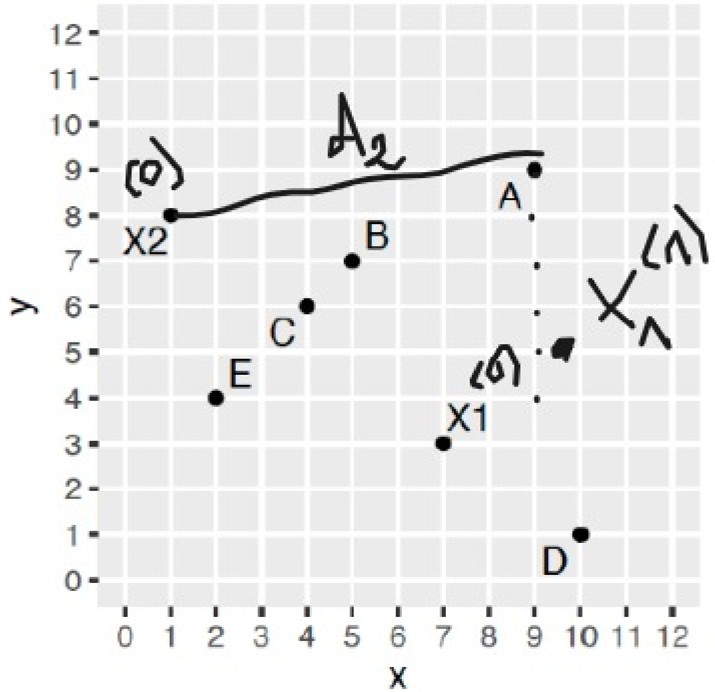
$d_2 = \sqrt{2^2 + 3^2} < \sqrt{3^2 + 3^2} = d_1$

$d_2 = \sqrt{1^2 + 8^2} \approx 8, \ldots$

$d_1 = \sqrt{2^2 + 6^2} \approx 6,3$

Problem 1 (2 credits)



The plot above displays a random initialization of a k-Means algorithm with k=2. X1, X2 are the randomly positioned centroids and A to E are the points of the 2-dimensional dataset to be clustered. Report the new positions of the centroids after the first iteration. Use the Euclidean distance. No standardization of the variables shall be applied. Justify your answer and provide relevant intermediate calculations. (1 point for the correct justification. If the justification is correct, 1 further point for the correct calculations.)

Cluster 1 D, A

Cluster 2 E, C, B

$$X_1^{(1)} = \frac{A + D}{2} = \frac{\begin{pmatrix} 9 \\ 9 \end{pmatrix} + \begin{pmatrix} 10 \\ 1 \end{pmatrix}}{2} = \frac{\begin{pmatrix} 19 \\ 10 \end{pmatrix}}{2} = \begin{pmatrix} 9,5 \\ 5 \end{pmatrix}$$

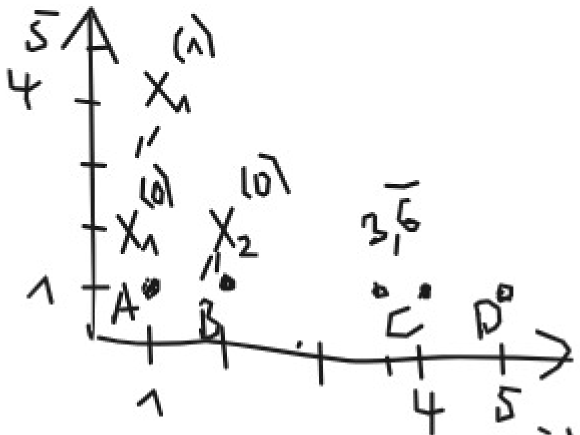
$$X_2^{(2)} = \frac{E + C + B}{3} = \frac{\begin{pmatrix} 2 \\ 4 \end{pmatrix} + \begin{pmatrix} 4 \\ 6 \end{pmatrix} + \begin{pmatrix} 5 \\ 7 \end{pmatrix}}{3} =$$

Problem 4 (2 credits)

Given the following data manually perform k-means clustering. Clearly state the assignment and centroids after each iteration. Start with the centroids (1, 1) and (2, 1).

```
##     x y
## 1:  1 1   A
## 2:  2 1   B
## 3:  4 1   C
## 4:  5 1   D
```

" $X_1^{(0)}$ " $X_2^{(0)}$

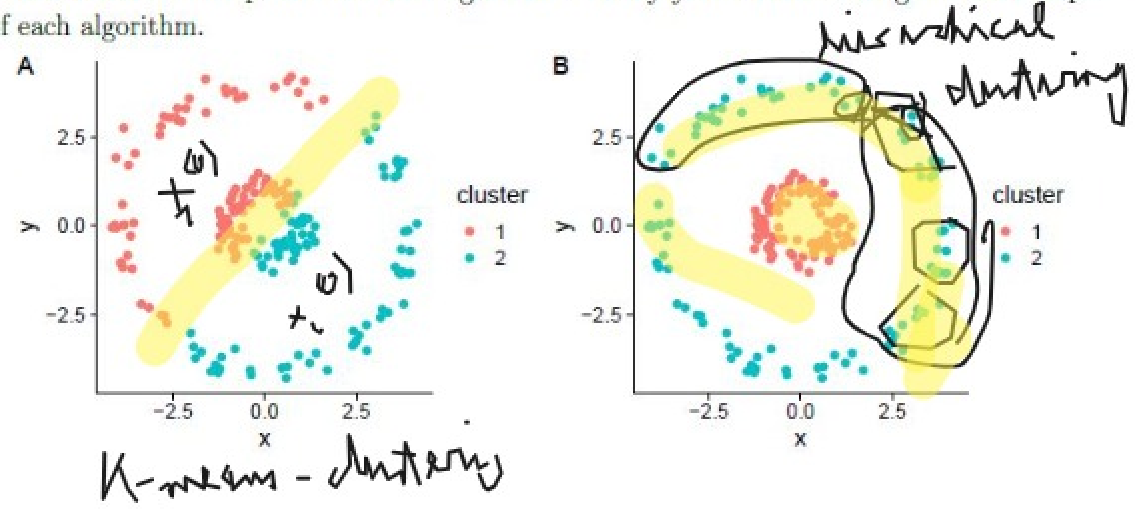


$$X_2^{(1)} = \frac{B+C+D}{3} = \frac{\binom{2}{1}+\binom{4}{1}+\binom{5}{1}}{3} = \frac{\binom{11}{3}}{3} = \binom{3,\bar{6}}{1}$$

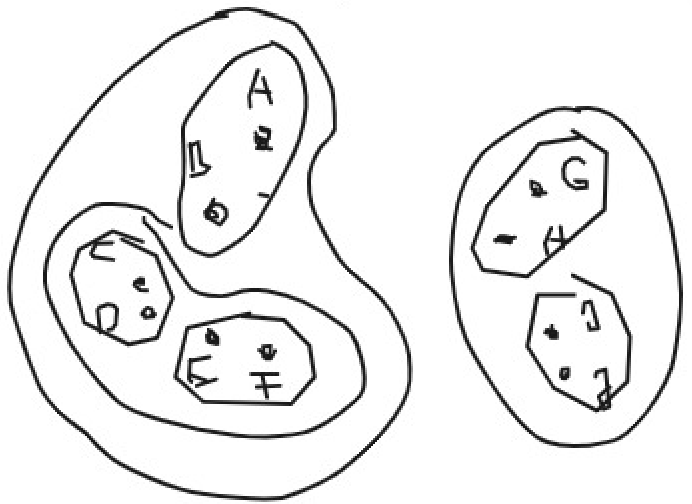
number of iteration ↓ k	$X_1^{(k)}$	$X_2^{(k)}$	Cluster 1	Cluster 2
0	$\binom{1}{1}$	$\binom{2}{1}$	A	B, C, D
1	$\binom{1}{1}$	$\binom{3,\bar{6}}{1}$	A, B	C, D
2	$\binom{1,5}{1}$	$\binom{4,5}{1}$	A, B	C, D
3	"	"	"	"

Problem 5 (3 credits)

You clustered a dataset using K-means clustering with $K = 2$ and hierarchical clustering using the single linkage rule. In the case of the hierarchical clustering the clustering tree was cut at a height which provides two clusters. Which of the two clustering results shown in Panel A and B corresponds to which algorithm? Justify your answer stating the main steps of each algorithm.



Beispiel für hierarchical clustering



1) Finde zwei Punkte/Punktgruppen, deren Distanz am geringsten

Beispiel: 10-mal Werfen einer Münze

p = W'keit für Kopf

$H_0: p \leq \frac{1}{2}$ $H_1: p > \frac{1}{2}$

$X \sim \text{Binom}(n,p)$

$\mathbb{P}(X=k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$

Resultat: Wenn _viele_ Köpfe kommen, wird H_0 abgelehnt.

K, K, K, Z, Z, K, K, K, K, K

Kann man H_0 zum Niveau $\alpha = 2\%$ abgelehnt werden?

X = Anzahl Köpfe $\sim \text{Binom}(\underset{10}{\underset{\shortparallel}{n}}, \underset{\frac{1}{2}}{\underset{\shortparallel}{p}})$ unter H_0

p-Wert = W'keit unter H_0 einen Wert zu erhalten, der mindestens so extrem ist wie der beobachtete Wert,

$$= \mathbb{P}_{H_0}(X \geq 8) = \mathbb{P}_{H_0}(X=8) + \mathbb{P}_{H_0}(X=9) + \mathbb{P}_{H_0}(X=10) =$$

$$= \binom{10}{8} \cdot \underbrace{\left(\frac{1}{2}\right)^8 \cdot \left(\frac{1}{2}\right)^2}_{=\left(\frac{1}{2}\right)^{10}} + \underset{10}{\underset{\shortparallel}{\binom{10}{9}}} \cdot \underbrace{\left(\frac{1}{2}\right)^9 \cdot \left(\frac{1}{2}\right)^1}_{=\left(\frac{1}{2}\right)^{10}} + \underset{1}{\underset{\shortparallel}{\binom{10}{10}}} \cdot \left(\frac{1}{2}\right)^{10} \cdot \underset{0}{\underset{\shortparallel}{\left(\frac{1}{2}\right)^0}}$$

$$= \left(\frac{1}{2}\right)^{10} \cdot \left(\binom{10}{8} + 10 + 1\right) \approx \underline{\underline{0{,}05}}$$

DU musst einen Zoom-Link schicken

$\Rightarrow$ p-Wert $>$ $\alpha = 0{,}02$

$\Rightarrow$ H_0 wird nicht abgelehnt

Resultat: KKKKKZZZZZ

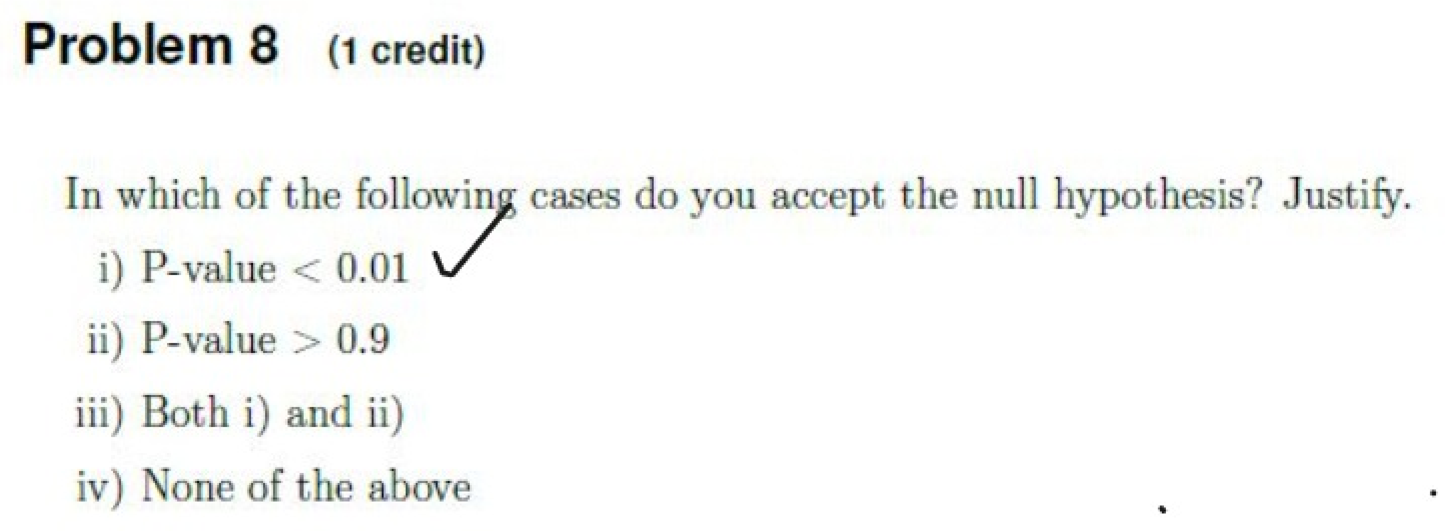
p-Wert $= \mathbb{P}_{H_0}(X \geq 5) \approx \frac{1}{2}$

Wir dürfen die Nullhypothese nicht ablehnen
Je größer der Wert, desto mehr spricht das für eine
wahre Nullhypothese

p-Wert $< \alpha$

$\Rightarrow$ H_0 wird abgelehnt

⇒ H_0 wird abge



Problem 8 (1 credit)

In which of the following cases do you accept the null hypothesis? Justify.

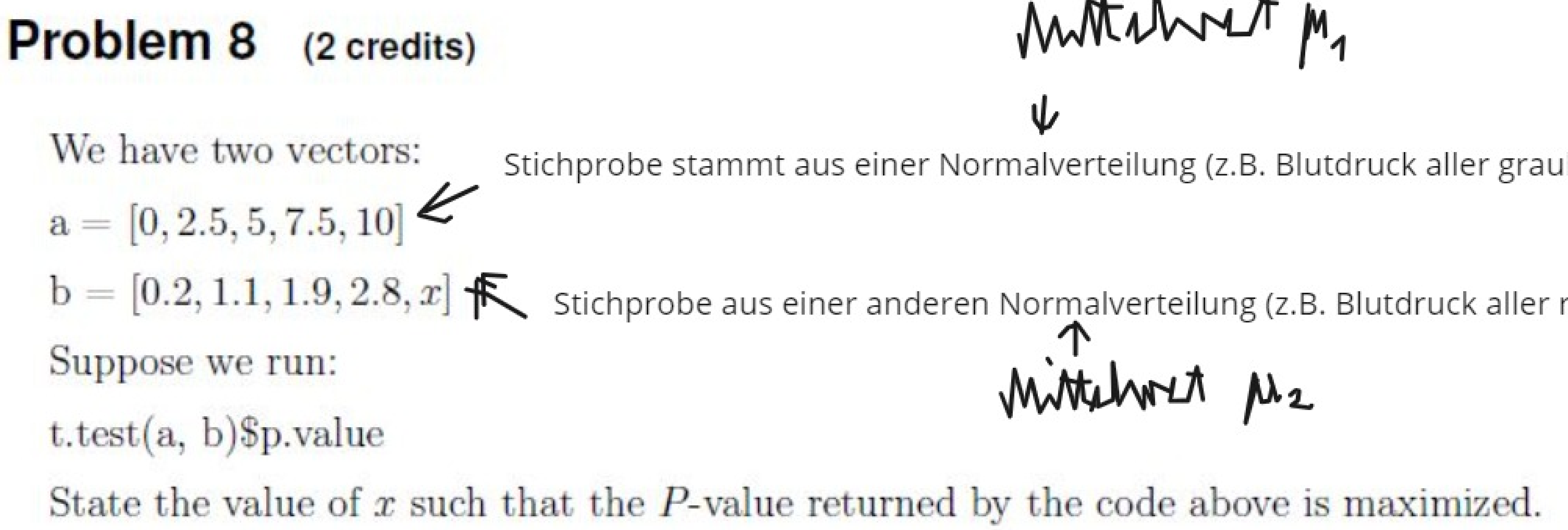
i) P-value < 0.01 ✓

ii) P-value > 0.9

iii) Both i) and ii)

iv) None of the above

Mittelwert μ_1



Problem 8 (2 credits)

We have two vectors:

a = [0, 2.5, 5, 7.5, 10] ← Stichprobe stammt aus einer Normalverteilung (z.B. Blutdruck aller grauhaarigen Deutschen)

b = [0.2, 1.1, 1.9, 2.8, x] ← Stichprobe aus einer anderen Normalverteilung (z.B. Blutdruck aller rothaarigen Deutschen)

Mittelwert μ_2

Suppose we run:

t.test(a, b)$p.value

State the value of x such that the P-value returned by the code above is maximized.

$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

Welcher Wert von x spricht am ehesten für H_0 (d.h. für maximalen p)?

$\bar{a} \overset{!}{=} \bar{b}$ x muss so gewählt werden, dass die Mittelwerte gleich sind

$$5 = \frac{6+x}{5}$$

$$x = 5 \cdot 5 - 6 = 25 - 6 = 19$$

$\rightarrow \boxed{x = 19}$