Biostatistik und Bioinformatik (WS 2018/19)

Klausur

20. Februar 2019

Name ..... **Matrikelnummer** .....

Studiengang .....

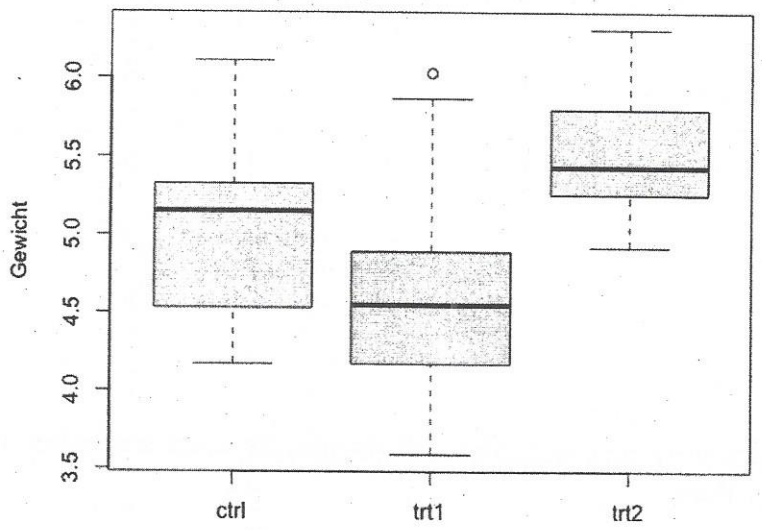
Wichtiger Hinweis!
Ohne vorherige Anmeldung in FlexNow kann die Klausur nicht gewertet werden.
Hiermit bestätige ich, dass ich für die Klausur in FlexNow angemeldet bin.

_____
Unterschrift

Punktzahl ..... **Note** .....

d) Wir haben mit einem Messinstrument 10000 Beobachtungen einer Variablen gemessen und diese durch einen Boxplot visualisiert. Wie würde sich die Höhe der Box ändern, wenn wir aus diesen 10000 Beobachtungen 1000 zufällig ziehen und hieraus die Box berechnen würden? Begründen Sie ihre Antwort. (1 Punkt)

e) In einem Experiment wurde das Wachstum einer Pflanzenart (Gewicht nach 3 Monaten) unter 3 verschiedenen Bedingungen (ctrl, trt1, trt2) gemessen (siehe nachfolgender Boxplot). Die Boxen von trt1 und trt2 überlappen sich nicht. Kann man daraus schließen, dass es mit hoher Wahrscheinlichkeit einen signifikanten Unterschied im Mittel der Verteilungen gibt? Begründen Sie ihre Antwort über die Antwort zur vorherigen Frage d) (1 Punkt)



Schließende Statistik (11 Punkte)

a) Die folgenden zwei Fragen beziehen sich auf das Experiment zum Pflanzengewicht im Teil „Deskriptive Statistik“. Um den Unterschied in ctrl und trt1 formal auf Signifikanz zu testen, wurde ein t-Test durchgeführt. Nennen Sie die vollständige Nullhypothese H0 für diesen Test, inklusive der Verteilungsannahme. (1 Punkt)

b) Das Ergebnis des t-Tests aus a) ist nachfolgend dargestellt. i) Ist das Ergebnis signifikant? ii) Wird die Nullhypothese abgelehnt? iii) Wie würden Sie das Ergebnis des Tests in einem Satz in ihrer Bachelorarbeit angeben? (1 Punkt)

```
> t.test(weight ~ group, data = PlantGrowth[PlantGrowth$group!= "trt2", ])

        Welch Two Sample t-test

data:  weight by group
t = 1.1913, df = 16.524, p-value = 0.2504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2875162  1.0295162
sample estimates:
mean in group ctrl mean in group trt1
             5.032              4.661
```

f) Erklären Sie den Unterschied zwischen Typ I Fehler und False Discovery Rate. (1 Punkt)

g) Nachfolgend die Ergebnisse einer linearen Regression in R. In dem Regressionsmodell wurde eine mögliche Abhängigkeit zwischen Lufttemperatur und Wind untersucht. Beantworten Sie die folgenden Fragen: i) Welche Form der Abhängigkeit wurde hier unterstellt? (Formel angeben) ii) Würde man aus den Ergebnissen schließen, dass es eine Abhängigkeit gibt? Woran sieht man das, und in welche Richtung geht die Abhängigkeit? (2 Punkte)

```
Call:
lm(formula = Temp ~ Wind, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-23.291  -5.723   1.709   6.016  19.199

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  90.1349     2.0522  43.921  < 2e-16 ***
Wind         -1.2305     0.1944  -6.331 2.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Versuchsplanung (6 Punkte)

a) Erklären Sie das Konzept eines Störfaktors. Warum sind Störfaktoren so gefährlich, d.h. welche Folgen kann ein unbeachteter Störfaktor in einer Datenanalyse haben? (1 Punkt) Wie kann man in einer praktischen Datenanalyse entscheiden, welche Variablen mögliche Störfaktoren sind, d.h. welche zwei Eigenschaften machen einen Störfaktor aus? (1 Punkt)

b) In einer Studie soll untersucht werden, ob der Verzehr von Kaviar die Wahrscheinlichkeit an Krebs zu erkranken beeinflusst. Hierzu wurden 10000 Probanden bzgl. ihres Kaviarkonsums und möglicher Krebserkrankungen befragt. Warum wäre die genetische Veranlagung wahrscheinlich KEIN Störfaktor bzgl. dieser Frage? (1 Punkt)

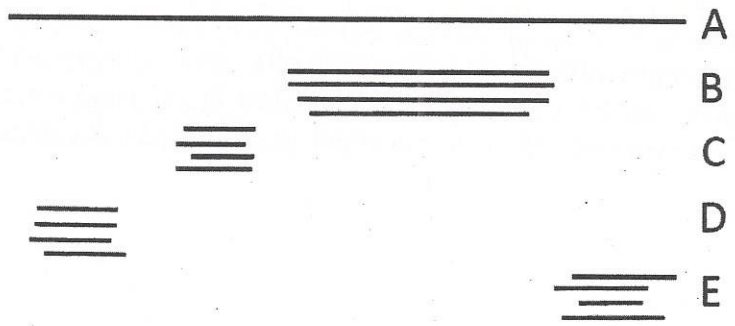
c) Was wäre ein möglicher Störfaktor bzgl. dieser Frage, und warum? (1 Punkt)

Bioinformatik, insgesamt 25 Punkte

Sequenzvergleich und Datenbanken (9 Punkte)

a) Konstruieren Sie einen Suffixbaum für die Sequenz A = TCCGT. Markieren Sie bitte die Wurzel. (1 Punkt)

b) Sie haben mithilfe des Programmes BLAST die Protein-Sequenz A untersucht. In der BLAST-Ausgabe finden Sie folgende Grafik; alle Treffer seien statistisch signifikant. Welche Strukturelemente repräsentieren die mit B - E markierten Sequenzen? Beschreiben Sie den Aufbau von A. (1 Punkt)



Multiple Sequenzalignments, Scoring-Systeme (3 Punkte)

a) Bei einem Proteindesign-Experiment müssen Sie ein Gln-Residuum ersetzen. Sie wollen einen möglichst konservativen Austausch vornehmen, d. h. eine Aminosäure einbauen, die Gln möglichst ähnlich ist. Benutzen Sie für Ihre Entscheidung die unten angegebene BLOSUM-Matrix und wählen sie zwei Aminosäuren, die hierfür in Frage kommen. Begründen Sie ihre Wahl. (1 Punkt)

A	Ala	4																		
R	Arg	-1	5																	
N	Asn	-2	0	6																
D	Asp	-2	-2	1	6															
C	Cys	0	-3	-3	-3	9														
Q	Gln	-1	1	0	0	-3	5													
E	Glu	-1	0	0	2	-4	2	5												
G	Gly	0	-2	0	-1	-3	-2	-2	6											
H	His	-2	0	1	-1	-3	0	0	-2	8										
I	Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
L	Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
K	Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
M	Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
F	Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
P	Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
S	Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
T	Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
W	Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Y	Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

b) Gegeben seien die folgenden Sequenzen: AGTT, GATT, AGTT, CTTG, TCTT. Wie lautet der $\log_{10}$-odds-score für G in der letzten Spalte des zugehörigen Profils? Geben Sie bitte die Zwischenschritte der Berechnung an. Skizzieren Sie für diese Spalte qualitativ ein Sequenzlogo. (2 Punkte)

Transkriptomik (2 Punkte)

a) Erläutern Sie, wie die Intensitäten eines Zweifarben-Experiments miteinander kombiniert werden. Entwickeln Sie eine Formel und liefern Sie bitte eine Begründung. Auf Probleme der Normalisierung müssen Sie nicht eingehen. (1 Punkt)

b) Beschreiben Sie den Aufbau der Gen-Ontologie in wenigen Sätzen. (1 Punkt)

Hidden-Markov-Modelle (4 Punkte)

a) In einem zeitweise unehrlichen Kasino werden zwei Münzen M_1 und M_2 im Wechsel verwendet. Bei Münze M_1 treten Kopf und Zahl jeweils mit gleicher Wahrscheinlichkeit auf, bei Münze M_2 ist $p(\text{Kopf}) = 1/10$. Wird Münze M_1 verwendet, so wird im folgenden Wurf mit $p = 0.95$ M_1 eingesetzt. Auf M_2 folgt mit $p = 0.75$ wiederum M_2. Zu Beginn werde mit $p = 0.8$ Münze M_2 gewählt. Zeichnen Sie ein Zustandsdiagramm und geben Sie sämtliche Wahrscheinlichkeiten an. (2 Punkte)

b) Unter Verwendung des oben beschriebenen Modells sind die Viterbivariablen für $t = i$ zu errechnen. Die Werte der Viterbivariablen für $t = i\text{-}1$ sind angegeben. Als nächstes Symbol x_i wird „Z" emittiert. Berechnen Sie für die zwei Zustände „M_1" und „M_2" den Wert der Viterbivariablen v_i. Tragen Sie in der Tabelle auch die Produktterme (Zahlenwerte) der zu vergleichenden Teilergebnisse (TE1, TE2) ein. (1 Punkt)

Zustände		**Viterbi-Variablen**	
		$t = i\text{-}1$	$t = i,\ x_i = \text{Z}$
M_1	....	0.2	TE1: TE2:
M_2	...	0.8	TE1: TE2:

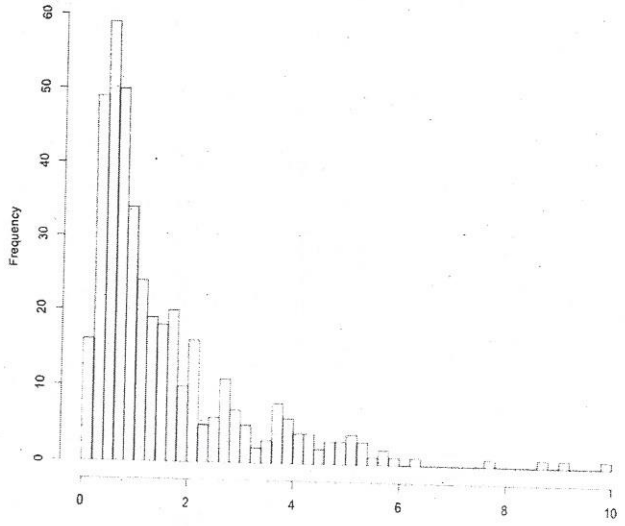
Biostatistik, insgesamt 25 Punkte

Deskriptive Statistik (8 Punkte)

a) Geben Sie die Standardabweichung und Schiefe (3. zentrales Moment) der Verteilung -2, 0, 2, 1 an. Hinweis: Formel angeben gibt Teilpunkte, falls Sie sich verrechnen. (1 Punkt)

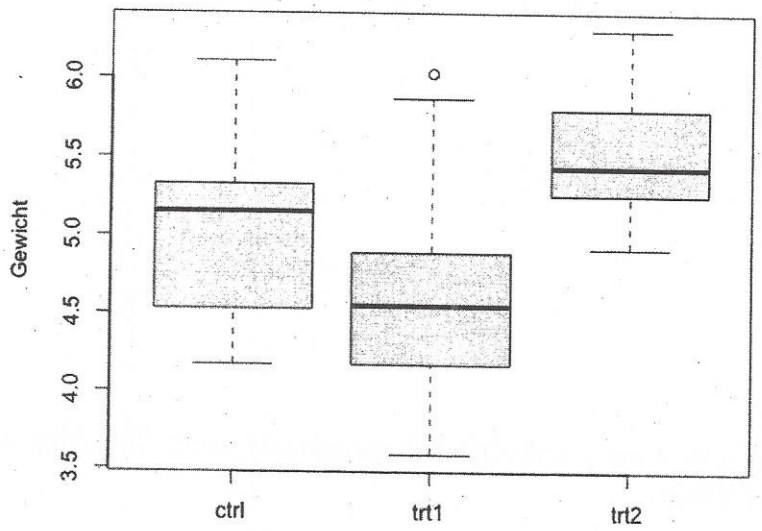
b) Bei der Berechnung der Kennzahlen einer Verteilung sehen Sie, dass der Mittelwert wesentlich kleiner ist als der Median. Was schließen Sie bzgl. der Schiefe der Verteilung? (1 Punkt)

c) Sie haben eine metrische Variable gemessen und die unten abgebildete Verteilung erhalten. Zeichnen sie die relevanten Kennzahlen für einen Boxplot in die Verteilung, und skizzieren Sie neben der Abbildung den Boxplot, der sich aus diesen Werten ergeben würde, inklusive Beschriftung der y-Achse mit ungefähren Werten. (1 Punkt)



d) Wir haben mit einem Messinstrument 10000 Beobachtungen einer Variablen gemessen und diese durch einen Boxplot visualisiert. Wie würde sich die Höhe der Box ändern, wenn wir aus diesen 10000 Beobachtungen 1000 zufällig ziehen und hieraus die Box berechnen würden? Begründen Sie ihre Antwort. (1 Punkt)

e) In einem Experiment wurde das Wachstum einer Pflanzenart (Gewicht nach 3 Monaten) unter 3 verschiedenen Bedingungen (ctrl, trt1, trt2) gemessen (siehe nachfolgender Boxplot). Die Boxen von trt1 und trt2 überlappen sich nicht. Kann man daraus schließen, dass es mit hoher Wahrscheinlichkeit einen signifikanten Unterschied im Mittel der Verteilungen gibt? Begründen Sie ihre Antwort über die Antwort zur vorherigen Frage d) (1 Punkt)



Schließende Statistik (11 Punkte)

a) Die folgenden zwei Fragen beziehen sich auf das Experiment zum Pflanzengewicht im Teil „Deskriptive Statistik". Um den Unterschied in ctrl und trt1 formal auf Signifikanz zu testen, wurde ein t-Test durchgeführt. Nennen Sie die vollständige Nullhypothese H0 für diesen Test, inklusive der Verteilungsannahme. (1 Punkt)

b) Das Ergebnis des t-Tests aus a) ist nachfolgend dargestellt. i) Ist das Ergebnis signifikant? ii) Wird die Nullhypothese abgelehnt? iii) Wie würden Sie das Ergebnis des Tests in einem Satz in ihrer Bachelorarbeit angeben? (1 Punkt)

```
> t.test(weight ~ group, data = PlantGrowth[PlantGrowth$group!= "trt2", ])

        Welch Two Sample t-test

data:  weight by group
t = 1.1913, df = 16.524, p-value = 0.2504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2875162  1.0295162
sample estimates:
mean in group ctrl mean in group trt1
             5.032              4.661
```

f) Erklären Sie den Unterschied zwischen Typ I Fehler und False Discovery Rate. (1 Punkt)

g) Nachfolgend die Ergebnisse einer linearen Regression in R. In dem Regressionsmodell wurde eine mögliche Abhängigkeit zwischen Lufttemperatur und Wind untersucht. Beantworten Sie die folgenden Fragen: i) Welche Form der Abhängigkeit wurde hier unterstellt? (Formel angeben) ii) Würde man aus den Ergebnissen schließen, dass es eine Abhängigkeit gibt? Woran sieht man das, und in welche Richtung geht die Abhängigkeit? (2 Punkte)

```
Call:
lm(formula = Temp ~ Wind, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-23.291  -5.723   1.709   6.016  19.199

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  90.1349     2.0522  43.921  < 2e-16 ***
Wind         -1.2305     0.1944  -6.331 2.64e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Versuchsplanung (6 Punkte)

a) Erklären Sie das Konzept eines Störfaktors. Warum sind Störfaktoren so gefährlich, d.h. welche Folgen kann ein unbeachteter Störfaktor in einer Datenanalyse haben? (1 Punkt) Wie kann man in einer praktischen Datenanalyse entscheiden, welche Variablen mögliche Störfaktoren sind, d.h. welche zwei Eigenschaften machen einen Störfaktor aus? (1 Punkt)

b) In einer Studie soll untersucht werden, ob der Verzehr von Kaviar die Wahrscheinlichkeit an Krebs zu erkranken beeinflusst. Hierzu wurden 10000 Probanden bzgl. ihres Kaviarkonsums und möglicher Krebserkrankungen befragt. Warum wäre die genetische Veranlagung wahrscheinlich KEIN Störfaktor bzgl. dieser Frage? (1 Punkt)

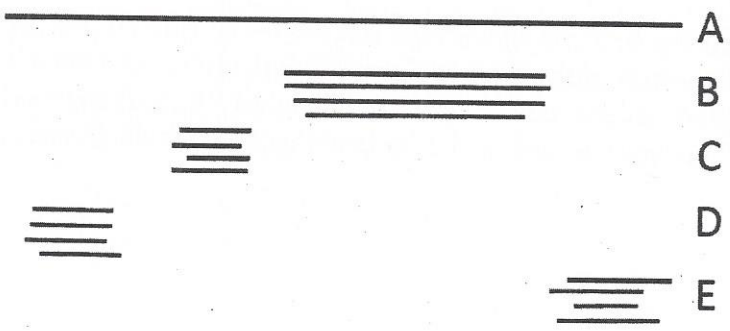
c) Was wäre ein möglicher Störfaktor bzgl. dieser Frage, und warum? (1 Punkt)

Bioinformatik, insgesamt 25 Punkte

Sequenzvergleich und Datenbanken (9 Punkte)

a) Konstruieren Sie einen Suffixbaum für die Sequenz A = TCCGT. Markieren Sie bitte die Wurzel. (1 Punkt)

b) Sie haben mithilfe des Programmes BLAST die Protein-Sequenz A untersucht. In der BLAST-Ausgabe finden Sie folgende Grafik; alle Treffer seien statistisch signifikant. Welche Strukturelemente repräsentieren die mit B - E markierten Sequenzen? Beschreiben Sie den Aufbau von A. (1 Punkt)



Multiple Sequenzalignments, Scoring-Systeme (3 Punkte)

a) Bei einem Proteindesign-Experiment müssen Sie ein Gln-Residuum ersetzen. Sie wollen einen möglichst konservativen Austausch vornehmen, d. h. eine Aminosäure einbauen, die Gln möglichst ähnlich ist. Benutzen Sie für Ihre Entscheidung die unten angegebene BLOSUM-Matrix und wählen sie zwei Aminosäuren, die hierfür in Frage kommen. Begründen Sie ihre Wahl. (1 Punkt)

A	Ala	4																		
R	Arg	-1	5																	
N	Asn	-2	0	6																
D	Asp	-2	-2	1	6															
C	Cys	0	-3	-3	-3	9														
Q	Gln	-1	1	0	0	-3	5													
E	Glu	-1	0	0	2	-4	2	5												
G	Gly	0	-2	0	-1	-3	-2	-2	6											
H	His	-2	0	1	-1	-3	0	0	-2	8										
I	Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
L	Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
K	Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
M	Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
F	Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
P	Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
S	Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
T	Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
W	Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Y	Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

b) Gegeben seien die folgenden Sequenzen: AGTT, GATT, AGTT, CTTG, TCTT. Wie lautet der $\log_{10}$-odds-score für G in der letzten Spalte des zugehörigen Profils? Geben Sie bitte die Zwischenschritte der Berechnung an. Skizzieren Sie für diese Spalte qualitativ ein Sequenzlogo. (2 Punkte)

Transkriptomik (2 Punkte)

a) Erläutern Sie, wie die Intensitäten eines Zweifarben-Experiments miteinander kombiniert werden. Entwickeln Sie eine Formel und liefern Sie bitte eine Begründung. Auf Probleme der Normalisierung müssen Sie nicht eingehen. (1 Punkt)

b) Beschreiben Sie den Aufbau der Gen-Ontologie in wenigen Sätzen. (1 Punkt)

Hidden-Markov-Modelle (4 Punkte)

a) In einem zeitweise unehrlichen Kasino werden zwei Münzen M_1 und M_2 im Wechsel verwendet. Bei Münze M_1 treten Kopf und Zahl jeweils mit gleicher Wahrscheinlichkeit auf, bei Münze M_2 ist $p(\text{Kopf}) = 1/10$. Wird Münze M_1 verwendet, so wird im folgenden Wurf mit $p = 0.95$ M_1 eingesetzt. Auf M_2 folgt mit $p = 0.75$ wiederum M_2. Zu Beginn werde mit $p = 0.8$ Münze M_2 gewählt. Zeichnen Sie ein Zustandsdiagramm und geben Sie sämtliche Wahrscheinlichkeiten an. (2 Punkte)

b) Unter Verwendung des oben beschriebenen Modells sind die Viterbivariablen für $t = i$ zu errechnen. Die Werte der Viterbivariablen für $t = i\text{-}1$ sind angegeben. Als nächstes Symbol x_i wird „Z" emittiert. Berechnen Sie für die zwei Zustände „M_1" und „M_2" den Wert der Viterbivariablen v_i. Tragen Sie in der Tabelle auch die Produktterme (Zahlenwerte) der zu vergleichenden Teilergebnisse (TE1, TE2) ein. (1 Punkt)

Zustände		**Viterbi-Variablen**	
		$t = i\text{-}1$	$t = i,\ x_i = \text{Z}$
M_1	....	0.2	TE1: TE2:
M_2	...	0.8	TE1: TE2: