Fos_HW_07_Group_II

1. & 2.

$\boxed{1}$

a) $$h_k = \frac{\left(\frac{\text{number of observations in bin 'k'}}{n} \right)}{\text{width bin 'k'}} = \frac{\frac{\sum_{i=1}^{n} \mathbb{1}(c_{k-1} < x_i \leq c_k)}{n}}{c_k - c_{k-1}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq c_k) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq c_{k-1})}{c_k - c_{k-1}} = \frac{\hat{F}_n(c_k) - \hat{F}_n(c_{k-1})}{c_k - c_{k-1}} \blacksquare$$

b) $$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \leq x) = \alpha \quad \Rightarrow \sum_{i=1}^{n} \mathbb{1}(x_i \leq x) = \alpha n$$

The equation shows, that this x allows summing up all values x_i until αn is reached

This corresponds to the alpha-quantile in Def 5.

$\boxed{2}$

a) $$\mathbb{E}[\hat{F}_n(x)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)\right] = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}(X_i \leq x)]$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{P}(X_i \leq x) \underset{\text{iid. assumption}}{=} \frac{1}{n}\, n \cdot \mathbb{P}(X_1 \leq x) = F(x) \blacksquare$$

b) $\mathrm{Cov}[\hat{F}_n(x), \hat{F}_n(y)]$

$$= \mathbb{E}[\hat{F}_n(x) \cdot \hat{F}_n(y)] - \mathbb{E}[\hat{F}_n(x)]\,\mathbb{E}[\hat{F}_n(y)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x) \cdot \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq y)\right] - F(x) \cdot F(Y)$$

$$= \mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}(X_i \leq x) \cdot \mathbb{1}(Y_j \leq x)\right] - F(X) \cdot F(Y) = \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[\mathbb{1}(X_i \leq x) \cdot \mathbb{1}(X_j \leq y)] - F(X)\,F(Y)$$

$$= \frac{1}{n^2} \cdot \left(\sum_{\substack{i=1 \\ i=j}}^{n} \mathbb{E}[\mathbb{1}(X_i \leq x, X_i \leq y)] + \sum_{\substack{i=1 \\ j=1 \\ i \neq j}}^{n} \mathbb{E}[X_i \leq x, X_j \leq y] \right) - F(x)\,F(Y) = \frac{1}{n^2} \left(\underbrace{\sum_{i=1}^{n} \mathbb{P}(X_i \leq \min(x_i, Y_i))}_{n \text{ times}} + \underbrace{\sum_{\substack{i,j=1 \\ i \neq j}}^{n} F(x) \cdot F(y)}_{n^2 - n} \right)$$

$$= \frac{1}{n} \cdot F(x \wedge y) + \frac{n^2 - n}{n^2} F(x)\,F(y) - F(x)\,F(y) = \frac{1}{n}\left(F(x \wedge y) - F(x)\,F(y)\right) \blacksquare$$

c) Since the correlation between $\hat{F}_n(x)$ and $\hat{F}_n(y)$ is not 0, they are obviously correlated.

3.

a) Let $x \in \mathbb{R}$ be fixed. Then x lies in exactly one class interval $I_k = (c_k, c_{k+1}]$.

We have:

$$nb\hat{f}_n(x) = \sum_{i=1}^{n} \mathbb{I}(X_i \in I_k)$$

This is just counting how many of the X_i fall into the interval I_k.

Each indicator $\mathbb{I}(X_i \in I_k)$ is Bernoulli distributed with probability $p_k = \mathbb{P}(X_i \in I_k)$.

Since the X_i are i.i.d., the sum of n independent Bernoulli(p_k) variables is binomial, so:

$$nb\hat{f}_n(x) \sim \mathrm{Bin}(n, p_k)$$

Mean:

$$\mathbb{E}[nb\hat{f}_n(x)] = np_k$$

$$\Rightarrow \mathbb{E}[\hat{f}_n(x)] = \frac{p_k}{b}$$

Variance:

$$\mathrm{Var}(nb\hat{f}_n(x)) = np_k(1 - p_k)$$

$$\Rightarrow \mathrm{Var}(\hat{f}_n(x)) = \frac{1}{n^2b^2} \cdot np_k(1 - p_k) = \frac{p_k(1 - p_k)}{nb^2}$$

b) From (a) we know $\mathbb{E}[\hat{f}_n(x)] = \frac{p_k}{b}$.

We can write p_k as:

$$p_k = \int_{c_k}^{c_{k+1}} f(t)\,dt$$

Since f is continuous, by the mean value theorem there exists $\xi_k \in (c_k, c_{k+1})$ with:

$$\int_{c_k}^{c_{k+1}} f(t)\,dt = f(\xi_k) \cdot b$$

So:

$$\mathbb{E}[\hat{f}_n(x)] = \frac{f(\xi_k) \cdot b}{b} = f(\xi_k)$$

Now as $b \to 0$, the interval I_k shrinks to the point x, so $\xi_k \to x$.

Since f is continuous:

$$\lim_{b \to 0} \mathbb{E}[\hat{f}_n(x)] = \lim_{b \to 0} f(\xi_k) = f(x)$$

$\square$

c)

$$MSE = \mathbb{E}[(\hat{f}_n(x) - f(x))^2] = \mathrm{Var}(\hat{f}_n(x)) + \Big(\mathbb{E}[\hat{f}_n(x)] - f(x)\Big)^2$$

From (a) we have:

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{p_k(1-p_k)}{nb^2}$$

From (b) we know $\mathbb{E}[\hat{f}_n(x)] = f(\xi_k)$ for some $\xi_k \in I_k$, so the bias is:

$$\mathrm{Bias} = f(\xi_k) - f(x)$$

Therefore:

$$\mathbb{E}[(\hat{f}_n(x) - f(x))^2] = \frac{p_k(1-p_k)}{nb^2} + (f(\xi_k) - f(x))^2$$

Taking the limit $b \to 0$ and $nb \to \infty$:

For the bias term: As $b \to 0$, we have $\xi_k \to x$, so by continuity of f:

$$(f(\xi_k) - f(x))^2 \to 0$$

For the variance term: Since $p_k = \int_{c_k}^{c_{k+1}} f(t)\,dt = f(\xi_k) \cdot b$, we get $p_k(1-p_k) \le p_k = f(\xi_k) \cdot b$.
So:

$$\frac{p_k(1-p_k)}{nb^2} \le \frac{f(\xi_k) \cdot b}{nb^2} = \frac{f(\xi_k)}{nb} \to 0$$

as $nb \to \infty$.

Therefore:

$$\lim_{\substack{b\to 0\\ nb\to\infty}} \mathbb{E}[(\hat{f}_n(x) - f(x))^2] = 0 \quad \square$$

d) We apply Markov's inequality to the random variable $(\hat{f}_n(x) - f(x))^2$.

Markov's inequality says $\mathbb{P}(Y \ge a) \le \frac{\mathbb{E}[Y]}{a}$ for $Y \ge 0$ and $a > 0$.

Let $Y = (\hat{f}_n(x) - f(x))^2$ and $a = \epsilon^2$. Then:

$$\mathbb{P}\left[(\hat{f}_n(x) - f(x))^2 \ge \epsilon^2\right] \le \frac{\mathbb{E}[(\hat{f}_n(x) - f(x))^2]}{\epsilon^2}$$

This is equivalent to:

$$\mathbb{P}\left[|\hat{f}_n(x) - f(x)| \ge \epsilon\right] \le \frac{\mathrm{MSE}}{\epsilon^2}$$

From (c) we showed that MSE $\to 0$ as $b \to 0$ and $nb \to \infty$.

Since $\epsilon > 0$ is fixed, we have:

$$\mathbb{P}\left[|\hat{f}_n(x) - f(x)| > \epsilon\right] \le \frac{\mathrm{MSE}}{\epsilon^2} \to 0$$

Therefore $\hat{f}_n(x) \xrightarrow{P} f(x)$, which means $\hat{f}_n(x)$ is a consistent estimator for $f(x)$. $\square$

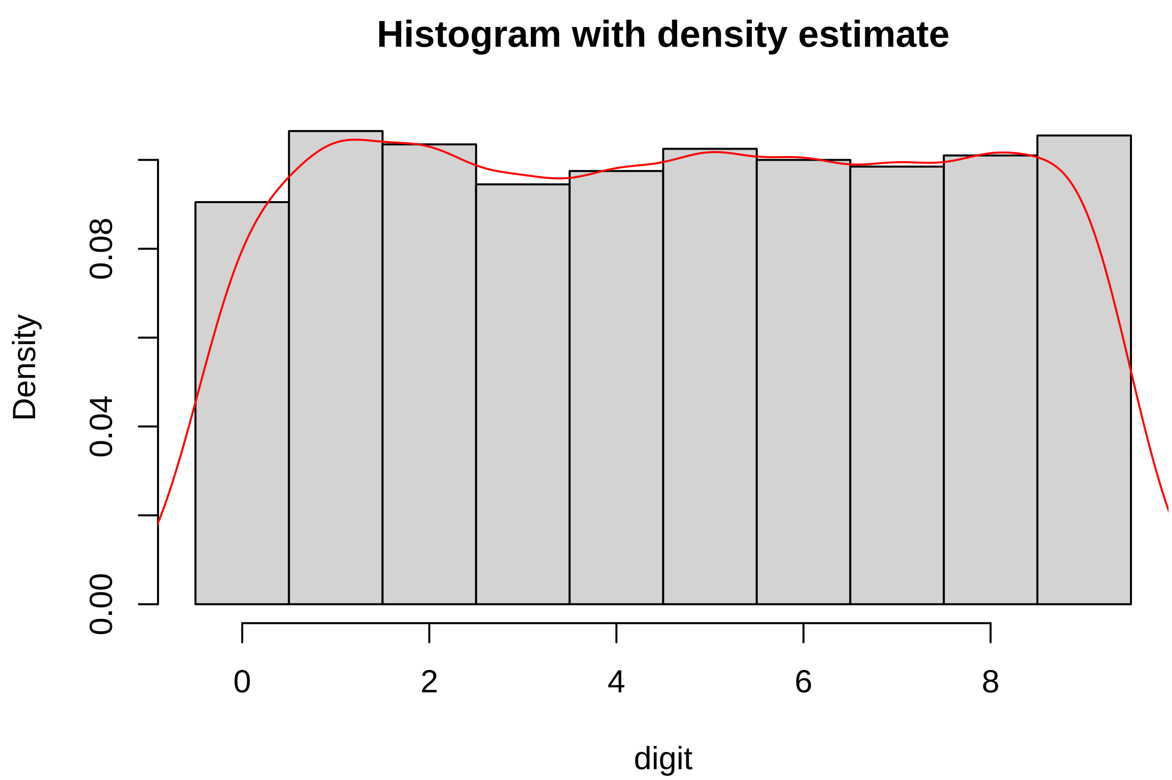
4.

```r
library("UsingR")
```

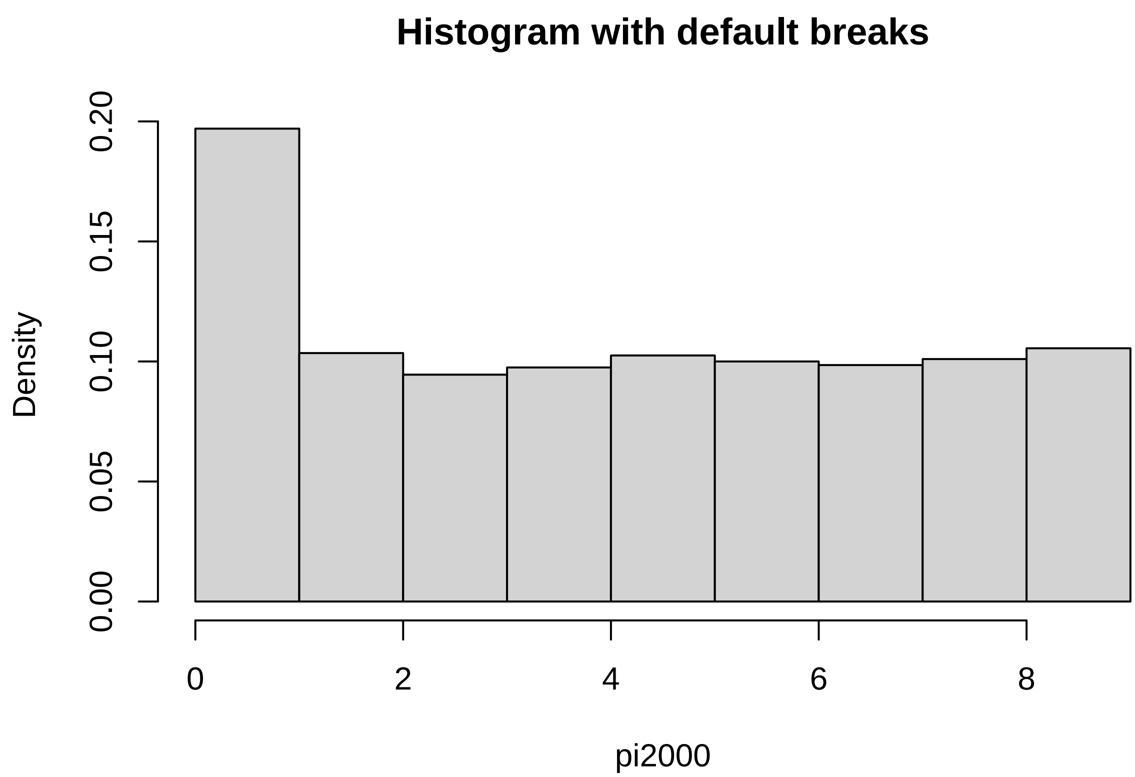
a)

```
## Lade nötiges Paket: MASS

## Lade nötiges Paket: HistData

## Lade nötiges Paket: Hmisc

##
## Attache Paket: 'Hmisc'

## Die folgenden Objekte sind maskiert von 'package:base':
##
##     format.pval, units
```

```r
d <- density(pi2000)
hist(pi2000, breaks = 0:10-0.5, prob=TRUE, xlab = "digit", main = "Histogram with density estimate")
lines(d, col = "red", lwd = 1)
```



```r
hist(pi2000, prob = TRUE, main = "Histogram with default breaks")
```



The argument `breaks = 0:10-0.5` centers the bins around the integer values 0-9, which makes sense since the data only contains digits 0-9. Without this, the default breaks might split the digits in awkward ways.

```r
table(pi2000)
```

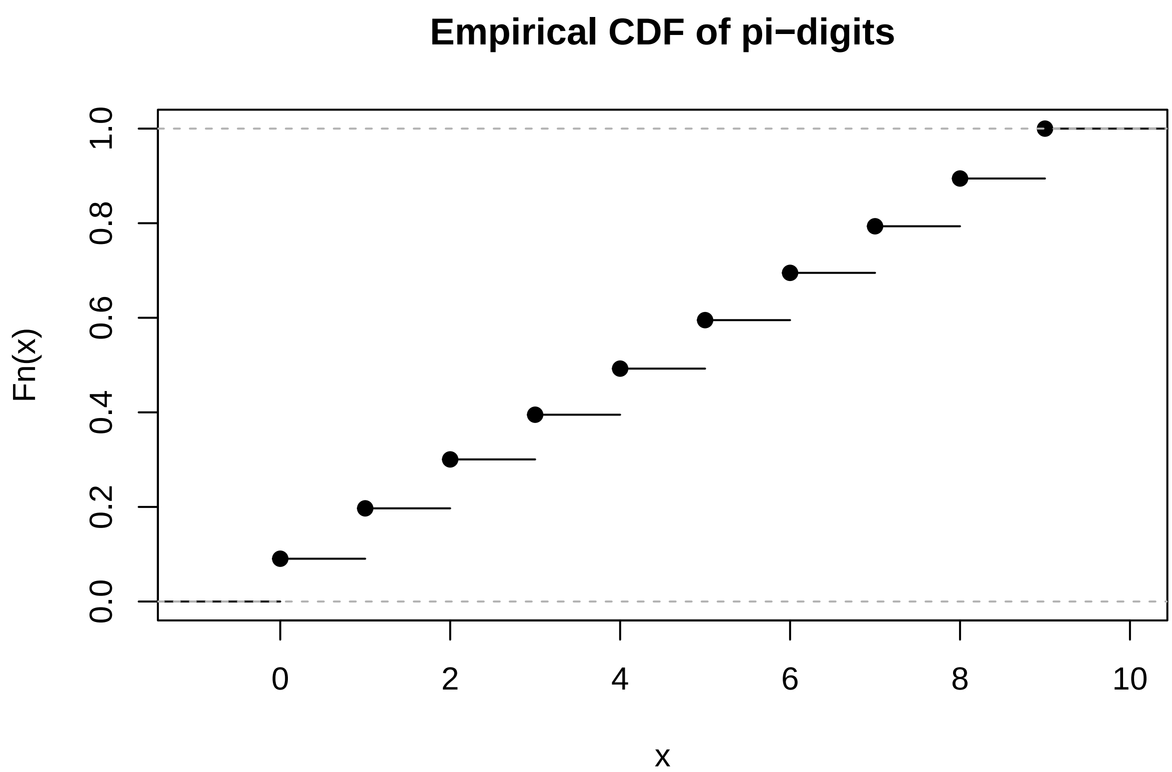
b)

```
## pi2000
##   0   1   2   3   4   5   6   7   8   9
## 181 213 207 189 195 205 200 197 202 211
```

```r
#cumsum(table(pi2000))

P <- ecdf(pi2000)
P(0.0)
```

```
## [1] 0.0905
```

```r
plot(P, main = "Empirical CDF of pi-digits")
```



c) I suspect a **uniform distribution** on $\{0, 1, \ldots, 9\}$. Each digit appears roughly equally often(around 200 times out of 2000), which is consistent with a discrete uniform distribution with $p = 1/10$ for each digit.

5.

Nr 5

a)

~~Rescue state~~	1 Class ~~rescued~~	2 Class	3 Class	Staff	total
r rescued	135	160	541	674	1510
d dead	202	125	180	211	718
total	337	285	721	885	2228

b) r := rescued d := dead 1 := 1 Class ... s := staff

$P(r\mid 1) = \frac{135}{337} \approx 0{,}4$ $P(r\mid 2) = \frac{160}{285} \approx 0{,}56$ $P(r\mid 3) = \frac{541}{721} \approx 0{,}75$

$P(r\mid s) = \frac{674}{885} \approx 0{,}76$

It seems like the „Higher" the Class the lower the surviving rate. ~~Which is interesting~~

c) $P(1) = \frac{337}{2228} \approx 0{,}15$ $P(2) \approx 0{,}13$ $P(3) \approx 0{,}32$

$P(S) \approx 0{,}40$ $P(r) \approx 0{,}68$ $P(d) \approx 0{,}32$

Each pair should be a multiplication

	1 classe	2 Classe	3 classe	Staff	
rescued	$P(r)P(1)\cdot Z$	$P(r)P(2)\cdot Z$	$P(r)P(3)Z$	$P(r)P(S)\cdot Z$	1510
dead	$P(d)P(1)Z$	$P(d)P(2)\cdot Z$	$P(d)P(3)Z$	$P(d)P(S)\cdot Z$	718
	337	285	721	885	$Z := 2228$

$\chi^2 =$ ~~0,35~~ 182,06

Cremer's V: 0,2859

```python
import numpy as np
import pandas as pd
from scipy.stats import chi2_contingency

data = {
    '1st Class': [135, 202],
    '2nd Class': [160, 125],
    '3rd Class': [541, 180],
    'Staff':     [674, 211]
}


observed_df = pd.DataFrame(data, index=['Rescued', 'Dead']).T
print(observed_df.T)
print("\n")


# Chi2-Statistik, p-Wert, Freiheitsgrade und die Erwarteten Häufigkeiten
chi2_stat, p_val, dof, expected = chi2_contingency(observed_df, correction=False)


expected_df = pd.DataFrame(expected, columns=['Rescued', 'Dead'], index=observed_df.index)

print(expected_df.round(2))
print("\n")


n = 2228
min_dim = min(observed_df.shape) - 1

cramers_v = np.sqrt(chi2_stat / (n * min_dim))

print("--- Statistik ---")
print(f"Chi-Quadrat-Wert: {chi2_stat:.4f}")
print(f"Gesamtzahl n:     {n}")
print(f"Cramer's V:       {cramers_v:.4f}")
```

```
         1st Class  2nd Class  3rd Class  Staff
Rescued        135        160        541    674
Dead           202        125        180    211


           Rescued    Dead
1st Class   228.40  108.60
2nd Class   193.16   91.84
3rd Class   488.65  232.35
Staff       599.80  285.20


--- Statistik ---
Chi-Quadrat-Wert: 182.0632
Gesamtzahl n:     2228
Cramer's V:       0.2859
```

So there is a association between travel class and rescue status but it could be stronger but it is not neglectabel

d)

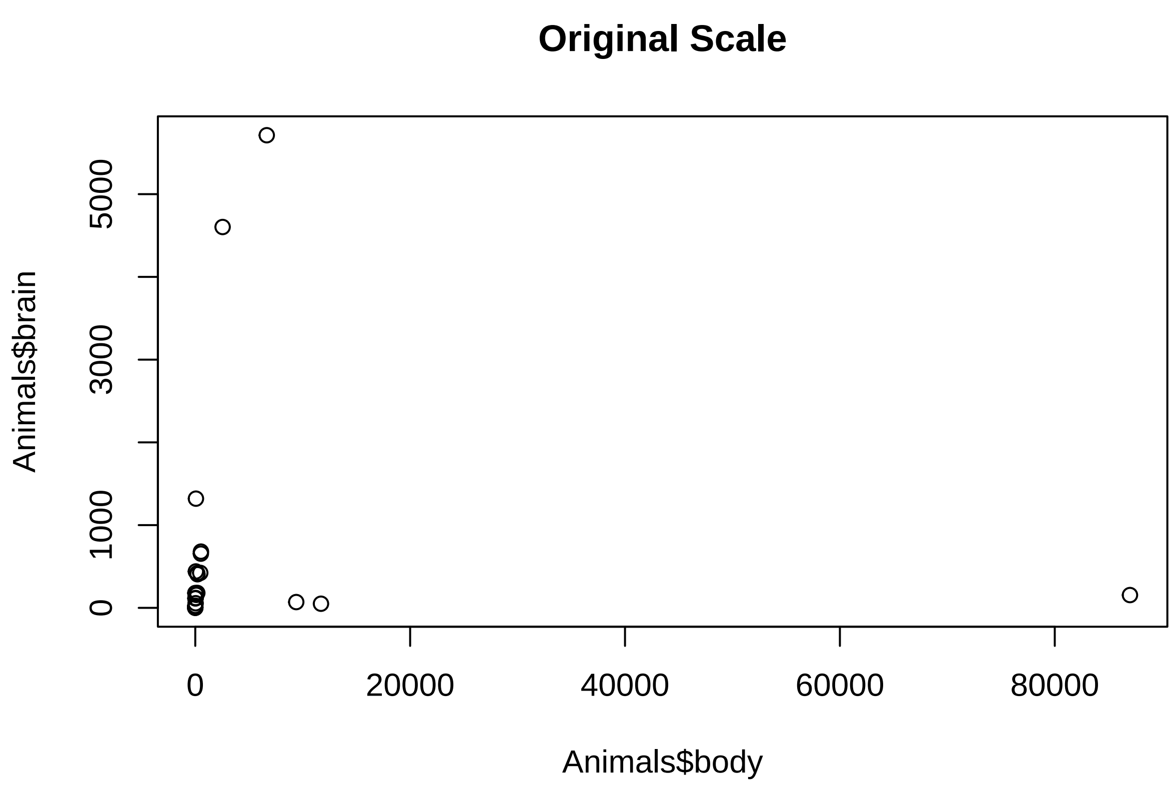
Conclusion: There is a depandancy between travel class and rescue status which may seem surprising but is shown by the data

6.

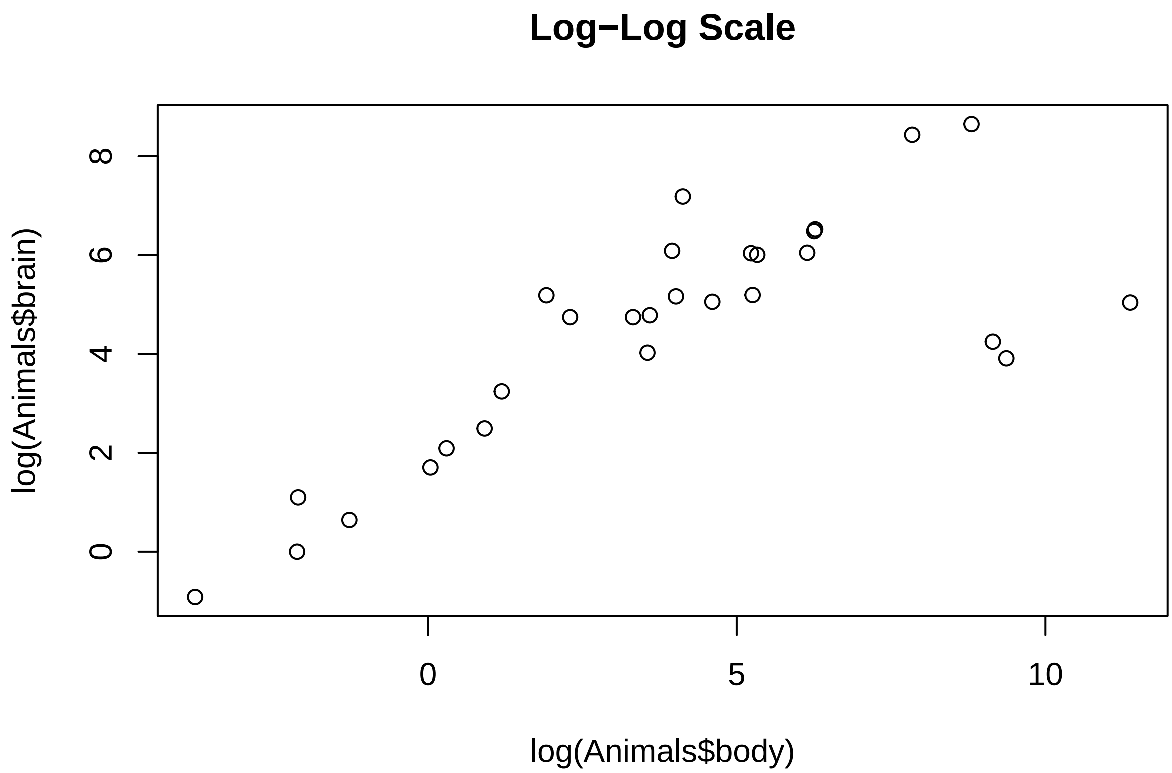
```r
library(MASS)
data(Animals)


#a)

cor(Animals$body, Animals$brain, method = "pearson")
```

```
## [1] -0.005341163
```

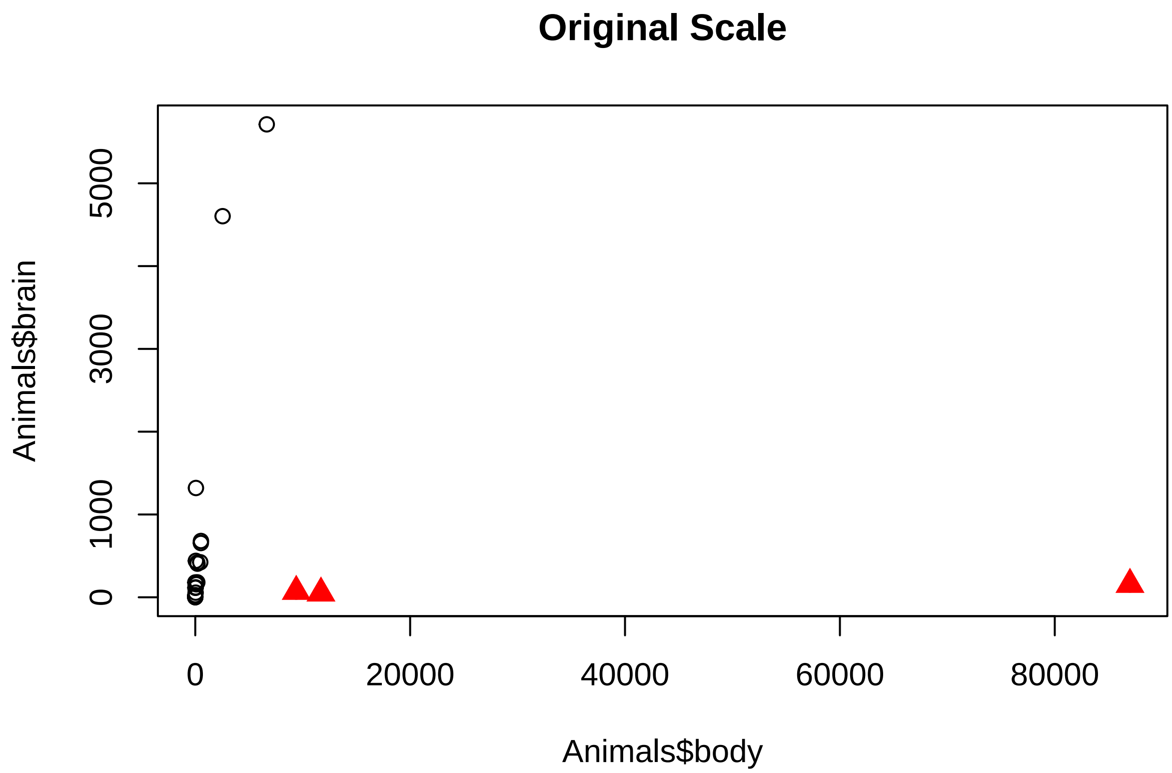
```r
plot(Animals$body, Animals$brain, main="Original Scale")
```




```r
plot(log(Animals$body), log(Animals$brain), main="Log-Log Scale")
```




```
#b)

dinos <- c("Brachiosaurus", "Dipliodocus", "Triceratops")
Animals_no_dino <- Animals[!(row.names(Animals) %in% dinos), ]
is_dino <- rownames(Animals) %in% dinos
plot(Animals$body, Animals$brain, main="Original Scale")
points(Animals$body[is_dino], Animals$brain[is_dino],
       col = "red", pch = 17, cex = 1.5)
```




```
cor(Animals_no_dino$body, Animals_no_dino$brain, method = "pearson")
```

```
## [1] 0.9318502
```

```
#c)
# Spearman mit Dinos

cor(Animals$body, Animals$brain, method = "spearman")
```

```
## [1] 0.7162994
```

```
# Spearman ohne Dinos

cor(Animals_no_dino$body, Animals_no_dino$brain, method = "spearman")
```

```
## [1] 0.9328717
```

It seams that the Spearmans rank is more robust to the presense of outliers as also mensiont in the lectures