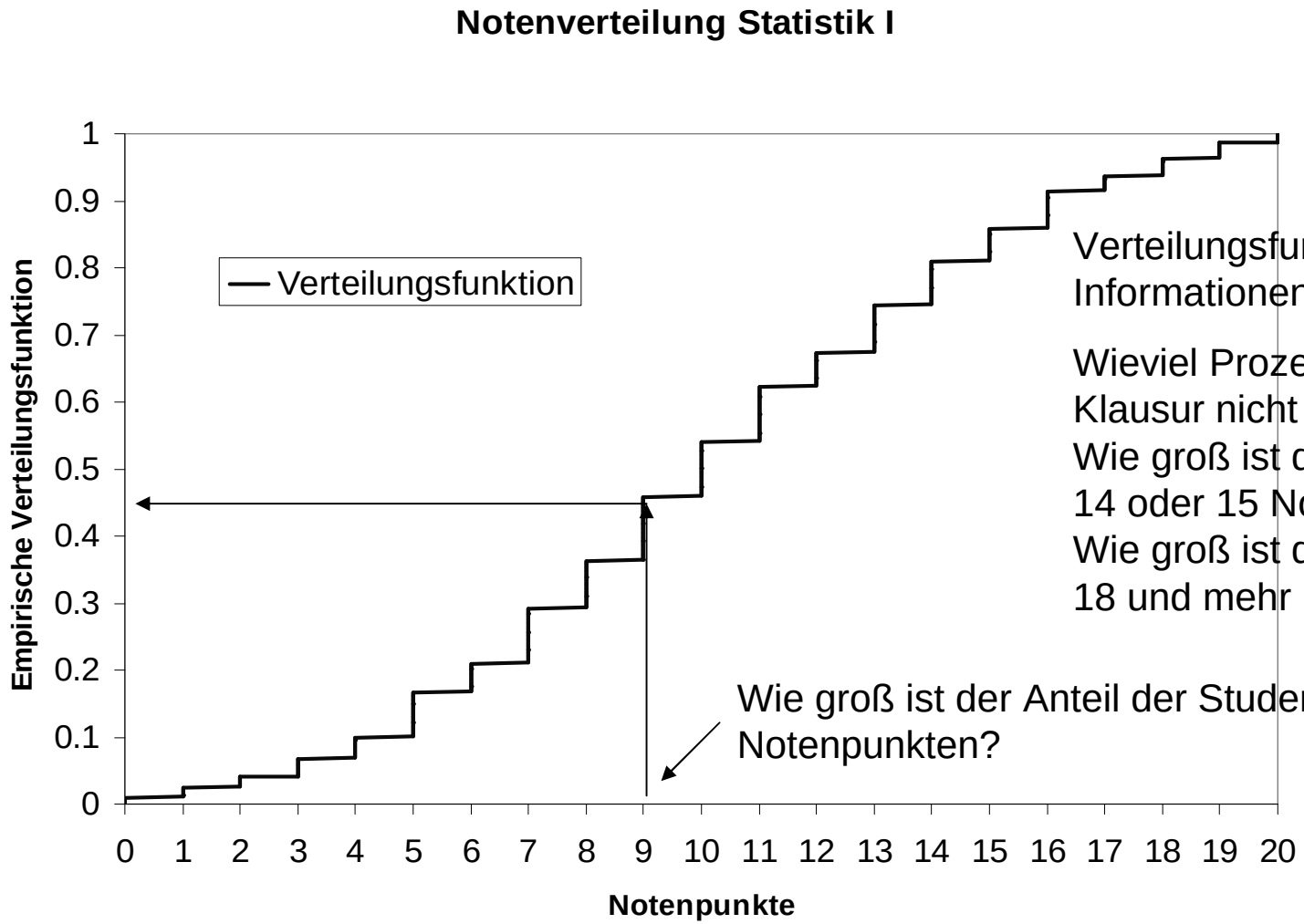
Explorative Datenanalyse bis Dezember in Schnelldurchlauf

Explorative Datenanalyse Teil 1:
Univariate Datensätze

Mit der empirischen Verteilungsfunktion können alle relevanten Informationen aus einem univariaten Datensatz geliefert werden

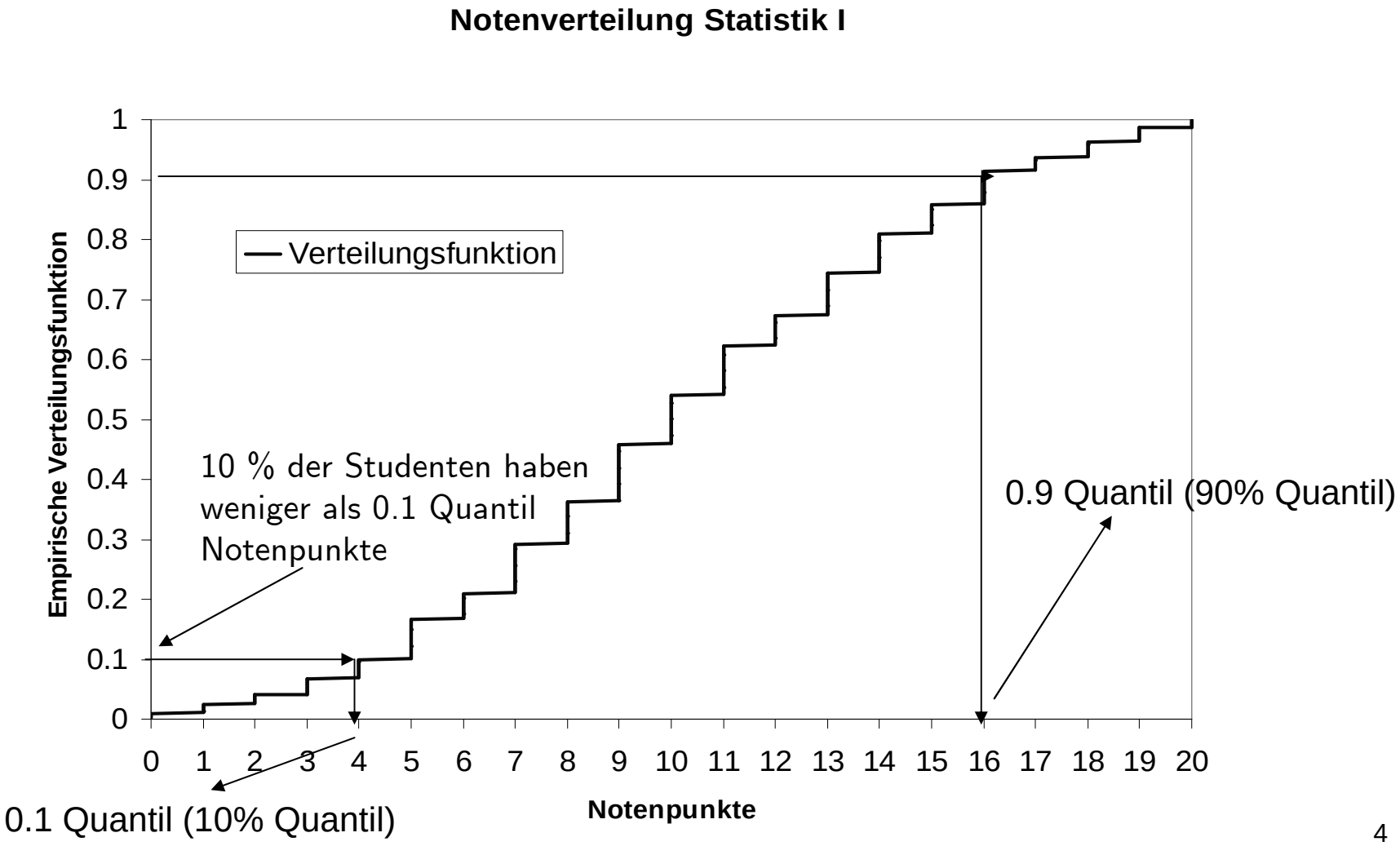


Verteilungsfunktion enthält alle relevanten Informationen

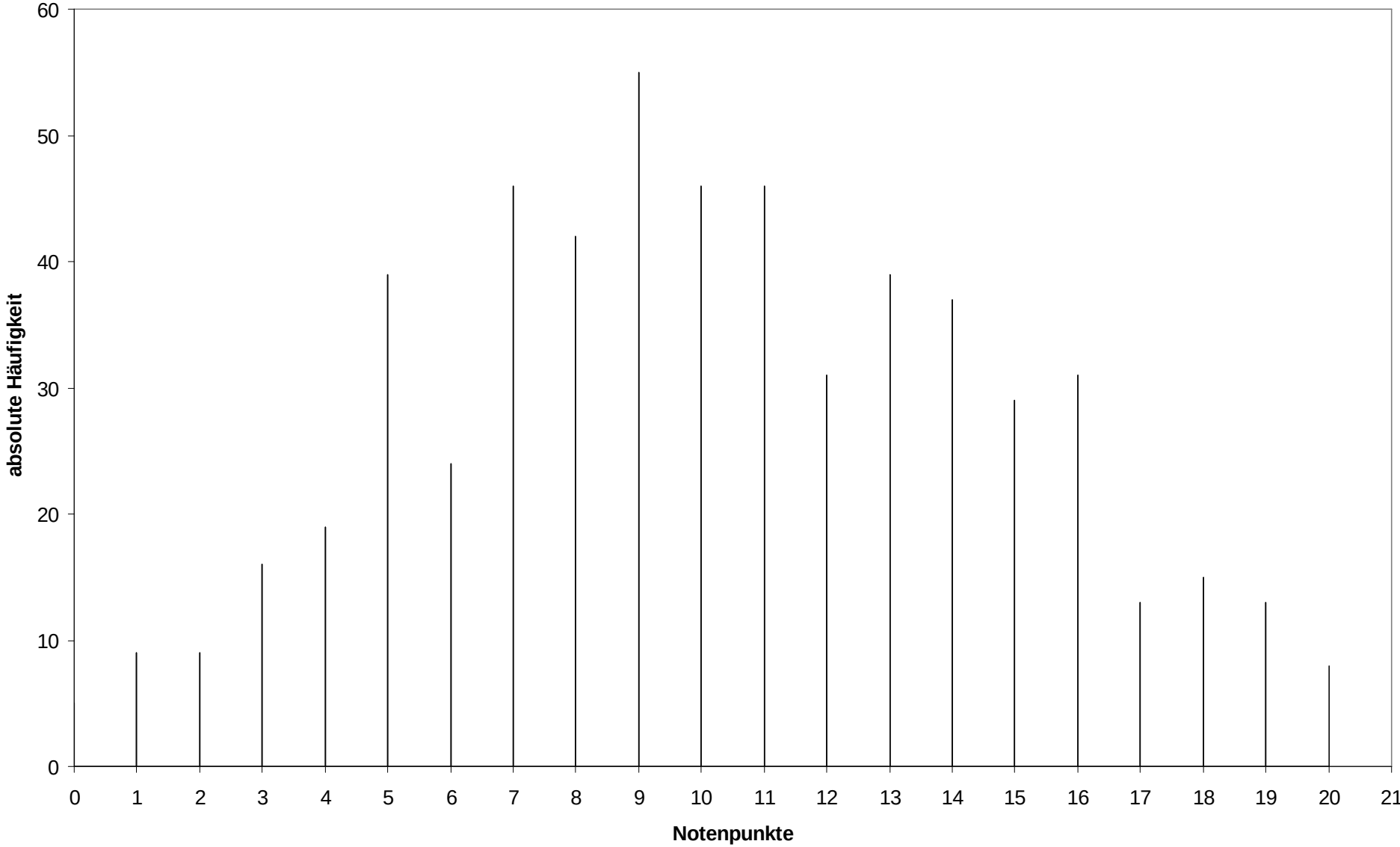
Wieviel Prozent der Studenten hat die Klausur nicht bestanden?
Wie groß ist der Anteil der Studenten mit 14 oder 15 Notenpunkten?
Wie groß ist der Anteil der Studenten mit 18 und mehr Punkten?

Wie groß ist der Anteil der Studenten mit 9 oder weniger Notenpunkten?

Die empirische Quantilsfunktion, die Umkehrfunktion der empirischen Verteilungsfunktion, ist wichtig für die Value-at-Risk Schätzung.

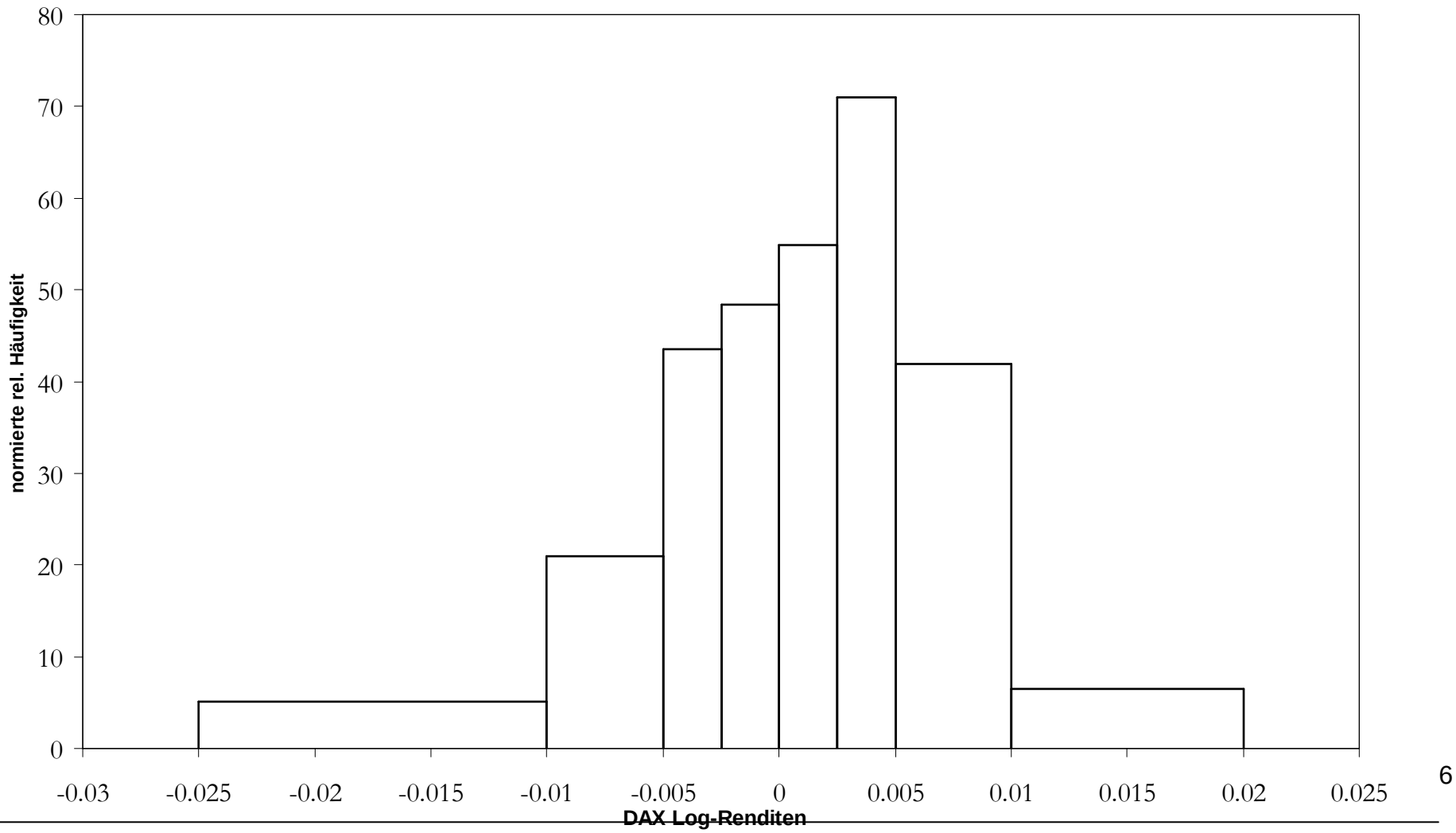


Auch das Stabdiagramm erhält alle Informationen der Daten. Sie verwenden es bei diskreten Daten

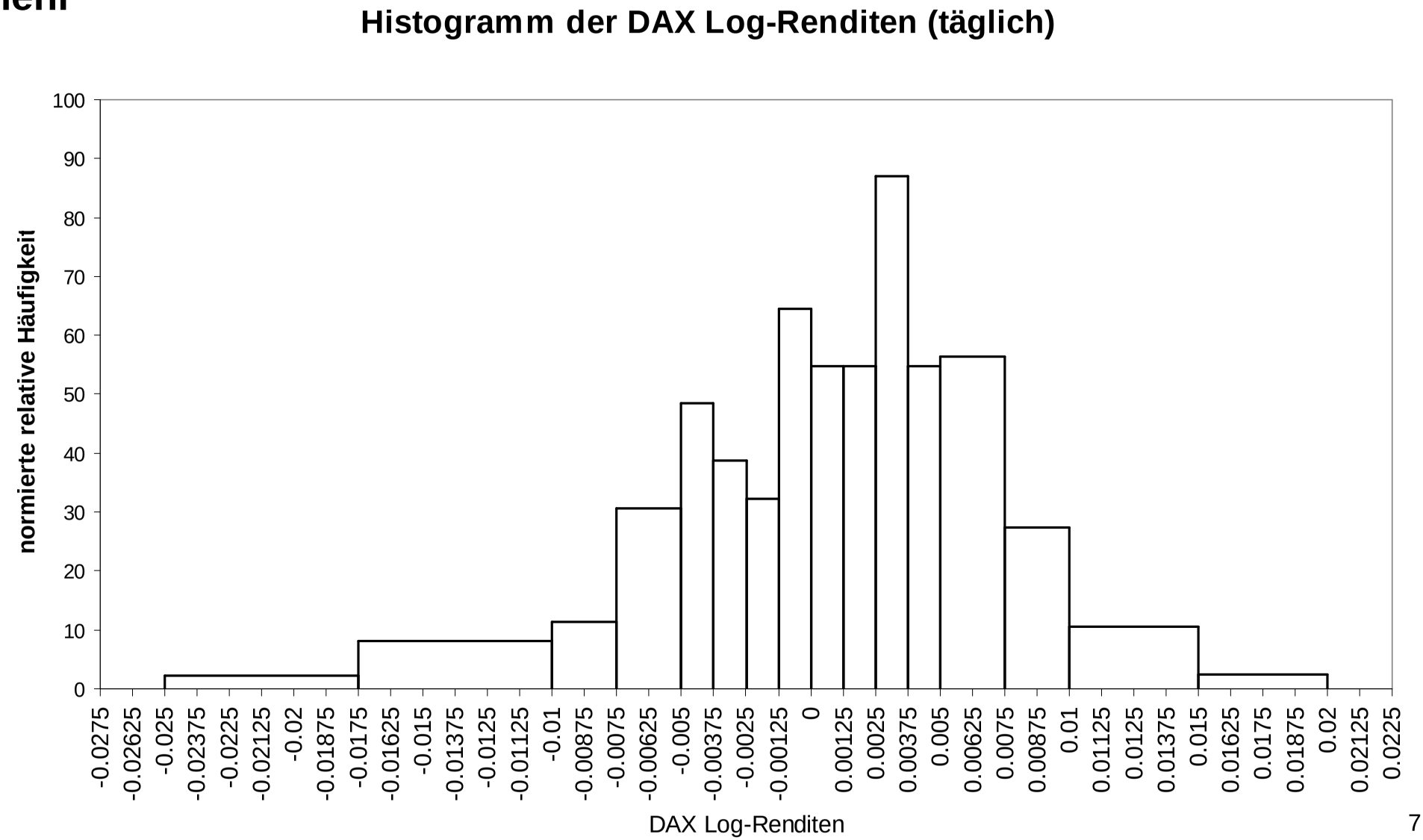


Das Histogramm verdeutlicht, auf welche Werte sich die Daten konzentrieren. Im Vergleich zur Verteilungsfunktion gehen aber Informationen verloren.

Histogramm der DAX Log-Renditen (täglich)



Mit der Wahl der Klassenbreiten verändern Sie das Aussehen des Histogramms. Eine rein „deskriptive“ Datenauswertung ist dies schon nicht mehr



Lageparameter oder Mittelwerte sind Werte, um die sich Daten gruppieren (Lage der Verteilung)

Arithmetisches Mittel

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{v=1}^{n} x_v$$

$$\overline{x} = \sum_{i=1}^{K} x_i h_i \text{ mit } h_i = \frac{n(X{=}x_i)}{n} \longrightarrow$$ Wenn X in K diskreten Ausprägungen auftritt

$\sum_{v=1}^{n}(x_v - a)^2$ Die Summe der quadrierten Abweichungen von einem Wert a wird minimal für

$a = \overline{x}$

Median $\tilde{x}$ ist das p=0.5 Quantil: Ordne die Daten der Größe nach an. Der Median ist der Wert in der Mitte, der „Zentralwert“

$\sum_{v=1}^{n} |x_v - a|$ Die Summe der absoluten Abweichungen von einem Wert a wird minimal für

$a = \tilde{x}$ Und dann gibt es noch den **Modus**, den häufigsten Wert, bzw. die modale Klasse

Einige wichtige Eigenschaften des arithmetischen Mittels sollten Sie sich stets vor Augen halten

Das arithmetische Mittel kann, muß aber nicht in den Daten vorkommen (Füße im Ofen, Kopf im Kühlschrank)

Das arithmetische Mittel bestimmt das Zentrum der Masseverteilung der Merkmalswerte

Ausreißer ziehen das arithmetische Mittel an!

Auch die empirische Varianz und die Schiefe- und Wölbungsmaße (empirische Momente) sind arithmetische Mittel

Jede Beobachtung geht mit dem gleichen Gewicht in die Berechnung ein (anders beim gewogenen arithmetischen Mittel

Aus dem Vergleich von arithmetischem Mittel und Median kann auf die Schiefe (rechts- oder links-schief) der Verteilung geschlossen werden

Bei rechts-schiefen Daten (Einkommensverteilung) ist das arithmetische Mittel stets größer als der Median (damit kann man Politik machen)

Das arithmetische Mittel sollte nur für metrisch skalierte Daten berechnet werden (Noten?)

Der Median als alternativer Lageparameter hat einige Vorteile gegenüber dem arithmetischen Mittel

Der Median ist nicht ausreißerempfindlich

Der Medianwert kommt in den Daten vor

Der Median liegt stets „in der Mitte der Daten“ (die Hälfte der Merkmalswerte sind größer/gleich, die andere Hälfte kleiner/gleich dem Median). Das muß beim arithmetischen Mittel nicht so sein.

Der Median kann auch für ordinalskalierte Daten berechnet werden und macht da auch Sinn

Mit dem geometrischen Mittel ermitteln wir durchschnittliche Wachstumsraten

Geometrisches Mittel

$$\sqrt[n]{(x_1 \cdot x_2 \cdot \ldots \cdot \ldots x_n)} =$$

Das arithmetische Mittel der logarithmierten Originaldaten

$$\exp\left(\frac{1}{n}\left(\ln x_1 + \ln x_2 + \ldots + \ln x_n\right)\right) = \exp\left(\frac{1}{n}\sum_{v=1}^{n} \ln x_v\right)$$

mit $x_1, x_2, \ldots$ Bruttowachstumsraten

Diese Berechnung ergibt die durchschnittliche Bruttowachstumsrate.

Wird die Nettowachstumsrate gesucht, ist noch die 1 abzuziehen.

Vorsicht bei der Ermittlung von arithmetischen Mitteln von Verhältniszahlen

Achtung, wenn der Zähler der zu mittelnden Verhältniszahl über die Beobachtungen nicht variiert!

Bsp. Geschwindigkeit (km/Std.) in 100 km Etappen
Effizienz (Stunden/produziertes Stück) an verschiedenen Tagen

dann:

Verwendung des harmonischen Mittels

oder

Nachdenken (Durchschnittsgeschwindigkeit)

oder

Umdrehen der Verhältniszahl und Anwendung des arithmetischen Mittels

Streungsparameter geben an, wie eng oder weit die Daten um die Lageparameter streuen

Empirische Varianz

$$s_X^2 = \frac{1}{n} \sum_{v=1}^{n} (x_v - \overline{x})^2$$

Die empirische Varianz ist ein arithmetisches Mittel!
(der quadrierten Abweichungen vom arithmetischen Mittel der Originaldaten)
und hat damit alle (negativen) Eigenschaften desselben

Empirische Standardabweichung

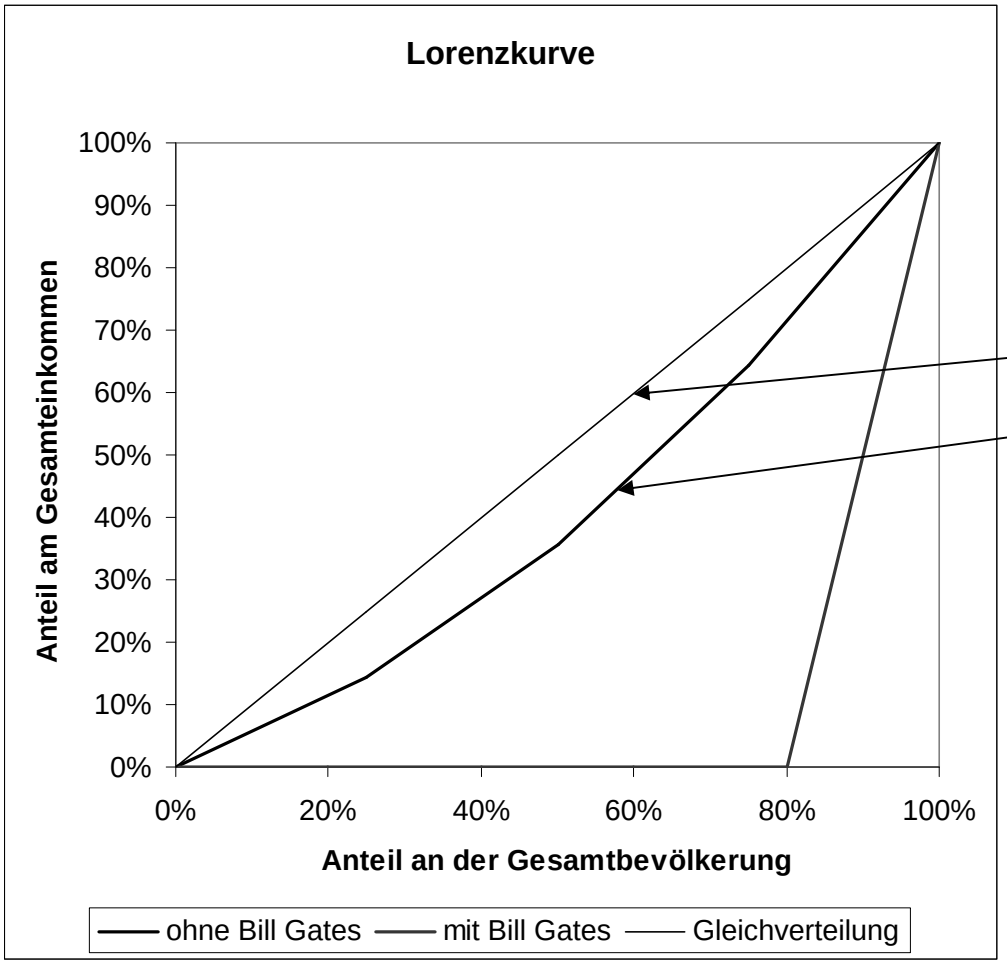
$$s_X = \sqrt{s_X^2}$$

Vorteil: Gleiche Dimension wie die Daten

Wenn X in K diskreten Ausprägungen auftritt, können Sie die empirische Varianz auch wie folgt berechnen:

$$s_X^2 = \sum_{i=1}^{K} (x_i - \overline{x})^2 \cdot h_i$$

Lorenzkurve und Gini-Koeffizient machen deutlich, wie stark die Daten auf einzelne Merkmalsträger konzentriert sind



Der Gini-Koeffizient ist das

Verhältnis der

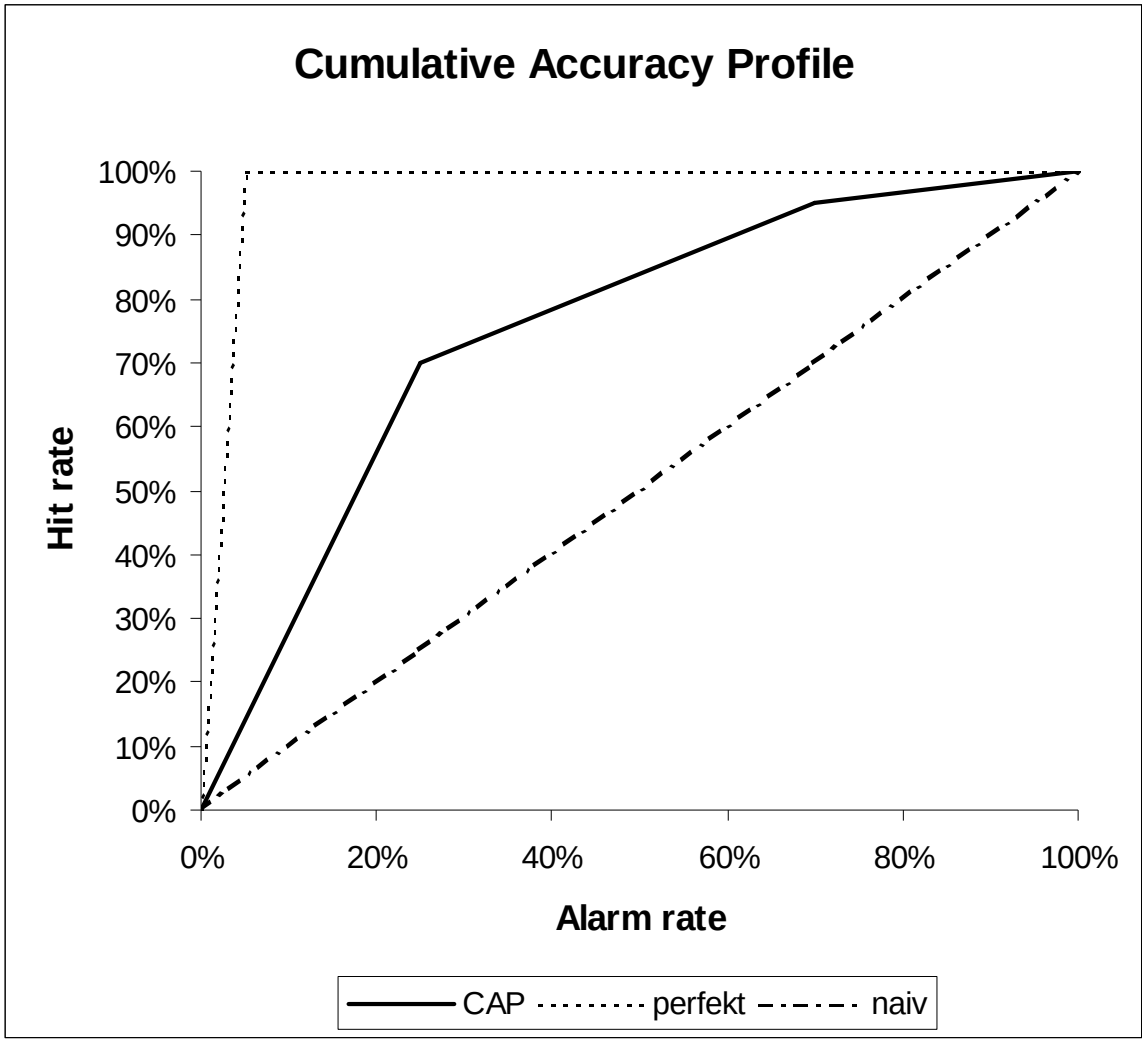
Fläche zwischen Gleichverteilungsgerade und Lorenzkurve und

der

Fläche zwischen Gleichverteilungsgerade und

Lorenzkurve bei maximaler Konzentration

Unter dem Namen CAP Kurve feiert die Lorenzkurve ein Comeback im Bereich der Kreditrisikomessung



Explorative Datenanalyse Teil 2:
Multivariate (bivariate) Datensätze und Regresssionsanalyse (K-Q-Methode)

Kontingenztabellen erlauben die detaillierteste Analyse eines bivariaten Datensatzes

Randhäufigkeiten (relative)

Gemeinsame relative Häufigkeiten

	h_ij	Y						
X		Lohnsteuer	Einkommensteuer	Umsatzsteuer	sonst.	h_i.	Y=hinterzogene Steuerart	
	Handelsb.	0.017	0.113	0.078	0.078	**0.287**	1=Lohnsteuer	
	Freie Ber.	0.226	0.096	0.130	0.078	**0.530**	2=Einkommensteuer	
	Fertigungsbetrieb	0.061	0.052	0.043	0.026	**0.183**	3=Umsatzsteuer	
	h_.j	**0.304**	**0.261**	**0.252**	**0.183**	**1.000**	4=Sonstiges	

Bedingte relative Häufigkeiten

Weichen bei Abhängigkeit voneinander ab

	h(x_i|y_j)	Y				
X		Lohnsteuer	Einkommensteuer	Umsatzsteuer	sonst.	
	Handelsb.	0.057	0.433	0.310	0.429	**0.287**
	Freie Ber.	0.743	0.367	0.517	0.429	**0.530**
	Fertigungsbetrieb	0.200	0.200	0.172	0.143	**0.183**
		1.000	1.000	1.000	1.000	**1.000**

	h(y_j|x_i)	Y			
X		Lohnsteuer	Einkommensteuer	Umsatzsteuer	sonst.
	Handelsb.	0.061	0.394	0.273	0.273
	Freie Ber.	0.426	0.180	0.246	0.148
	Fertigungsbetrieb	0.333	0.286	0.238	0.143

Nur wenn die statistischen Variablen unabhängig sind, kann von den Randverteilungen auf die gemeinsame Häufigkeit geschlossen werden

h_ij	zahlungsfähig	nicht zahlungsfähig		h_i.
GmbH	0.224	0.011		**0.235**
OHG	0.056	0.003		**0.059**
KG	0.112	0.006		**0.118**
Einzelunternehmer	0.560	0.028		**0.588**
h_.j	**0.952**	**0.048**		

h(x_i|y_j)	zahlungsfähig	nicht zahlungsfähig
GmbH	0.235	0.235
OHG	0.059	0.059
KG	0.118	0.118
Einzelunterr	0.588	0.588

Bei Unabhängigkeit der beiden Merkmale kann aus den Randverteilungen auf die gemeinsame Verteilung geschlossen werden

Bei Unabhängigkeit entsprechen die Randverteilungen den bedingten Verteilungen

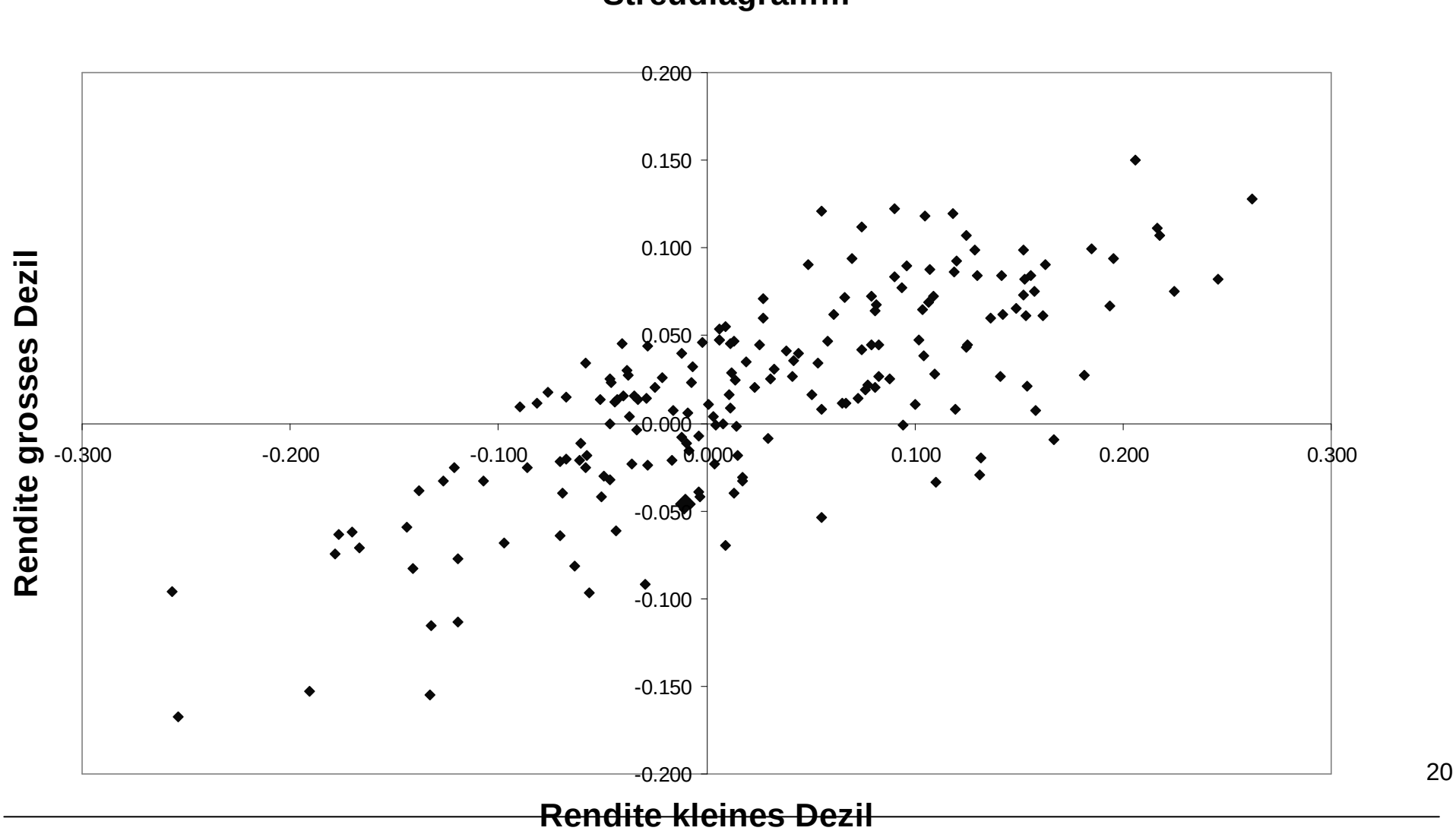
h(y_j|x_i)	zahlungsfähig	nicht zahlungsfähig
GmbH	0.952	0.048
OHG	0.952	0.048
KG	0.952	0.048

Kontingenztabellen erlauben die detaillierteste Analyse eines bivariaten Datensatzes, bei metrischen Daten verwischt die Klassenbildung aber Informationen

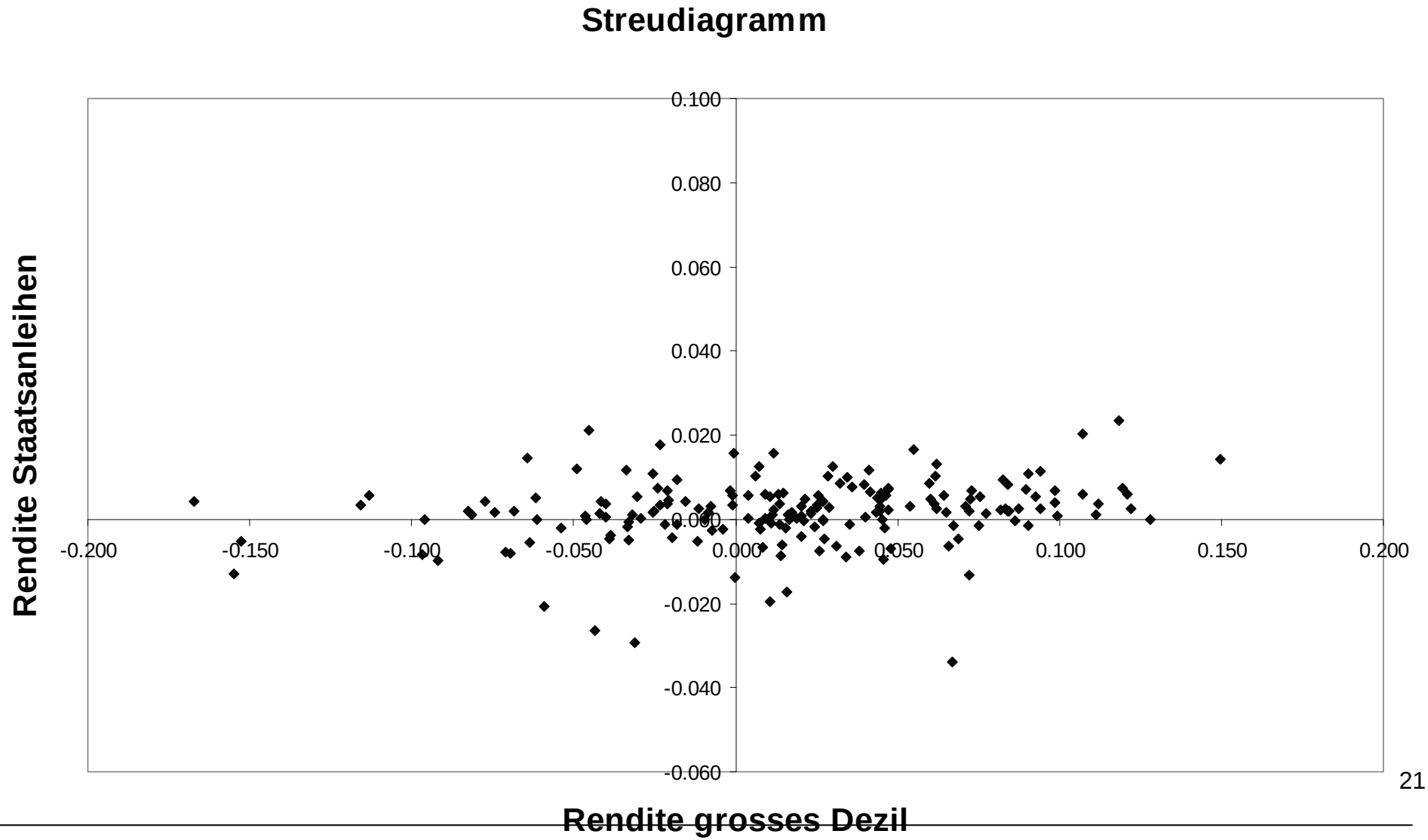
Die Klasseneinteilung mag sinnvoll sein, ist aber willkürlich

		log Rendite grösstes Dezil						
		<-0.10	-0.10 bis unter -0.05	-0.05 bis unter 0	0 bis unter 0.05	0.05 bis unter 0.10	> 0.10	**Zeilensumme**
log Rendite kleinstes Dezil	<-0.10	5	8	4	0	0	0	17
	-0.10 bis unter -0.05	0	4	9	6	0	0	19
	-0.05 bis unter 0	0	2	18	21	0	0	41
	0 bis unter 0.05	0	1	9	19	5	0	34
	0.05 bis unter 0.10	0	1	1	16	9	3	30
	> 0.10	0	0	4	10	25	7	46
	Spaltensumme	5	16	45	72	39	10	**187**

Die Messung des Zusammenhangs von zwei statistischen Variablen wird für viele wirtschaftswissenschaftliche Fragestellungen benötigt



Die Messung des Zusammenhangs von zwei statistischen Variablen wird für viele wirtschaftswissenschaftliche Fragestellungen benötigt



Empirische Kovarianz und Korrelationskoeffizient messen den linearen Zusammenhang von zwei metrischen Variablen

Auch die empirische Kovarianz ist ein arithmetisches Mittel
(des Produktes der Abweichungen von jeweiligen arithmetischen Mittel
der Originaldaten Y und X)

$$c_{XY} = \frac{1}{n} \sum_{v=1}^{n} (x_v - \overline{x})(y_v - \overline{y})$$

$$\rho_{XY} = \frac{c_{XY}}{s_X s_Y}$$

Der (Bravais) Pearson Korrelationskoeffizient ergibt sich aus dem Verhältnis von Kovarianz und dem Produkt der Standardabweichungen der beiden statistischen Variablen
Er kann Werte von -1 (perfekte negative Korrelation) bis 1 (perfekte positive Korrelation) annehmen

Empirische Kovarianz und Korrelationskoeffizient sind wichtigsten Maße zur Messung eines linearen Zusammenhangs zweier metrischer Variablen

Empirische Kovarianz und Korrelation messen den „Gleichklang“ zweier metrischer Variablen. Von Kausalität ist nicht die Rede

Häufige Fehler sind:

Eine dritte (oder mehrere) Hintergrundvariable begründen den beobachteten Zusammenhang (Bsp. Studiendauer und Einstiegsgehalt)

Die Hintergrundvariable ist der Zufall (Zahl der Störche und Geburtenzahl)

Die Kausalrichtung wird verdreht (Läuse und Körpertemperatur)

Post hoc ergo propter hoc-Fehler (weil etwas vorher geschah war es die Ursache dessen, was später geschah)

Korrelation ist nicht gleich Kausalität

Empirische Kovarianz und Korrelation messen den „Gleichklang“ zweier metrischer Variablen. Von Kausalität ist nicht die Rede

Häufige Fehler sind:

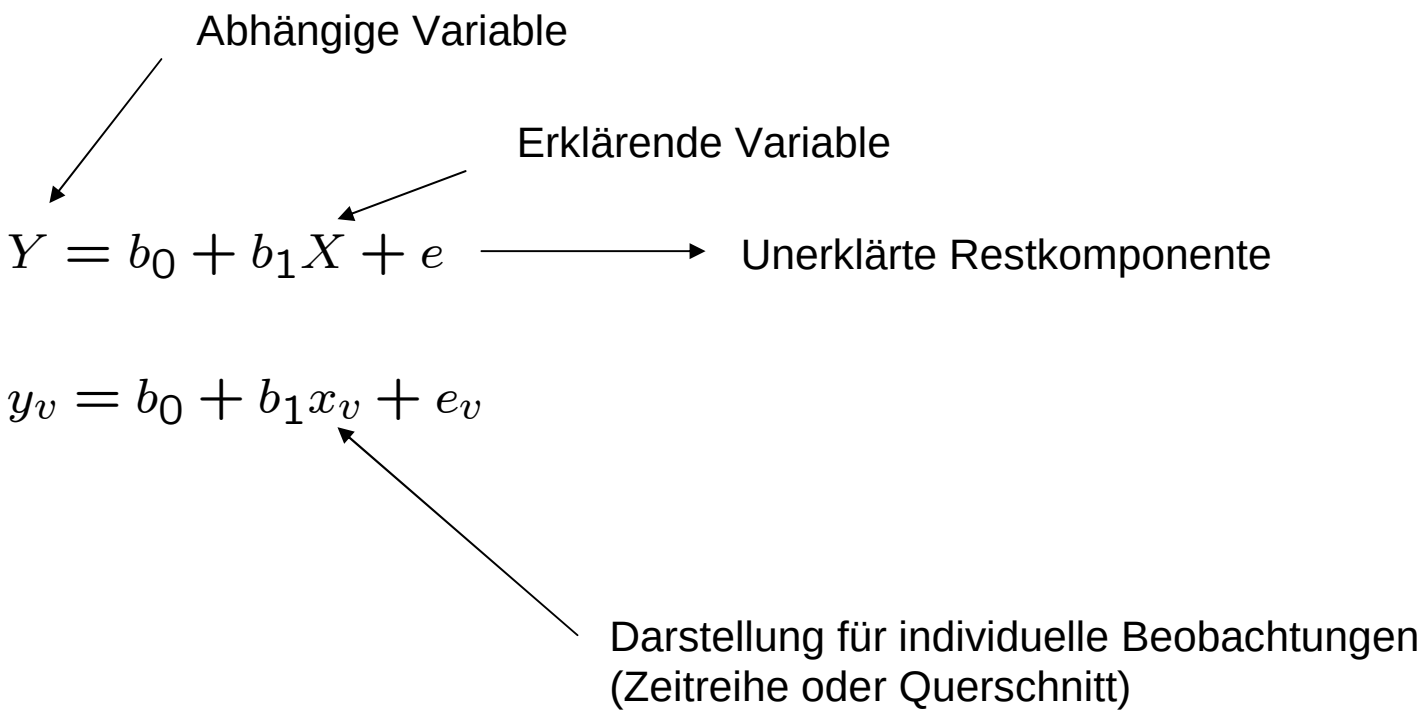
Eine dritte (oder mehrere) Hintergrundvariable begründen den beobachteten Zusammenhang (Studiendauer und Einstiegsgehalt)

Die Hintergrundvariable ist der Zufall (Zahl der Störche und Geburtenzahl)

Die Kausalrichtung wird verdreht (Läuse und Körpertemperatur)

Post hoc ergo propter hoc -Fehler (weil etwas vorher geschah war es die Ursache dessen, was später geschah)

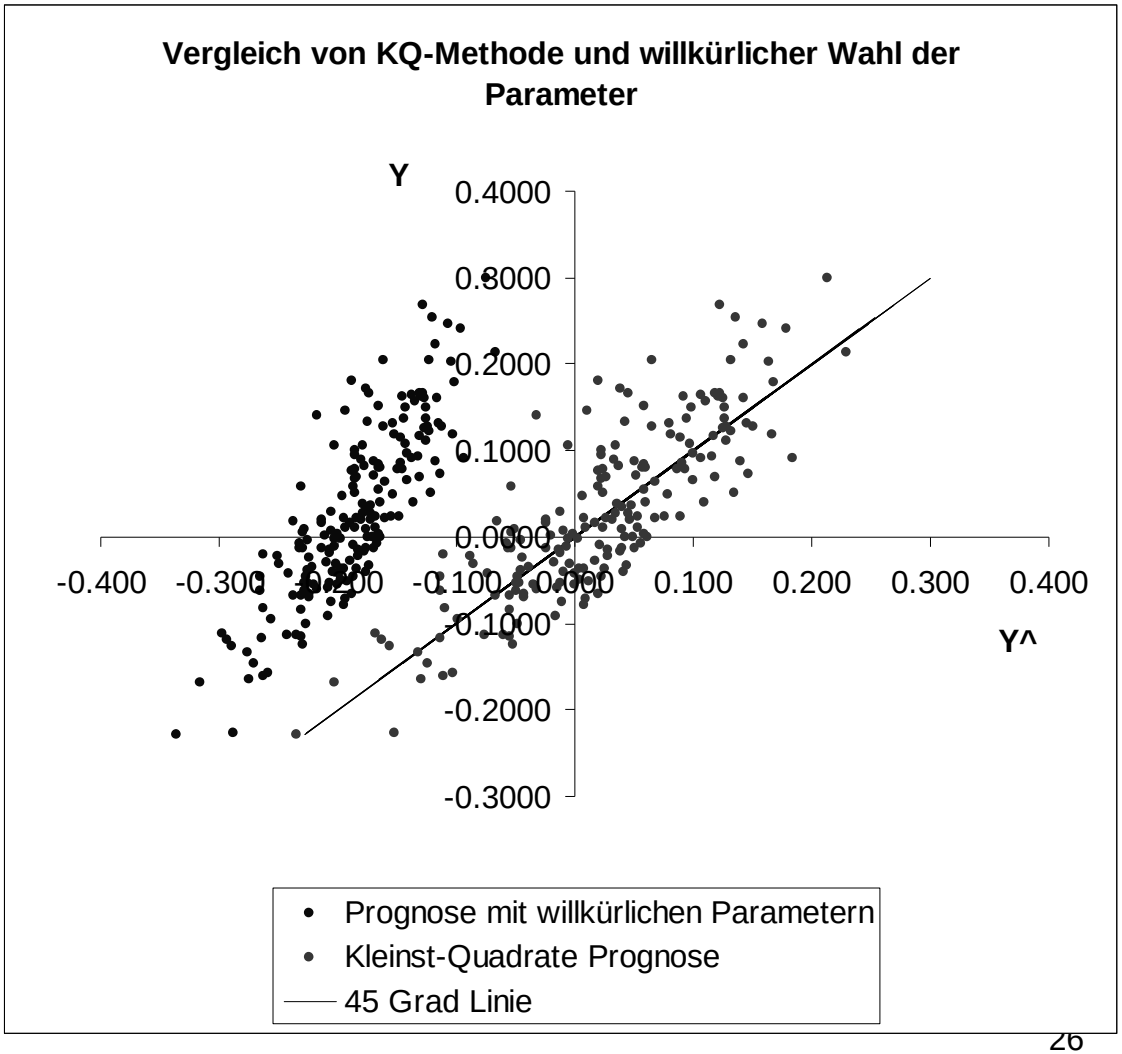
In der linearen Einfachregression wird eine abhängige Variable Y von einer erklärenden Variablen X und einer unerklärten Restkomponente linear beeinflußt



Die unbekannten Parameter schätzen wir durch die Minimierung der Summe der quadrierten Abweichungen von beobachteten und prognostiziertem Wert der abhängigen Variablen (Kleinst-Quadrate-Methode)

$$\arg\min_{\{\hat{b}_0, \hat{b}_1\}} \sum_{v=1}^{n} (y_v - \hat{y}_v)^2$$

$$\hat{y}_v = \hat{b}_0 + \hat{b}_1 x_v$$



Schätzer für die unbekannten Parameter ergeben sich aus den Bedingungen erster Ordnung für ein Minimum der Kleinst-Quadrate (K-Q)-Zielfunktion

Empirische Kovarianz erklärende und anhängige Variable

$$\hat{b}_1 = \frac{\frac{1}{n}\sum_{v=1}^{n}(x_v-\overline{x})(y_v-\overline{y})}{\frac{1}{n}\sum_{v=1}^{n}(x_v-\overline{x})^2} = \frac{c_{XY}}{s_X^2}$$

Empirische Varianz erklärende Variable

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}$$

Arithmetische Mittel von erklärender und abhängiger Variable

Auch die Schätzung der Parameter einer nicht-linearen Beziehung Y=f(X,U)
ist mit der K-Q-Methode möglich, wenn die Beziehung linearisiert werden kann
(Logarithmierung).
Verwenden Sie dann die transformierten Originaldaten zur Parameterschätzung

Wir verwenden das Bestimmtheitsmaß zur Beurteilung der Güte einer K-Q-Schätzung

Das Bestimmheitsmaß ist zwischen 0 und 1 definiert.
Es entspricht dem quadrierten Korrelationskoeffizient von abhängiger Variable und Prognosewerten

Empirische Varianz der Prognosewerte (erklärte Varianz)

$$B \text{ (oder } R^2) = \frac{\frac{1}{n}\sum_{v=1}^{n}(\widehat{y}_v - \overline{y})^2}{\frac{1}{n}\sum_{v=1}^{n}(y_v - \overline{y})^2} = \frac{s^2_{\widehat{Y}}}{s^2_Y}$$

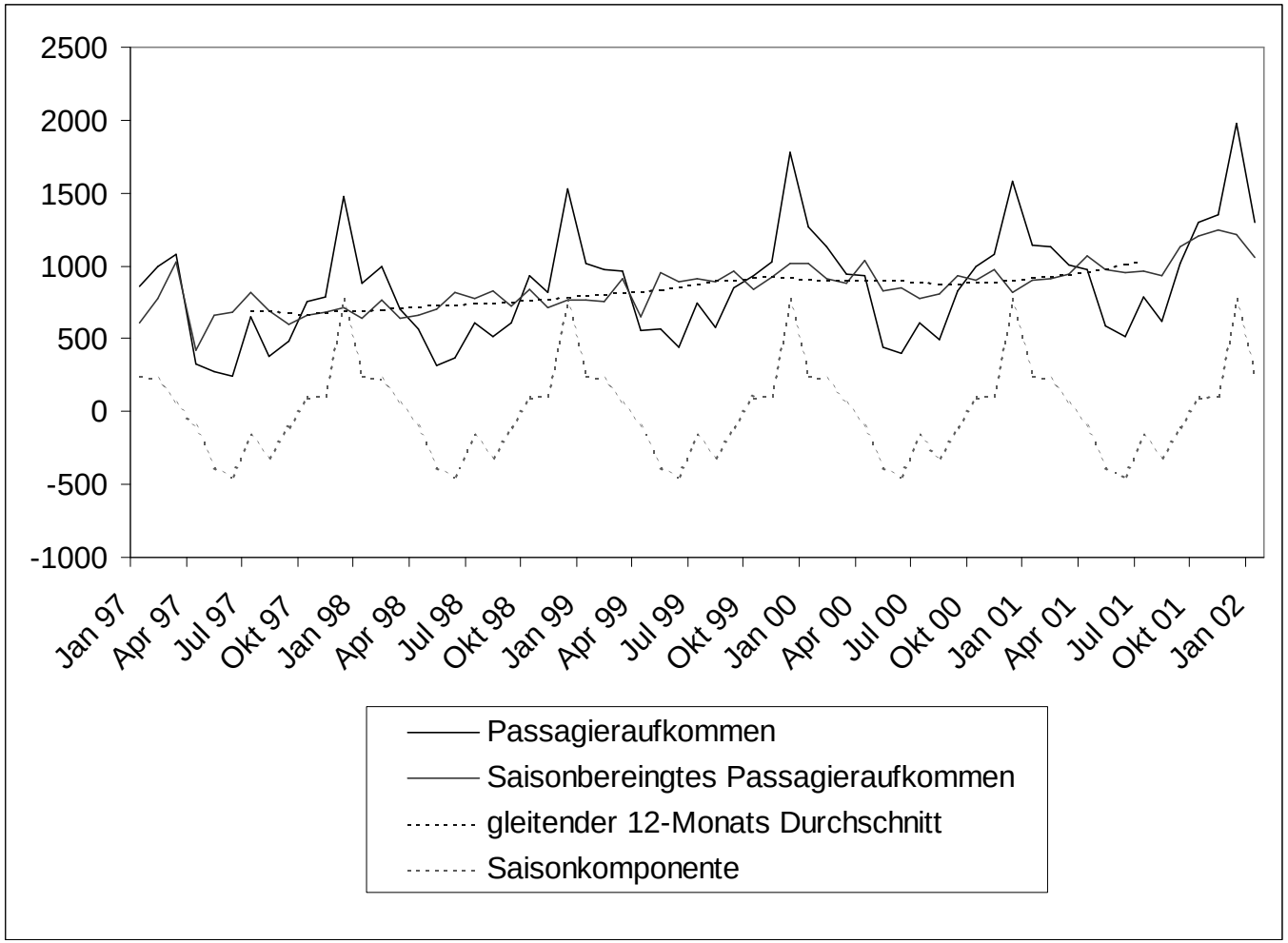
$$\overline{y} = \overline{\widehat{y}}$$

Empirische Varianz der abhängigen Variable (gesamte Varianz)

Das sollten zeigen können (und ein paar andere Resultate auch)

Explorative Datenanalyse Teil 3
Deskriptive Zeitreihenanalyse und Wirtschaftsstatisik (Indizes)

Die Zerlegung in Trend, zyklische, saisonale und Zufallskomponente ist oft durch Ansicht einer Zeitreihe erkennbar



Zur Schätzung der Trendkomponente bietet sich die K-Q-Methode an

Die Zeit t ist die erklärende Variable in der Regressionsgleichung

$$y_t = b_0 + b_1 t + e_t$$

In Ihren Daten kodieren Sie in t=1,2,...,T um

$$\arg\min_{\{\hat{b}_0, \hat{b}_1\}} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2$$

$$\hat{y}_t = \hat{b}_0 + \hat{b}_1 t$$

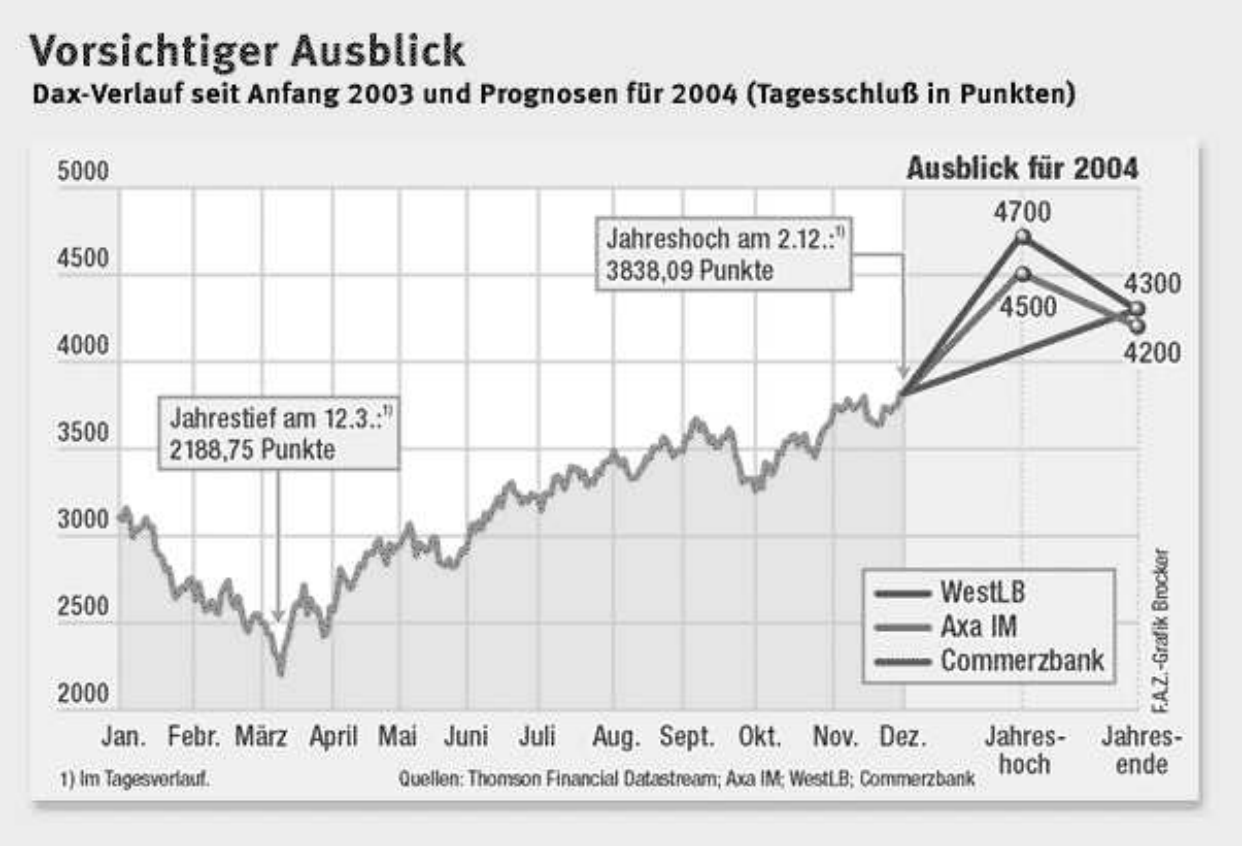
$$\hat{b}_1 = \frac{\sum_{t=1}^{T} (t - \bar{t})(y_t - \bar{y})}{\sum_{t=1}^{T} (t - \bar{t})^2}$$

Standard-K-Q-Ergebnisse/Formeln

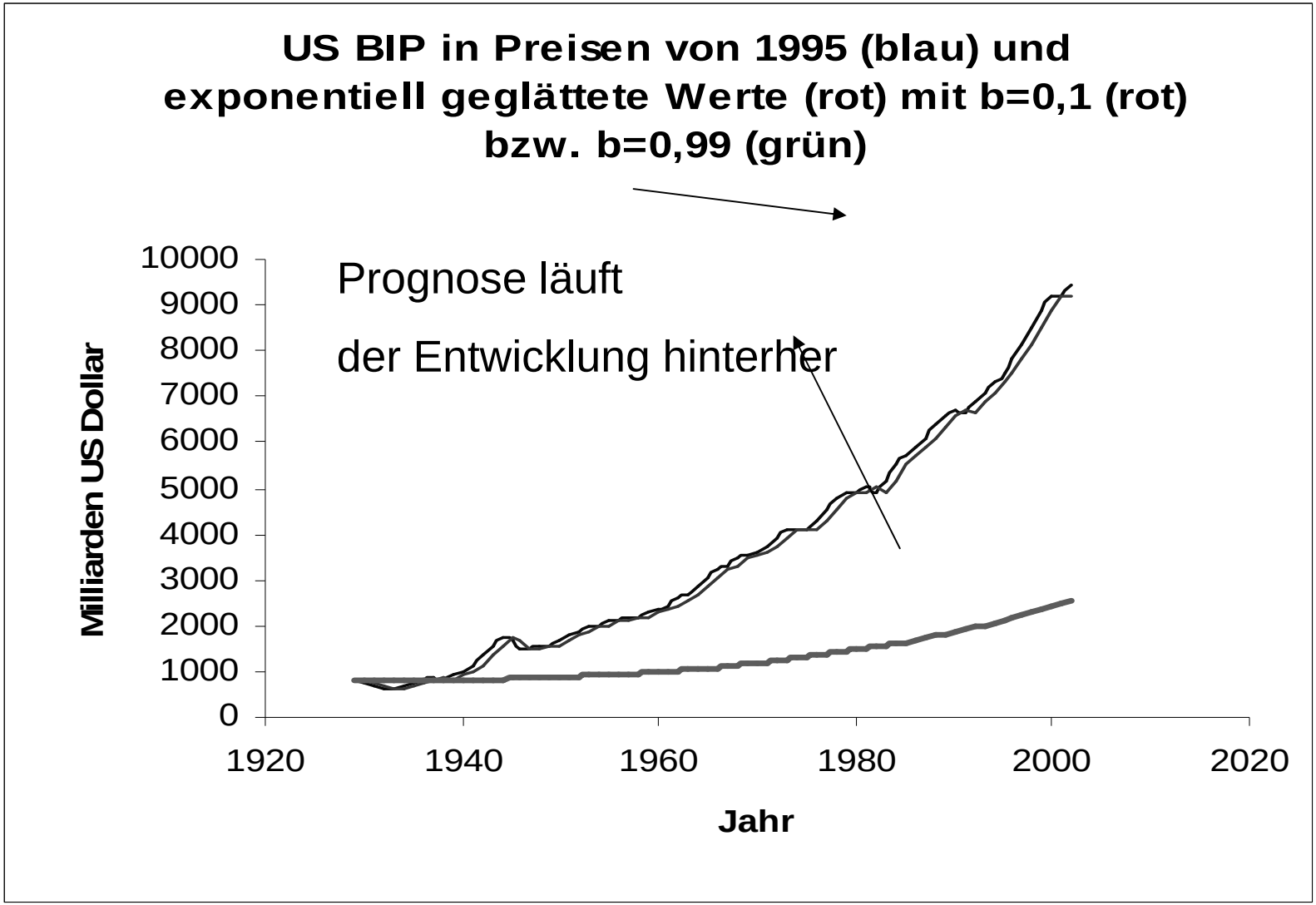
$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{t}$$

Wenn Sie einen exponentiellen Trend unterstellen, logarithmieren Sie einfach die zu modellierende Variable und wenden auf die transformierten Daten die K-Q-Methode an

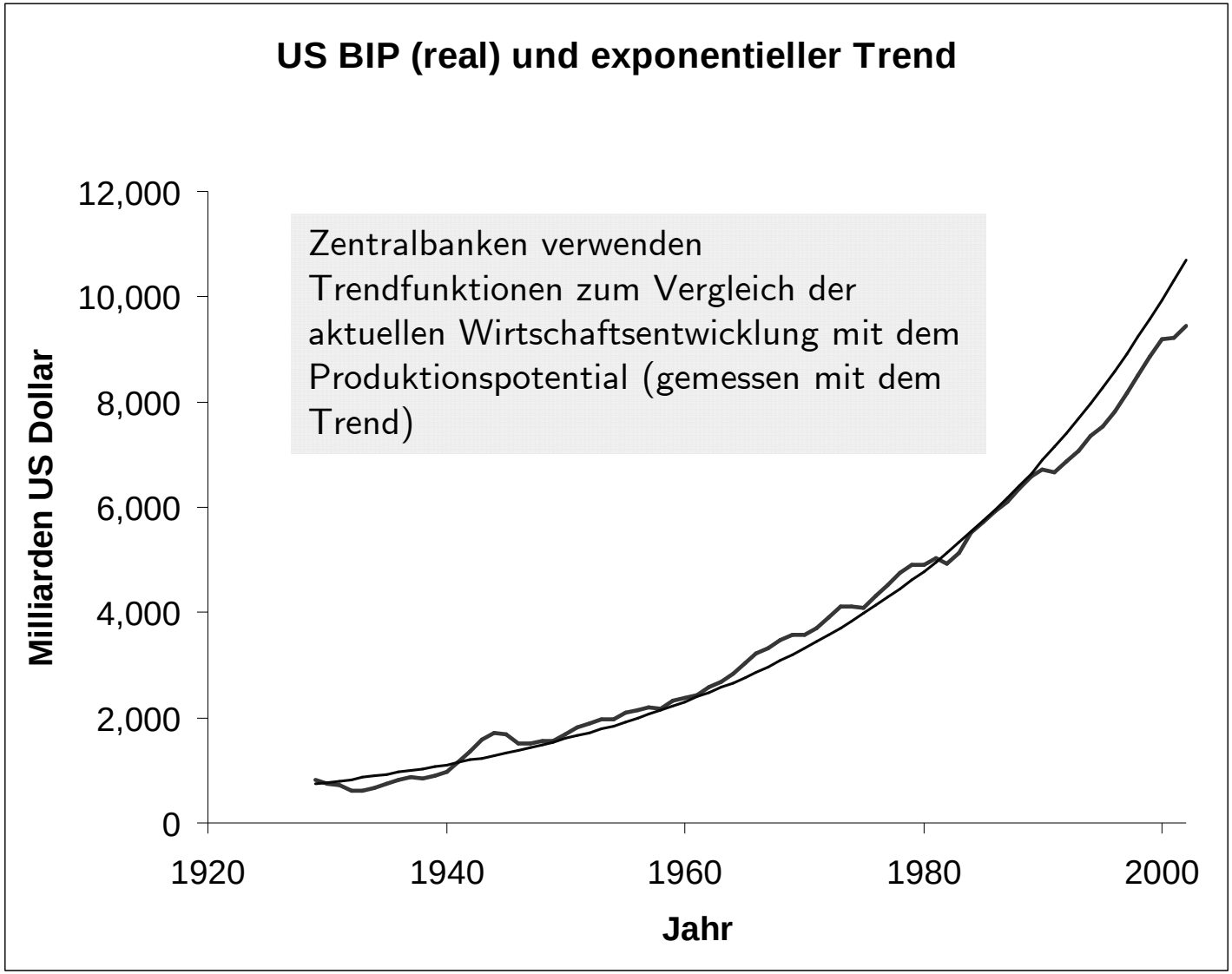
Prognosen mit Trendfunktionen sind einfach, Sie sollten aber damit vorsichtig sein (Russells Huhn)



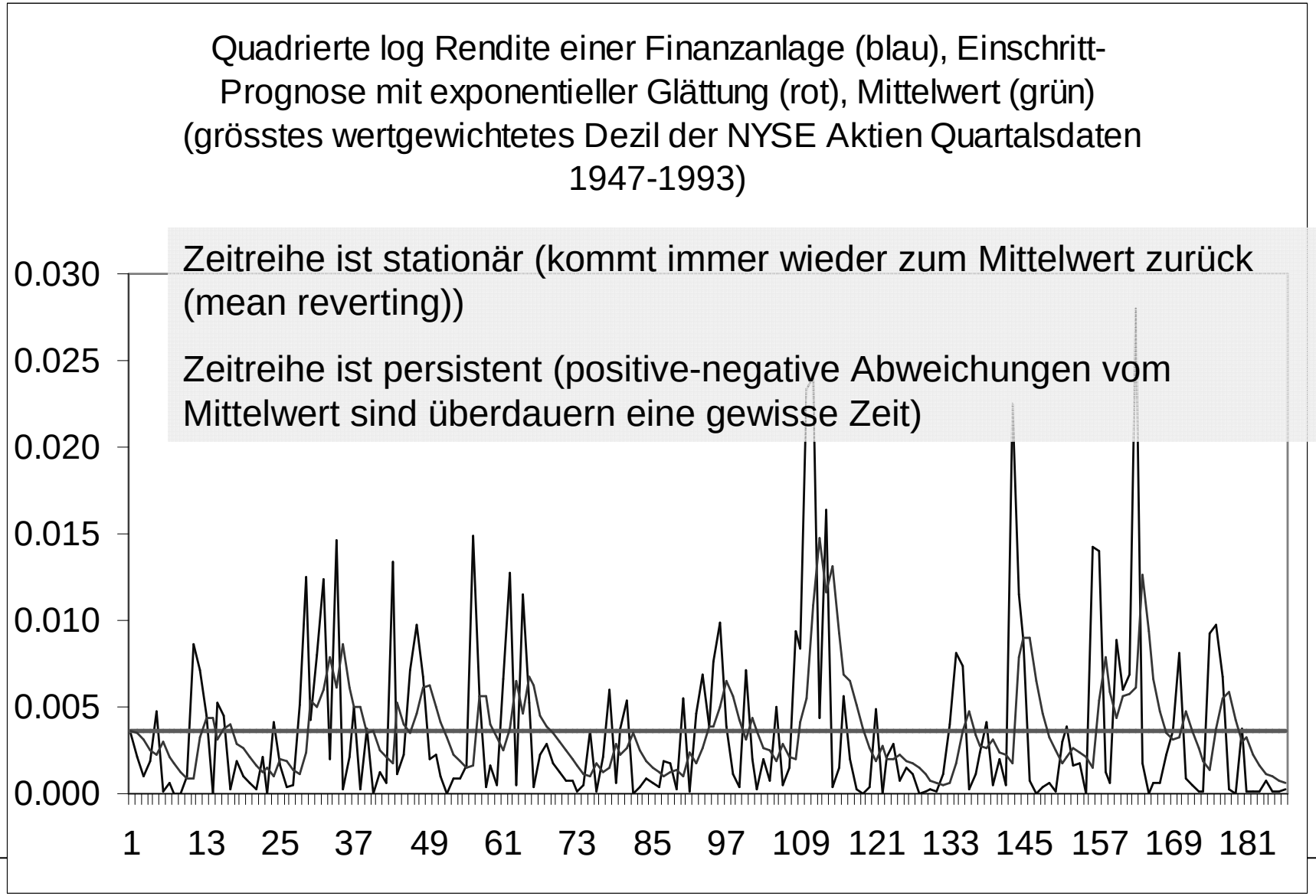
Die exponentielle Glättung wird zur Ermittlung der glatten Komponente oder Einschritt-Prognosen verwendet. Die Daten dürfen aber keinen Trend aufweisen



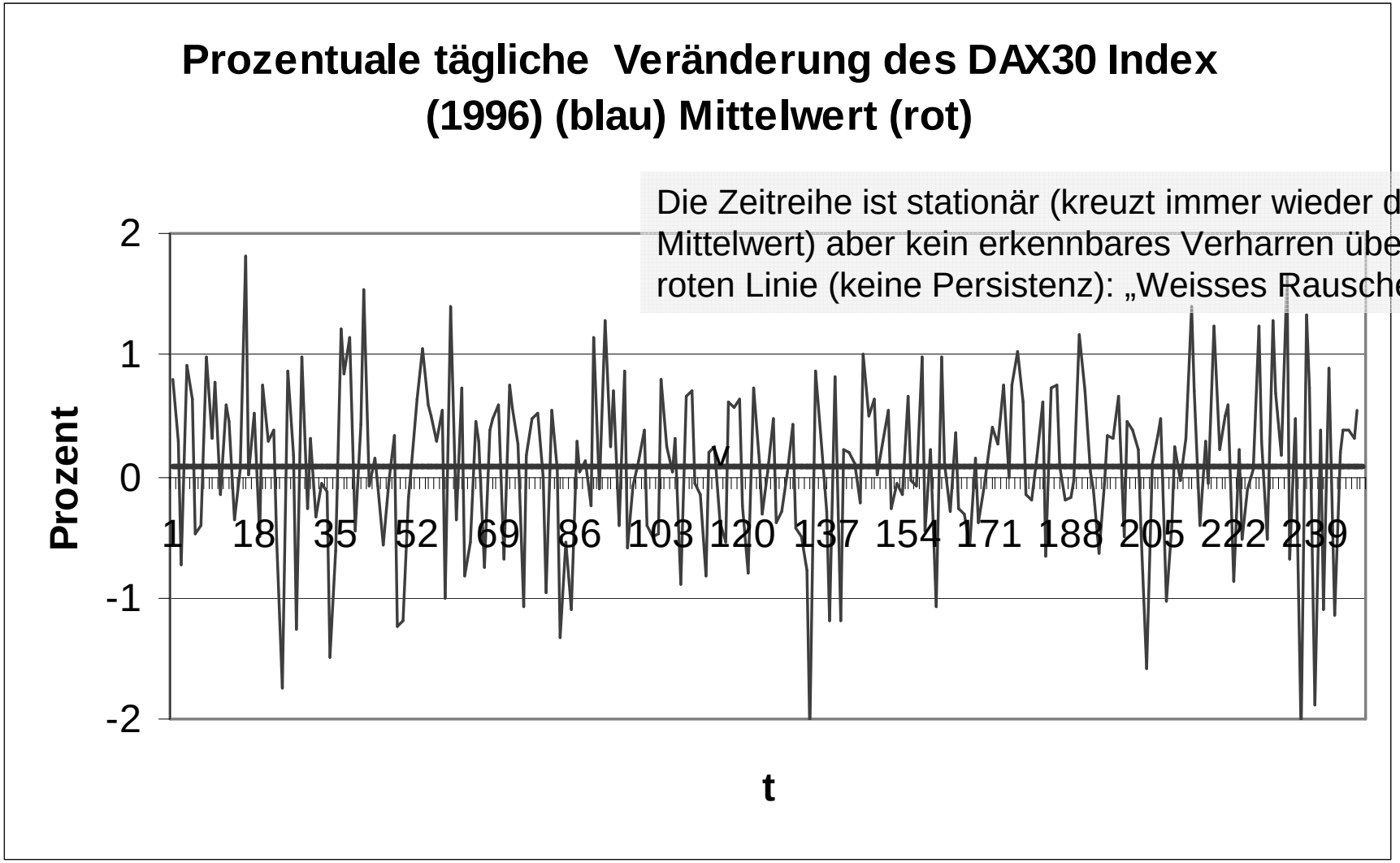
Hier wird das reale US-BIP mit einer exponentiellen Trendfunktion „modelliert“.



Für „persistente“ stationäre Daten (kein Trend) macht eine Prognose mit exponentieller Glättung mehr Sinn

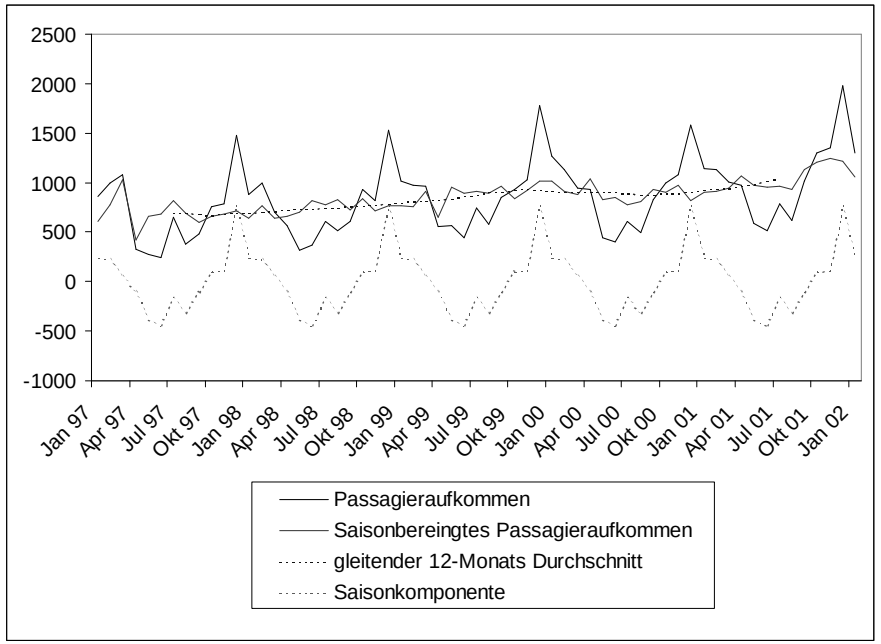


Renditezeitreihen sind ein Beispiel für eine stationäre, aber nicht persistente Zeitreihe (und damit nicht prognostizierbar)



Die Ermittlung der glatten Komponente mit gleitenden Durchschnitten ist einfach. Die Methode wird zur Saisonbereinigung verwendet

DATUM	paxe	gleitender 12-Monats Durchschnitt	y(i, j)-y'(i, j)
1997-01-01	857		
1997-02-01	996		
1997-03-01	1082		
1997-04-01	327		
1997-05-01	270		
1997-06-01	240		
1997-07-01	650	693.13	-43.13
1997-08-01	374	694.08	-320.08
1997-09-01	484	678.13	-194.13
1997-10-01	758	672.29	85.71
1997-11-01	785	684.29	100.71
1997-12-01	1481	691.83	789.17
1998-01-01	884	695.58	188.42
1998-02-01	992	699.63	292.38
1998-03-01	703	710.79	-7.79
2000-10-01	993	887.38	105.63
2000-11-01	1077	894.54	182.46
2000-12-01	1576	904.92	671.08
2001-01-01	1143	917.21	225.79
2001-02-01	1134	930.17	203.83
2001-03-01	1001	943.54	57.46
2001-04-01	970	964.46	5.54
2001-05-01	584	988.92	-404.92
2001-06-01	513	1017.21	-504.21
2001-07-01	790	1040.42	-250.42
2001-08-01	619		
2001-09-01	1019		
2001-10-01	1302		
2001-11-01	1355		
2001-12-01	1977		
2002-01-01	1299		



	Saisonkomponente
January	244.56
February	221.80
March	59.18
April	-98.31
May	-386.42
June	-445.27
July	-169.03
August	-316.85
September	-111.95
October	94.95
November	105.48
December	762.86

	Saisonbereingtes Passagieraufkommen
1997-01-01	612.44
1997-02-01	774.20
1997-03-01	1022.82
1997-04-01	425.31
1997-05-01	656.42
1997-06-01	685.27
1997-07-01	819.03
1997-08-01	690.85
1997-09-01	595.95
1997-10-01	663.05
1997-11-01	679.52
1997-12-01	718.14

Preisindizes werden für die Messung der Geldwertstabilität und für Kaufkraftvergleiche verwendet

Preisindizes sollen die „allgemeine“ Preisentwicklung messen, nicht die einzelner Güter

Bei der Berechnung eines Paasche Preisindex verändert sich der Warenkorb von Periode zu Periode, bei der Berechnung des Laspeyres Index bleibt der Basiswarenkorb konstant.

Die Formeln für den Paasche- und den Laspeyres-Index müssen Sie können!

Inflationsraten werden aus prozentualen Veränderungen von Preisindizes berechnet

Das statistische Bundesamt berechnet Laspeyres Preisindizes für ausgewählte (repräsentative) Warenkörbe

Zum Deflationieren nominaler Größen (BIP) braucht es einen passenden Paasche-Index. Gibt es den nicht, behilft man sich mit einem Laspeyeres Index.

Auch der Deutsche Aktienindex (DAX) ist ein Laspeyres Preisindex

Bei der Messung der Geldwertstabilität, bei Kaufkraftvergleichen und bei der Deflationierung sind einige Probleme der Indexbildung zu beachten

Qualitätsverbesserungen von Gütern werden bei der Ermittlung von Preisindizes nicht berücksichtigt.

Warenkörbe veralten und damit wird die Aussagekraft eines Laspeyres Preisindex zweifelhaft.

Durch die Veränderung des Basiswarenkorbes sind lange Reihen von Preisindizes (und damit Inflationsraten) nur durch (mitunter) heroische Annahmen zu ermitteln.

Da einige Güter in der Vergangenheit gar nicht verfügbar waren, ist die Ermittlung eines Paasche Index mitunter gar nicht möglich.

Wie steht es um die Repräsentativität des verwendeten Warenkorbes?

Zur Deflationierung benötigt man Paasche Indizes, aber meist sind nur Laspeyres Indizes verfügbar

Eigentlich bräuchte jeder seinen „persönlichen“ Preisindex

Insbesondere in der Europäischen Union machen regional verschiedene Warenkörbe viel Sinn (Bedeutung Heizöl in Sizilien und in Finnland)

Die Umbasierung und Rückrechnung von Indizes ist mit einfachen Dreisatz-Rechnungen möglich

Jahr	I 80,t	I 85,t					
1976	85						
1977	88.6						
1978	91						
1979	94.8						
1980	100		Rückrechnung: Rückgerechnete Veränderungen des neuen Index				
1981	106.3		sollen den relativen Veränderungen des alten Index entsprechen				
1982	111.9						
1983	115.6						
1984	118.4						
1985	121	100.00					
1986		99.9	Weiterführung des alten Index: relative Veränderungen des weitergeführten				
1987		100.1	Index sollen den relativen Veränderungen des neuen Index entsprechen				
1988		101.4					
1989		104.2					

Jahr	I 80,t	I 85,t
1976	85	70.2
1977	88.6	73.2
1978	91	75.2
1979	94.8	78.3
1980	100	82.6
1981	106.3	87.9
1982	111.9	92.5
1983	115.6	95.537
1984	118.4	97.9
1985	121	100
1986	120.879	99.9
1987	121.121	100.1
1988	122.694	101.4
1989	126.082	104.2