SE Geographie und Ökonomie

Einheit 2:
Univariate Datenanalyse: Deskriptive Statistik

Bernhard Schmidpeter

bernhard.schmidpeter@jku.at

Institut für Volkswirtschaftslehre

SoSe 2024

(Fiktiver) Gemeindefinanzbericht

	A	B	C	D	E	F	G	H
1	Region_ID	Verwaltungseinheit	Ausgaben Verwaltung	Ausgaben öff. Güter	Einnahmen Gewerbesteuer	Einnahmen Umverteilung	Überschuss	
2	101	1	0.90	4.70	1.62	1.48	-2.50	
3	101	2	0.66	5.19	1.16	1.17	-3.52	
4	103	1	0.73	4.89	2.20	1.19	-2.24	
5	103	2	0.09	5.88	1.43	1.17	-3.36	
6	104	1	0.63	3.54	1.01	0.39	-2.77	
7	104	2	2.28	4.82	14.48	1.68	9.06	
8	105	1	0.27	1.28	2.35	0.52	1.33	
9	105	2	0.40	5.53	3.39	1.99	-0.54	
10	106	1	1.40	7.70	2.14	1.12	-5.84	
11	106	2	0.66	8.84	3.33	1.13	-5.04	
12	107	1	0.50	-159.35	-4.88	1.24	155.22	
13	107	2	0.21	-29.71	-1.79	0.33	28.04	
14	108	1	0.54	5.07	1.49	1.39	-2.74	
15	108	2	0.70	5.24	1.23	1.09	-3.61	
16	109	1	0.88	8.62	1.62	1.07	-6.82	
17	109	2	0.40	9.65	3.55	1.69	-4.81	
18	110	1	1.46	3.80	2.36	1.02	-1.89	
19	110	2	19.35	116.30	32.97	2.82	-99.86	
20	120	1	0.90	8.67	-1.11	0.86	-9.81	
21	120	2	2.90	7.98	9.10	1.92	0.14	
22	131	1	0.23	4.70	2.92	0.63	-1.38	
23	131	2	-0.02	4.70	2.46	0.31	-1.91	
24	132	1	0.11	0.90	1.83	0.05	0.87	
25	132	2	0.05	2.49	1.49	-0.15	-1.20	
26	133	1	1.14	4.71	3.40	1.91	-0.54	
27	133	2	0.41	4.78	1.26	0.45	-3.49	
28	139	1	0.33	7.43	0.74	0.42	-6.60	
29	139	2	0.29	5.61	-0.36	0.32	-5.93	
30	141	1	0.72	4.42	1.84	0.84	-2.46	
31	141	2	0.47	2.99	-0.96	0.78	-3.64	
32	143	1	1.08	8.55	1.78	0.91	-6.93	
33	143	2	-0.38	6.66	1.86	0.64	-3.78	
34								

Deskriptive Statistik

- Rohdaten und assozierten 'Urlisten' der enthaltenen Merkmäler sind oft unübersichtlich
- Die Daten sollten deshalb in einer übersichtlichen Form dargestellt werden
  - Verteilung eines Merkmals in den Daten
  - Durchschnitt und Streuung eines Merkmals
- Um zu einer übersichtlichen Form zu gelangen, werden Daten of verdichtet
  - Durch verdichten gehen Informationen verloren
  - Durch verdichten können bestimmte Zusammenhänge suggeriert werden

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)**
- Beschreibung einer Variable durch **Lage- und Streuungsmaße**
- Sie sind sich bewusst, wie unterschiedliche Darstellungen der Daten einen unterschiedlichen Zusammenhang suggerieren können

Häufigkeitsverteilung

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

Bezeichnungen:

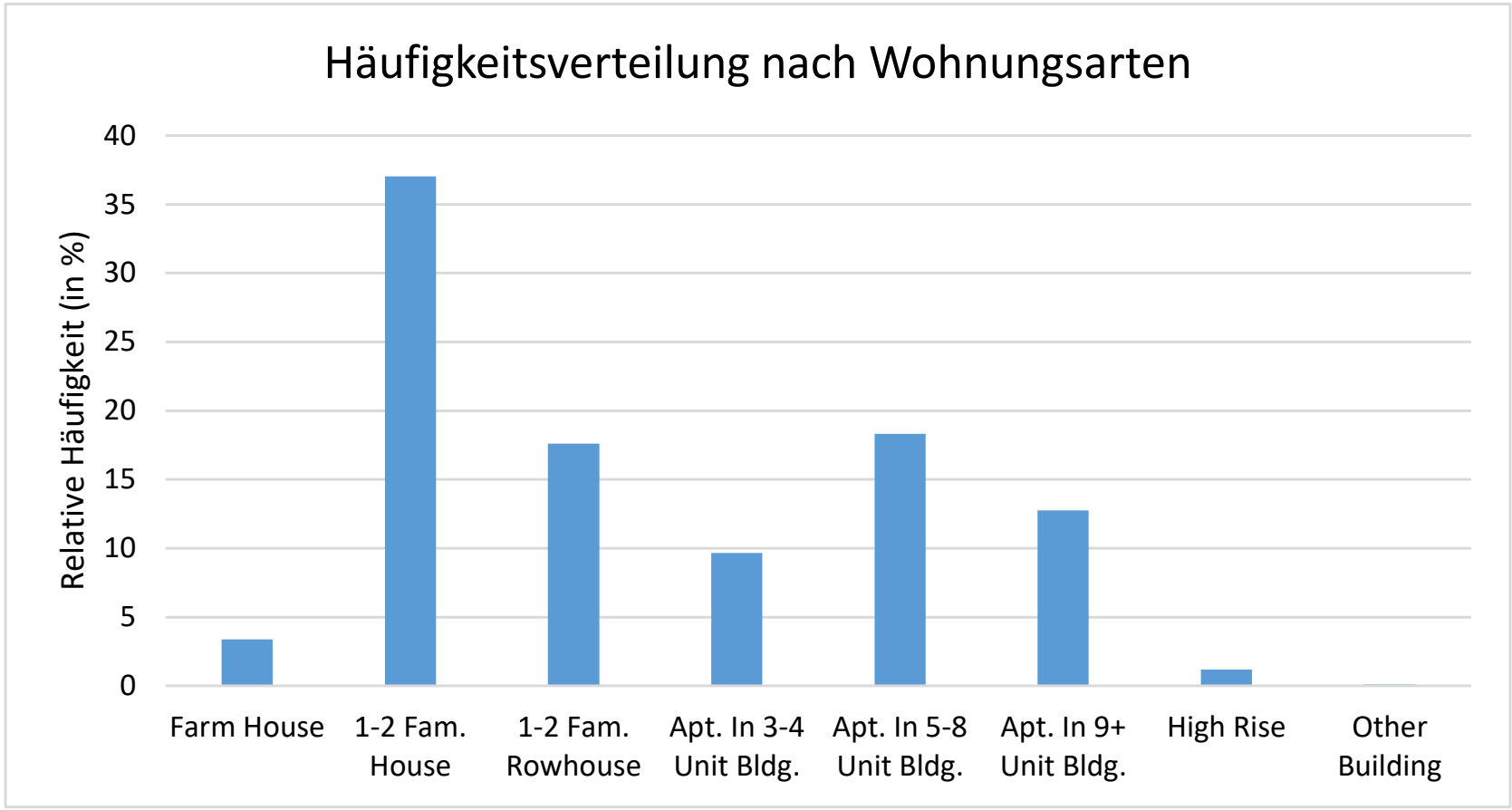
N	Untersuchungsumfang (Population von Interesse)
n	Stichprobenumfang (Unsere "Daten")
r	Anzahl an verschiedenen Ausprägungen (Variablen)
x_m	Ausprägung, $m = 1, ..., r$
h_m	absolute Häufigkeit der Ausprägung x_m
p_m	$= h_m/n$ relative Häufigkeit der Ausprägung x_m
P_m	$= 100 \cdot p_m$ relative Häufigkeit 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.

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.
- Unterschiedliche Intervallbreiten können zu unterschiedlichen Wahrnehmungen der Daten führen

Häufigkeitsverteilung: Histogramm (Wohnungsgröße)

Verteilung der Wohnungsgröße:

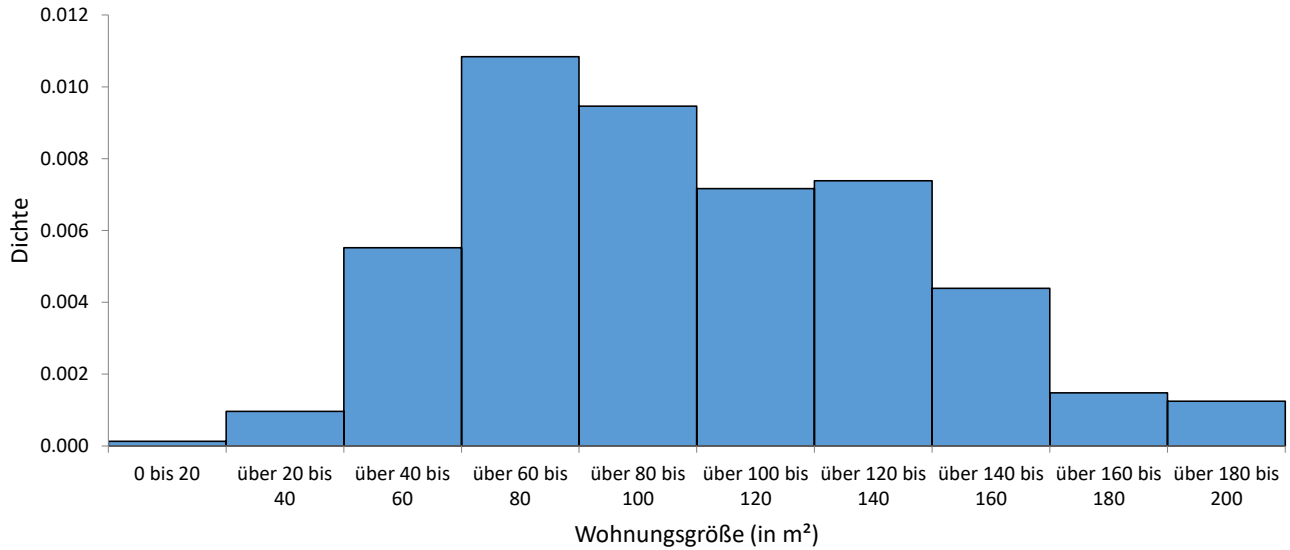
Klasse m	Ausprägung x_m	Absolute Häufigk. h_m	Relative Häufigk. p_m	Relative Häufigk. P_m (in %)	Dichte $f_m = p_m/d_m$
1	0 m^2 bis 20 m^2	14	0.003	0.3	0.00013
2	über 20 m^2 bis 40 m^2	104	0.019	1.9	0.00096
3	über 40 m^2 bis 60 m^2	597	0.110	11.0	0.00552
4	über 60 m^2 bis 80 m^2	1,173	0.217	21.7	0.01084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

Histogramm: Wohnungsgröße (grafische Darstellung)

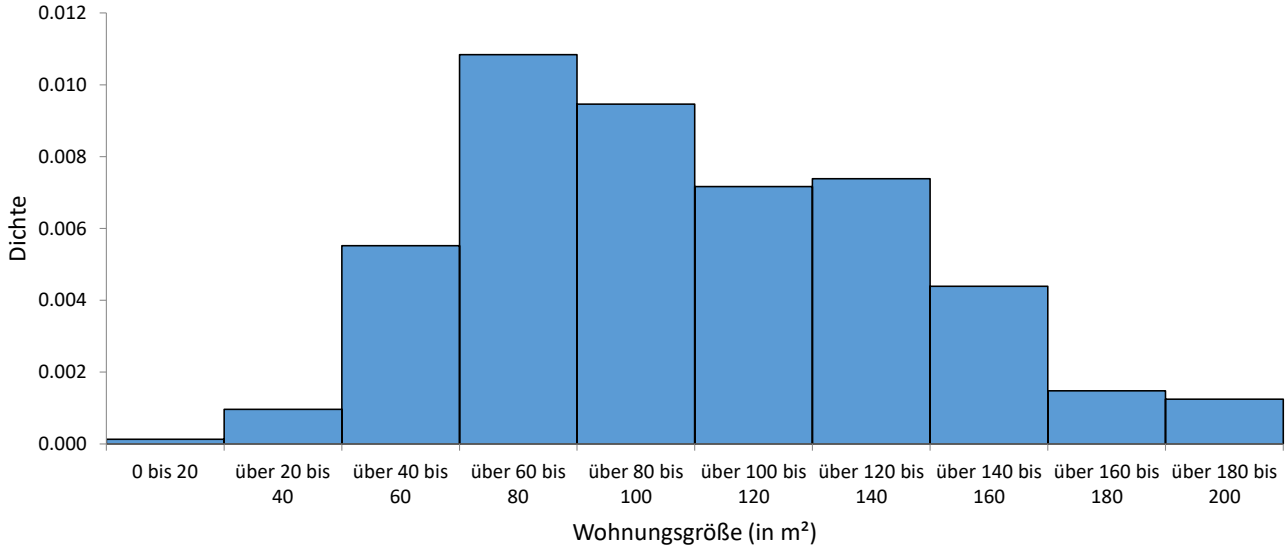
- Eine Tabellendarstellung kann oft unübersichtlich sein und eine graphische Darstellung wird bevorzugt

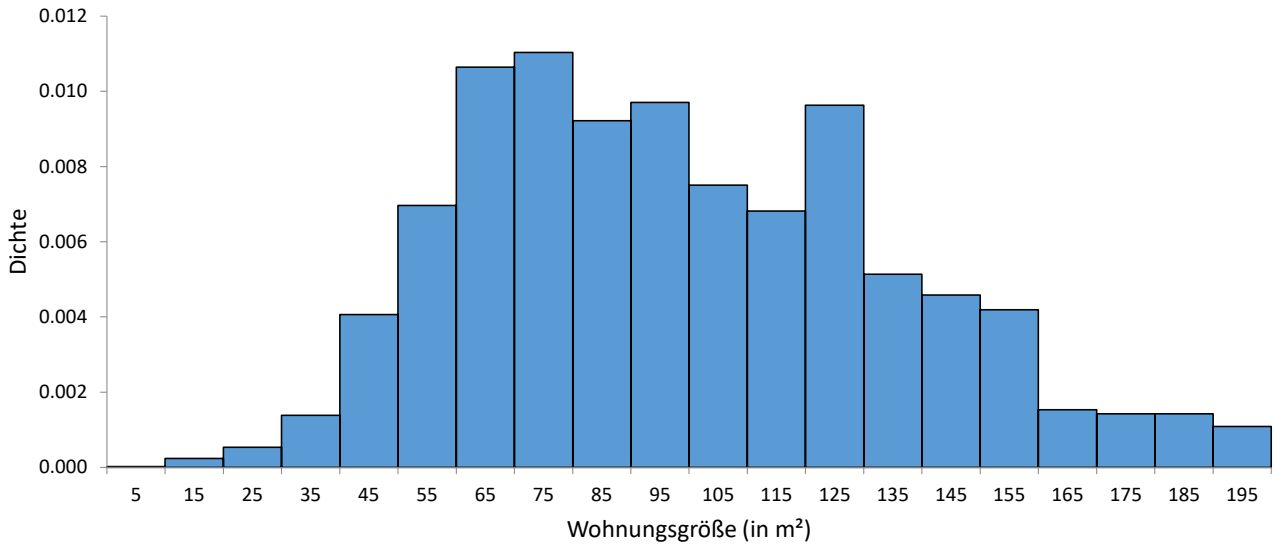
- Ein Histogramm ist für **metrische stetige Merkmale** geeignet, deren Ausprägungen in **Intervalle** zusammengefasst wurden.
  - Auf der x-Achse die **Ausprägungen** aufgetragen
  - Auf der y-Achse die **Dichten** f_m aufgetragen
  - Wenn alle Intervall **gleich** breit sind, so kann man anstatt der Dichte die **Häufigkeit** verwenden, was einfacher zu interpretieren ist

Histogramm: Wohnungsgröße (grafische Darstellung)

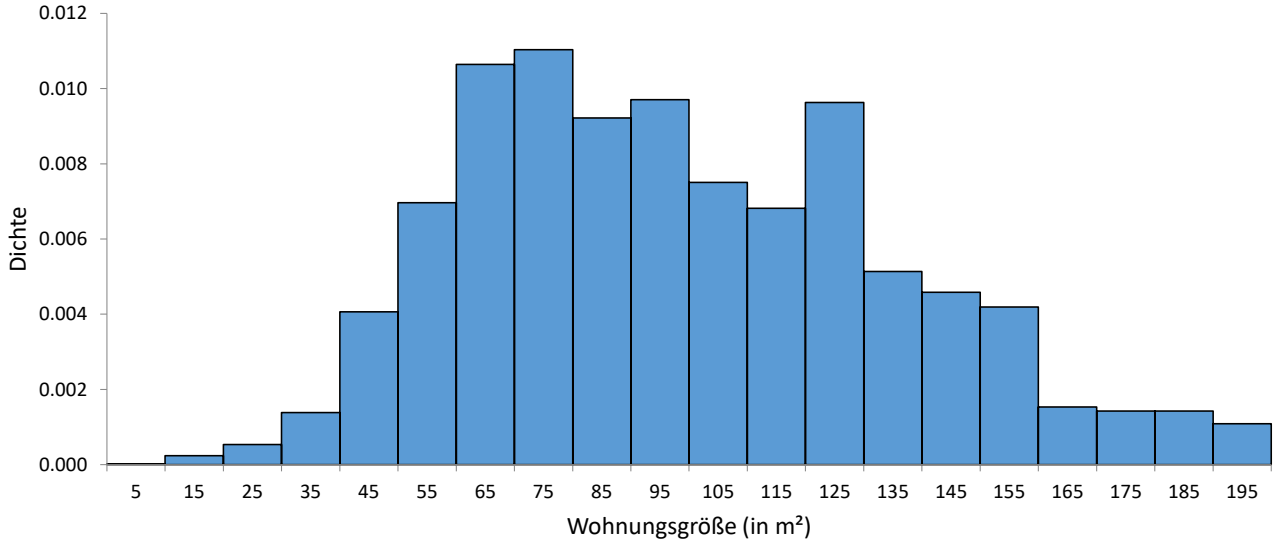


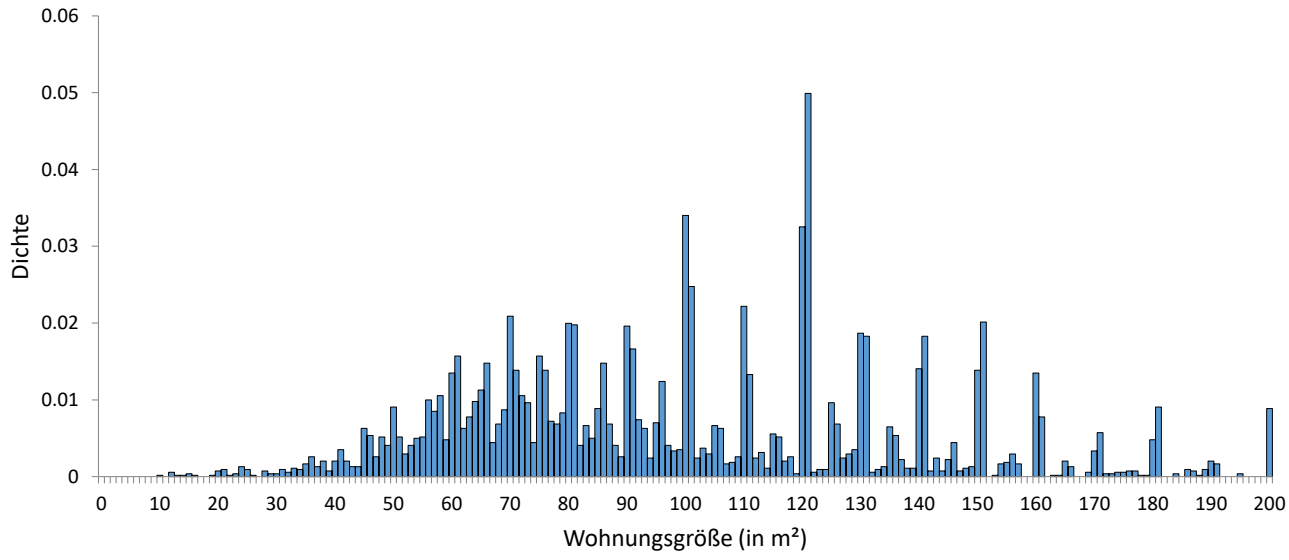
Histogramm: verschiedene Intervallbreiten (1)





Histogramm: verschiedene Intervallbreiten (2)





EXCEL Add-In Analysefunktionen

- Eine Häufigkeitstabelle kann in Excel auch über **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Histogramm** erstellt werden
- Dabei werden lediglich absolute Häufigkeiten h_m ausgewiesen. Die relativen Häufigkeiten p_m und die Dichte f_m muss selbständig berechnet werden.
- Wird **Diagrammdarstellung** angehakt, wird ein Stabdiagram bzw. ein Histogramm ausgegeben, welches noch formatiert werden muss. Zum Beispiel, wird hier die absoluten Häufigkeiten aufgetragen.
- Mit der Analyse Funktion können nur **nur numerische Ausprägungen** verarbeitet werden
  - Zum Beispiel ist es nicht möglich, die Namensverteilung in einer Klasse darstellen zu lassen

Alternative Vorgehensweise in EXCEL

- EXCEL-Befehl **HÄUFIGKEIT**: 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.
- EXCEL-Befehl **ZÄHLENWENN**: Die absoluten Häufigkeiten einzelner Merkmale werden abgezählen. Dieser Befehl kann **auch nicht-numerische Informationen** verarbeiten.
  - Flexibel aber arbeitsaufwendig
  - Alle benötigten Größen müssen selbstständig berechnet werden

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 (z.B. Mittelwert)
2. **Streuungsmaße:** geben an, wie weit die Daten von einander oder von einer Lagemaßzahl abweichen (z.B. Varianz)

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, $\bar{x}$)

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

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

$$\bar{x} = \frac{1}{N} \sum_{m=1}^{r} x_m h_m = \sum_{m=1}^{r} x_m p_m \tag{2}$$

berechnet werden. (Formel 1 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 Formel 2 verwendet werden.
- **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:

- **EXCEL-Befehl:** MEDIAN (auch QUANTIL.INKL möglich, siehe nächste Seiten)
- Für **ordinale und metrische Merkmale** geeignet. Ungeeignet für nominale Merkmale.

Lagemaße: Mittelwert vs Median

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.

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.

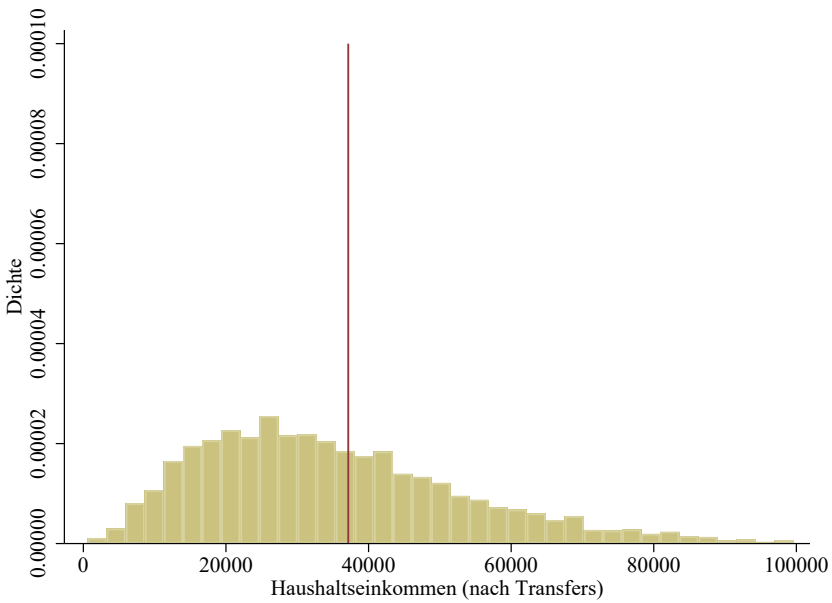
$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{wenn } N \cdot \alpha \text{ keine ganze Zahl ist} \\ & k \text{ ist dann die auf } N \cdot \alpha \text{ folgende ganze Zahl} \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{wenn } N \cdot \alpha \text{ eine ganze Zahl ist} \\ & \text{dann ist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

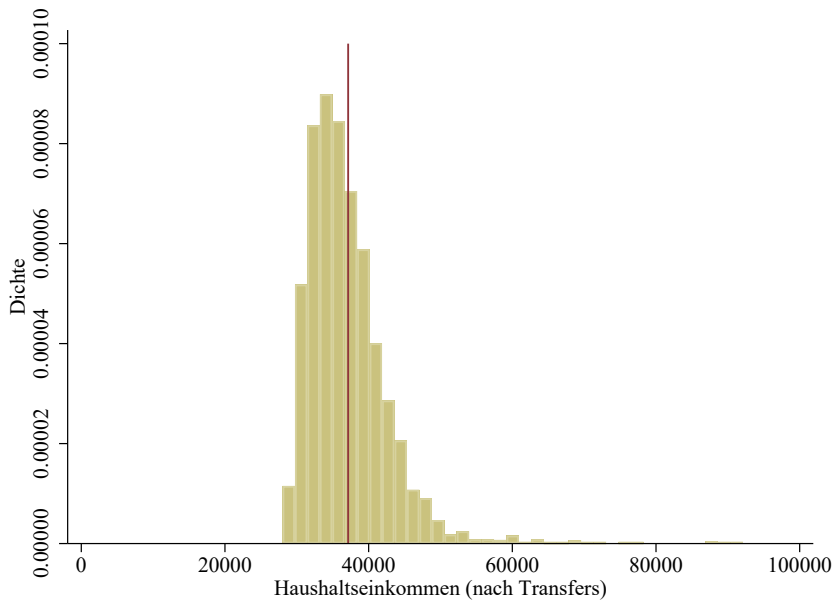
Spezialfälle:

- **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)

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.

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

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

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

Streuungsmaße (2)

Anmerkungen:

- **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:

- **Varianz:** VAR.P für Grundgesamtheit
- **Standardabweichung:** STABW.N für Grundgesamtheit

Wichtig: 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)

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!