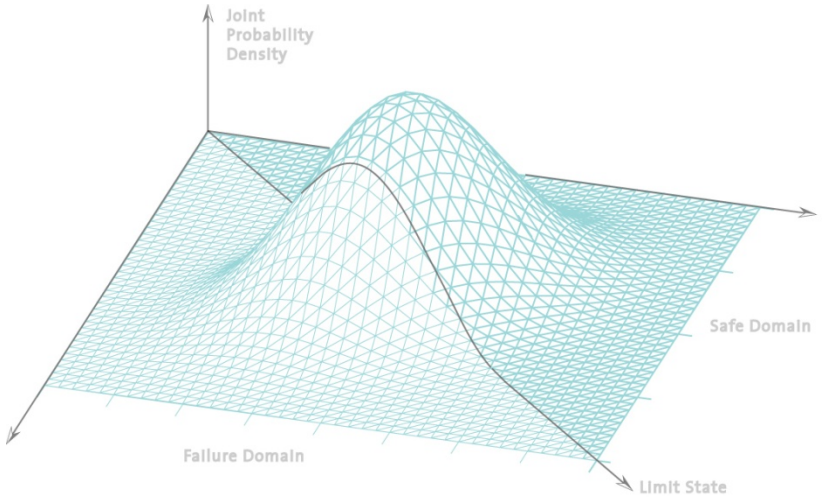


Statistik und Wahrscheinlichkeitsrechnung

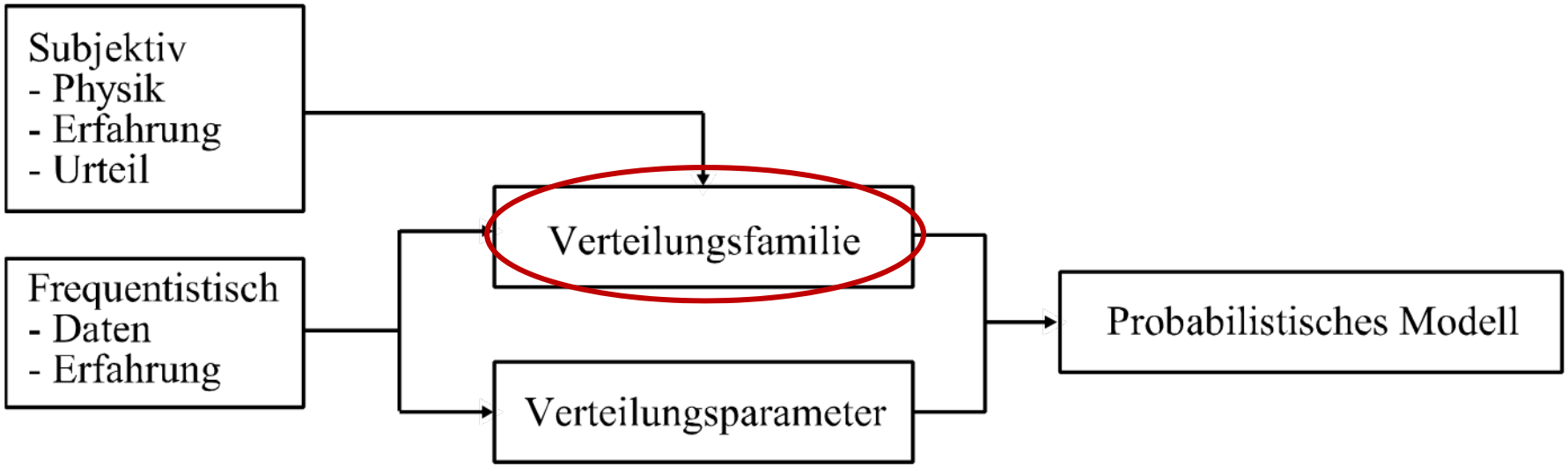
Übung 9

Inhalt der heutigen Übung

- Gemeinsames Lösen der Übungsaufgaben
  - E.5: Wahrscheinlichkeitspapier
  - E.14: Regressionsanalyse
  - E.9: Konfidenzintervalle
- Vorstellen der Gruppenaufgabe E.1: Hypothesentest

Wahrscheinlichkeitspapier

Problemstellung:



Wahrscheinlichkeitspapier

Wir wollen wissen, ob eine Stichprobe mit einer bestimmten Verteilungsfamilie beschrieben werden kann.

Hierfür kann für eine erste Abschätzung ohne Bestimmung der Verteilungs-parameter eine Überprüfung mit Hilfe von einem Wahrscheinlichkeitspapier durchgeführt werden.

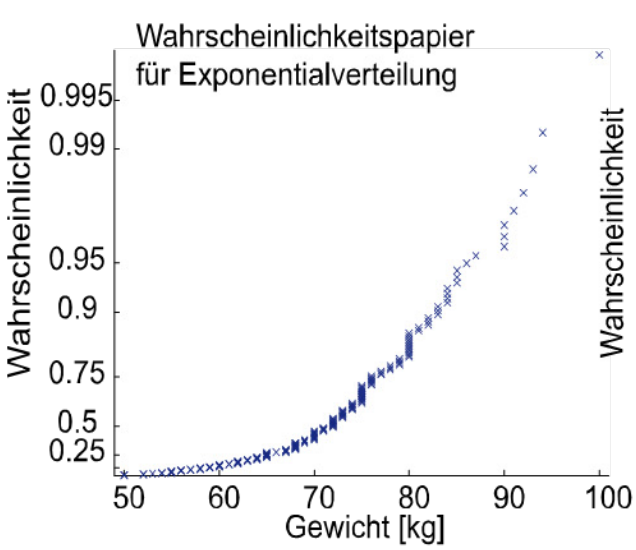
- ✓ Wenn die Punkte der beobachteten Werte auf dem Wahrscheinlichkeitspapier in etwa auf einer Geraden liegen, dann kann die Verteilungsfamilie akzeptiert werden.

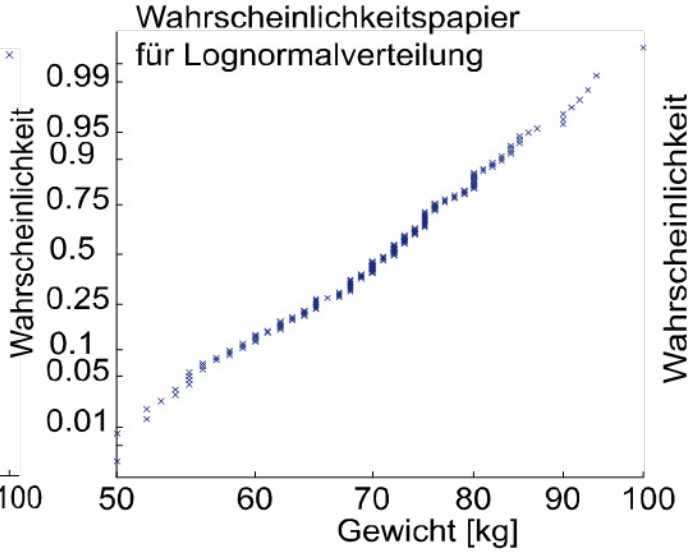
Wie kann eine Punktmenge auf einer Linie liegen?

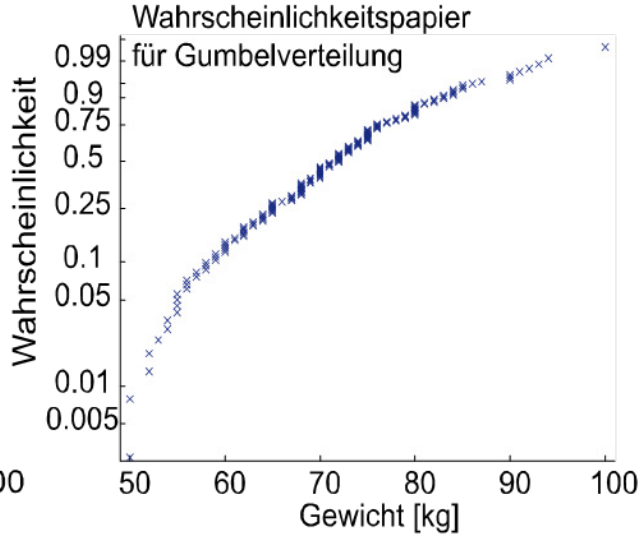
→ Linearisierung der y-Achse für die betrachtete Verteilungsfamilie

Wahrscheinlichkeitspapier

Ein Beispiel: Zur Beschreibung des Körpergewichtes der Studierenden aus der Statistik-Vorlesung wird eine geeignete Verteilungsfamilie gesucht.







→ Das Körpergewicht der Studierenden folgt am ehesten einer Lognormalverteilung.

Aufgabe E.5

Aus Verkehrszählungen liegt eine Datenserie vor, die den täglichen Verkehrsfluss in der Rosengartenstrasse in Zürich beschreibt.

a) Erstelle ein Wahrscheinlichkeitspapier für folgende Dichtefunktion

$$f_X(x) = \begin{cases} \dfrac{2}{10000^2}\,x & 0 \le x \le 10000 \\ 0 & \text{sonst} \end{cases}$$

b) Überprüfe mit Hilfe des erstellten Wahrscheinlichkeitspapiers, ob der tägliche Verkehrsfluss mit dieser Verteilung beschrieben werden kann.

Beobachtung	Anzahl Fahrzeuge
1	3600
2	4500
3	5400
4	6500
5	7000
6	7500
7	8700
8	9000
9	9500

Aufgabe E.5 – Lösung

Wahrscheinlichkeits-dichtefunktion

$$f_X(x) = \begin{cases} \frac{2}{10000^2} x & 0 \le x \le 10000 \\ 0 & \text{sonst} \end{cases}$$

Wahrscheinlichkeits-verteilungsfunktion

$$F_X(x) = \begin{cases} 0 & 0 \le x \\ \left(\frac{x}{10000}\right)^2 & 0 < x \le 10000 \\ 1 & x > 10000 \end{cases}$$

Linearisierung

$$F_X(x) = \left(\frac{x}{10000}\right)^2 \Leftrightarrow \sqrt{F_X(x)} = \frac{x}{10000}$$

Aufgabe E.5 – Lösung

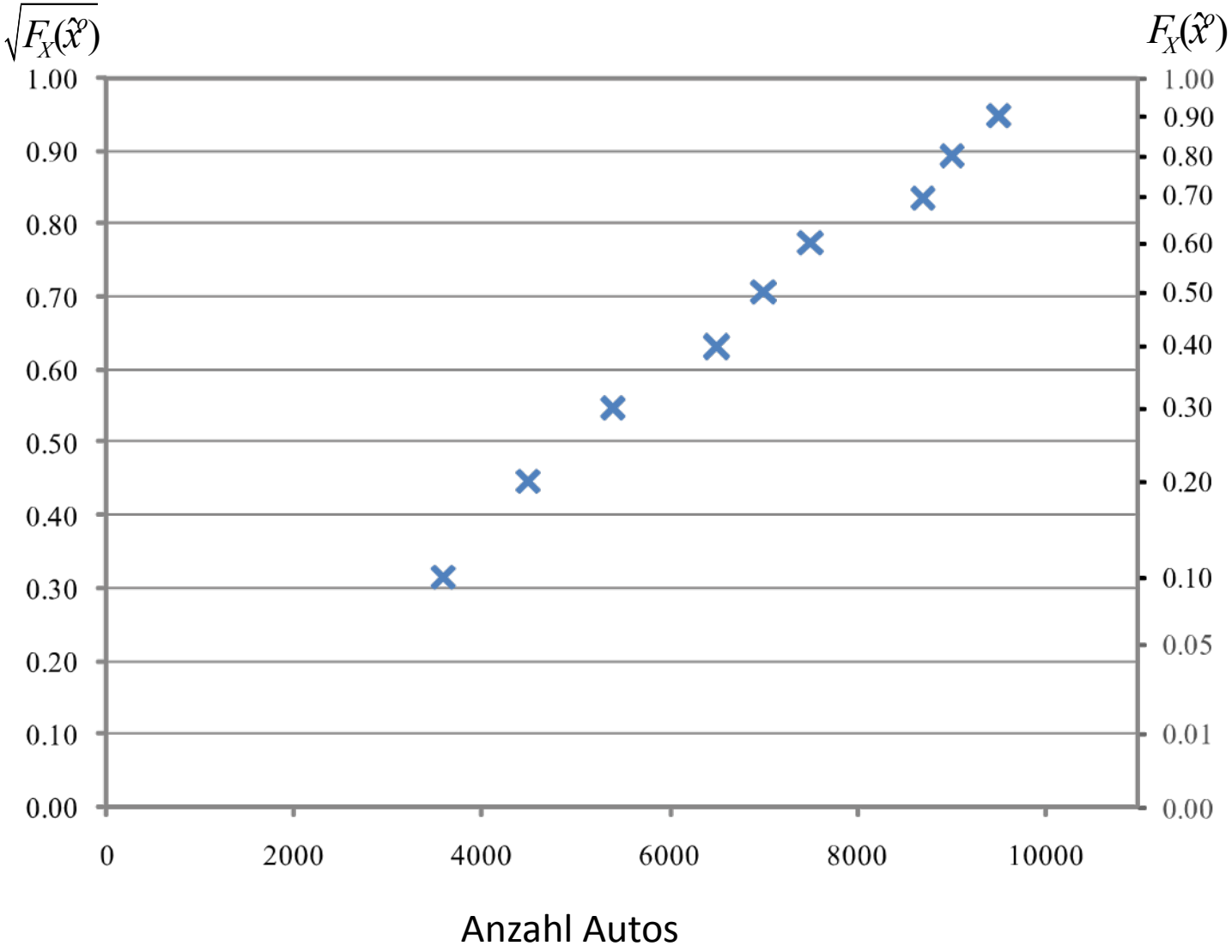
Eine linearisierte y-Achse erhalten wir, indem wir für alle Quantile in bestimmten Abständen die linearisierte Form berechnen - beispielsweise zu jedem Datenpunkt.

Berechnung des Quantilindex:

$$\upsilon = \frac{i}{n+1} = F_X(\hat{x}^o)$$

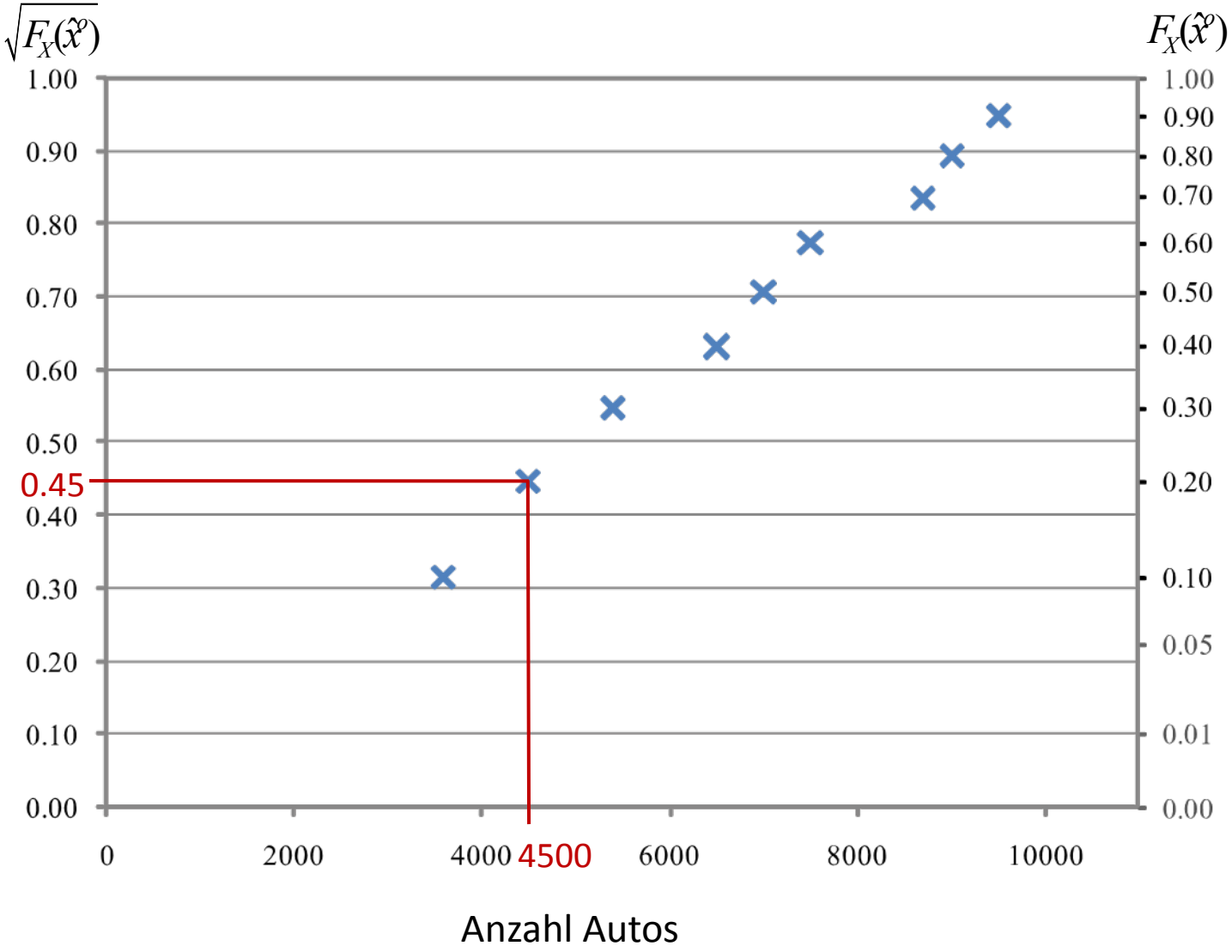
Rang i	Anzahl Autos (x-Achse)	$F_X(\hat{x}^o)$	$\sqrt{F_X(\hat{x}^o)}$
		0	0
1	3600	0.1	0.32
2	4500	0.2	0.45
3	5400	0.3	0.55
4	6500	0.4	0.63
5	7000	0.5	0.71
6	7500	0.6	0.77
7	8700	0.7	0.84
8	9000	0.8	0.89
9	9500	0.9	0.95
		1.0	1.0

Aufgabe E.5 – Lösung



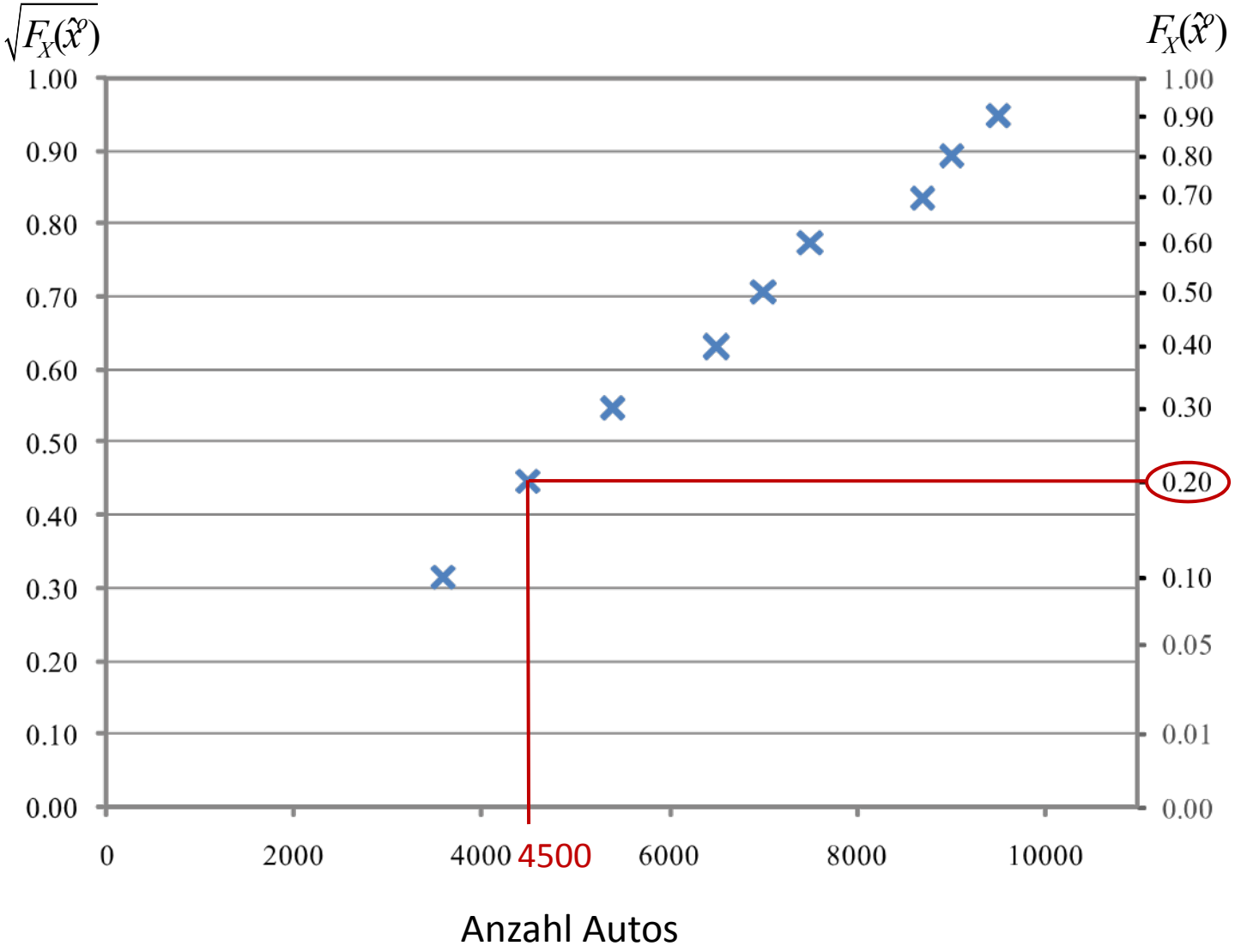
Anzahl Autos	$F_X(\hat{x}^o)$	$\sqrt{F_X(\hat{x}^o)}$
	0	0
3600	0.1	0.32
4500	0.2	0.45
5400	0.3	0.55
6500	0.4	0.63
7000	0.5	0.71
7500	0.6	0.77
8700	0.7	0.84
9000	0.8	0.89
9500	0.9	0.95
	1.0	1.0

Aufgabe E.5 – Lösung



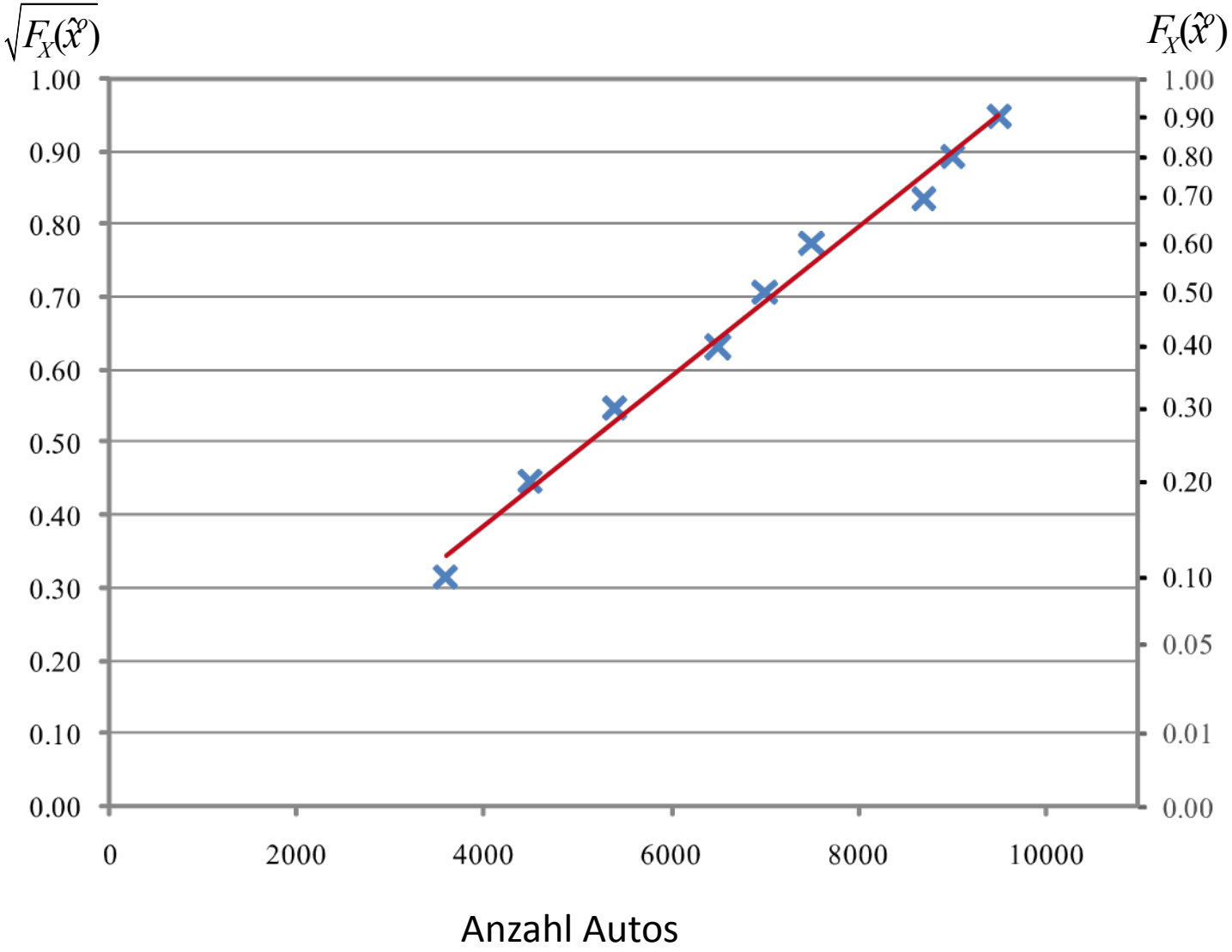
Anzahl Autos	$F_X(\hat{x}^o)$	$\sqrt{F_X(\hat{x}^o)}$
	0	0
3600	0.1	0.32
4500	0.2	0.45
5400	0.3	0.55
6500	0.4	0.63
7000	0.5	0.71
7500	0.6	0.77
8700	0.7	0.84
9000	0.8	0.89
9500	0.9	0.95
	1.0	1.0

Aufgabe E.5 – Lösung



Anzahl Autos	$F_X(\hat{x}^o)$	$\sqrt{F_X(\hat{x}^o)}$
	0	0
3600	0.1	0.32
4500	0.2	0.45
5400	0.3	0.55
6500	0.4	0.63
7000	0.5	0.71
7500	0.6	0.77
8700	0.7	0.84
9000	0.8	0.89
9500	0.9	0.95
	1.0	1.0

Aufgabe E.5 – Lösung

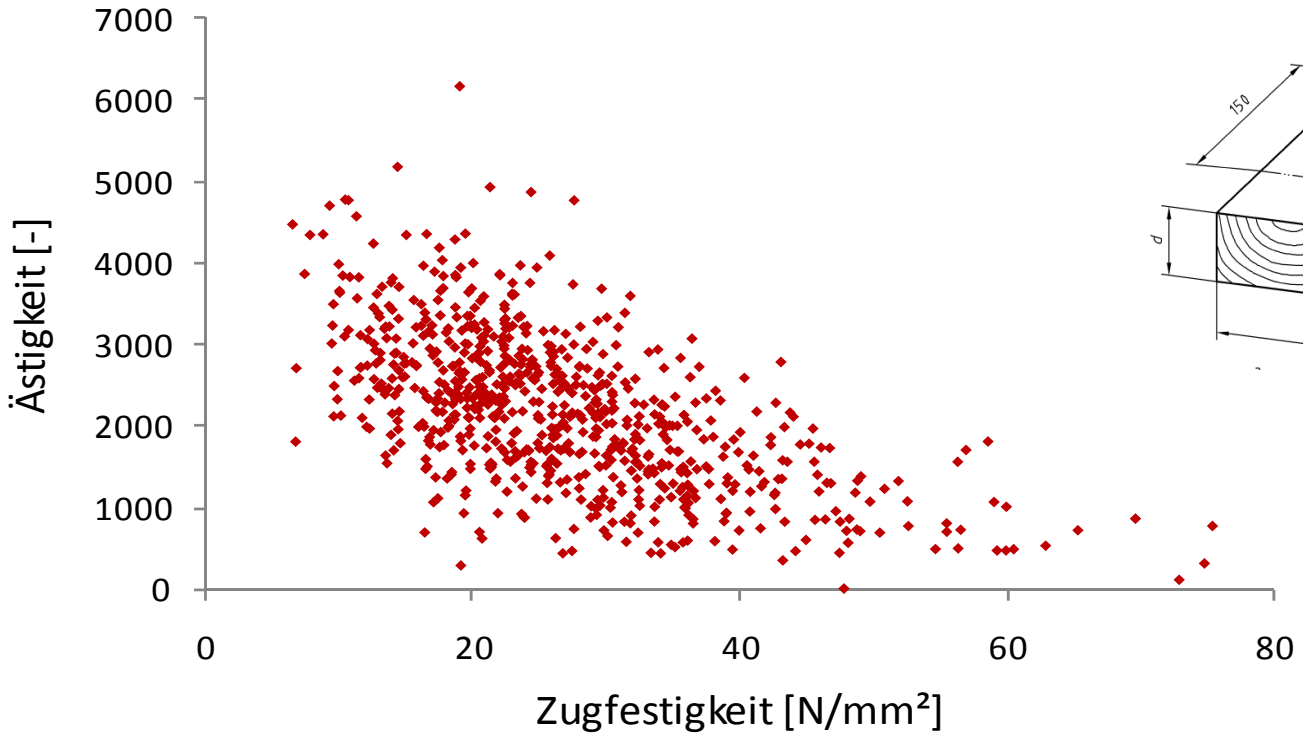


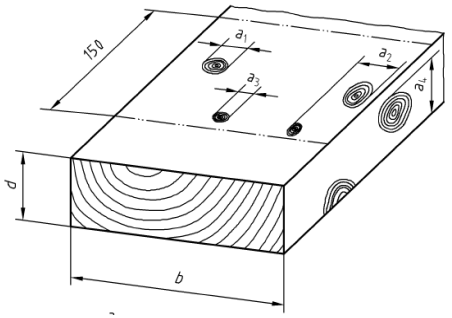
Anzahl Autos	$F_X(\hat{x}^o)$	$\sqrt{F_X(\hat{x}^o)}$
	0	0
3600	0.1	0.32
4500	0.2	0.45
5400	0.3	0.55
6500	0.4	0.63
7000	0.5	0.71
7500	0.6	0.77
8700	0.7	0.84
9000	0.8	0.89
9500	0.9	0.95
	1.0	1.0

(Lineare) Regression

Problemstellung:

Der funktionale Zusammenhang zwischen zwei Zufallsvariablen soll bestimmt werden. → Im Falle der linearen Regression eine Gerade.

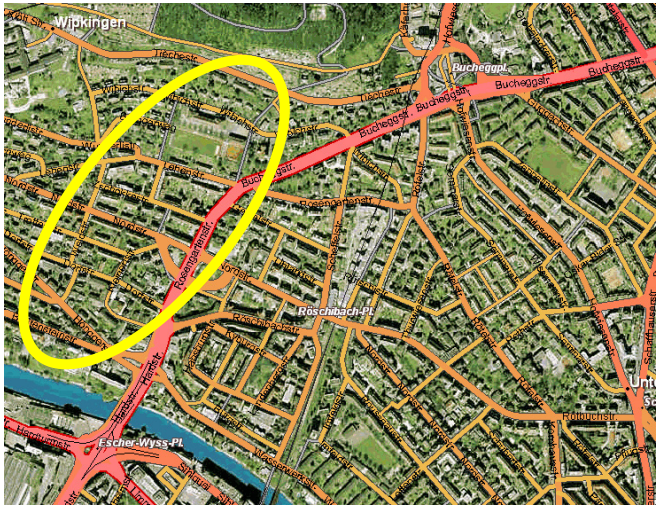




Aufgabe E.14

Datum	Richtung 1	Richtung 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313

Uns wurden Verkehrsdaten der Rosengartenstrasse in Zürich zur Auswertung übergeben. Richtung 1 gibt die Verkehrsbelastung zum Bucheggplatz, Richtung 2 die Belastung zum Escher-Wyss-Platz an.



Aufgabe E.14

Der Zusammenhang der Verkehrsdaten für die beiden Richtungen soll mit Hilfe einer linearen Regression bestimmt werden.

a) Erstelle ein Regressionmodell für die Abhängigkeit zwischen den Daten der beiden Richtungen: Bestimme die Regressions-koeffizienten sowie die Varianz des Residualwertes und der geschätzten Modellparameter. Verwende hierzu nur die ersten 10 Datenpaare.

b) Das Regressionsmodell soll nun mit neuen Daten aktualisiert werden. Bestimme die a posteriori Regressionskoeffizienten und quantifiziere die Unsicherheit des aktualisierten Regressionsmodells. Verwende hierzu die zweite Hälfte des Datensatzes.

Aufgabe E.14

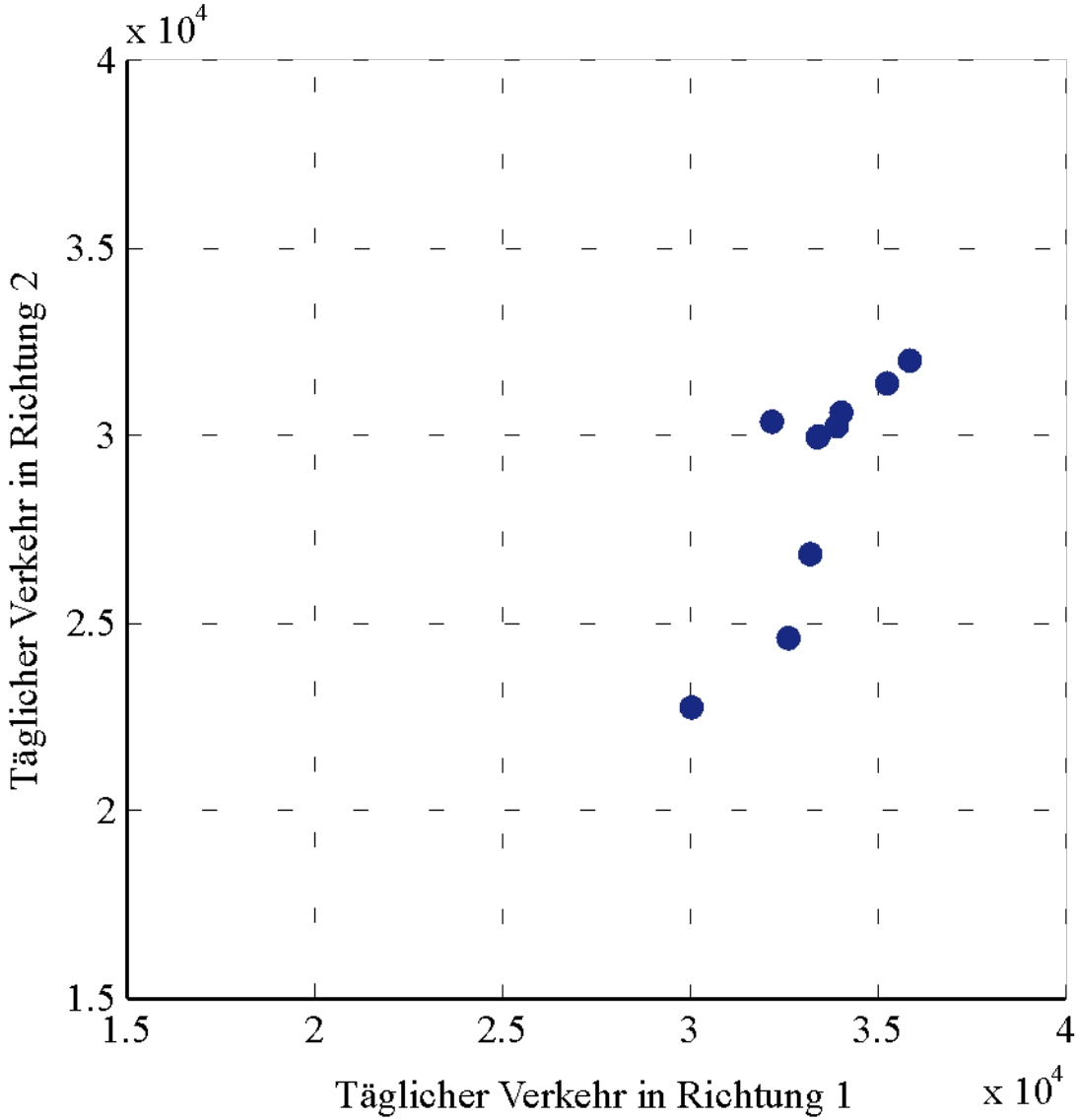
Der Zusammenhang der Verkehrsdaten für die beiden Richtungen soll mit Hilfe einer linearen Regression bestimmt werden.

Regressionsgerade:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

x_i Verkehr Richtung 1

y_i Verkehr Richtung 2



Aufgabe E.14

a) Erstelle ein Regressionmodell für die Abhängigkeit zwischen den Daten der beiden Richtungen: Bestimme die Regressions-koeffizienten sowie die Varianz des Residualwertes und der geschätzten Modellparameter.

Methode der kleinsten Quadrate: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} \left(\hat{y}_i - \left(\beta_0 + \beta_1 \hat{x}_i \right) \right)^2$$

$$0 = \sum_{i=1}^{n} \left(\hat{y}_i - \beta_0 - \beta_1 \hat{x}_i \right)^2 \frac{\partial}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left(\hat{y}_i - \beta_0 - \beta_1 \hat{x}_i \right) \hat{x}_i \quad (1)$$

$$0 = \sum_{i=1}^{n} \left(\hat{y}_i - \beta_0 - \beta_1 \hat{x}_i \right)^2 \frac{\partial}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(\hat{y}_i - \beta_0 - \beta_1 \hat{x}_i \right) \quad (2)$$

Aufgabe E.14

Gleichung (2) umformen:

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i = \overline{y} - \beta_1 \overline{x}$$

In Gleichung (1) einsetzen:

$$\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)\hat{x}_i - \beta_1 \sum_{i=1}^{n}\left(\hat{x}_i - \overline{x}\right)\hat{x}_i$$

$$\Rightarrow \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i\hat{x}_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i}{\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i^2 - \overline{x}\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i\hat{x}_i - \overline{y}\,\overline{x}}{\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i^2 - \overline{x}^2} = \frac{s_{XY}}{s_X^2}$$

Aufgabe E.14

Mit den Verkehrsdaten vom 1. bis zum 10. April 2001 ergeben sich die folgenden Werte:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i = 33'377 \qquad \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i^2 = 1'116'363'757$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = 28'883 \qquad \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{y}_i = 967'681'266$$

$$\Rightarrow \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \hat{x}_i - \bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i^2 - \bar{x}^2} = \frac{967'681'266 - 28'883 \cdot 33'377}{1'116'363'757 - (33'377)^2} = 1.554$$

$$\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x} = 28'883 - 1.554 \cdot 33'377 = -22'986$$

Aufgabe E.14

Die Regressionsgerade hat die folgende Form:

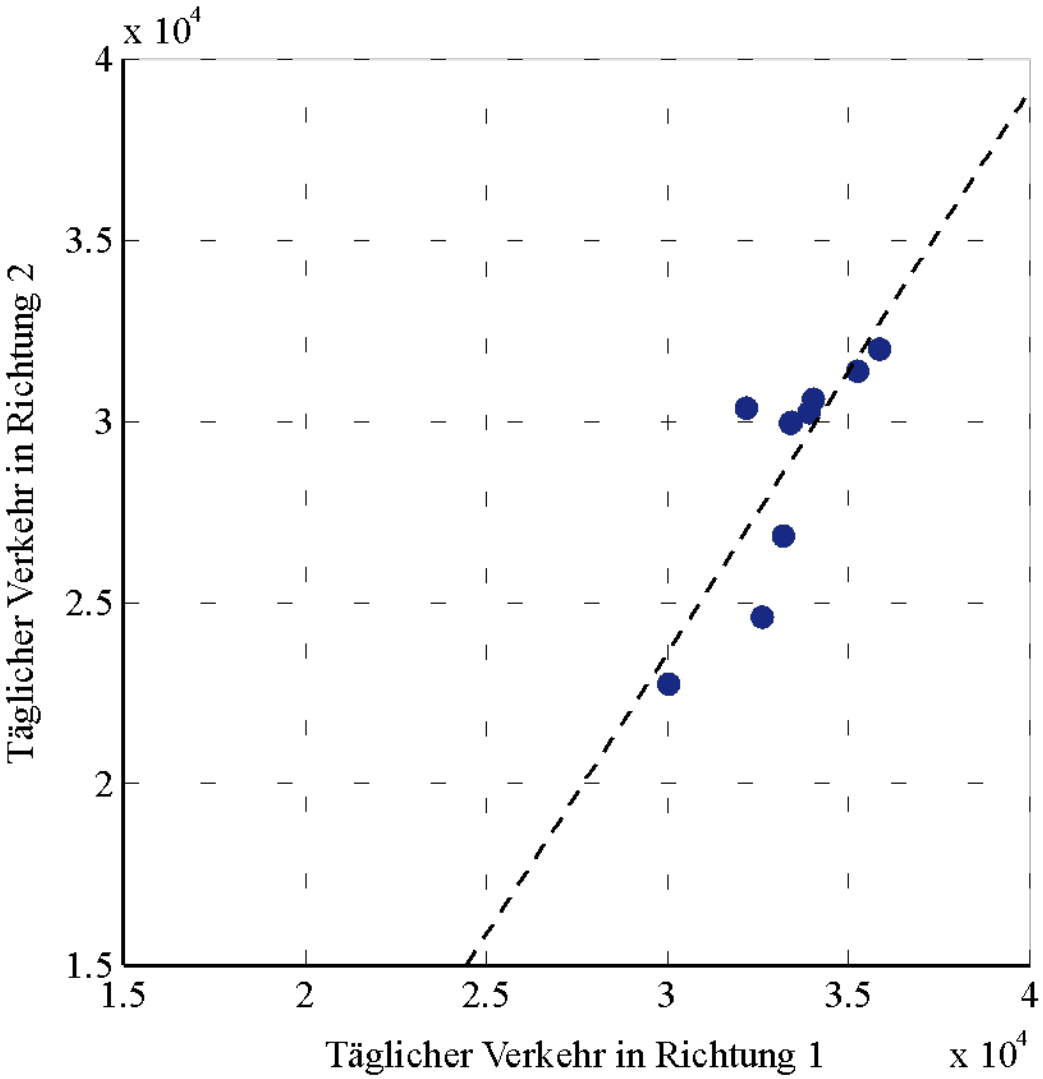
$$y_i = -22'986 + 1.554 x_i + \varepsilon_i$$

Der Zusammenhang ist jedoch nicht deterministisch:

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma_\varepsilon)$$

$$\varepsilon \sim N(0, \sigma_\varepsilon)$$

Residualwert, Fehler



Aufgabe E.14

Anhand der Summe der quadratischen Fehler kann man beurteilen, wie gut das Regressionsmodell die Daten abbilden kann.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \left(\beta_0 + \beta_1 \hat{x}_i \right) \right)^2 = 28'814'936$$

Berechnung der Standardabweichung des Fehlers:

$$\sigma_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-k}} = \sqrt{\frac{28'814'936}{10-2}} = \sqrt{3'601'867} = 1897.9$$

Anzahl Daten

Anzahl Parameter

Aufgabe E.14

Auf das gleiche Ergebnis kommt man mit der Matrix-Schreibweise:

Schätzung der Regressionskoeffizienten:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \left(\hat{\mathbf{X}}^T \hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}^T \hat{\mathbf{y}}$$

Quantifizierung der statistischen Unsicherheit:

$$\sigma_\varepsilon^2 = \left(\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\right)^T \left(\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\right) / \left(n - k\right)$$

$$Cov\left(\boldsymbol{\beta}\right) = \sigma_\varepsilon^2 \mathbf{V}_\beta \qquad \mathbf{V}_\beta = \left(\hat{\mathbf{X}}^T \hat{\mathbf{X}}\right)^{-1}$$

$$\hat{\mathbf{X}} = \begin{pmatrix} 1 & 32618 \\ 1 & 33380 \\ 1 & 34007 \\ 1 & 33888 \\ 1 & 35237 \\ 1 & 35843 \\ 1 & 33197 \\ 1 & 30035 \\ 1 & 32158 \\ 1 & 33406 \end{pmatrix} \qquad \hat{\mathbf{y}} = \begin{pmatrix} 24609 \\ 29965 \\ 30629 \\ 30263 \\ 31405 \\ 31994 \\ 26846 \\ 22762 \\ 30366 \\ 29994 \end{pmatrix}$$

Aufgabe E.14

$$\hat{\mathbf{X}}^T\hat{\mathbf{X}} = \begin{pmatrix} 1 & \cdots & 1 \\ \hat{x}_1 & \cdots & \hat{x}_n \end{pmatrix} \begin{pmatrix} 1 & \hat{x}_1 \\ \vdots & \vdots \\ 1 & \hat{x}_i \\ \vdots & \vdots \\ 1 & \hat{x}_n \end{pmatrix} = \begin{pmatrix} n \cdot 1^2 & \sum_{i=1}^{n} \hat{x}_i \\ \sum_{i=1}^{n} \hat{x}_i & \sum_{i=1}^{n} \hat{x}_i^2 \end{pmatrix} = \begin{pmatrix} 10 & 333'769 \\ 333'769 & 11'163'637'569 \end{pmatrix}$$

$$\mathbf{V}_\beta = \left(\hat{\mathbf{X}}^T\hat{\mathbf{X}}\right)^{-1} = \begin{pmatrix} 10 & 333'769 \\ 333'769 & 11'163'637'569 \end{pmatrix}^{-1} = \begin{pmatrix} 47.58 & -1.42 \cdot 10^{-3} \\ -1.42 \cdot 10^{-3} & 4.26 \cdot 10^{-8} \end{pmatrix}$$

Aufgabe E.14

$$\hat{\mathbf{X}}^T\hat{\mathbf{y}} = \begin{pmatrix} 1 & \cdots & 1 \\ \hat{x}_1 & \cdots & \hat{x}_n \end{pmatrix} \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \hat{y}_i \\ \sum_{i=1}^{n} \hat{x}_i \hat{y}_i \end{pmatrix} = \begin{pmatrix} 288'833 \\ 9'676'812'660 \end{pmatrix}$$

Nun lassen sich die Regressionskoeffizienten bestimmen:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \left(\hat{\mathbf{X}}^T\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}^T\hat{\mathbf{y}}$$

$$= \begin{pmatrix} 47.58 & -1.42\cdot 10^{-3} \\ -1.42\cdot 10^{-3} & 4.26\cdot 10^{-8} \end{pmatrix} \begin{pmatrix} 288'833 \\ 9'676'812'660 \end{pmatrix} = \begin{pmatrix} -22'986 \\ 1.554 \end{pmatrix}$$

Aufgabe E.14

Die Fehlervarianz berechnet sich wie zuvor:

$$\sigma_\varepsilon^2 = \frac{1}{n-k}\left(\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\right)^T\left(\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\right) = \frac{1}{n-k}\sum_{i=1}^{n}\left(\hat{y}_i - \left(\beta_0 + \beta_1\hat{x}_i\right)\right)^2$$

$$= 3'601'867$$

Hieraus ergibt sich die Kovarianz-Matrix der Modellparameter:

$$Cov\left(\boldsymbol{\beta}\right) = \sigma_\varepsilon^2\mathbf{V}_\beta = 3'601'867\begin{pmatrix} 10 & 333'769 \\ 333'769 & 11'163'637'569 \end{pmatrix}$$

$$= \begin{pmatrix} 1.71\cdot 10^8 & -5.12\cdot 10^3 \\ -5.12\cdot 10^3 & 0.154 \end{pmatrix}$$

Aufgabe E.14

Regressionsmodell:

$$Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma_\varepsilon)$$

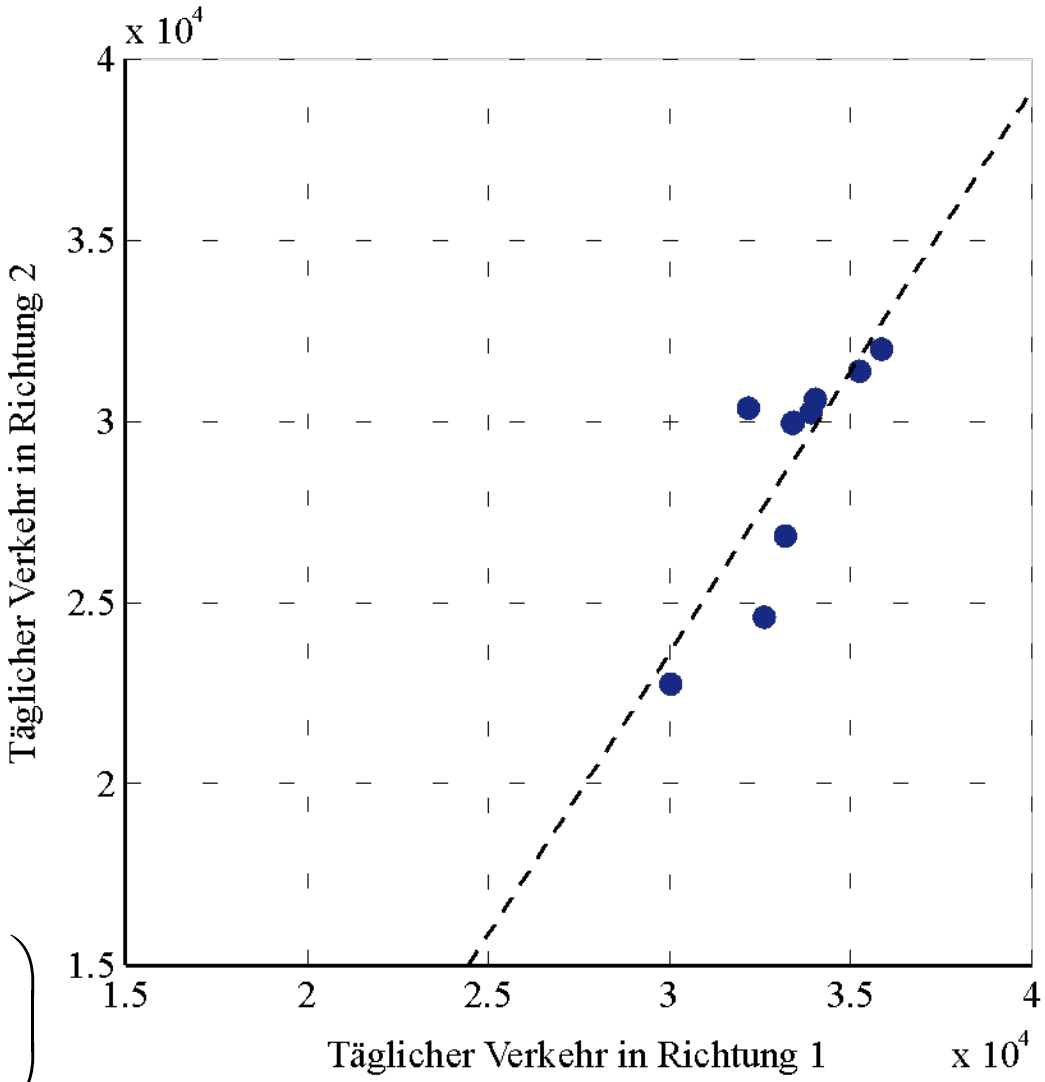
Regressionskoeffizienten:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} -22'986 \\ 1.554 \end{pmatrix}$$

Unsicherheit des Modells:

$$\sigma_\varepsilon = 1897.9$$

$$Cov(\boldsymbol{\beta}) = \begin{pmatrix} 1.71 \cdot 10^8 & -5.12 \cdot 10^3 \\ -5.12 \cdot 10^3 & 0.154 \end{pmatrix}$$



Aufgabe E.14

b) Das Regressionsmodell soll nun mit neuen Daten aktualisiert werden. Bestimme die a posteriori Regressionskoeffizienten und quantifiziere die Unsicherheit des aktualisierten Regressionsmodells. Verwende hierzu die zweite Hälfte des Datensatzes.

$$\boldsymbol{\beta}'' = \mathbf{V}_\beta''\left(\left(\mathbf{V}_\beta'\right)^{-1}\boldsymbol{\beta}' + \hat{\mathbf{X}}_n{}^T\hat{\mathbf{y}}_n\right) \qquad \left(\mathbf{V}_\beta''\right)^{-1} = \left(\mathbf{V}_\beta'\right)^{-1} + \hat{\mathbf{X}}_n^T\hat{\mathbf{X}}_n$$

A posteriori Modell — A priori Modell — neue Daten

Die Informationen aus dem bestehenden a priori Regressionsmodell und den neuen Daten werden für das a posteriori Modell kombiniert.

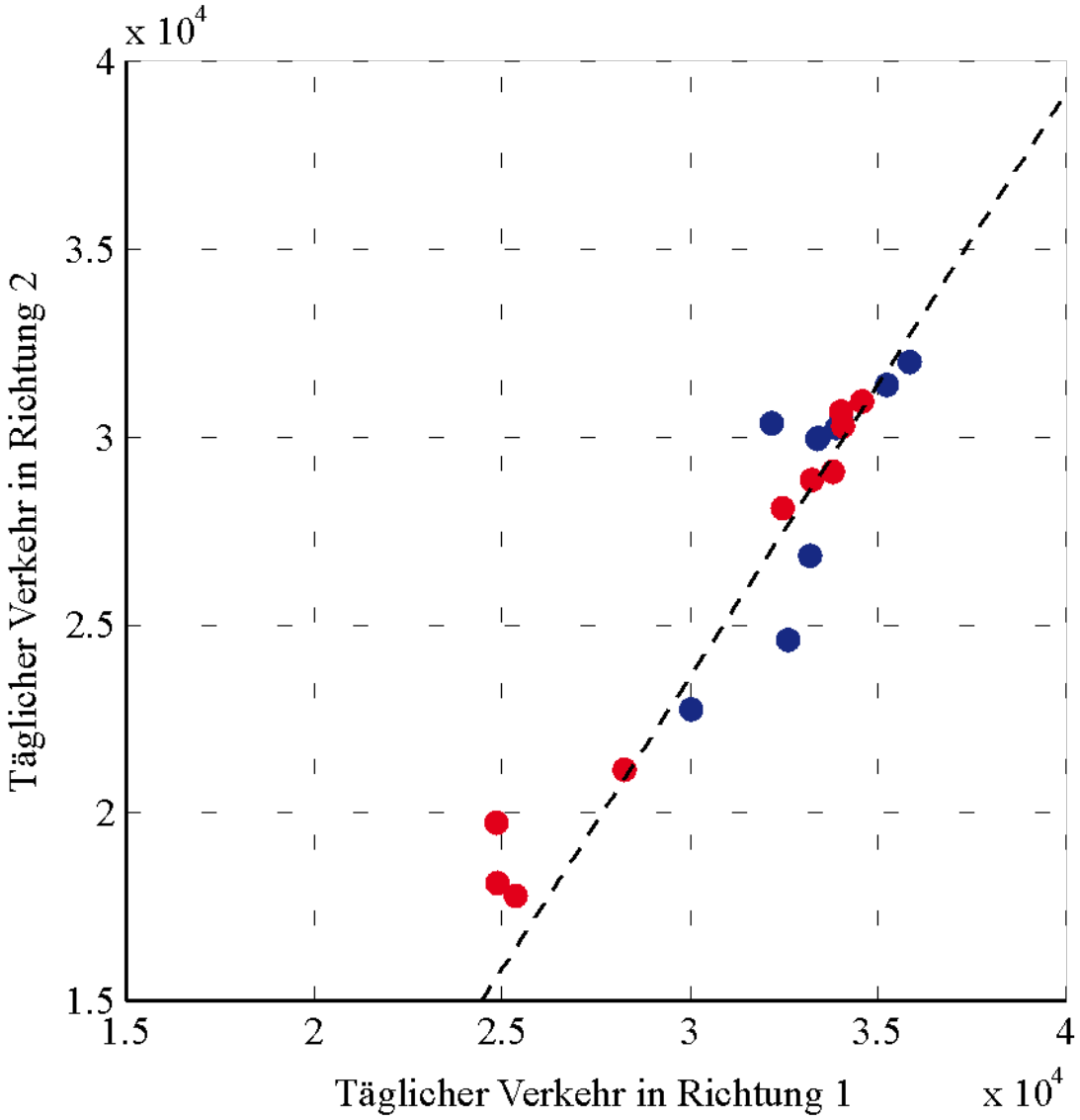
Aufgabe E.14

Das a priori Modell aus Teilaufgabe a) soll mit neuen Daten aktualisiert werden.

Das a posteriori Modell wird folgendermassen bestimmt:

$$\boldsymbol{\beta}'' = \mathbf{V}_\beta''\left(\left(\mathbf{V}_\beta'\right)^{-1}\boldsymbol{\beta}' + \hat{\mathbf{X}}_n^T\hat{\mathbf{y}}_n\right)$$

$$\left(\mathbf{V}_\beta''\right)^{-1} = \left(\mathbf{V}_\beta'\right)^{-1} + \hat{\mathbf{X}}_n^T\hat{\mathbf{X}}_n$$



Aufgabe E.14

Das a priori Modell aus Teilaufgabe a) soll mit neuen Daten aktualisiert werden.

Das a posteriori Modell wird folgendermassen bestimmt:

$$\boldsymbol{\beta}'' = \mathbf{V}_\beta'' \left(\left(\mathbf{V}_\beta' \right)^{-1} \boldsymbol{\beta}' + \hat{\mathbf{X}}_n^T \hat{\mathbf{y}}_n \right)$$

$$\left(\mathbf{V}_\beta'' \right)^{-1} = \left(\mathbf{V}_\beta' \right)^{-1} + \hat{\mathbf{X}}_n^T \hat{\mathbf{X}}_n$$

ACHTUNG:

Im Skript sind diese Gleichungen falsch angegeben – Korrekturseiten sind auf der Homepage

(betrifft Gleichungen E.45 und E.46, sowie das Beispiel E.4)

Aufgabe E.14

Zunächst berechnen wir $\hat{\mathbf{X}}_n^T \hat{\mathbf{X}}_n$ und $\hat{\mathbf{X}}_n^T \hat{\mathbf{y}}_n$ für die neuen Daten:

$$\hat{\mathbf{X}}_n = \begin{pmatrix} 1 & 34576 \\ 1 & 34013 \\ 1 & 24846 \\ 1 & 28252 \\ 1 & 25365 \\ 1 & 24862 \\ 1 & 32472 \\ 1 & 33245 \\ 1 & 33788 \\ 1 & 34076 \end{pmatrix} \qquad \hat{\mathbf{y}}_n = \begin{pmatrix} 30958 \\ 30680 \\ 19735 \\ 21145 \\ 17805 \\ 18123 \\ 28117 \\ 28858 \\ 29080 \\ 30313 \end{pmatrix}$$

$$\hat{\mathbf{X}}_n^T \hat{\mathbf{X}}_n = \begin{pmatrix} n \cdot 1^2 & \sum_{i=1}^n \hat{x}_i \\ \sum_{i=1}^n \hat{x}_i & \sum_{i=1}^n \hat{x}_i^2 \end{pmatrix} = \begin{pmatrix} 10 & 305'495 \\ 305'495 & 9'491'848'963 \end{pmatrix}$$

$$\hat{\mathbf{X}}_n^T \hat{\mathbf{y}}_n = \begin{pmatrix} \sum_{i=1}^n \hat{y}_i \\ \sum_{i=1}^n \hat{x}_i \hat{y}_i \end{pmatrix} = \begin{pmatrix} 254'814 \\ 7'991'745'111 \end{pmatrix}$$

Aufgabe E.14

Die neuen Daten werden mit dem a priori Modell kombiniert:

$$\left(\mathbf{V}_\beta''\right)^{-1} = \overbrace{\left(\mathbf{V}_\beta'\right)^{-1}}^{\text{A priori}} + \overbrace{\hat{\mathbf{X}}_n^T\hat{\mathbf{X}}_n}^{\text{neu}} = \begin{pmatrix} 10 & 333'769 \\ 333'769 & 11'163'637'569 \end{pmatrix} + \begin{pmatrix} 10 & 305'495 \\ 305'495 & 9'491'848'963 \end{pmatrix}$$

$$= \begin{pmatrix} 20 & 639'264 \\ 639'264 & 20'655'486'532 \end{pmatrix}$$

$$\mathbf{V}_\beta'' = \begin{pmatrix} 20 & 639'264 \\ 639'264 & 20'655'486'532 \end{pmatrix}^{-1} = \begin{pmatrix} 4.64 & -1.44\cdot10^{-4} \\ -1.44\cdot10^{-4} & 4.49\cdot10^{-9} \end{pmatrix}$$

Nun können die a posteriori Regressionskoeffizienten bestimmt werden:

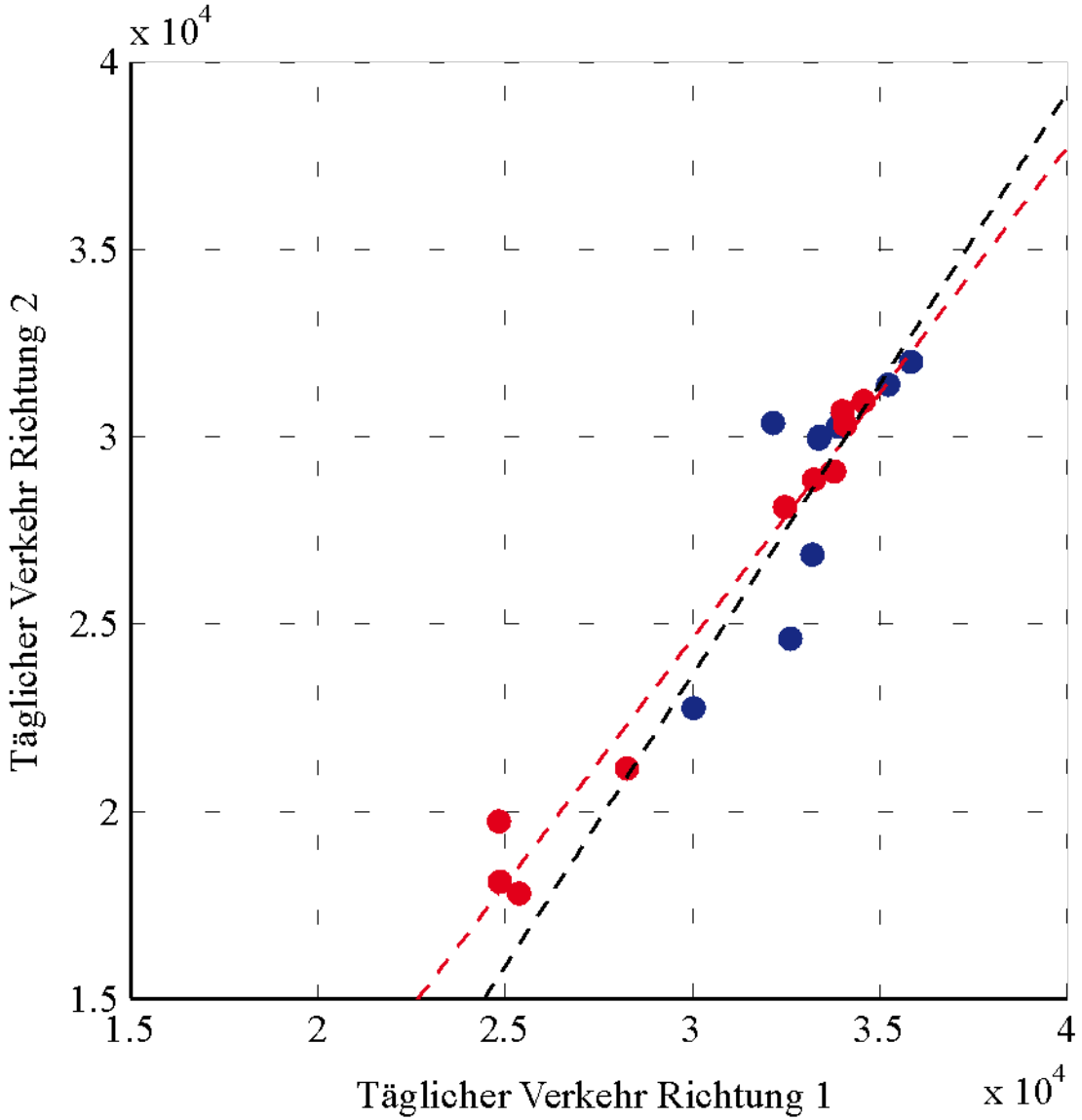
$$\boldsymbol{\beta}'' = \mathbf{V}_\beta''\left(\left(\mathbf{V}_\beta'\right)^{-1}\boldsymbol{\beta}' + \hat{\mathbf{X}}_n^T\hat{\mathbf{y}}\right) = \begin{pmatrix} -14'733 \\ 1.311 \end{pmatrix} = \begin{pmatrix} \beta_0'' \\ \beta_1'' \end{pmatrix}$$

Aufgabe E.14

Das a priori Modell wurde mit den neuen Daten aktualisiert.

Die a posteriori Regressionsgerade hat die folgende Form:

$$y_i = -14'733 + 1.311 x_i + \varepsilon_i$$



Stichprobenstatistiken

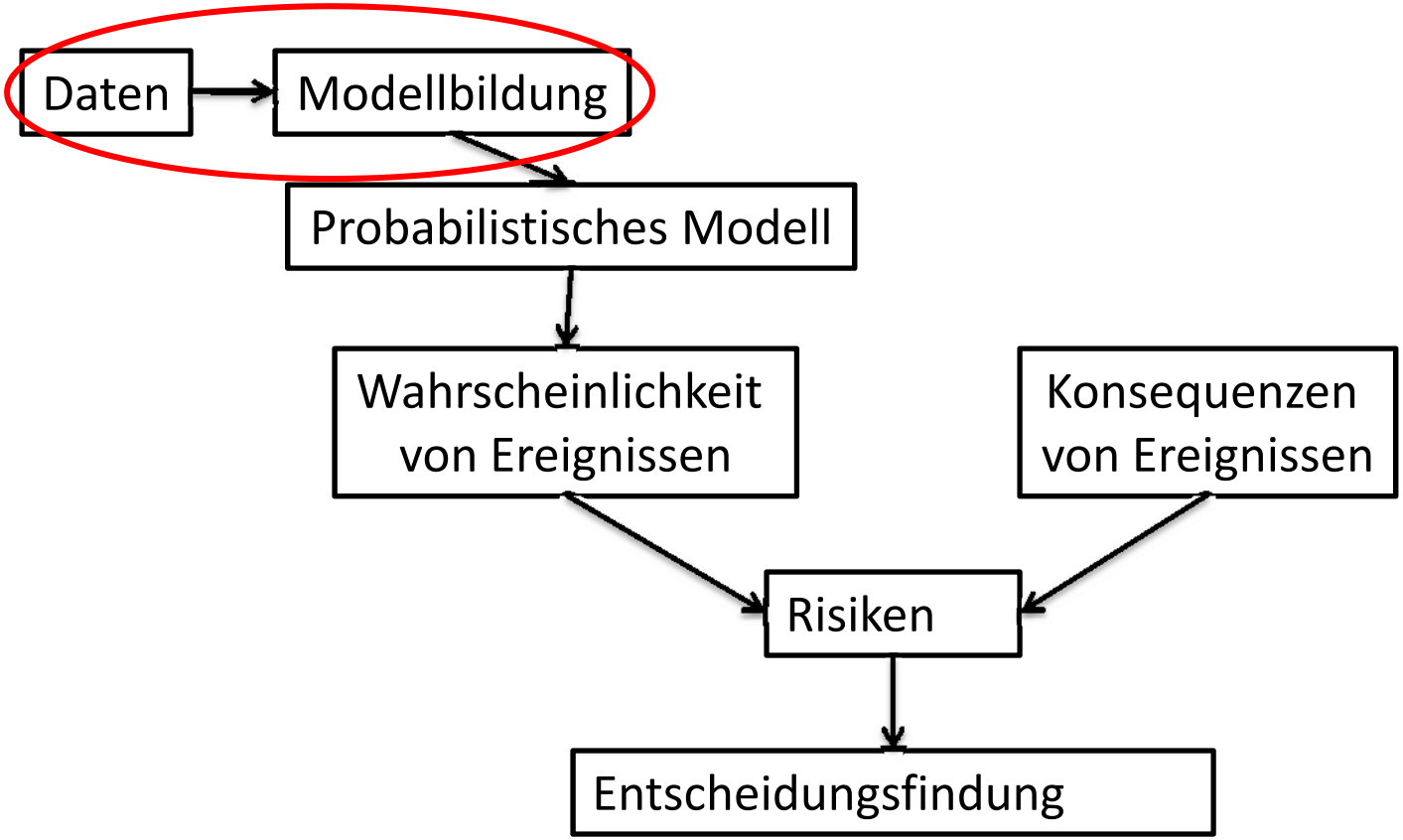
Problemstellung

Wir wollen die Lage und Variabilität von Grundgesamtheiten beschreiben,

z.B. das Körpergewicht aller Studierenden an Schweizer Hochschulen im 2. Semester.

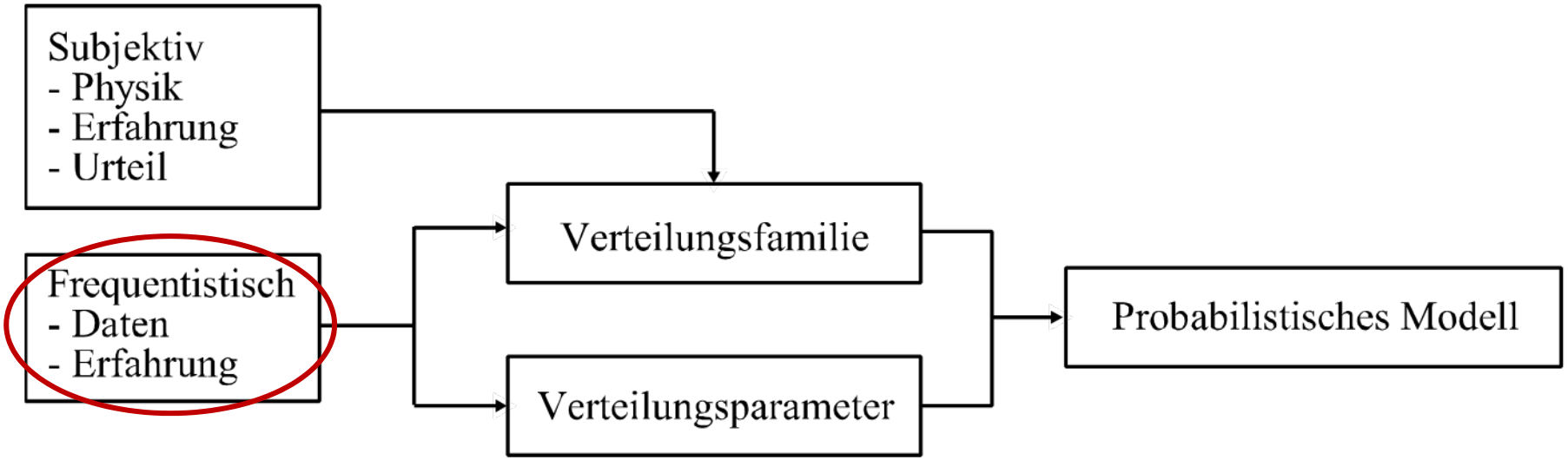
Stichprobenstatistiken

Problemstellung



Stichprobenstatistiken

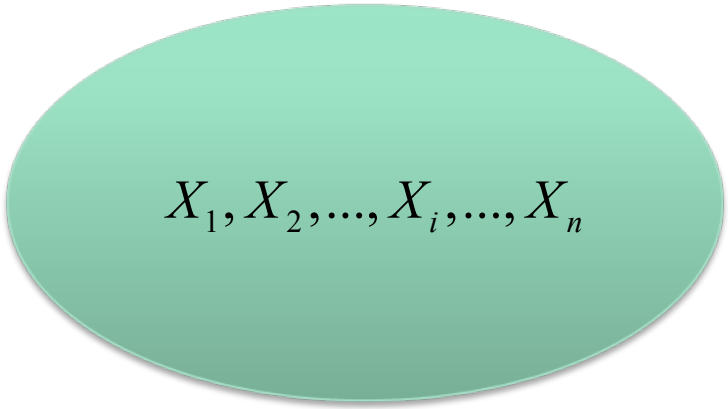
Problemstellung



Stichprobenstatistiken

Wir beschreiben das Körpergewicht mit einer Zufallsvariablen. Die Zufallsvariable hat einen Mittelwert und eine Standardabweichung.

Die Grundgesamtheit besteht aus einer Sequenz identisch verteilter Zufallsvariablen X_i .

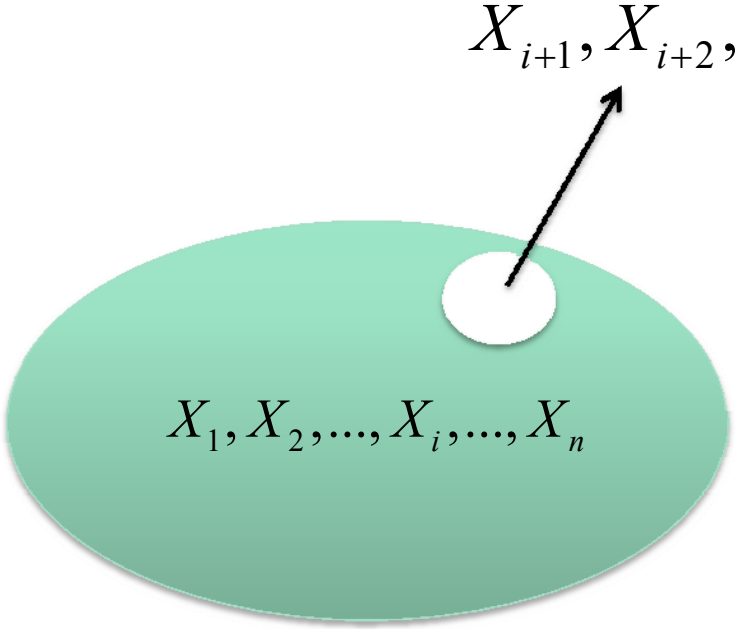


Stichprobenstatistiken

Wir wollen nun den Mittelwert und die Standardabweichung abschätzen – anhand einer Stichprobe mit Umfang n = 10.

$$E\left[\bar{X}\right] = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$Var\left[\bar{X}\right] = \frac{1}{n}\sigma_X^2$$



$$E\left[S^2_{erwartungstreu}\right] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sigma_X^2$$

$$Var\left[S^2_{erwartungstreu}\right] = \frac{2(n-1)}{n^2}\sigma_X^4$$

Aufgabe E.9

Von allen Studenten der ETH Zürich wurde das Gewicht in Kilogramm gemessen. Die Entnahme einer Stichprobe $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)$ ergibt folgende Werte: $95, 77, 83, 71$ [kg].

Das Gewicht der Studenten kann als Zufallsvariable X modelliert werden. Die Standardabweichung $\sigma_X = 9$ [kg] sei bekannt.

a. Schätze anhand der Stichprobe den Erwartungswert $E(\overline{X})$ und die Varianz $Var(\overline{X})$ des Stichprobenmittelwertes $\overline{X}$.
b. Stelle die Streuung des Mittelwertes grafisch dar.
c. Ermittle den Bereich, in dem der wahre Mittelwert μ_X mit 95% Konfidenz zu erwarten ist.

Aufgabe E.9

a. Schätze anhand der Stichprobe den Erwartungswert $E(\overline{X})$ und die Varianz $Var(\overline{X})$ des Stichprobenmittelwertes $\overline{X}$.

$$\sigma_X = 9 \text{ [kg]}$$

$$E\left[\overline{X}\right] = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{4}326 = 81.5$$

$$Var\left[\overline{X}\right] = \frac{1}{n}\sigma_X^2 = \frac{1}{4}81 = 20.25$$

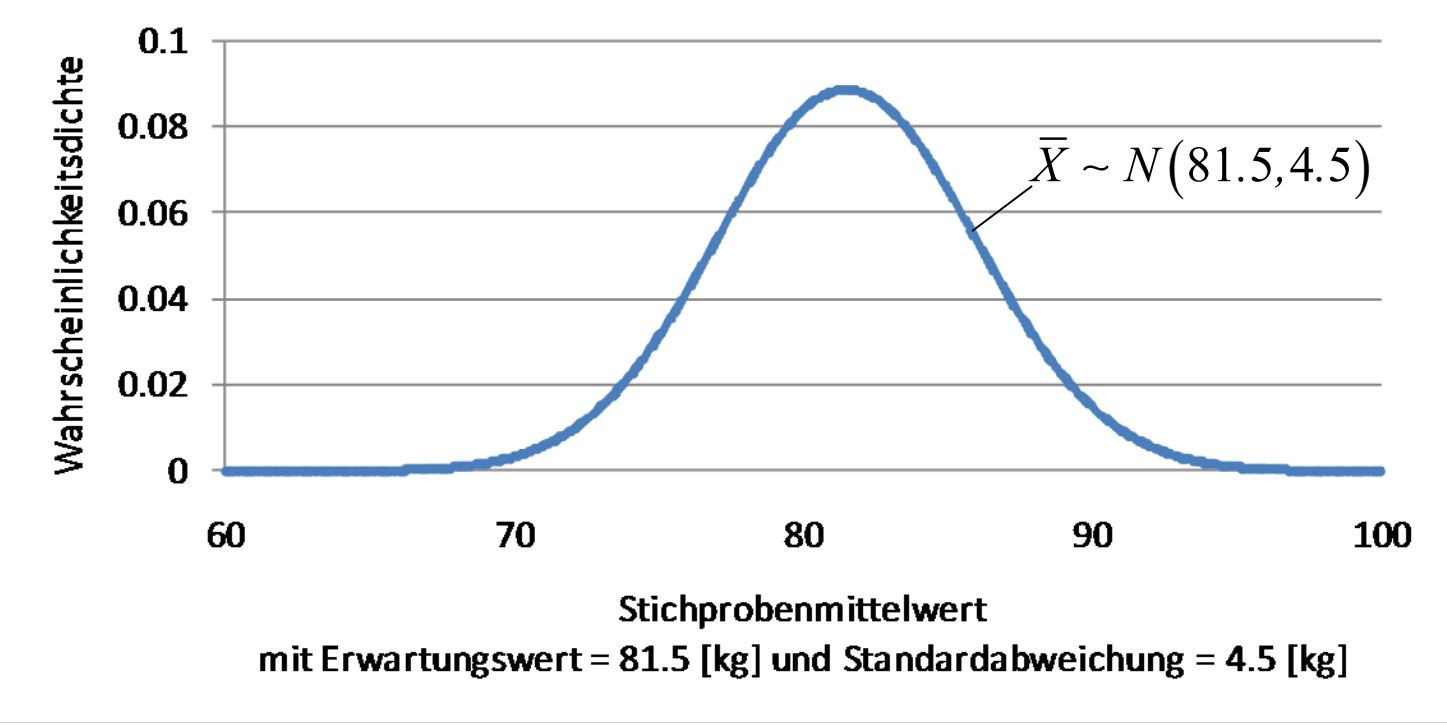
Aufgabe E.9

b. Stelle die Streuung des Mittelwertes grafisch dar.

Annahme: sei $\bar{X}$ normalverteilt

$$E\left[\bar{X}\right] = 81.5$$

$$Var\left[\bar{X}\right] = 20.25$$

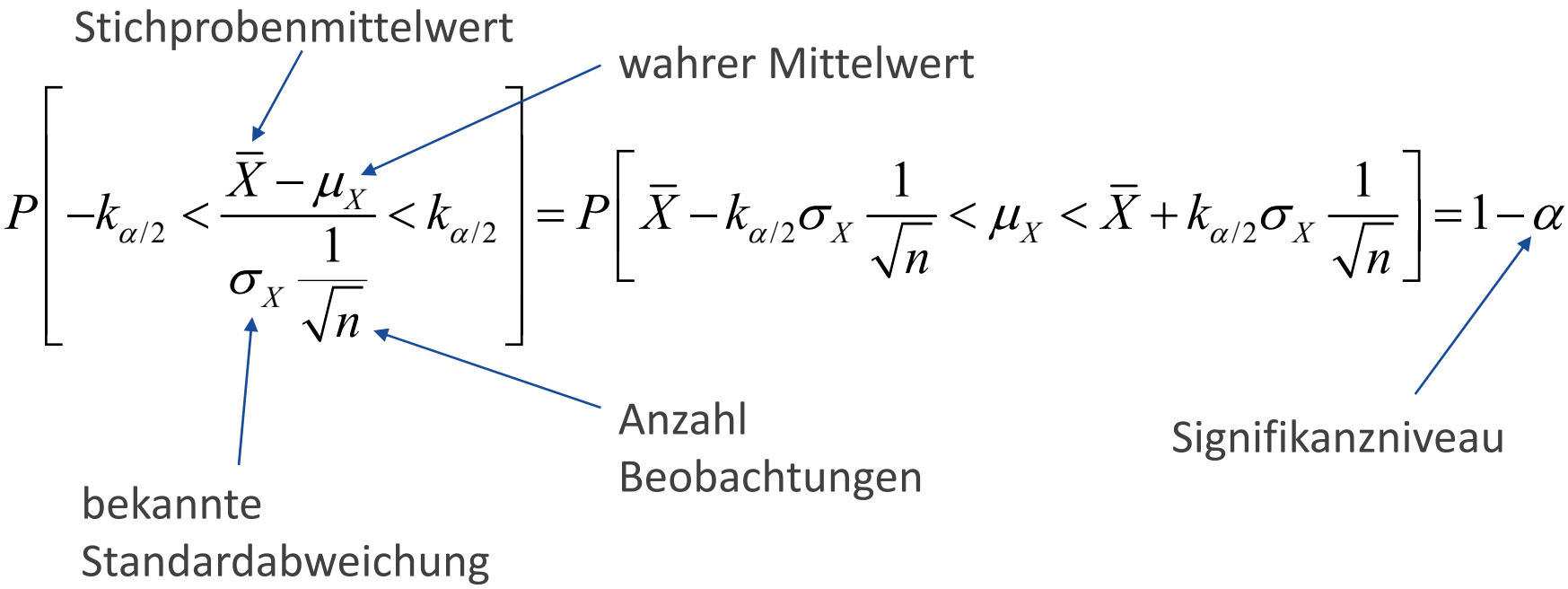


Aufgabe E.9

$\bar{x} = 81.5 \quad \sigma_X = 9$

$n = 4 \quad \alpha = 0.05$

c. Ermittle den Bereich, in dem der wahre Mittelwert μ_X mit 95% Konfidenz zu erwarten ist.



Aufgabe E.9

$\bar{x} = 81.5 \quad \sigma_X = 9$

$n = 4 \quad \alpha = 0.05$

c. Ermittle den Bereich, in dem der wahre Mittelwert μ_X mit 95% Konfidenz zu erwarten ist.

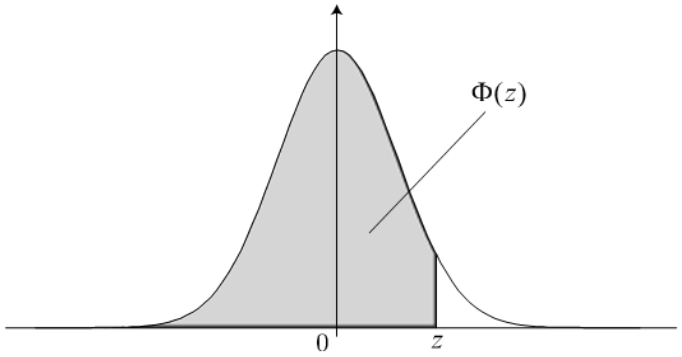
$$P\left[\bar{X} - k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}} < \mu_X < \bar{X} + k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}}\right] = 1 - \alpha$$

$$P\left[81.5 - k_{\alpha/2} 9 \frac{1}{\sqrt{4}} < \mu_X < 81.5 + k_{\alpha/2} 9 \frac{1}{\sqrt{4}}\right] = 0.95$$

$$k_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \Phi^{-1}\left(1 - \frac{0.05}{2}\right) = \Phi^{-1}(0.975) = TABELLE...$$

Aufgabe E.9

Skript Anhang T1



Wahrscheinlichkeitsdichtefunktion der Standardnormalverteilung.

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.31	0.6217	0.81	0.7910	1.31	0.9049	1.81	0.9649		
0.32	0.6255	0.82	0.7939	1.32	0.9066	1.82	0.9656		
0.33	0.6293	0.83	0.7967	1.33	0.9082	1.83	0.9664		
0.34	0.6331	0.84	0.7995	1.34	0.9099	1.84	0.9671		
0.35	0.6368	0.85	0.8023	1.35	0.9115	1.85	0.9678		
0.36	0.6406	0.86	0.8051	1.36	0.9131	1.86	0.9686		
0.37	0.6443	0.87	0.8078	1.37	0.9147	1.87	0.9693		
0.38	0.6480	0.88	0.8106	1.38	0.9162	1.88	0.9699		
0.39	0.6517	0.89	0.8133	1.39	0.9177	1.89	0.9706		
0.40	0.6554	0.90	0.8159	1.40	0.9192	1.90	0.9713		
0.41	0.6591	0.91	0.8186	1.41	0.9207	1.91	0.9719		
0.42	0.6628	0.92	0.8212	1.42	0.9222	1.92	0.9726		
0.43	0.6664	0.93	0.8238	1.43	0.9236	1.93	0.9732		
0.44	0.6700	0.94	0.8264	1.44	0.9251	1.94	0.9738		
0.45	0.6736	0.95	0.8289	1.45	0.9265	1.95	0.9744		
0.46	0.6772	0.96	0.8315	1.46	0.9279	1.96	0.9750		
0.47	0.6808	0.97	0.8340	1.47	0.9292	1.97	0.9756		
0.48	0.6844	0.98	0.8365	1.48	0.9306	1.98	0.9761		
0.49	0.6879	0.99	0.8389	1.49	0.9319	1.99	0.9767		

Aufgabe E.9

$$\bar{x} = 81.5 \quad \sigma_X = 9$$
$$n = 4 \quad \alpha = 0.05$$

c. Ermittle den Bereich, in dem der wahre Mittelwert μ_X mit 95% Konfidenz zu erwarten ist.

$$P\left[81.5 - k_{\alpha/2} 9\frac{1}{\sqrt{4}} < \mu_X < 81.5 + k_{\alpha/2} 9\frac{1}{\sqrt{4}}\right] = 0.95$$

$$k_{\alpha/2} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right) = \Phi^{-1}\left(1-\frac{0.05}{2}\right) = \Phi^{-1}(0.975) = 1.96$$

$$P\left[81.5 - 1.96 \cdot 9\frac{1}{\sqrt{4}} < \mu_X < 81.5 + 1.96 \cdot 9\frac{1}{\sqrt{4}}\right] = 0.95$$

$$\underline{\underline{P[72.68 < \mu_X < 90.32] = 0.95}}$$

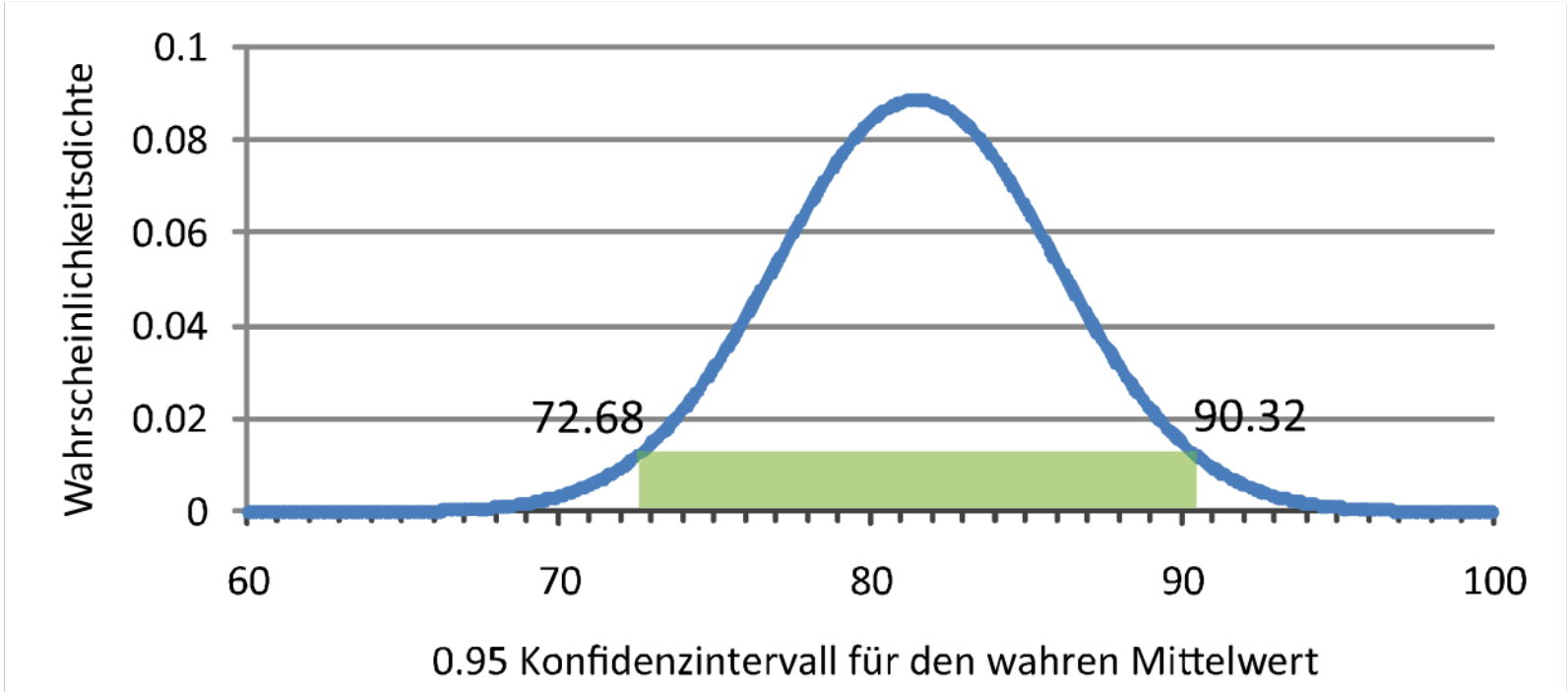
Aufgabe E.9

$\bar{x} = 81.5 \quad \sigma_X = 9$

$n = 4 \quad \alpha = 0.05$

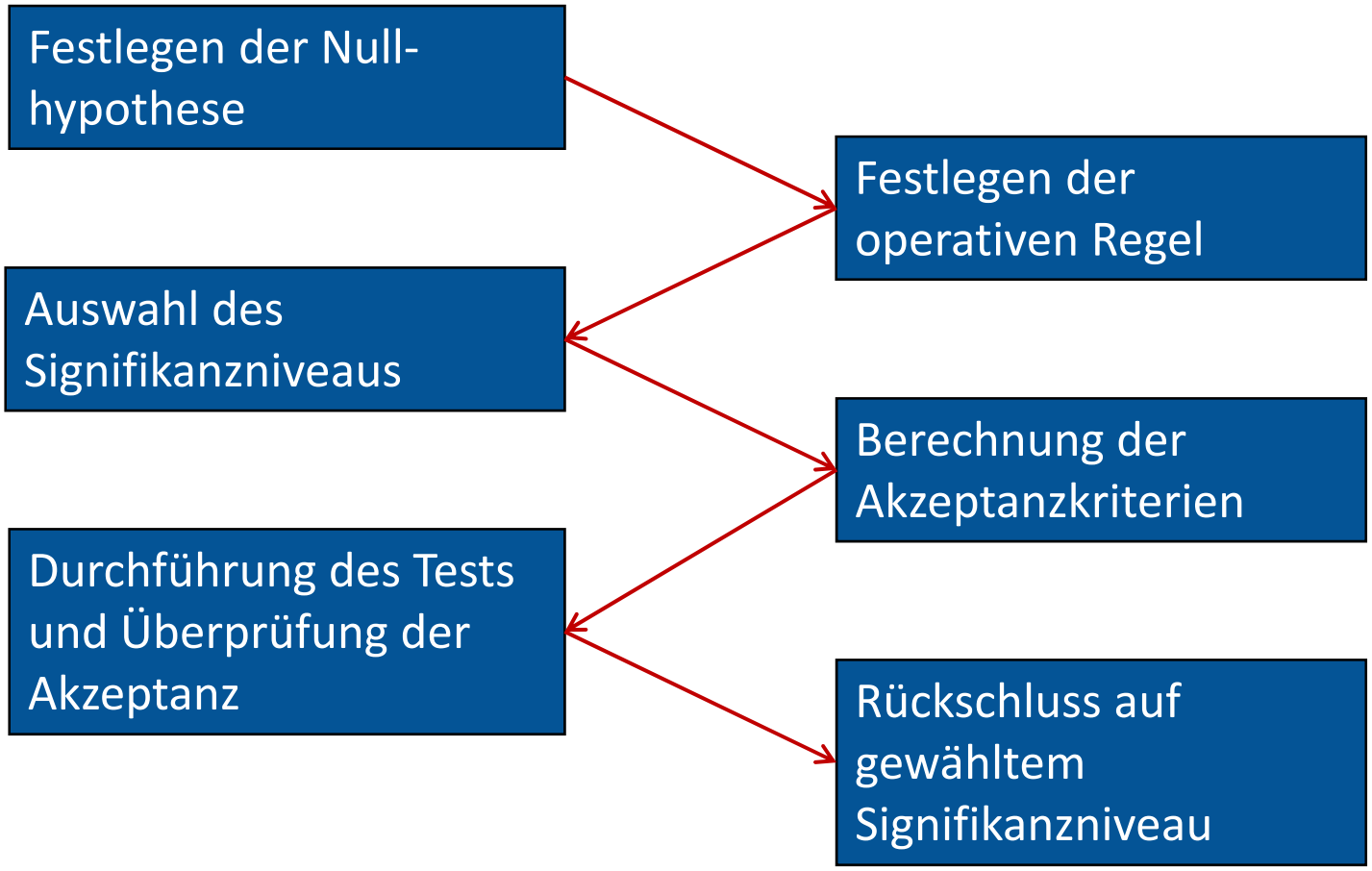
$$P\left[72.68 < \mu_X < 90.32\right] = 0.95$$

Das 0.95-Konfidenzintervall für den wahren Mittelwert liegt zwischen 72.68 kg und 90.32 kg.



Testen von Hypothesen

Generelles Vorgehen beim Hypothesentest



Konfidenzintervall

- Stichprobenstatistiken beinhalten Unsicherheiten – die Angabe von Intervallen ist notwendig, in denen ein Parameter in $(1-\alpha)\%$ der Fälle liegt.

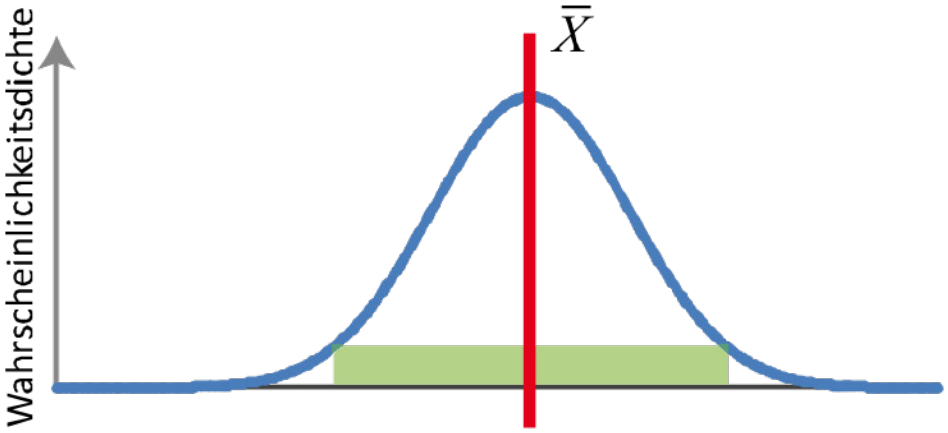
Hypothesentest

- Dient der Aussage, ob aufgrund vollzogener Beobachtungen ein Wert signifikant vom angenommenen Wert der Grundgesamtheit abweicht.

- Konfidenzintervalle können auch als Grundlage für Hypothesentests dienen:
  Testen wir z.B. die Nullhypothese $\mu_X = 23.7$, dann ist das $(1-\alpha)$ Kriterium für das Verwerfen der Nullhypothese, dass das Konfidenzintervall den Wert 23.7 nicht enthält.

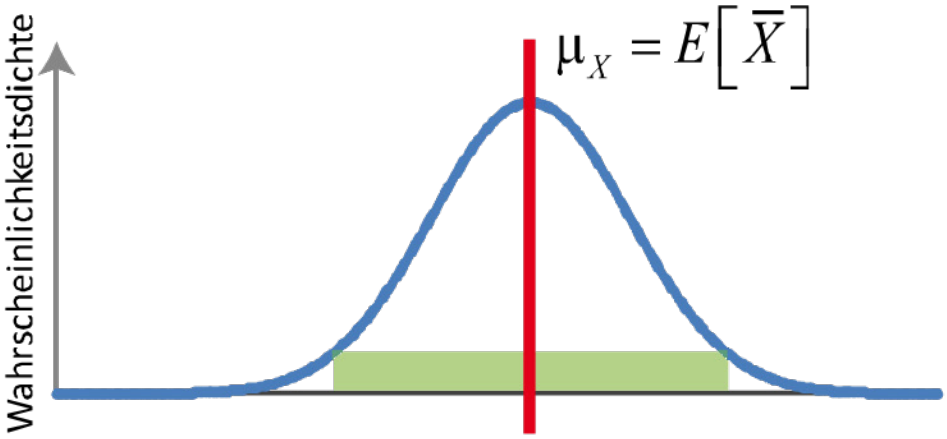
Konfidenzintervall  Hypothesentest

$$P\left[\bar{X}-\Delta<\mu_X<\bar{X}+\Delta\right]=1-\alpha$$



Konfidenzintervall für den wahren Mittelwert

$$P\left[\mu_X-\Delta<\bar{X}<\mu_X+\Delta\right]=1-\alpha$$



Intervall für den Stichprobenmittelwert
zum Prüfen einer Hypothese

Gruppenaufgabe E.1

Um die Qualität des Betons auf einer Baustelle zu prüfen, wird die Druckfestigkeit des hergestellten Betons getestet.

Erfahrungsgemäss ist bekannt, dass die Druckfestigkeit einer Normalverteilung folgt und die Varianz der Druckfestigkeit für diese Betonsorte 16.36 [MPa^2] ist.

Akzeptanzkriterium für die Qualität des Betons auf der Baustelle ist, dass der Mittelwert des Betons gleich $30 \pm \Delta \ [\mathrm{MPa}]$ ist. Dies wird täglich am jeweils hergestellten Beton gemessen. Um die Homogenität der Verarbeitung zu gewährleisten, sind sowohl kleinere wie auch grössere Werte nicht akzeptabel.

Gruppenaufgabe E.1

Nummer der Probe (i)	Druckfestigkeit [MPa]
1	24.4
2	26.5
3	27.8
4	29.2
5	39.2
6	37.8
7	35.1
8	30.8
9	30.3
10	39.7
11	38.4
12	33.3
13	33.5
14	28.1
15	34.6

Aus einer Tagesproduktion werden 15 Proben entnommen und ihre Druckfestigkeit getestet (siehe Tabelle)

Kann die Qualität des Betons akzeptiert werden?

Teste die Hypothese jeweils für ein Signifikanzniveau von 10 % und 1 %.

Vielen Dank für eure Aufmerksamkeit!

kraemer@ibk.baug.ethz.ch
HIL E 13.1