Obergrenze des m-ten Intervalls.
- $h_m = h(e_{m-1} < x \leq e_m)$ ist die absolute Häufigkeit des Intervalls $I_i = (e_{m-1}, e_m]$.
- $d_m = e_{m-1} - e_m$ ist die Intervallbreite.
- Die Dichte $f_m = p_m / d_m$ ist der Quotient aus relativer Häufigkeit $p_m = h_m / n$ und Intervallbreite d_m.
- Es empfiehlt sich (außer in Ausnahmefällen), für alle Intervalle die gleichen Intervallbreite zu wählen.

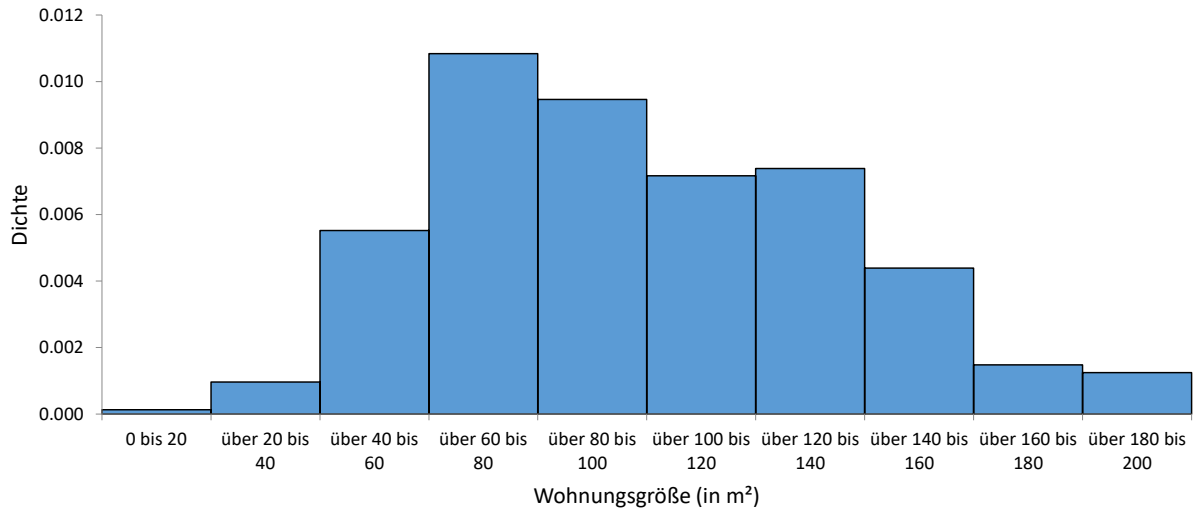
Häufigkeitsverteilung: Histogramm (Wohnungsgröße)

Verteilung der Wohnungsgröße:

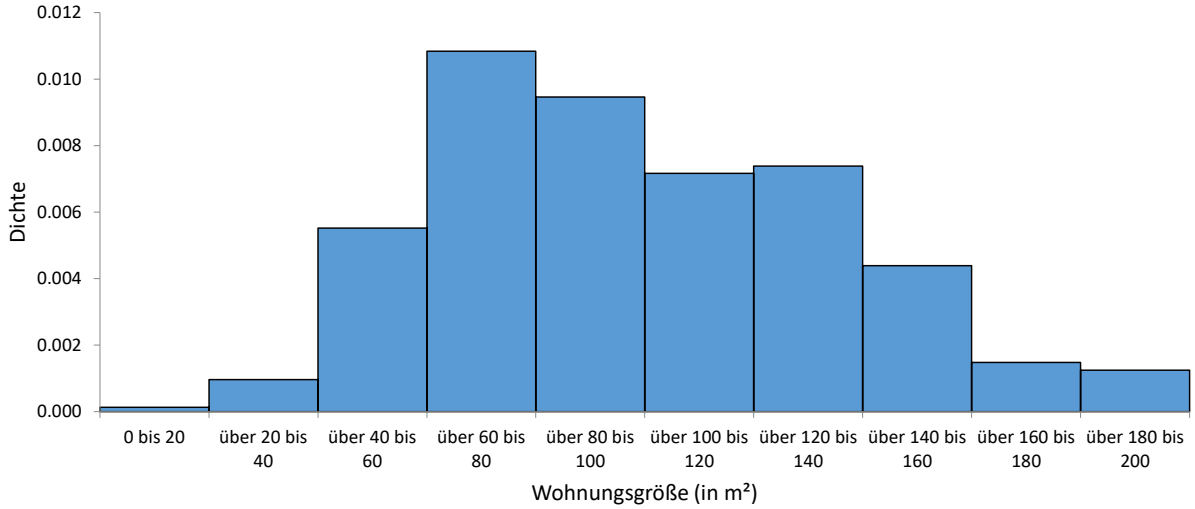
Klasse m	Ausprägung x_m	Absolute Häufigk. h_m	Relative Häufigk. p_m	Relative Häufigk. P_m (in %)	Dichte $f_m = p_m/d_m$
1	0 m² bis 20 m²	14	0.003	0.3	0.00013
2	über 20 m² bis 40 m²	104	0.019	1.9	0.00096
3	über 40 m² bis 60 m²	597	0.110	11.0	0.00552
4	über 60 m² bis 80 m²	1,173	0.217	21.7	0.01084
5	über 80 m² bis 100 m²	1,024	0.189	18.9	0.00946
6	über 100 m² bis 120 m²	775	0.143	14.3	0.00716
7	über 120 m² bis 140 m²	799	0.148	14.8	0.00738
8	über 140 m² bis 160 m²	475	0.088	8.8	0.00439
9	über 160 m² bis 180 m²	160	0.030	3.0	0.00148
10	über 180 m² bis 200 m²	135	0.025	2.5	0.00125
$(r =)$ 11	über 200 m²	154	0.028	2.8	
Summe	$(n =)$	5,410	1	100	

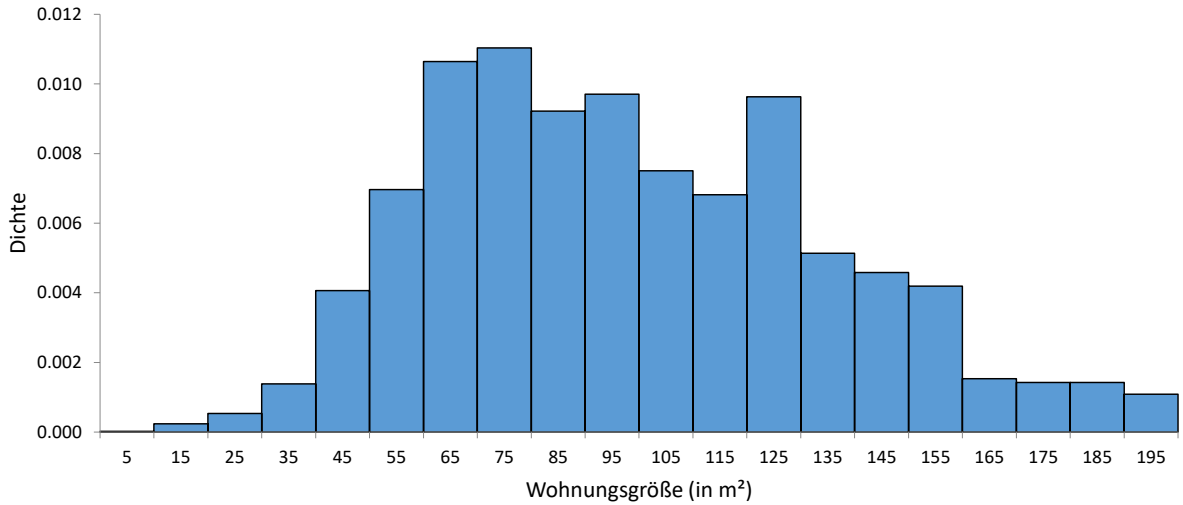
Histogramm: Wohnungsgröße (grafische Darstellung)

Ein Histogramm ist für **metrische stetige Merkmale** geeignet, deren Ausprägungen in **Intervalle** zusammengefasst wurden. Bei einem Histogramm werden auf der x-Achse die **Ausprägungen** und auf der y-Achse die **Dichten** f_m aufgetragen. In dieser Darstellungsform sind die relativen Häufigkeiten als Flächen sichtbar.
Bei **gleich breiten Intervallen** ist es zulässig, **statt der Dichten die Häufigkeiten** aufzutragen. Dies liegt daran, dass man bei der Betrachtung eines Histogramms automatisch die Verhältnisse der Flächen wahrnimmt, und nicht die Relationen der Rechteckshöhen.

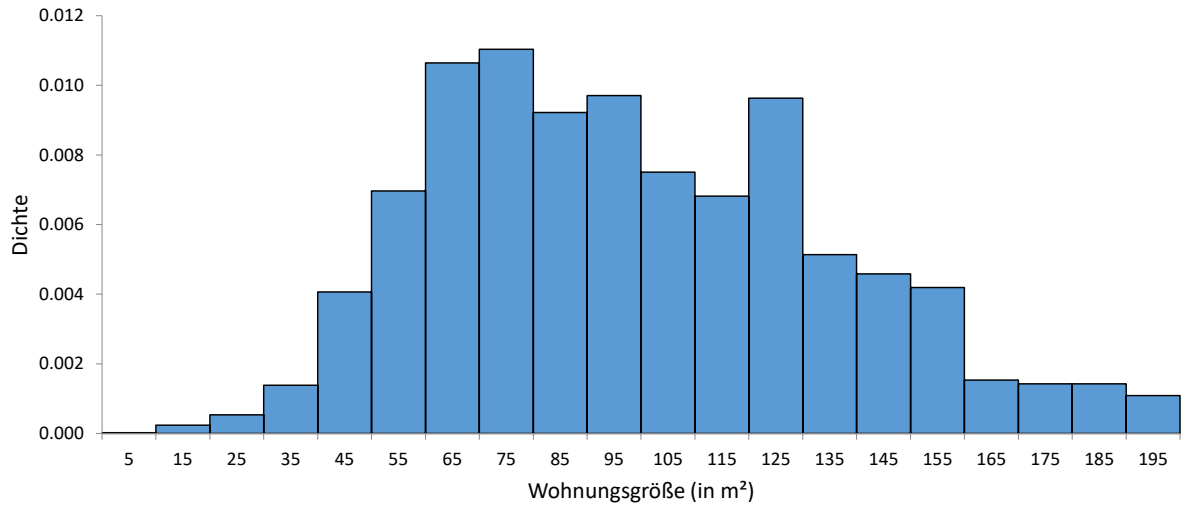


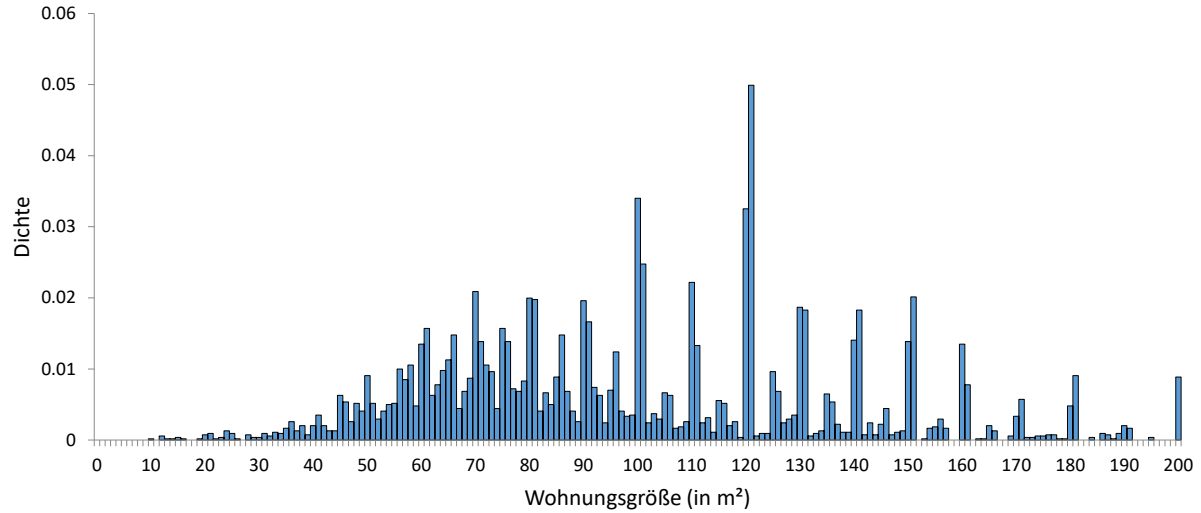
Histogramm: verschiedene Intervallbreiten (1)





Histogramm: verschiedene Intervallbreiten (2)





EXCEL Add-In Analysefunktionen

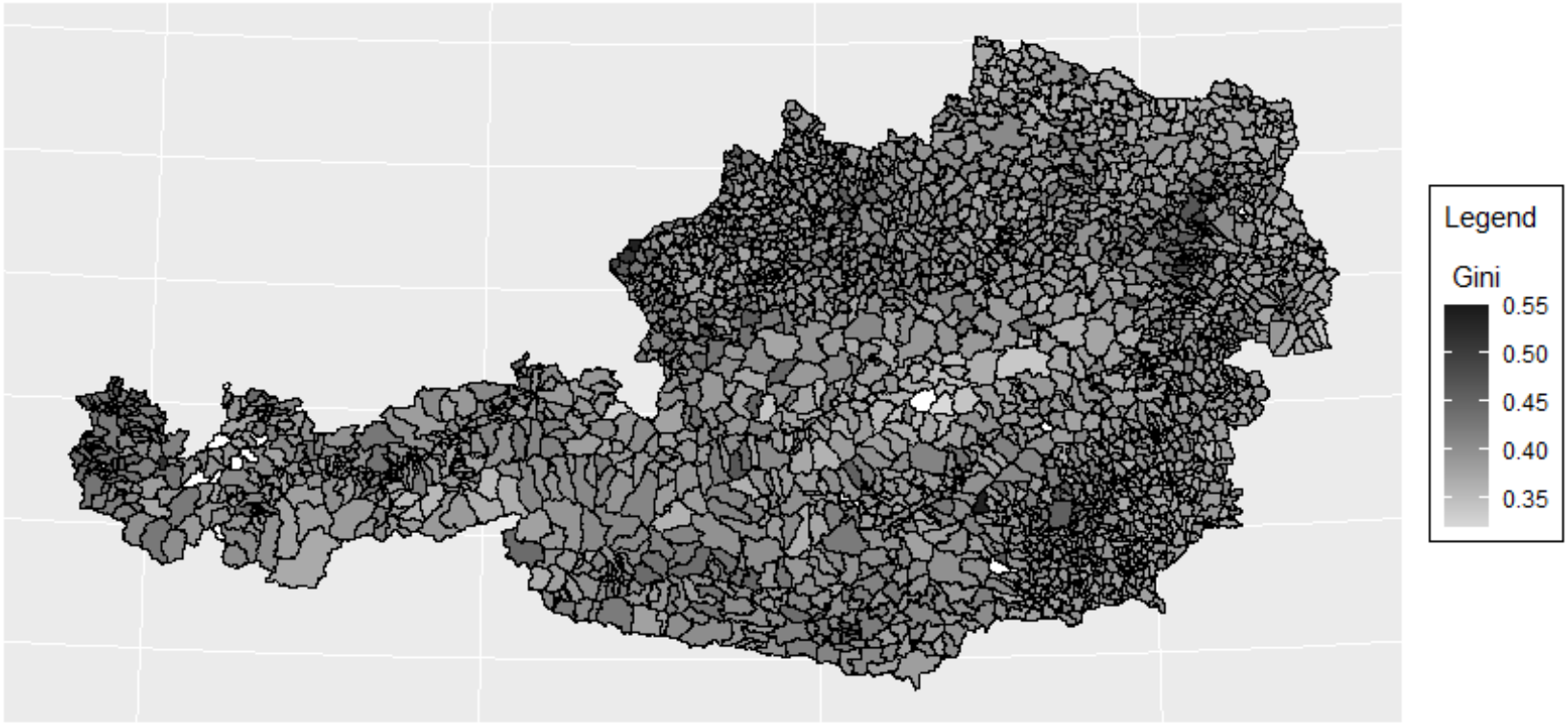
Eine Häufigkeitstabelle kann in Excel auch über **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Histogramm** erstellt werden, wobei hier lediglich absolute Häufigkeiten h_m ausgewiesen werden. Die relativen Häufigkeiten p_m und die Dichte f_m muss selbständig berechnet werden.

- Wird **Diagrammdarstellung** angehakt, wird ein Stabdiagram bzw. ein Histogramm ausgegeben. Histogramme müssen noch gesondert formatiert werden, da in EXCEL die absoluten Häufigkeiten aufgetragen werden (und nicht die Dichte), und weil zwischen den Balken ein Freiraum gelassen wird (was nicht sinnvoll ist).
- Mit diesem Befehl können **nur numerische Ausprägungen** verarbeitet werden können.

Alternativ dazu kann auch der **EXCEL-Befehl** HÄUFIGKEIT verwendet werden. Dabei handelt es sich um eine sog. Matrix-Formel, die die absoluten Häufigkeiten als einspaltige Matrix zurück gibt. Man muss daher den gesamten Ausgabebereich formatieren, und die Eingabe nicht nur mit *Enter*, sonder mit *Strg + Umschalt + Enter* bestätigen.
Alternativ kann man mit dem Befehl ZÄHLENWENN die absoluten Häufigkeiten einzelner Merkmale abzählen. Dieser Befehl kann **auch nicht-numerische Informationen** verarbeiten.

Darstellung als Karte (1)

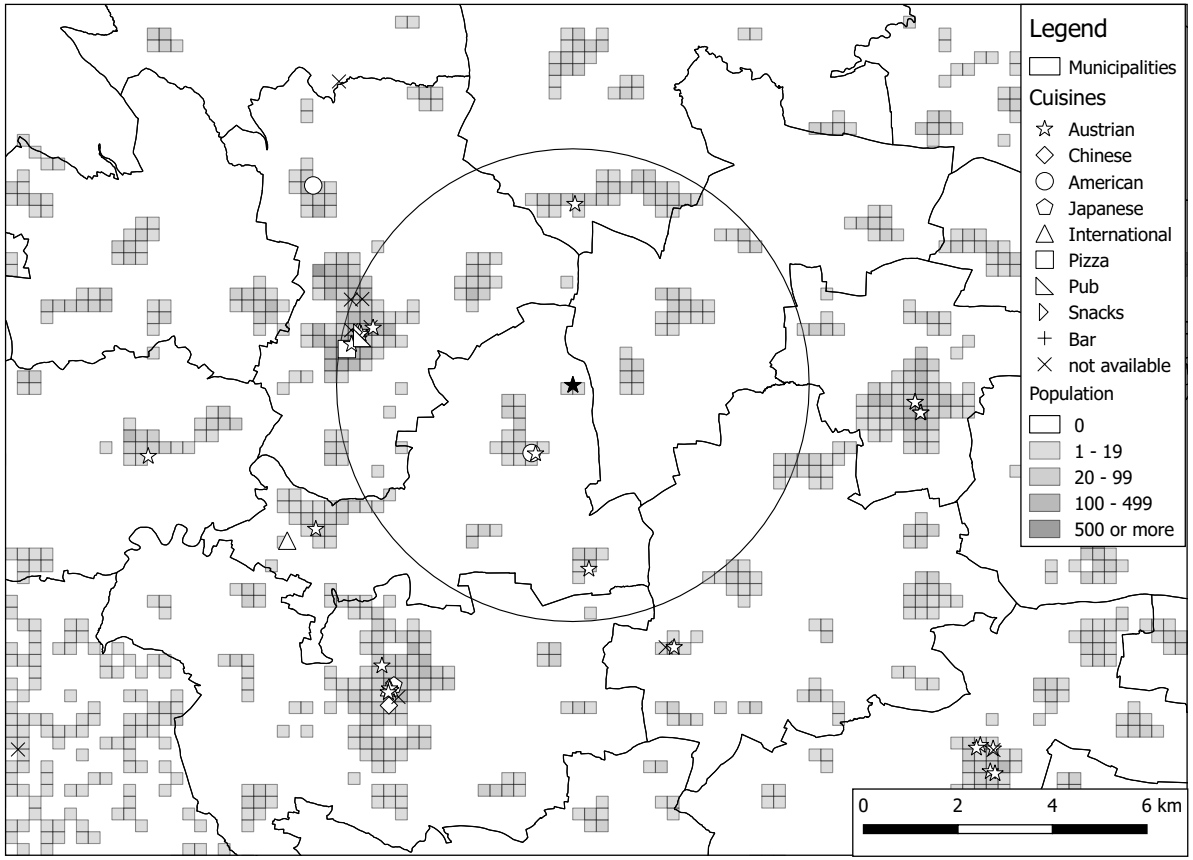
Räumliche Daten können auch mittels Karten dargestellt werden. Hier sehen Sie die Einkommensungleichheit auf Ebene der Gemeinden in Österreich:



Datenquelle: Statistik Austria (2013). Eigene Darstellung.

Darstellung als Karte (2)

Die Darstellung als Karte ist nicht auf administrative Gebietseinheiten beschränkt:



Datenquelle: www.tripadvisor.com, www.restauranttester.at (zwischen November 2016 und März 2017 gesammelt) and Statistik Austria (2014). Eigene Darstellung.

Maßzahlen für eindimensionale Verteilung

Manchmal ist man an **Informationen** über ein Merkmal in sehr **komprimierter Form** interessiert. Spezifische Maßzahlen beinhalten möglichst viel Information über die Daten in einer **einzigen Zahl**. Man unterscheidet:

1. **Lagemaße:** spiegeln das Zentrum der Verteilung wider
2. **Streuungsmaße:** geben an, wie weit die Daten von einander oder von einer Lagemaßzahl abweichen

Manche Maßzahlen sind nicht für alle Skalenniveaus sinnvoll:

Merkmalsausprägungen	Unterscheiden	Ordnen	Summen / Differenzen	Quotienten
Nominal	Ja	Nein	Nein	Nein
Ordinal	Ja	Ja	Nein	Nein
Metrisch				
Intervallskaliert	Ja	Ja	Ja	Nein
Verhältnisskaliert	Ja	Ja	Ja	Ja

Lagemaße: Arithmetisches Mittel

Arithmetisches Mittel (Mittelwert, Durchschnitt, $\bar{x}$)

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

Liegen nur r verschiedene Ausprägungen vor, kann der Mittelwert vereinfachend auch mit

$$\bar{x} = \frac{1}{n} \sum_{m=1}^{r} x_m h_m = \sum_{m=1}^{r} x_m p_m$$

berechnet werden. (Die erste Formel behält aber weiterhin Gültigkeit.)

Hinweise:

- Ausschließlich für **metrische Merkmale** geeignet. Ungeeignet für nominale und ordinale Merkmale.
- Bei intervallskalierten Merkmalen werden als Ausprägungen die Intervallmitten verwendet. Hier muss die zweite Formel verwendet werden.
- Die Berechnung des Mittelwertes bei dichotomen Merkmalen ergibt den Anteil der 1-Kodierungen (*wenn* das Merkmal mit 0 und 1 kodiert sind).
- **EXCEL-Befehl:** MITTELWERT

Lagemaße: Median

Der **Median** $\tilde{x}_{0,5}$ ist der mittlere Wert einer geordneten Datenreihe. Mindestens 50 % der Objekte haben eine Ausprägung, die höchstens so groß ist wie der Median, und mindestens 50 % der Objekte haben eine Ausprägung, die mindestens so groß ist wie der Median.
Wenn $x_{(i)}$ die i-te Stelle einer geordneten Datenreihe ist, dann ist der Median:

$$\tilde{x}_{0,5} = \begin{cases} x_{\frac{n+1}{2}} & \text{wenn } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{wenn } n \text{ gerade} \end{cases}$$

Hinweis:

- Für **ordinale und metrische Merkmale** geeignet. Ungeeignet für nominale Merkmale.

Beispiel:

- 3 Personen sind 150 cm, 160 cm und 200 cm groß. Die Personen sind *durchschnittlich* 170 cm groß (**arithmetisches Mittel**). Die *durchschnittliche Person* der Gruppe ist 160 cm groß (**Median**).
- Wird die Stichprobe um eine 4. Person ergänzt, die 170 cm groß ist, bleibt der Mittelwert unverändert, während der Median auf 165 cm steigt.
- **EXCEL-Befehl:** MEDIAN (auch QUANTIL.INKL möglich, siehe nächste Seite)

Quantil

Quantile (auch Perzentile, $\tilde{x}_\alpha$) sind Ausprägungen von quantitativen Variablen, die **geordnete Datenreihen** in Gruppen unterteilen, so dass ein bestimmter Anteil (oder Prozentsatz) über und ein bestimmter Anteil unter dem Quantil liegt. Das α-Quantil ist jener Wert $\tilde{x}_\alpha$, für den mindestens der Anteil α der Daten kleiner oder gleich $\tilde{x}_\alpha$ und mindestens der Anteil $1-\alpha$ der Daten größer oder gleich $\tilde{x}_\alpha$ ist.

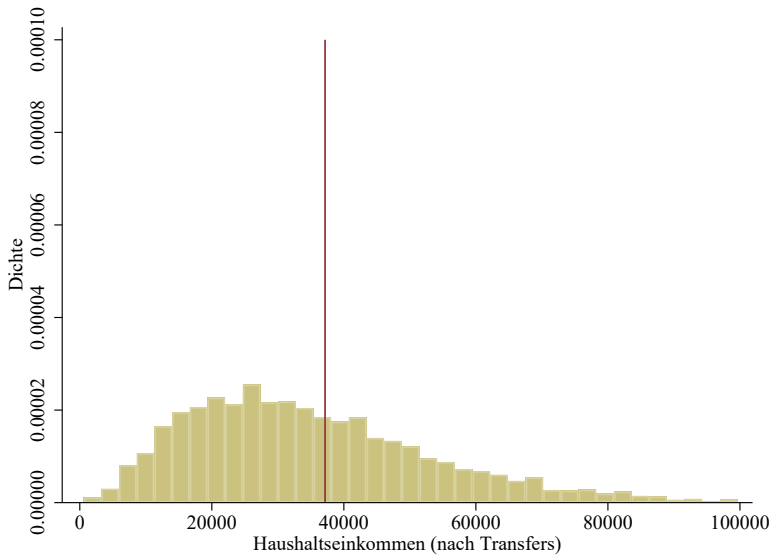
$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{wenn } n \cdot \alpha \text{ keine ganze Zahl ist} \\ & k \text{ ist dann die auf } n \cdot \alpha \text{ folgende ganze Zahl} \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{wenn } n \cdot \alpha \text{ eine ganze Zahl ist} \\ & \text{dann ist } k = n \cdot \alpha \end{cases}$$

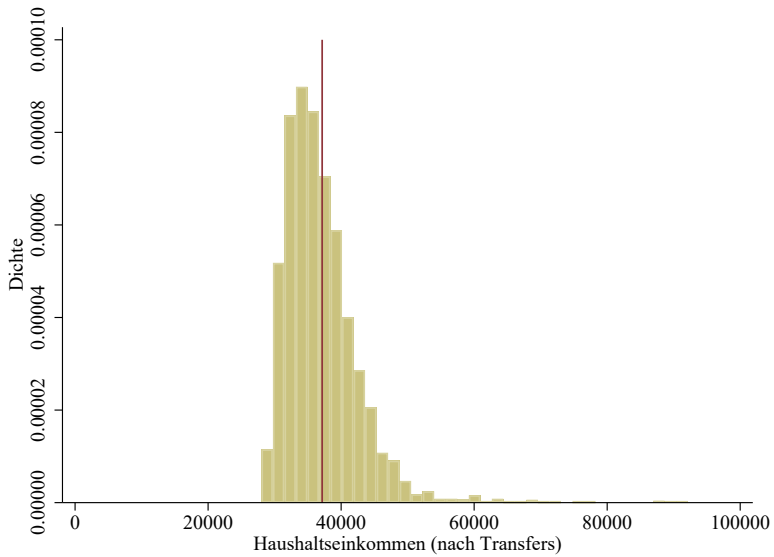
Spezialfälle:

- **Median**: 0,5-Quantil
- **Quartile**: $\tilde{x}_{0,25}$, $\tilde{x}_{0,5}$ (= Median) und $\tilde{x}_{0,75}$ teilen Daten in 4 gleich große Gruppen.
- **EXCEL-Befehl:** QUANTIL.INKL

Streuungsmaße: Motivation

Abbildungen zeigen Histogramme zu tatsächlichem (links) und modifiziertem (rechts) Haushaltseinkommen. Das durchschnittliche Haushaltseinkommen (Mittelwert) beträgt in beiden Fällen 37,150 Euro.





Streuungsmaße (1)

Spannweite (Range, Wertebereich, R) gibt den Abstand zwischen der größten und der kleinsten vorkommenden Ausprägung eines Merkmals an:

$$R = x_{max} - x_{min}$$

Die wichtigste Streuungskennzahl ist die **Varianz** (s^2), die das arithmetische Mittel der quadrierten Abstände der Datenpunkte zum Mittelwert ist. Ausgehend von der Varianz werden weitere Streuungsmaße wie die **Standardabweichung** (s) oder der **Variationskoeffizient** (V) berechnet. Wenn alle N Erhebungseinheiten der Grundgesamtheit beobachtet werden, können die Maßzahlen wie folgte berechnet werden:

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$s = +\sqrt{s^2}$$

$$V = \frac{s}{\bar{x}}$$

Streuungsmaße (2)

Anmerkungen:

- **Spannweite und Varianz** (und somit Standardabweichungen und Variationskoeffizient) sind nur für **metrische Merkmale** geeignet, nicht für nominale oder ordinale Merkmale.
- Die **Maßeinheit der Varianz ist quadratisch**, die Standardabweichung und die Spannweite werden in der gleichen Maßeinheit wie die Messwerte angegeben, der **Variationskoeffizient** besitzt keine Maßeinheit, ist also **dimensionslos**.
- Beispiel:

		HH-Einkommen in Euro	HH-Einkommen in 1,000 Euro	Verhältnis
Untersuchungsumfang	n	5,407	5,407	1
Mittelwert	$\bar{x}$	37,149.97	37.15	1,000
Varianz	s^2	714,384,481.32	714.38	1,000,000
Minimum	x_{min}	583.00	0.58	1,000
Maximum	x_{max}	507,369.00	507.37	1,000
Standardabweichung	s	26,727.97	26.73	1,000
Variationskoeffizient	V	0.72	0.72	1

Streuungsmaße (2)

EXCEL-Befehle:

- **Spannweite:** muss über die Befehle MAX und MIN (für die größte und kleinste Ausprägung) berechnet werden.
- **Varianz:** VAR.P
- **Standardabweichung:** STABW.N

Handelt es sich bei einem Datensatz nur um eine Stichprobe (mit Umfang $n < N$), dann muss die **korrigierte Varianz** $\hat{s}^2$ und die **korrigierte Standardabweichung** $\hat{s}$ berechnet werden (weil mit $n = 1$ $\hat{s}^2$ nicht berechnet werden kann):

$$\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\hat{s} = +\sqrt{\hat{s}^2}$$

Die **EXCEL-Befehle** sind VAR.S (korrigierte Varianz) und STABW.S (korrigierte Standardabweichung). Wenn der Stichprobenumfang n groß ist, ist der Unterschied allerdings vernachlässigbar, da $\frac{1}{n-1} \approx \frac{1}{n}$.

EXCEL Add-In Analysefunktionen

Eine Berechnung der Lage und Streuungsmaße ist in Excel auch über **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Populationskenngrößen** $\rightarrow$ **Statistische Kenngrößen** möglich. Hierbei wird auf die korrigierte Varianz bzw. die korrigierte Standardabweichung zurückgegriffen.

- Mit diesem Befehl können **nur numerische Ausprägungen** verarbeitet werden können.
- Wenn die Merkmale numerisch sind, werden **alle Populationskenngrößen** ausgewiesen, selbst dann, wenn einzelne Maßzahlen **nicht sinnvoll** sind!