SE Greographie und Ökonomie

Einheit 4:
Bivariate Datenanalyse

Dieter Pennerstorfer

dieter.pennerstorfer@jku.at

Institut für Volkswirtschaftslehre

WS 2021/22

Motivation

- Üblicherweise ist man nicht an einem Merkmal interessiert, sondern an mehreren Merkmalen bzw. **Zusammenhängen zwischen verschiedenen Merkmalen**.
- Wir untersuchen in dieser Einheit den **Zusammenhang zwischen zwei Merkmalen**. Zusammenhänge zwischen zwei Merkmalen (etwa: Unterschiede zwischen Gruppen) sind auch von wirtschaftspolitischem Interesse.
- **Anwendungsbeispiele:**
  - Arbeiten Frauen häufiger Teilzeit als Männer?
  - Verdienen auch Vollzeit beschäftigte Frauen weniger als (Vollzeit beschäftigte) Männer?
  - Wohnen reichere Leute in größeren Wohnungen?

Lernziele der Einheit 4

- Sie können bestimmen, **ob es einen Zusammenhang** zwischen zwei Merkmalen gibt.
- Abhängig davon, ob es sich um **diskrete oder stetige Merkmale** handelt, wissen Sie, **welche Maßzahlen geeignet** sind, einen möglichen Zusammenhang zwischen zwei Merkmalen abzubilden.
- Die können diese Maßzahlen **in EXCEL berechnen** und richtig **interpretieren**.
- Sie können beurteilen, ob dieser **Zusammenhang statistisch signifikant** ist. Das bedeutet, dass es diesen Zusammenhang mit hoher Wahrscheinlichkeit **auch in der Grundgesamtheit** gibt.

Zusammenhänge zwischen zwei Variablen

Was bedeutet Zusammenhang?

- Merkmal x **beeinflusst** Merkmal y: Wenn ich meine Arbeitszeit erhöhe, dann steigt mein Jahreseinkommen.
- Merkmal x und Merkmal y hängen zusammen (d.h. sie sind **nicht unabhängig voneinander**): Personen mit höheren Einkommen wohnen in größeren Wohnungen. Kausalität in beide Richtungen denkbar. Möglich, dass eine Dritte Variable (z.B.: Vermögen der Eltern) beide Variablen beeinflusst.
- Merkmal x beinhaltet Informationen über Merkmal y: Leute mit längeren Beinen sind üblicherweise größer.
- Üblicherweise sind wir an **kausalen Wirkungszusammenhängen** interessiert. Das ist aber oft sehr schwierig festzustellen (wir werden später darauf zurückkommen).

Wie der Zusammenhang zwischen zwei Variablen untersucht werden kann, hängt davon ab, ob es sich um **diskrete oder stetige metrische Merkmale** handelt:

		Variable 2	
		diskret	stetig
Variable 1	diskret	Kreuztabelle	Mittelwertvergleich
	stetig	Mittelwertvergleich	Korrelation

Zusammenhang zwischen diskretem und stetigem Merkmal: Mittelwertvergleich

		Variable 2	
		diskret	stetig
Variable 1	diskret	Kreuztabelle	Mittelwertvergleich
	stetig	Mittelwertvergleich	Korrelation

Mittelwertvergleich: Beispiel

Haben Frauen und Männer (diskretes Merkmal) im Durchschnitt unterschiedliche (Haushalts-)Einkommen (stetiges Merkmal)?

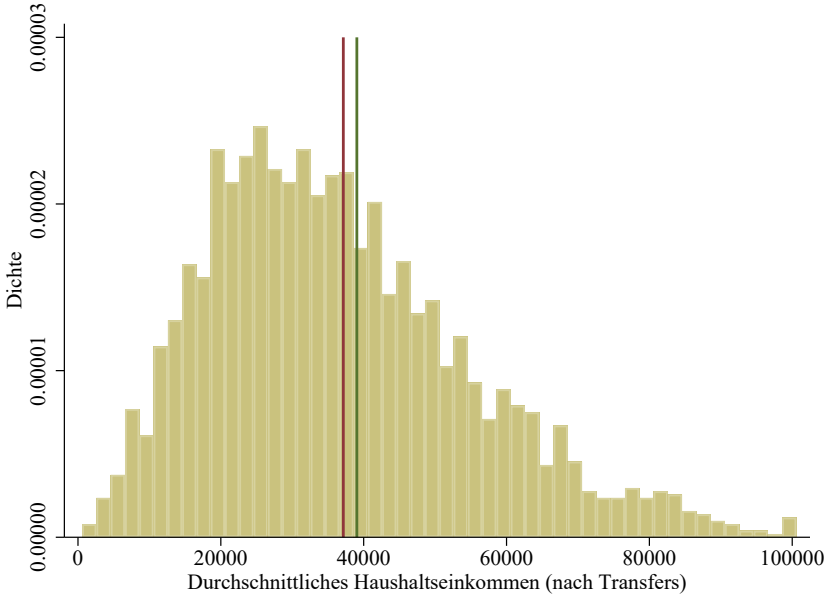
- **Nullhypothese** (H_0): **In der Grundgesamtheit** gibt es **keinen Unterschied** im Haushaltseinkommen zwischen den beiden Gruppen: $\mu_M = \mu_F$ (allgemein: $\mu_1 = \mu_2$)
- **Alternativhypothese** (H_1): **In der Grundgesamtheit** gibt es **einen Unterschied** im Haushaltseinkommen zwischen den beiden Gruppen: $\mu_M \neq \mu_F$ (allgemein: $\mu_1 \neq \mu_2$)
- Wir führen dazu einen **zweiseitigen Zweistichproben-t-Test** für **unabhängige Stichproben** durch.

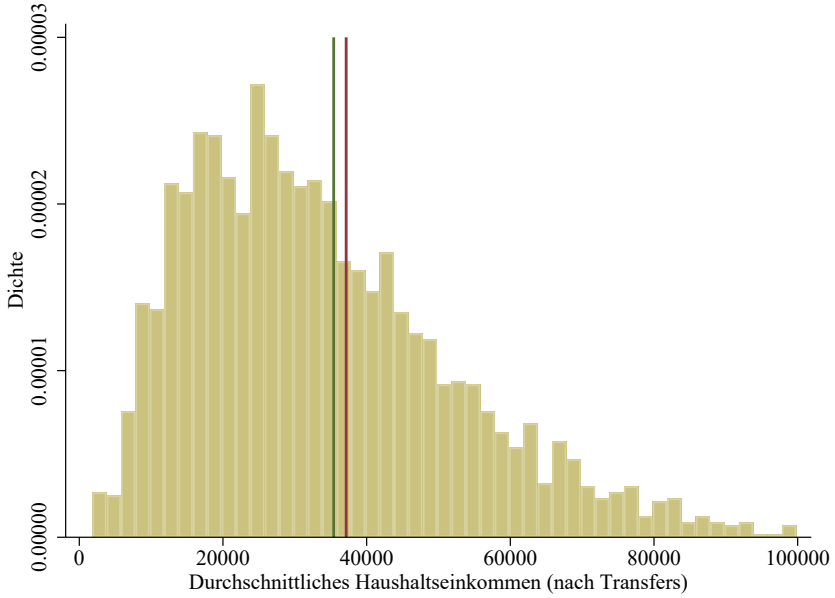
  Die Nullypothese wird daher verworfen, wenn Männer ein signifikant höheres Haushaltseinkommen haben als Frauen **oder** wenn Männer ein signifikant niedrigeres Haushaltseinkommen haben als Frauen.

	n	$\overline{x}$	s	x_{min}	x_{max}
HH-Einkommen (alle)	5,407	37,149.97	26,727.97	583	507,369
HH-Einkommen Männer	2,583	39,044.55	28,397.41	583	507,369
HH-Einkommen Frauen	2,824	35,417.08	24,983.54	1,809	507,369

(Haushalts-)Einkommensverteilung von Männern und Frauen

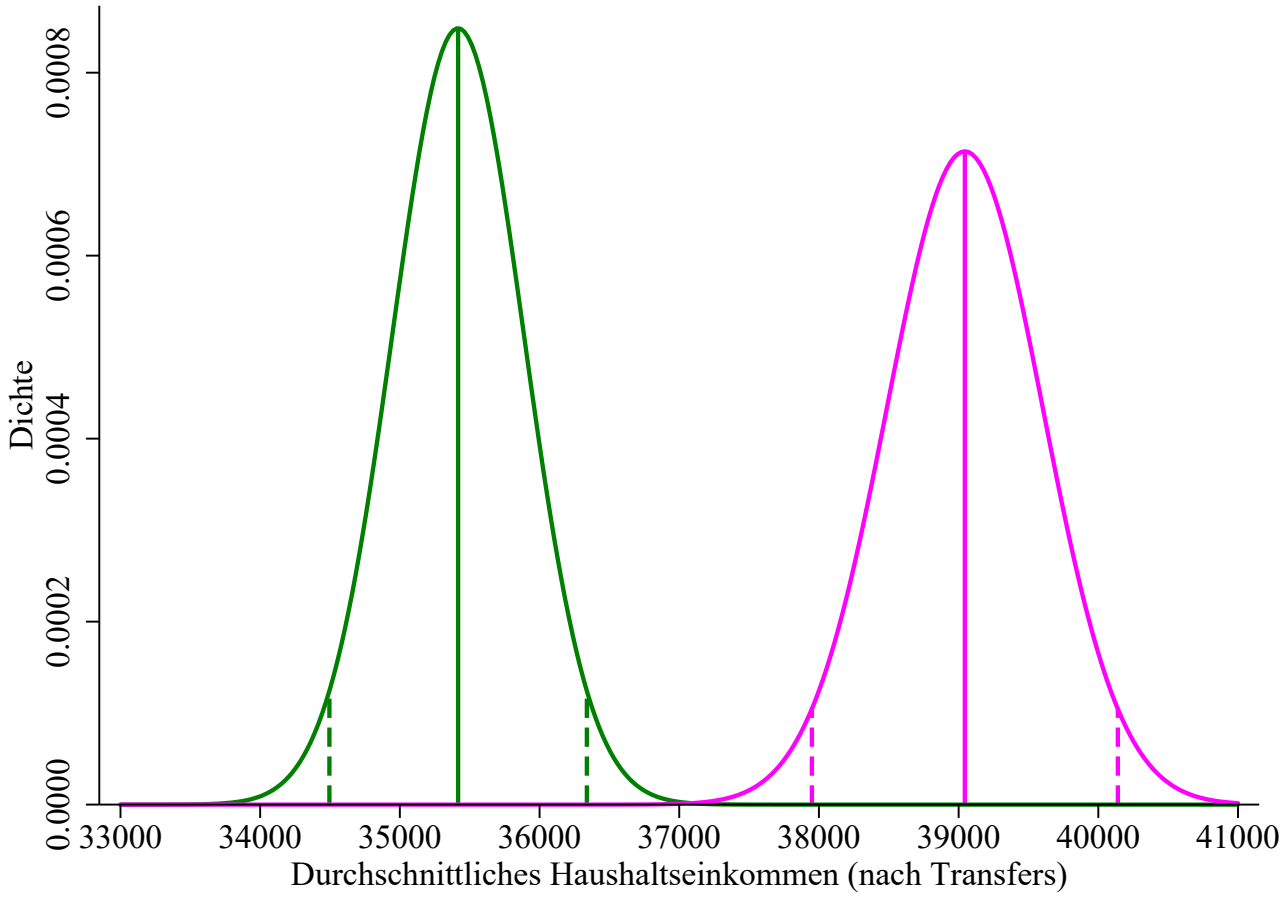
Histogramm der Haushaltseinkommen von Männern (links) und Frauen (rechts). Die rote Linie markiert das durchschnittliche Haushaltseinkommen für die gesamte Stichprobe, die grünen Linien das durchschnittliche Haushaltseinkommen der Gruppe (Männer bzw. Frauen).





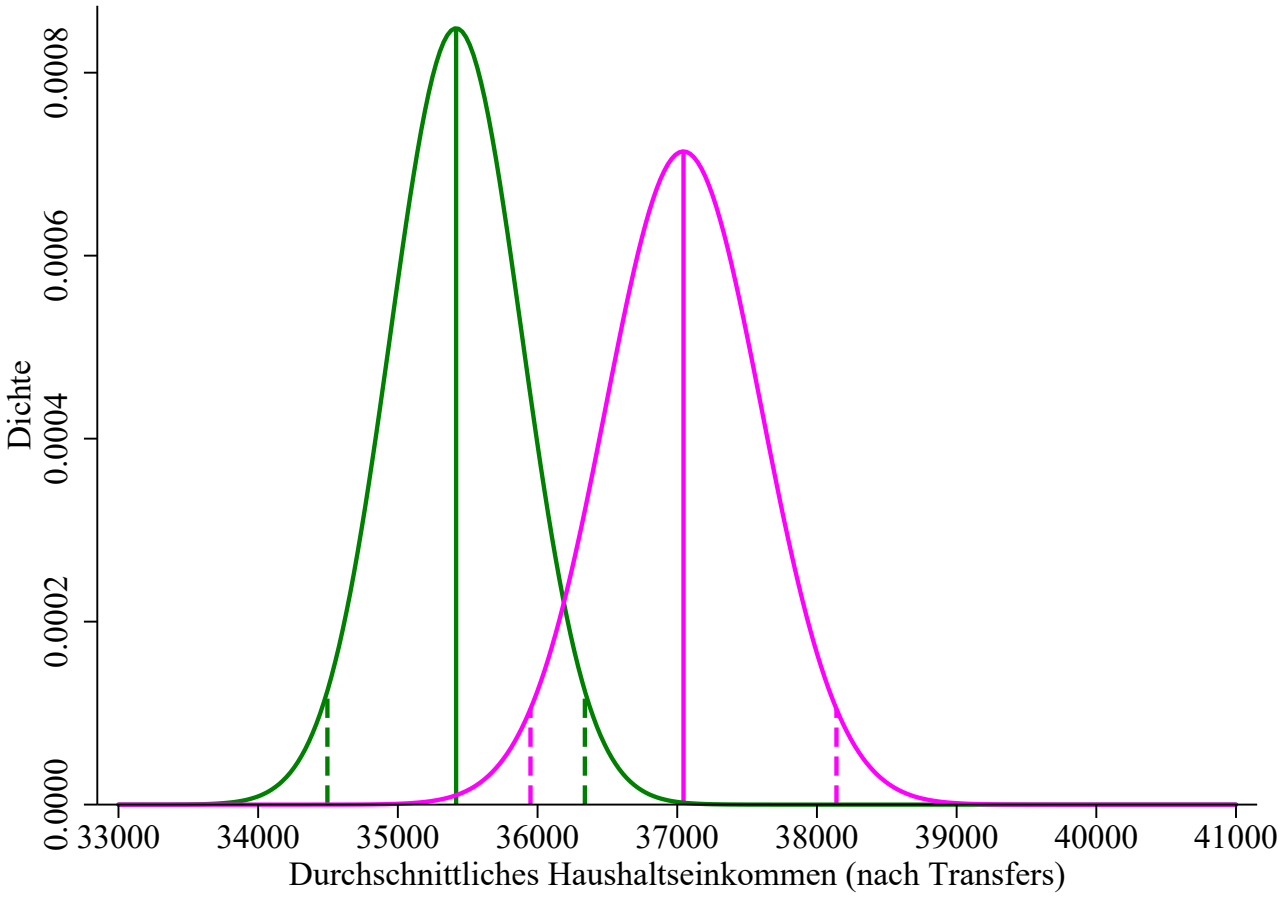
Verteilung der Stichprobenmittelwerte

Konfidenzintervall für Frauen (grün) und Männer (rosa). Eindeutig, da sich Konfidenzintervalle nicht überschneiden: Nullhypothese kann verworfen werden.



Verteilung der Stichprobenmittelwerte

Konfidenzintervall für Frauen (grün) und Männer (rosa). Hier ist die Situation weniger eindeutig:



Zweiseitiger Zweistichproben-t-Test

Nullhypothese (H_0): $\mu_M = \mu_F \;\Rightarrow\; \mu_M - \mu_F = 0$
Alternativhypothese (H_1): $\mu_M \neq \mu_F \;\Rightarrow\; \mu_M - \mu_F \neq 0$
Die Mittelwerte der Grundgesamtheit μ_M und μ_F sind nicht beobachtbar, wir müssen uns mit einem Vergleich der Stichprobenmittelwerte $\bar{x}_M$ und $\bar{x}_F$ begnügen.

- Wir **wissen von letzter Einheit**, dass:
  - $\bar{x}_M \sim N(\mu_M, \sigma_M^2/n_M)$
  - $\bar{x}_F \sim N(\mu_F, \sigma_F^2/n_F)$
- Summen und Differenzen von normalverteilten Zufallsvariablen folgen ebenfalls einer Normalverteilung:
  - $\bar{x}_M + \bar{x}_F \sim N(\mu_M + \mu_F, \sigma_M^2/n_M + \sigma_F^2/n_F)$
  - $\bar{x}_M - \bar{x}_F \sim N(\mu_M - \mu_F, \sigma_M^2/n_M + \sigma_F^2/n_F)$

  Wenn σ_M^2 und σ_F^2 bekannt sind, kann das $(1-\alpha)$ Konfidenzintervall wie folgt berechnet werden:

$$\overline{x}_M - \overline{x}_F - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}} \leq \underbrace{\mu_M - \mu_F}_{=0 \text{ unter } H_0} \leq \overline{x}_M - \overline{x}_F + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}$$

Zweiseitiger Zweistichproben-t-Test

Standardisierte Teststatistik:

$$\overline{x}_M - \overline{x}_F - u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}} \leq 0 \leq \overline{x}_M - \overline{x}_F + u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}$$

$$u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}} \leq \overline{x}_M - \overline{x}_F \leq +u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}$$

$$-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{x}_M - \overline{x}_F}{\sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}} \leq +u_{1-\frac{\alpha}{2}}$$

$$\left| \frac{\overline{x}_M - \overline{x}_F}{\sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_F^2}{n_F}}} \right| \leq u_{1-\frac{\alpha}{2}}$$

Wenn σ_M und σ_F nicht bekannt sind: (üblicherweise der Fall)

$$\text{t-Wert} = \left| \frac{\overline{x}_M - \overline{x}_F}{\sqrt{\frac{\hat{s}_M^2}{n_M} + \frac{\hat{s}_F^2}{n_F}}} \right| \leq t_{n_M+n_F-2;1-\frac{\alpha}{2}} = \text{kritischer Wert } c \text{ der Verteilung}$$

mit $\hat{s}_M^2/n_M = \hat{se}_M^2$ und $\hat{s}_F^2/n_F = \hat{se}_F^2$.

Beispiel

	n	$\overline{x}$	s	x_{min}	x_{max}
HH-Einkommen (alle)	$5,407$	$37,149.97$	$26,727.97$	583	$507,369$
HH-Einkommen Männer	$2,583$	$39,044.55$	$28,397.41$	583	$507,369$
HH-Einkommen Frauen	$2,824$	$35,417.08$	$24,983.54$	$1,809$	$507,369$

Standardisierte Teststatistik:
H_0 wird nicht verworfen, wenn:

$$\text{t-Wert} = \left| \frac{\overline{x}_M - \overline{x}_F}{\sqrt{\frac{\hat{s}_M^2}{n_M} + \frac{\hat{s}_F^2}{n_F}}} \right| \leq t_{n_M+n_F-2;1-\frac{\alpha}{2}} = \text{ kritischer Wert } c$$

$$\text{t-Wert} = \left| \frac{39,045 - 35,417}{\sqrt{\frac{28,397^2}{2,583} + \frac{24,984^2}{2,824}}} \right| = \frac{3,627}{730} = 4.97$$

$$\text{kritischer Wert } c = t_{n_M+n_F-2;1-\frac{\alpha}{2}} = t_{5,405;0.975} = 1.96$$

$$\text{t-Wert} = 4.97 > 1.96 = \text{ kritischer Wert } c$$

$\rightarrow$ **Nullhypothese wird** zugunsten der Alternativhypothese **verworfen**.
$\rightarrow$ Es ist sehr wahrscheinlich, dass sich in der Grundgesamtheit die Haushaltseinkommen von Männern und Frauen unterscheiden.

Einseitiger Zweistichproben-t-Test

- **Jedenfalls**
  - Berechnung einer standardisierten Teststatistik (t-Wert):

$$\text{t-Wert} = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \right|$$

  - Ermittlung des kritischen Werts c der Verteilung: $c = t_{n_1+n_2-2;1-\alpha}$

- Fall 1: $H_0: \ \mu_1 \leq \mu_2$; $H_1: \ \mu_1 > \mu_2$
  - H_0 wird (zugunsten von H_1) verworfen, wenn

$$\text{t-Wert} = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \right| \geq t_{n_1+n_2-2;1-\alpha} \ \textbf{und} \ \overline{x}_1 > \overline{x}_2$$

- Fall 2: $H_0: \ \mu_1 \geq \mu_2$; $H_1: \ \mu_1 < \mu_2$
  - H_0 wird (zugunsten von H_1) verworfen, wenn

$$\text{t-Wert} = \left| \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}} \right| \geq t_{n_1+n_2-2;1-\alpha} \ \textbf{und} \ \overline{x}_1 < \overline{x}_2$$

Umsetzung in EXCEL

Es bietet sich an, die Variablen der beiden Gruppen in getrennte Spalten zu kopieren. Alle notwendigen Befehle habe wir in Einheit 3 schon besprochen.
Alternativ: **Daten** → **Analyse** → **Datenanalyse** → **Zweistichproben t-Test: Unterschiedlicher Varianzen**

	hhinc Männer	hhinc Frauen
Mittelwert	39044.5451	35417.079
Varianz	806412884	624177386
Beobachtungen	2583	2824
Hypothetische Differenz der Mittelwerte	0	
Freiheitsgrade (df)	5165	
t-Statistik	4.96761202	
P(T<=t) einseitig	3.4984E-07	
Kritischer t-Wert bei einseitigem t-Test	1.6451487	
P(T<=t) zweiseitig	6.9968E-07	
Kritischer t-Wert bei zweiseitigem t-Test	1.96042339	

Anmerkung: Wenn sich die Varianzen stark unterscheiden (Faustregel: die Varianz einer Gruppe ist mehr als doppelt so groß wie die Varianz einer anderen Gruppe), dann sollten die Freiheitsgrade angepasst werden. Das Datenanalyse-Tool in Excel macht das automatisch, wodurch ein geringfügig anderer kritischer t-Wert berechnet wird. Für unsere Zwecke spielt das aber keine Rolle.

Zusammenhang zwischen zwei diskreten Merkmalen: Kreuztabelle

		Variable 2	
		diskret	stetig
Variable 1	diskret	Kreuztabelle	Mittelwertvergleich
	stetig	Mittelwertvergleich	Korrelation

Kreuztabelle

Eine **Kreuztabelle** (auch Kontingenztabelle oder Kontingenztafel) weist die absoluten oder die relativen Häufigkeiten aller Kombinationen der Merkmalsausprägungen von zwei diskreten Merkmalen aus (Merkmale müssen nicht notwendigerweise dichotom sein):

		Variable 2		
		$y = 1$	$y = 2$	Summe
Variable 1	$x = 1$	h_{11}	h_{12}	h_{1+}
	$x = 2$	h_{21}	h_{22}	h_{2+}
	Summe	h_{+1}	h_{+2}	n

Bezeichnungen:

h_{ij}	absolute Häufigkeit der Kombination $x = x_i$ und $y = y_j$
n	Stichprobenumfang, $n = \sum_i \sum_j h_{ij}$
$p_{ij} = h_{ij}/n$	relative Häufigkeit der Kombination $x = x_i$ und $y = y_j$
$P_{ij} = p_{ij} \cdot 100$	relative Häufigkeit der Kombination $x = x_i$ und $y = y_j$ in %
$h_{i+}(p_{i+})$	Zeilensummen, Randhäufigkeiten des Merkmals x
$h_{+j}(p_{+j})$	Spaltensummen, Randhäufigkeiten des Merkmals y

Beispiel: Erwerbsstatus nach Geschlecht

Grundgesamtheit: Alle regulär Erwerbstätigen in Deutschland
Stichprobe: alle regulär Erwerbstätigen im GSOEP

Kreuztabelle: absolute Häufigkeiten

		Erwerbsstatus		
		Vollzeit	Teilzeit	Summe
Geschlecht	Männer	1,346	59	1,405
	Frauen	695	540	1,235
	Summe	2,041	599	2,640

Kreuztabelle: relative Häufigkeiten

		Erwerbsstatus		
		Vollzeit	Teilzeit	Summe
Geschlecht	Männer	0.510	0.022	0.532
	Frauen	0.263	0.205	0.468
	Summe	0.773	0.227	1.000

Randverteilungen: Gibt Auskunft über die Verteilung eines Merkmals, ohne die Verteilung des anderen Merkmals zu berücksichtigen.

Bedingte Wahrscheinlichkeiten

Ist der Anteil der Vollzeitbeschäftigten bei Männern höher als bei Frauen?

Bezeichnung:

$h_{ij}/h_{i+} = p_{ij}/p_{i+}$	bedingte relative Häufigkeit der Ausprägung y_j des Merkmals y bei gegebener Ausprägung x_i des Merkmals x

Bedingte relative Häufigkeiten

		Erwerbsstatus		
		Vollzeit	Teilzeit	Summe
Geschlecht	Männer	0.510/0.532	0.022/0.532	0.532/0.532
	Frauen	0.263/0.468	0.205/0.468	0.468/0.468

Bedingte relative Häufigkeiten

		Erwerbsstatus		
		Vollzeit	Teilzeit	Summe
Geschlecht	Männer	0.958	0.042	1.000
	Frauen	0.563	0.437	1.000

Stärke des Zusammenhangs

Idee: Um die Stärke des Zusammenhangs zu beurteilen, soll die **beobachtete Verteilung** mit jener Verteilung verglichen werden, die ich **erwarten** würden, wenn die beiden Merkmale keinen Zusammenhang aufweisen.

Bezeichnungen:

$p^e_{ij} = p_{i+} \cdot p_{+j}$	erwartete (**e** für expected) relative Häufigkeit von $x = x_i$ und $y = y_j$ bei Unabhängigkeit von x und y
$h^e_{ij} = p^e_{ij} \cdot n$ $= (h^o_{i+} \cdot h^o_{+j})/n$	erwartete absolute Häufigkeit dieser Kombination bei Unabhängigkeit von x und y
h^o_{ij}	beobachtete (**o** für observed) absolute Häufigkeit dieser Kombination

Erwartete und beobachtete Häufigkeiten

Kreuztabelle: relative Häufigkeiten

		Erwerbsstatus		
		Vollzeit	Teilzeit	Summe
Geschlecht	Männer	$p^o_{11} = 0.510$ $p^e_{11} = 0.411$	$p^o_{12} = 0.022$ $p^e_{12} = 0.121$	$p^o_{1+} = 0.532$
	Frauen	$p^o_{21} = 0.263$ $p^e_{21} = 0.362$	$p^o_{21} = 0.205$ $p^e_{22} = 0.106$	$p^o_{2+} = 0.468$
	Summe	$p^o_{+1} = 0.773$	$p^o_{+1} = 0.227$	1.000

Kreuztabelle: absolute Häufigkeiten

		Erwerbsstatus		
		Vollzeit	Teilzeit	Summe
Geschlecht	Männer	$h^o_{11} = 1{,}346$ $h^e_{11} = 1{,}086$	$h^o_{12} = 59$ $h^e_{12} = 319$	$h^o_{1+} = 1{,}405$
	Frauen	$h^o_{21} = 695$ $h^e_{21} = 955$	$h^o_{21} = 540$ $h^e_{22} = 280$	$h^o_{2+} = 1{,}235$
	Summe	$h^o_{+1} = 2{,}041$	$h^o_{+1} = 599$	2,640

Maßzahlen

Das **Assoziationsmaß Chi-Quadrat** χ^2_{err} (auch: Pearson's χ^2) mit

$$\chi^2_{err} = \sum_i \sum_j \frac{(h^o_{ij} - h^e_{ij})^2}{h^e_{ij}}$$

misst den Zusammenhang zwischen zwei diskreten Merkmalen.

Da das Assoziationsmaß χ^2_{err} mit dem Stichprobenumfang steigt, bietet sich das **Cramersche Assoziationsmaß V** an:

$$V = \sqrt{\frac{\chi^2_{err}}{n \cdot (min(r,s) - 1)}}$$

r gibt die Anzahl der Merkmalsausprägungen des Merkmals x an.
s gibt die Anzahl der Merkmalsausprägungen des Merkmals y an.
Es gilt $0 \leq V \leq 1$

Maßzahlen: Beispiel

$$\chi^2_{err} = \sum_i \sum_j \frac{(h^o_{ij} - h^e_{ij})^2}{h^e_{ij}} = \frac{(1{,}346-1{,}086)^2}{1{,}086} + \frac{(59-319)^2}{319} + \frac{695-955)^2}{955} + \frac{(540-280)^2}{280} \approx 585$$

$$V = \sqrt{\frac{\chi^2}{n \cdot (min(r,s)-1)}} = \sqrt{\frac{585}{2{,}640 \cdot (min(2,2)-1)}} = \sqrt{\frac{585}{2{,}640}} \approx 0.471$$

Interpretationshilfe für das Cramersche Assoziationsmaß V:

$V = 0$	kein Zusammenhang
$0 < V \leq 0.3$	schwacher Zusammenhang
$0.3 < V \leq 0.7$	mittlerer Zusammenhang
$0.7 < V < 1$	starker Zusammenhang
$V = 1$	vollständiger Zusammenhang

Zusammenhang oder Zufall? χ^2-Test auf Unabhängigkeit

Nullhypothese (H_0): Es gibt **keinen Zusammenhang** zwischen Geschlecht und Erwerbsstatus **in der Grundgesamtheit**
Alternativhypothese (H_1): Es gibt **einen Zusammenhang** zwischen Geschlecht und Erwerbsstatus **in der Grundgesamtheit**

Intuition:
Ein Zusammenhang in der Grundgesamtheit ist dann wahrscheinlich, wenn der Zusammenhang in der Stichprobe groß ist (V liegt deutlich über 0) und wenn der Stichprobenumfang groß ist (kleine Unsicherheit).

Teststrategie zielt auf χ^2_{err} ab:
H_0 wird nicht verworfen, wenn gilt:

$$\chi^2_{(r-1)(s-1);1-\alpha} \geq \chi^2_{err} = \sum_i \sum_j \frac{(h^o_{ij} - h^e_{ij})^2}{h^e_{ij}},$$

H_0 wird mit einer Irrtumswahrscheinlichkeit α verworfen, wenn gilt:

$$\chi^2_{(r-1)(s-1);1-\alpha} < \chi^2_{err} = \sum_i \sum_j \frac{(h^o_{ij} - h^e_{ij})^2}{h^e_{ij}},$$

wobei $(r-1) \cdot (s-1)$ die Zahl der Freiheitsgrade bezeichnet.

χ^2-Test: Beispiel

Teststrategie zielt auf χ^2_{err} ab:
H_0 wird nicht verworfen, wenn gilt:

$$\chi^2_{err} = \sum_i \sum_j \frac{(h^o_{ij} - h^e_{ij})^2}{h^e_{ij}} \leq \chi^2_{(r-1)(s-1);1-\alpha}$$

H_0 wird mit einer Irrtumswahrscheinlichkeit α verworfen, wenn gilt:

$$\chi^2_{err} = \sum_i \sum_j \frac{(h^o_{ij} - h^e_{ij})^2}{h^e_{ij}} > \chi^2_{(r-1)(s-1);1-\alpha}$$

$\rightarrow$ auch hier wird eine standardisierte Teststatistik mit einem kritischen Wert der Verteilung verglichen!

Beispiel:

$$\chi^2_{err} = 585 > 3.84 = \chi^2_{1;0.95} = \chi^2_{(2-1)(2-1);1-0.05}$$

$\rightarrow$ **Die Nullhypothese** wird sehr deutlich **verworfen**: In der Grundgesamtheit gibt es (höchstwahrscheinlich) einen Zusammenhang zwischen Geschlecht und Erwerbsstatus.

Umsetzung in EXCEL

1. Erstellung der **Kreuztabelle der beobachteten** absoluten **Häufigkeiten**:
   - Hier bietet sich der **EXCEL-Befehl** ZÄHLENWENNS (S am Ende beachten!) an, wo die Zahl an Beobachtungen gezählt wird, die mehrere Bedingungen erfüllt (etwa *Männer* und *Vollzeitbeschäftigt*)
   - Nachdem alle absoluten Häufigkeiten h_{ij}^{o} erstellt wurden, können die Randverteilungen ermittelt werden.
2. Erstellung der **Kreuztabelle der erwarteten** absoluten **Häufigkeiten**:
   - Mit den Randverteilungen können die erwarteten absoluten Häufigkeiten h_{ij}^{e} berechnet werden, und in eine neue Kreuztabelle eingetragen werden.
3. Die einfachste Möglichkeit besteht darin, mit dem **EXCEL-Befehl** CHIQU.TEST den p-Wert von χ^2_{err} zu bestimmen. Dazu muss χ^2_{err} gar nicht berechnet werden, sondern es ist ausreichend, die beobachteten und die erwarteten Häufigkeiten zu markieren. Ist dieser **p-Wert kleiner als** die angenommene Irrtumswahrscheinlichkeit α, dann wird die **Nullhypothese** H_0 **verworfen**.

Alternativ kann χ^2_{err} berechnet werden und mit dem kritischen $\chi^2_{(r-1)(s-1);1-\alpha}$-Wert verglichen werden. Den kritischen Wert erhält man mit **EXCEL-Befehl** CHIQU.INV$(1-\alpha;\ (r-1)(s-1))$.

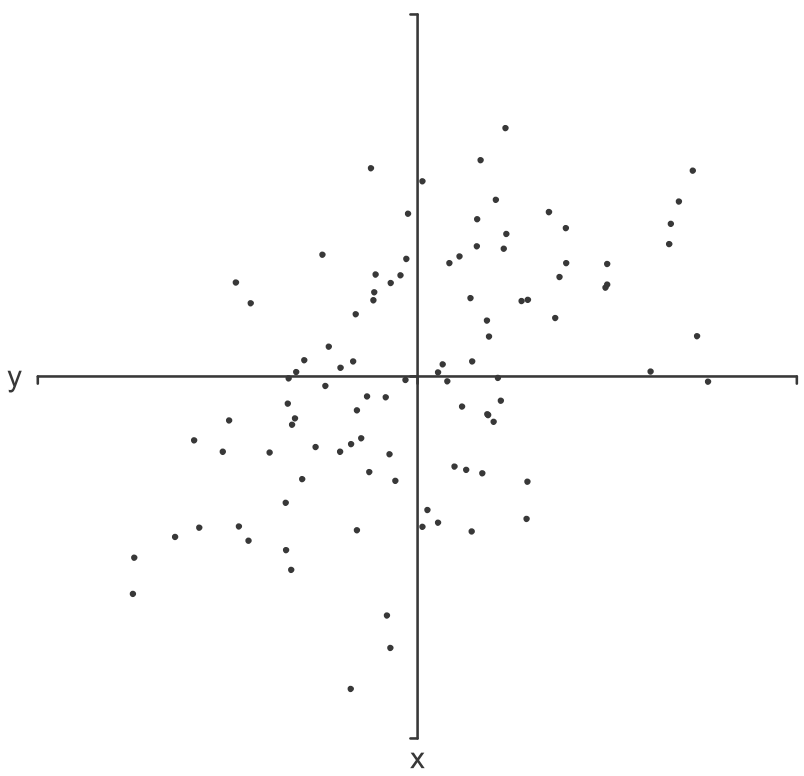
Zusammenhang zwischen zwei metrischen Variablen: Korrelation

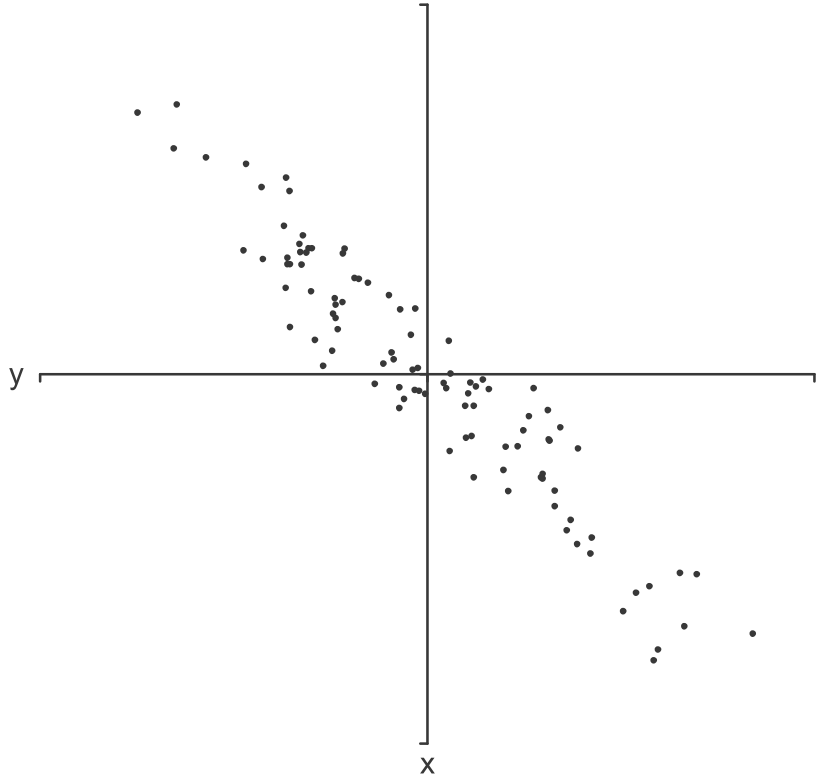
		Variable 2	
		diskret	stetig
Variable 1	diskret	Kreuztabelle	Mittelwertvergleich
	stetig	Mittelwertvergleich	Korrelation

Der **(Bravais-Pearson-)Korrelationskoeffizient** ρ gibt an, (i) ob der Zusammenhang zwischen zwei metrischen Variablen x und y **positiv oder negativ** ist, und (ii) wie ähnlich dieser Zusammenhang einem linearen Zusammenhang ist.

Grafische Darstellung: Streudiagramm

Streudiagramm: Ein Streudiagramm ist eine grafische Darstellung eines zweidimensionalen metrischen Merkmals. Dabei wird jeder Erhebungseinheit der zugehörige Datenpunkt in einem Koordinatensystem zugeordnet. Streudiagramme erleichtern das Auffinden von Zusammenhängen.





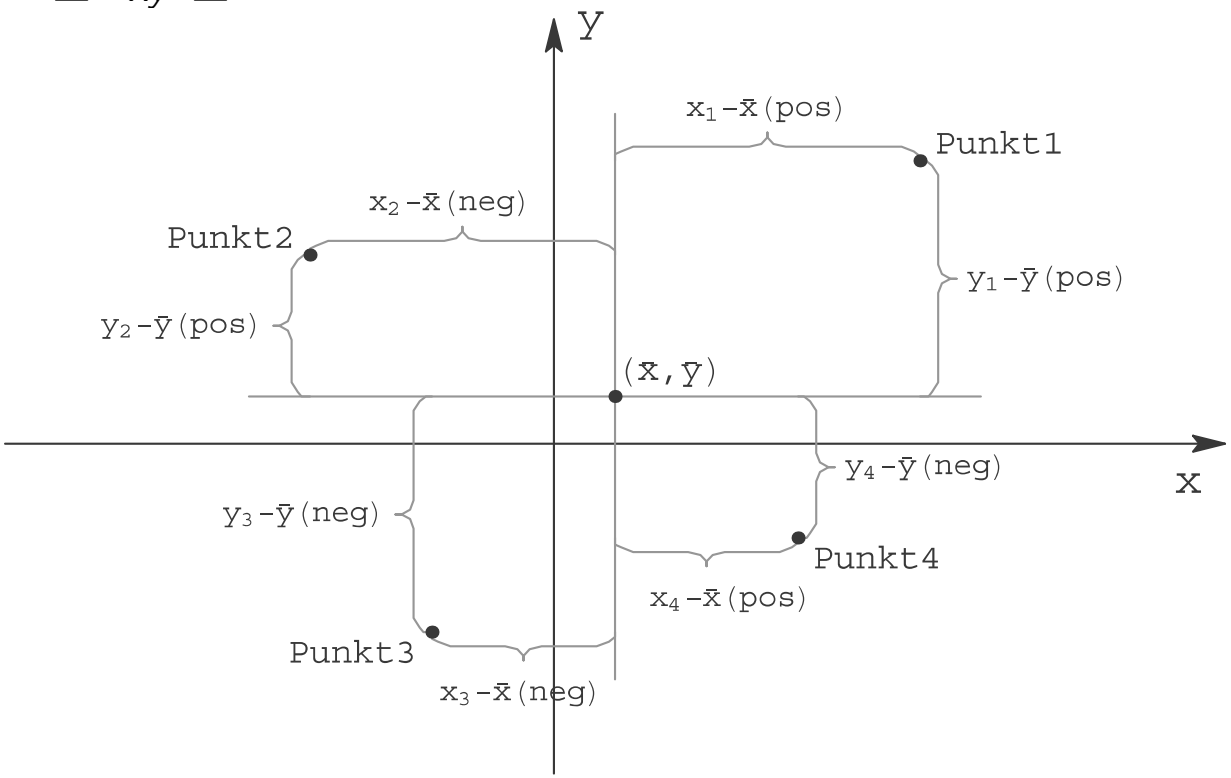
(siehe Duller, Abbildung 7.2)

Kovarianz als Ausgangspunkt

Die **Kovarianz** zu den Merkmalen x und y einer Stichprobe ist gegeben durch:

$$\hat{s}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i \cdot y_i - \overline{x} \cdot \overline{y})$$

Wobei gilt: $-\infty \leq \hat{s}_{xy} \leq \infty$



(siehe Duller, Abbildung 7.1)

Korrelationskoeffizient: Berechnung

Der **Korrelationskoeffizient** stellt ein standardisiertes Maß zur Messung eines linearen Zusammenhangs zwischen zwei metrischen Merkmalen x und y dar:

$$\rho = \frac{\hat{s}_{xy}}{\hat{s}_x \cdot \hat{s}_y}$$

wobei $\hat{s}_x$ und $\hat{s}_y$ die Standardabweichung der Merkmale x und y darstellt.
Es gilt: $-1 \leq \rho \leq 1$

Anmerkungen:

- Es ist auch möglich, ein Konfidenzintervall von ρ zu berechnen und zu beurteilen, ob ρ statistisch signifikant von 0 verschieden ist. Da die statistische Signifikanz selten ausgewiesen wird und relativ aufwendig zu berechnen ist, wird dieses Thema nicht besprochen.
- Wenn die Grundgesamtheit (statt einer Stichprobe) vorliegt, ist die Korrektur (indem durch $n-1$ statt n dividiert wird) bei Berechnung von Kovarianz und Standardabweichung nicht notwendig. In der Praxis ist der Unterschied aber meist gering und von geringer Relevanz.
- Bei der Berechnung des Korrelationskoeffizienten ρ ist es ohnehin unerheblich, ob diese Korrektur vorgenommen wird.

Korrelationskoeffizient: Interpretation

Interpretation:
$\rho > 0$ gleichsinniger (positiver) linearer Zusammenhang
$\rho = 0$ kein linearer Zusammenhang
$\rho < 0$ gegensinniger (negativer) linearer Zusammenhang

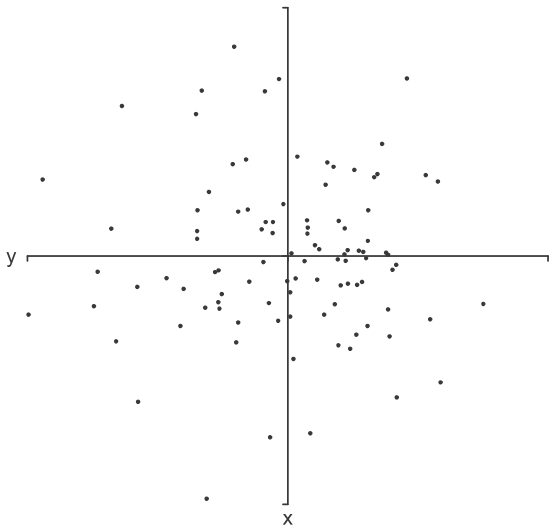
Die Richtung des Zusammenhanges ergibt sich aus dem **Vorzeichen**.

Interpretationshilfe für den Korrelationskoeffizienten:

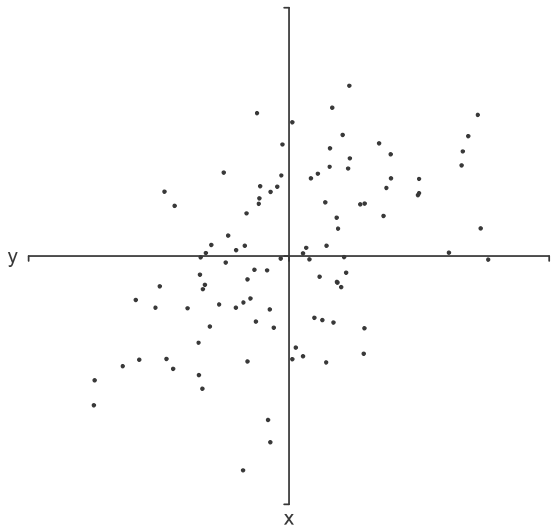
$\rho = 0$	kein Zusammenhang
$0 < \lvert\rho\rvert \leq 0.3$	schwacher Zusammenhang
$0.3 < \lvert\rho\rvert \leq 0.7$	mittlerer Zusammenhang
$0.7 < \lvert\rho\rvert < 1$	starker Zusammenhang
$\lvert\rho\rvert = 1$	vollständiger Zusammenhang

Anmerkung: ρ^2 entspricht dem Bestimmtheitsmaß (R^2) bei einer linearen Einfachregression. Das bedeutet, dass ein Anteil ρ^2 der Variation eines Merkmals durch die Variation des zweiten Merkmals erklärt werden kann. Wir werden später genauer darauf zurückkommen.

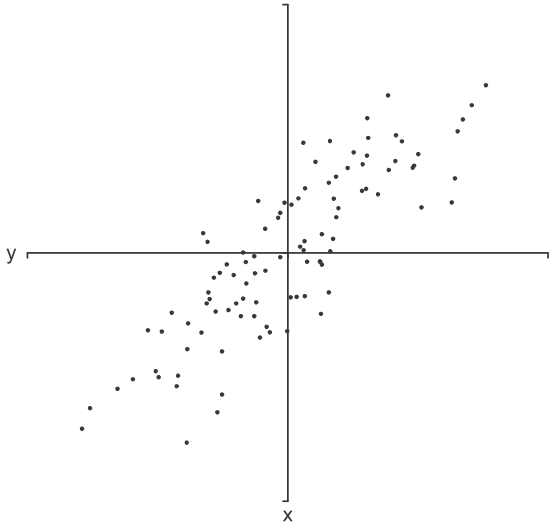
Korrelationskoeffizient: Beispiele



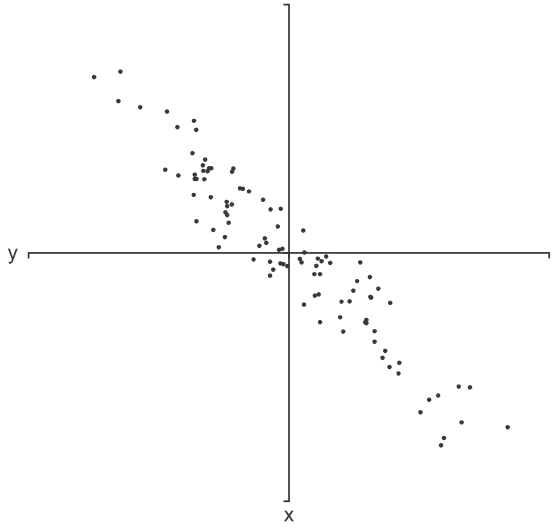
Korrelation $\rho = 0$



Korrelation $\rho = 0,5$



Korrelation $\rho = 0,85$



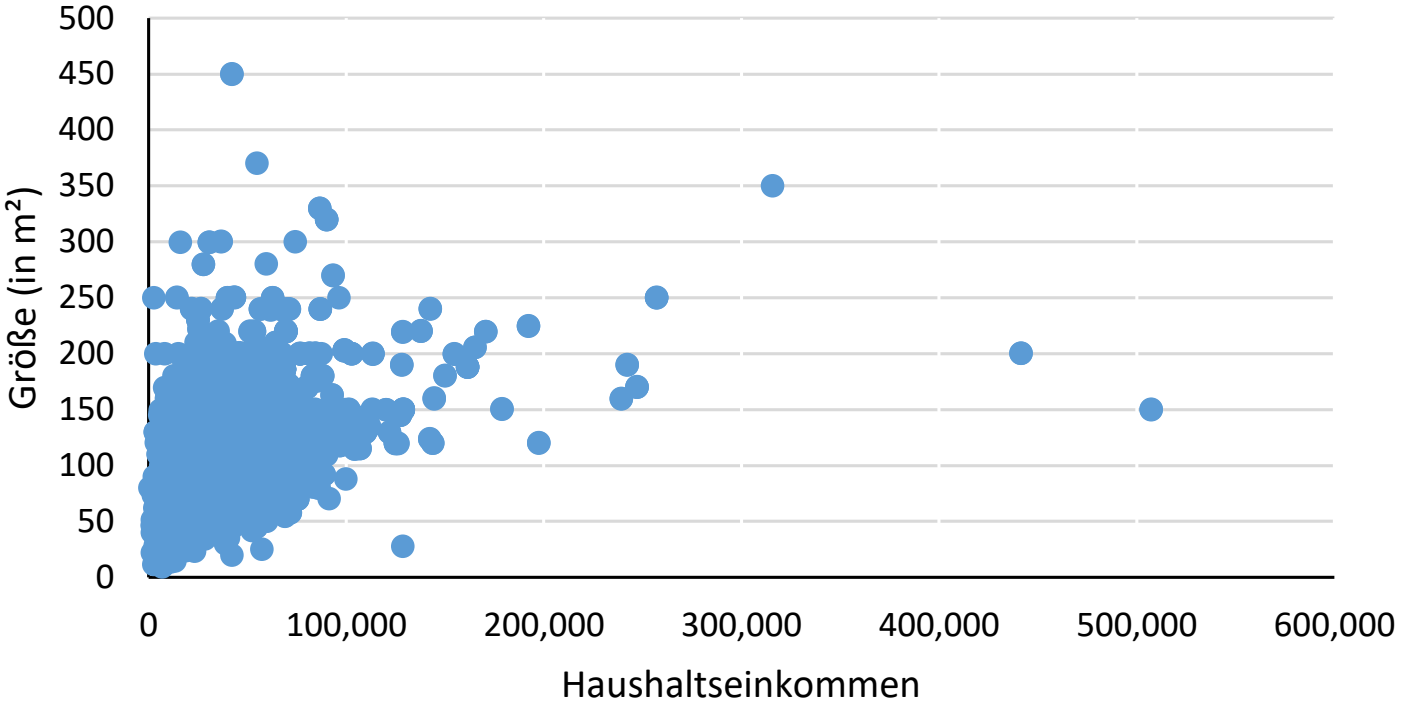
Korrelation $\rho = -0,95$

(siehe Duller, Abbildung 7.2)

Umsetzung in EXCEL (1)

Streudiagramm:

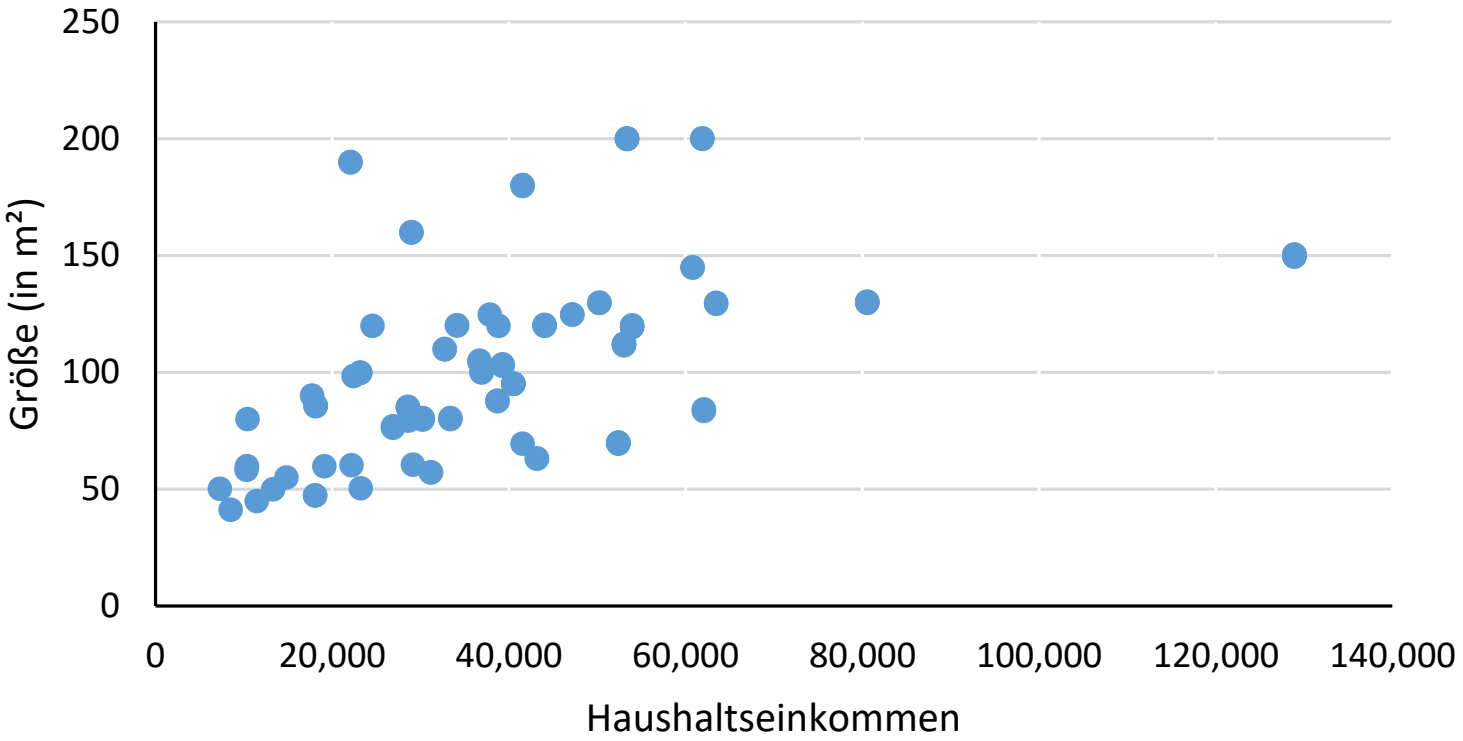
Mit **Einfügen** → **Diagramme** → **Punkt (XY)** kann ein **Streudiagramm** eingefügt werden. Streudiagramme sind oft unübersichtlich. Das kann damit zusammenhängen, (i) dass es **Ausreißer** gibt, und/oder (ii) dass der Stichprobenumfang n sehr groß ist (und es daher sehr viele Punkte im Streudiagramm gibt). Man kann auch nur eine (Zufalls-) Auswahl der Stichprobe für das Streudiagramm verwenden.



Umsetzung in EXCEL (2)

Streudiagramm:

Mit **Einfügen** → **Diagramme** → **Punkt (XY)** kann ein **Streudiagramm** eingefügt werden. Streudiagramm sind oft unübersichtlich. Das kann damit zusammenhängen, (i) dass es **Ausreißer** gibt, oder (ii) dass der Stichprobenumfang n sehr groß ist (und es daher sehr viele Punkte im Streudiagramm gibt). Man kann auch nur eine (Zufalls-) Auswahl der Stichprobe für das Streudiagramm verwenden.



Umsetzung in EXCEL (3)

- **Kovarianz:**
  Mit den **Excel-Befehlen** KOVARIANZ.S und KOVARIANZ.P kann die **Kovarianz** zwischen zwei Merkmalen einer Stichprobe (".S") bzw. der Grundgesamtheit (".P") berechnet werden.
- **Korrelation:**
  Mit den **Excel-Befehl** KORREL kann der **Korrelationskoeffizient** von zwei Merkmalen berechnet werden.
  - Hier ist eine Unterscheidung in Stichprobe und Grundgesamtheit nicht notwendig, da sich $1/(n-1)$ bzw. $1/N$ wegkürzt.
- Alternativ kann **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Kovarianz** bzw. **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Korrelation** verwendet werden.
  - Vorteil: Gibt eine formatierte Tabelle zurück. Es können auch Kovarianzen bzw. Korrelationskoeffizienten von mehreren Merkmalen (bzw.: von mehreren Merkmals-Paaren) berechnet werden.
  - Nachteil: "Eingabebereich muss ein zusammenhängender Bezug sein" (d.h. die Variablen müssen nebeneinander stehen).

Umsetzung in EXCEL (4)

Varianz-Kovarianz-Matrix:

	persnr	*ybirth*	*income*	*hhinc*	*size*	*Größe (in m2)*
persnr	9, 588, 648, 723, 916					
ybirth	−11, 977, 831	329				
income	206, 952, 278	52, 743	1, 400, 149, 494			
hhinc	−4, 608, 430, 303	74, 305	479, 177, 712	714, 252, 359		
size	−99, 474, 892	680	3, 073, 259	5, 895, 795	211, 692	
Größe (in m2)	−9, 241, 443	63	285, 513	547, 733	19, 667	1, 827

Korrelationsmatrix:

	persnr	*ybirth*	*income*	*hhinc*	*size*	*Größe (in m2)*
persnr	1					
ybirth	−0.213	1				
income	0.002	0.081	1			
hhinc	−0.056	0.153	0.483	1		
size	−0.070	0.081	0.178	0.479	1	
Größe (in m2)	−0.070	0.081	0.178	0.479	1	1