Nachholklausur Statistik
Lösungshinweise

Prüfungsdatum: 12. Januar 2018 – Prüfer: Etschberger, Ivanov, Jansen, Wesp, Wins
Studiengang: IM, BW, Inf und W-Inf
Punkte: 18, 11, 11, 17, 11, 11, 11 ; Summe der Punkte: 90

Aufgabe 1 — 18 Punkte

Eine Befragung unter 100 Studierenden ergab für das Merkmal

$X :=$ „*wöchentliche Vor- und Nachbearbeitungszeit zur Statistikveranstaltung (in Stunden)*“

die empirische Verteilungsfunktion F mit dem folgenden Funktionsterm:

$$F(x) = \begin{cases} 0 & \text{falls} \quad x < 3 \\ 0.10 & \text{falls} \quad 3 \leqq x < 6 \\ 0.30 & \text{falls} \quad 6 \leqq x < 7 \\ 0.65 & \text{falls} \quad 7 \leqq x < 9 \\ 0.85 & \text{falls} \quad 9 \leqq x < 12 \\ \boxed{0.95} & \text{falls} \quad \boxed{12 \leqq x < 15} \\ 1 & \text{falls} \quad \boxed{15} \leqq x \end{cases}$$

Zudem sei die Spannweite von X mit $\mathrm{SP}(X) = 12$ sowie $\sum_{i=1}^{100} x_i = 770$ bekannt.

a) Ergänzen Sie die fehlenden Angaben von $F(x)$ oben in den leeren Feldern.

*Hinweis: Falls Sie Teilaufgabe a) nicht lösen konnten (und nur dann!), rechnen Sie bitte weiter mit den **falschen** letzten beiden Zeilen der Definition von $F(x)$ gemäß nebenstehenden Angaben*

$$F(x) = \begin{cases} \vdots & \vdots \\ 0.99 & \text{falls } 12 \leqq x < 24 \\ 1.00 & \text{falls } 24 \leqq x \end{cases}$$

Bestimmen Sie für das Merkmal X

b) den Modus und das arithmetische Mittel,

c) die absolute Häufigkeit $h(9)$,

d) die relative Häufigkeit $f(8)$ und die kumulierte relative Häufigkeit $F(8)$,

e) das 25%-Quantil,

f) die erste Knickstelle der Lorenzkurve,

Die individuellen Vor- und Nachbearbeitungszeiten der Studierenden seien nun in `R` in einem Vektor `x` gespeichert. Geben Sie jeweils einen oder zwei `R`-Befehle an:

g) Zur Bestimmung der Tabelle der relativen Häufigkeiten der Merkmalsausprägungen.

```
prop.table(table(x)) oder table(x)/length(x)
```

h) Zur Visualisierung der empirischen Verteilungsfunktion als Funktionsgraph.

```
plot(ecdf(x))
```

i) Zur Berechnung des normierten Gini-Koeffizienten.

```
library(ineq); Gini(x, cor=TRUE)
```

j) Für ein Merkmal Y, dessen Ausprägungen in *R* im Objekt *y* gespeichert wurden, sei nun folgender *R*-Output gegeben. Zeichnen Sie den zugehörigen Boxplot.

```
table(y)

## y
##  0  1  2  3  4  6
##  1 13  7  1  2  1

quantile(y, probs=c(0.25, 0.5, 0.75), type=2)

## 25% 50% 75%
##   1   1   2
```

Lösungshinweis:

a) siehe oben

b) $x_{\text{mod}} = 7, \qquad \bar{x} = 7.7$

c) $h(9) = 20$

d) $f(8) = 0$
$F(8) = F(7) = 0.65$

e) $x_{0.25} = 6$

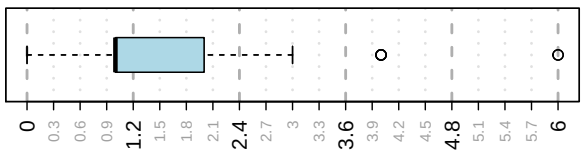
f) $u_1 = 0.1, v_1 = \frac{30}{770}$

g) siehe oben

h) siehe oben

i) siehe oben

j)
```
boxplot(y)
```



Aufgabe 2 — 11 Punkte

In den beiden R-Vektoren (jeweils mit der Länge n) und den Bezeichnern `bitcoin` bzw. `gold` sind für die Tage $1, \ldots, n$ die täglichen Euro-Preise für Bitcoins bzw. Gold gespeichert.

a) Geben Sie einen R-Befehl für die Berechnung eines geeigneten Korrelationskoeffizienten für den Zusammenhang zwischen den Preisen für Bitcoins und Gold an.

b) Geben Sie einen R-Befehl an, der ein Streuungsdiagramm für bitcoin und gold erstellt.

Es wird vermutet, dass der Bitcoin-Preis kausal vom Goldpreis abhängt. Um diesen Zusammenhang zu analysieren, wird mittels einfacher linearer Regression ein Modell geschätzt.

c) Geben Sie einen R-Befehl an, der dieses Modell schätzt und im Objekt `regression` speichert.

d) Geben Sie einen R-Befehl an, der in das Streuungsdiagramm aus b) die geschätzte Regressionsgerade aus c) einzeichnet.

e) Nehmen Sie an, das geschätzte Modell lautet

$$\hat{y} = -3 \cdot x + 4000\,.$$

Welcher Goldpreis führt in diesem Modell zu einem Bitcoin-Preis von 1000?

f) Der Determinationskoeffizient für das geschätzte Modell aus e) ist nun mit 20 % gegeben. Bestimmen Sie dazu rechnerisch den Korrelationskoeffizienten nach Bravais-Pearson.

1000 Anleger wurden gefragt, ob sie einen Teil ihres Vermögens in Bitcoin oder Gold angelegt haben. Aus der Umfrage ergab sich folgende Tabelle:

		Bitcoins	
		ja	nein
Gold	ja	1	500
	nein	450	49

g) Wie hoch ist der prozentuale Anteil der befragten Personen, die ihr Vermögen weder in Bitcoins noch in Gold angelegt haben?

h) Wie hoch ist der Anteil der Bitcoin-Anleger unter den Befragten, die einen Teil ihres Vermögens auch in Gold angelegt haben?

i) Berechnen Sie den normierten Kontingenzkoeffizienenten, wenn $\chi^2 = 817$.
(*Hinweis: Sie dürfen den Wert von χ^2 verwenden, er muss nicht nachgerechnet werden*)

Lösungshinweis:

a) `cor(bitcoin, gold)`

b) `regression = lm(bitcoin ~ gold)`

c) `plot(bitcoin, gold)`

d) `abline(regression)`

e) 1000

a) Da Zusammenhang negativ und $R^2 = 0.2 \Rightarrow r = -\sqrt{0.2} \approx -0.4472$

b) 0.049

c) $1/45 \approx 0.0022$

d) $K = \sqrt{\frac{817}{1000+817}} \approx 0.67$, $K^* = \sqrt{2} \cdot K \approx 0.9484$

Aufgabe 3 — 11 Punkte

Beim TicTacToe füllen zwei Spieler abwechselnd ein 3×3 großes Quadrat mit einem **x** (Spieler 1) oder einem **o** (Spieler 2). Gewonnen hat der Spieler, der es zuerst schafft eine Zeile, eine Spalte oder eine Diagonale mit seinem Symbol zu füllen.

1	2	3
4	5	6
7	8	9

a) Wie viele Möglichkeiten gibt es, das TicTacToe-Spielfeld mit 9 beliebigen Symbolen aus {**x**, **o**} zu füllen? Dabei sind auch Konfigurationen zugelassen, die nicht bei einem Spiel auftreten können (z.B. nur **x** auf dem Spielbrett).

b) Wie viele Möglichkeiten gibt es in einem echten Spiel? Beachten Sie, dass sowohl Spieler 1 als auch Spieler 2 beginnen kann und dass das Spiel immer gespielt wird bis alle 9 Felder gefüllt sind.

c) Xaver und Otto spielen TicTacToe. Momentan sind die beiden in einer Spielsituation wie in der Abbildung rechts. Dabei sind die numerierten Felder noch frei. Berechnen Sie die Wahrscheinlichkeit für einen Sieg von Xaver, einen Sieg von Otto oder ein Unentschieden je mit Hilfe eines Baumdiagramms, falls

o	2	**o**
4	**x**	**x**
x	8	**o**

(1) Xaver (**x**) begonnen hat.

(2) Otto (**o**) begonnen hat.

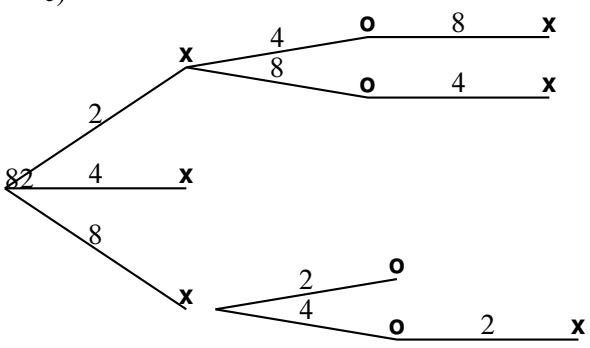
Gehen Sie dabei davon aus, dass die Wahrscheinlichkeit für das Befüllen eines Feldes $p = \frac{1}{n}$ für jedes Feld ist, wobei n die Anzahl der noch freien Felder ist (das heißt alle noch möglichen Felder sind gleich wahrscheinlich!)

Lösungshinweis:

a) Es gibt 9 Felder mit Je 2 Möglichkeiten, also $2^9 = 512$ Möglichkeiten.

b) Bei einem realen Spiel sind entweder 4 Felder mit **x** und 5 Felder mit **o** belegt oder umgekehrt. Es ergeben sich also $\binom{9}{4} + \binom{9}{5} = 126 + 126 = 252$

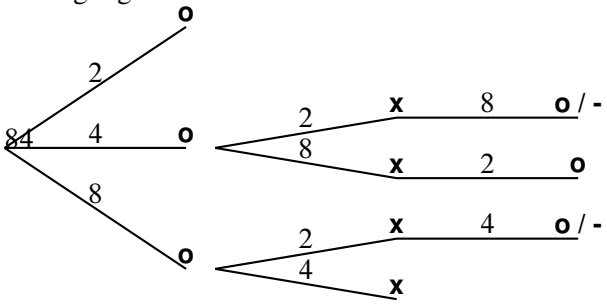
c)



Mit Ausnahme des (8, 2)-Astes gewinnt immer der **x**-Spieler. Es gibt keine Partien, die Unentschieden ausgehen. Damit ergibt sich:

$$P(\mathbf{x}\text{ gewinnt}) = 1/6 + 1/6 + 1/3 + 1/6 = 5/6$$
$$P(\mathbf{o}\text{ gewinnt}) = 1/6$$

Analog ergibt sich für den anderen Fall:



$$P(\mathbf{x}\text{ gewinnt}) = 1/6$$
$$P(\mathbf{o}\text{ gewinnt}) = 1/3 + 1/6 = 1/2$$
$$P(\text{unentschieden}) = 1/6 + 1/6 = 1/3$$

Aufgabe 4 — 17 Punkte

Die meisten Tachometer von Autos sind nicht sehr genau und zeigen etwas mehr als die tatsächlich gefahrene Geschwindigkeit an. Gehen Sie davon aus, dass die angezeigte Geschwindigkeit X bei einer tatsächlichen Geschwindigkeit von 100 km/h einer normalverteilten Zufallsvariable $X \sim N(\mu, \sigma)$ entspricht.

Von einem Autozulieferer werden die Tachometer so kalibriert, dass bei einer tatsächlichen Fahrtgeschwindigkeit von 100 km/h

- bei 1 % der Autos weniger als 100 km/h sowie
- bei 5 % der Autos mehr als 120 km/h angezeigt werden.

Gehen Sie im Folgenden von Geschwindigkeitsmessungen eines zufällig ausgewählten Tachometers bei einer tatsächlichen Fahrtgeschwindigkeit von 100 km/h aus.

a) Berechnen Sie die Standardabweichung σ sowie den Erwartungswert μ von X.

b) Wie groß ist die Wahrscheinlichkeit für eine Messung mit mehr als 110 km/h?

c) Bestimmen Sie das 99 %-Quantil $x_{0,99}$ der Verteilung. Was bedeutet diese Zahl bezogen auf die Geschwindigkeitsmessung?

d) Berechnen Sie das 2-σ-Intervall von X. Wie groß ist die Wahrscheinlichkeit, dass eine Messung innerhalb dieses Intervalls liegt?

e) Es werden 50 Tachometer getestet. Wie groß sind für das arithmetische Mittel aller 50 Messungen Erwartungswert und Standardabweichung? Geben Sie auch für die gemessene Durchschnittsgeschwindigkeit den 2-σ-Bereich an.

Lösungshinweis:

a) $\Phi\left(\frac{\mu - 100}{\sigma}\right) = 0.99 \quad \text{und} \quad \Phi\left(\frac{120 - \mu}{\sigma}\right) = 0.95$
$\Leftrightarrow \mu - 100 \approx 2.33 \cdot \sigma \quad \text{und} \quad 120 - \mu \approx 1.64 \cdot \sigma$
$\Rightarrow \sigma \approx \frac{120 - 100}{2.33 + 1.64} \approx 5.0377834 \quad \text{und} \quad \mu \approx 2.33 \cdot \sigma + 100 \approx 111.7380353$

b) $P(X \geqq 110) \approx 0.6368307$

c) $x_{0,99} \approx 123.4576719$. Auf lange Sicht höchstens 123.46 km/h gemessen bei 99 % der Tachometer.

d) $I_{2\sigma,X} = [\mu \pm 2\sigma] = [101.66, 121.81]$. $P(|X - \mu| < 2\sigma) = 2 \cdot \Phi(2) - 1 \approx 0.9544997$.

e) $\overline{X} = \frac{1}{n} \sum_{i=1}^{50} X_i \quad \text{mit} \quad X_i \sim N(\mu, \sigma) \Rightarrow \overline{X} \sim N(\mu; \frac{\sigma}{\sqrt{50}} \sim N(111.738; 0.712)$
$\Rightarrow I_{2\sigma,X} = [\mu \pm 2\sigma] = [110.313, 113.163]$.
Die Wahrscheinlichkeit ist für jeden 2-σ-Bereich gleich groß, siehe Teilaufgabe d).

Aufgabe 5 — 11 Punkte

Gegeben seien die Graphen der Funktionen

$$f, g : \mathbb{R} \to \mathbb{R}$$

in Abbildung 1 bzw. Abbildung 2.
Dazu werden folgende 11 Aussagen formuliert:

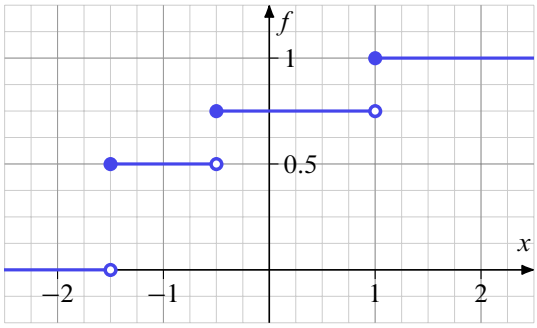


Abbildung 1: Graph von f

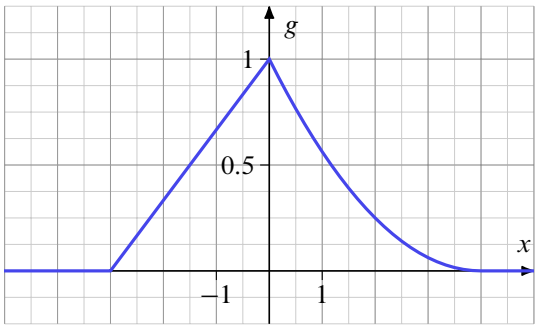


Abbildung 2: Graph von g

a) f ist eine Dichtefunktion.

b) g ist eine Dichtefunktion.

c) f ist eine Verteilungsfunktion.

d) g ist eine Verteilungsfunktion.

Falls f eine Verteilungsfunktion zur Zufallsvariable X wäre, würde gelten:

e) $P(X \leqq -1)$ ist nicht definiert, da X zwischen 0 und 1 liegen muss.

f) $P(X = 1) = 0.75$

g) $P(0 \leqq X \leqq 1) = 0.75$

h) $P(-1 \leqq X \leqq 1) = 0.75$

i) $P(X \leqq 1) = 0.75$

j) $P(X \leqq 2 \,|\, X \geqq 1)$ kann man nicht berechnen.

k) $P(X \leqq 0 \,|\, X \geqq -1) = 0.5$.

Entscheiden Sie für jede der 11 Aussagen, ob sie richtig oder falsch ist und begründen Sie Ihre Entscheidung jeweils.

(Hinweis: Für eine Antwort ohne Begründung gibt es keine Punkte, auch für eine richtige Antwort mit falscher Begründung gibt es keine Punkte)

Lösungshinweis:

a) Falsch: Fläche unter Graph von f ist größer als 1.

b) Falsch: Fläche unter Graph von g ist größer als 1.

c) Richtig: f ist monoton steigend, $\lim_{x\to-\infty} f(x) = 0, \lim_{x\to\infty} f(x) = 1$.

d) Falsch: g ist nicht monoton.

e) Falsch: Natürlich kann eine Zufallsvariable Werte kleiner 0 annehmen.

f) Falsch: $P(X = 1) = f(1) = 0.25$.

g) Falsch: $P(0 \leqq X \leqq 1) = f(1) = 0.25$.

h) Falsch: $P(-1 \leqq X \leqq 1) = f(-0.5) + f(1) = 0.25 + 0.25 = 0.5$.

i) Falsch: $P(X \leqq 1) = f(-1.5) + f(-0.5) + f(1) = 1$.

j) Falsch: $P(X \leqq 2 \,|\, x \geqq 1) = \dfrac{P(1 \leqq X \leqq 2)}{P(x \geqq 1)} = \dfrac{f(1)}{f(1)} = 1$.

k) Richtig: $P(X \leqq 0 \,|\, X \geqq -1) = \dfrac{P(X \leqq 0 \cap X \geqq -1)}{P(X \geqq -1)} = \dfrac{P(-1 \leqq X \leqq 0)}{P(X \geqq -1)} = \dfrac{0.25}{0.5} = 0.5$.

Aufgabe 6 — 11 Punkte

Die meisten Tachometer von Autos sind nicht sehr genau. Es soll im Folgenden die durchschnittliche gemessene Geschwindigkeit bei einer tatsächlich gefahrenen Geschwindigkeit von 100 km/h mit einer einfachen Stichprobe von 18 Autos geschätzt werden.

Gehen Sie davon aus, dass die angezeigte Tachometer–Geschwindigkeit bei einer tatsächlichen Fahrtgeschwindigkeit von 100 km/h einer normalverteilten Zufallsvariable $X \sim N(\mu, \sigma)$ entspricht.

Die Messwerte sind im Einzelnen:

Stichprobenelement Nr.	1	2	3	4	5	6	7	8	9
gemessene Geschwindigkeit [in km/h]	115	108	109	110	109	109	112	110	111
Stichprobenelement Nr.	10	11	12	13	14	15	16	17	18
gemessene Geschwindigkeit [in km/h]	115	111	116	115	111	114	111	109	110

a) Bestimmen Sie ein Konfidenzintervall für μ zu einem Konfidenzniveau von 95 %.

R b) Die Messwerte aus der Stichprobe seien in der R-Variable `tachometer` gespeichert. Schreiben Sie in folgendes Kästchen *einen* R-Befehl, mit dem man das Konfidenzintervall aus Teilaufgabe a) berechnen kann.

```
t.test(x, conf.level = 0.95)
```

R c) Sie wollen statistisch testen, ob die gemessene Geschwindigkeit aller Autos der Grundgesamtheit durchschnittlich größer als `mu0` $= 110$ km/h ist. Die Ausgabe eines entsprechenden Tests in R sieht aus wie folgt:

```
t.test(tachometer, mu = mu0, alternative = "greater")

##
##  One Sample t-test
##
## data:  tachometer
## t = 2.3348, df = 17, p-value = 0.01604
## alternative hypothesis: true mean is greater than 110
## 95 percent confidence interval:
##  110.3541      Inf
## sample estimates:
## mean of x
##  111.3889
```

Erklären Sie das Ergebnis des Tests (ohne selbst zu rechnen), wenn zu einem Signifikanzniveau $\alpha = 0.01$ getestet werden soll.

Lösungshinweis:

a) $c = x_{0.975} = 2.11, \overline{x} = 111.389, s = 2.524 \quad \Rightarrow \quad \left[\overline{x} \pm \frac{sc}{\sqrt{n}}\right] = [110.134;\ 112.644]$

b) siehe oben

c) Nullhypothese (Tacho-Geschw. ist durchschn. gleich 110) kann mit Gegenhypothese (größer als 110) nicht verworfen werden, da `p-value` > 0.05.

Aufgabe 7 — 11 Punkte

Zur Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ sei der Funktionsterm $f(x,y)$ unbekannt. Über die partiellen Ableitungen von f weiß man allerdings, dass

- $f_x(x,y)$ einem Polynom zweiten Grades in x ohne Abhängigkeit von y sowie
- $f_y(x,y)$ einem Polynom zweiten Grades in y ohne Abhängigkeit von x entspricht.

a) Begründen Sie, wie viele kritische Punkte (Nullstellen des Gradienten) f maximal besitzen kann.

b) In Abbildung 3 ist der Graph von f_x dargestellt. Geben Sie den Funktionsterm $f_x(x,y)$ an.
Hinweis: Sie dürfen benutzen, dass für die Position des Minimums von f_x gilt:

$$f_x(-0.5, y) = -2.25\,.$$

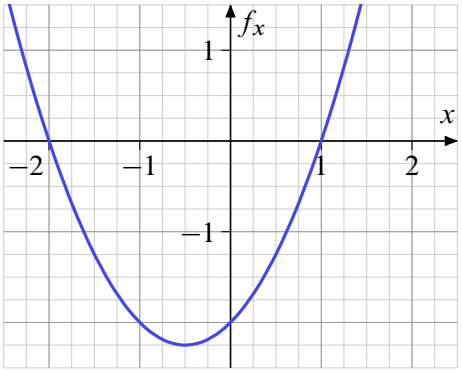


Abbildung 3: Graph von f_x

c) Der Funktionsterm von f_y lautet

$$f_y(x,y) = -y^2 + 4y - 3\,.$$

Zeichnen Sie den Graphen der Funktion f_y in das Koordinatensystem in Abbildung 4 ein.

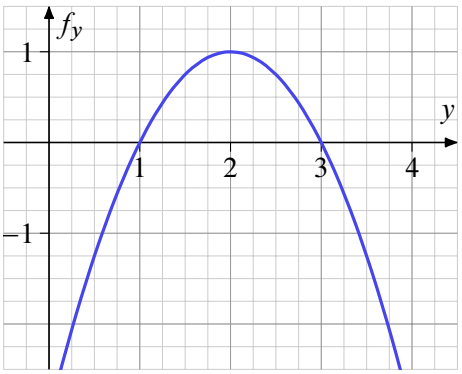


Abbildung 4: Graph von f_y

d) Bestimmen Sie sämtliche kritischen Punkte von f.

e) Ermitteln Sie die Hessematrix $\mathrm{H}_f(x,y)$ von f.

f) Bestimmen Sie die Art des Extremums für den Punkt $P(1,1)$. Falls Sie die Hesse-Matrix nicht bestimmen konnten, gehen Sie von der folgender (falscher) Matrix aus:

$$\mathrm{H}_f(x,y) = \begin{pmatrix} e^{x^2+y^2} & 1 \\ 1 & (x^4+y^2) \end{pmatrix}.$$

Lösungshinweis:

a) Es können maximal vier kritische Punkte auftreten. Jede Parabel kann höchstens zwei (verschiedene) reelle Nullstellen besitzen.

b) Der Funktionsterm für $f_x(x,y)$ lässt sich anhand der Nullstellen im Graph bestimmen. Es gilt somit $f_x(x,y) = A(x-(-2))\cdot(x-1) = A\cdot(x^2+x-2)$.
Für A gilt (Einsetzen des Minimums): $f_x(-0.5,y) = -\frac{5}{4}$ gelten muss.

c) $\partial_y f(x,y) = -y^2 + 4y - 3 = -(y-1)(y-3)$.
Der Scheitel dieser Funktion liegt bei $y = 2$.

d) Mit den Nullstellen von f_x, f_y ergeben sich die kritischen Punkte: $(-2,1)$, $(-2,3)$, $(1,1)$, $(1,3)$.

e) $H_f(x,y) = \begin{pmatrix} 2x+1 & 0 \\ 0 & -2y+4 \end{pmatrix}$.

f) Einsetzen: $H_f(1,1) = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$. Die Matrix ist also positiv definit.
Damit ist $(x,y) = (1,1)$ ein lokales Minimum.