Statistik II

Paul Strimtu 3898312
Jakob Striegel 4351490

Abgabe 5: 27 Mai 2024

1 Maximum-Likelihood-Schätzung (analytisch)

Sei $f_\theta(x) = \frac{\theta^6 x^5}{120} e^{-\theta x}, x > 0$ die Dichte der gegebenen Verteilung zu 100 Tauchgängen von Kegelrobben ($i \in \{1, \cdots, 100\}$ und $n = 100$).

Allgemein

$$\Rightarrow \mathcal{L}(\theta) = f_\theta(x_1, \cdots, x_n) \overset{iid}{=} \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{\theta^6 x_i^5}{120} e^{-\theta x_i} = \left(\frac{\theta^6}{120}\right)^n \left(\prod_{i=1}^{n} x_i\right)^5 exp\left(-\theta \sum_{i=1}^{n} x_i\right)$$

$$\Rightarrow \ell(\theta) = \log(\mathcal{L}(\theta)) = 6n\log(\theta) - n\log(120) + 5\sum_{i=1}^{n} x_i - \theta \sum_{i=1}^{n} x_i$$

$$\Rightarrow \ell'(\theta) = \frac{6n}{\theta} - \sum_{i=1}^{n} x_i$$

$$\Rightarrow \ell'(\theta) = 0 \Leftrightarrow \frac{1}{\theta} = \frac{\sum_{i=1}^{n} x_i}{6n} \Leftrightarrow \theta = \frac{6n}{\sum_{i=1}^{n} x_i} \Leftrightarrow \theta = \frac{6n}{n\bar{x}} \Rightarrow \hat{\theta} = \frac{6}{\bar{x}}$$

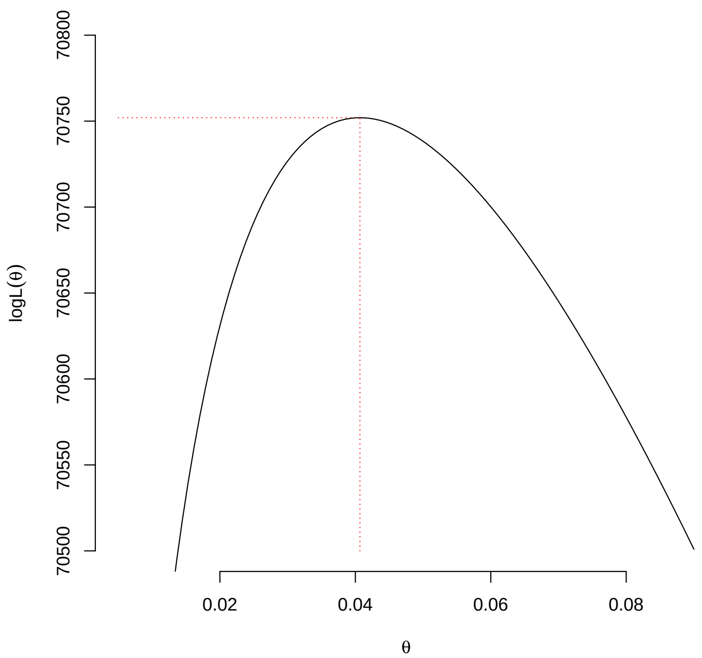
Der allgemeine ML-Schätzer lautet $\hat{\theta} = \frac{6}{\bar{x}}$.

Konkret

Für die konkrete Berechnung werden $n = 100$ und $\hat{x}$ als arithmetisches Mittel der gegebenen Daten genutzt.

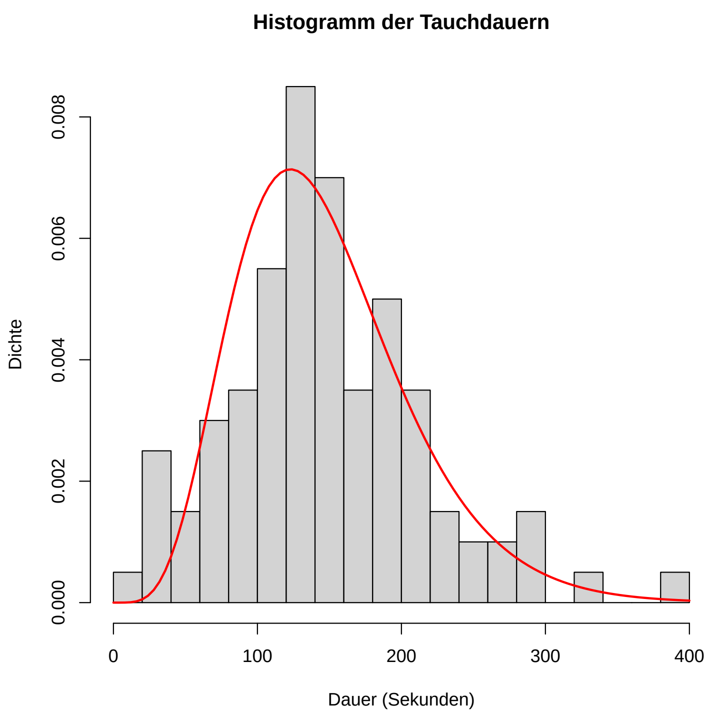
```r
dauer <- read.csv("kegelrobbe.csv")$x
n <- length(dauer)
sum_dauer <- sum(dauer)

#Log-Likelihood Funktion
logL <- function(theta){6*n*log(theta)-n*log(120)+5*sum_dauer-theta*sum_dauer}

theta <- seq(0.005,0.09,length=100) # x Werte liste
logL_theta <- rep(NA,100)           # y Werte liste "leer"
for (k in 1:100){
  logL_theta[k] <- logL(theta[k]) #rechnet y Werte aus
}
```

```r
tH <- 6/mean(dauer); tH #mgl auch 6*n/sum(dauer)

## [1] 0.04067686

plot(theta,logL_theta,type="l",xlab=expression(theta),
     ylab=expression(logL(theta)),bty="n", ylim = c(70500,70800))
segments(x0=tH,y0=70500,x1=tH,y1=logL(tH),col="red",lty="dotted")
segments(x0=min(theta),y0=logL(tH),x1=tH,y1=logL(tH),col="red",lty="dotted")
```



Der ML-Schätzer für die gegebenen Daten liegt rechnerisch und grafisch bei ca. $\hat{\theta} = 0.0407$ Mithilfe des ML-Schätzers kann nun eine angepasste Dichtefunktion erstellt werden, die die gegebenen Daten gut abbildet. Das kann man im folgenden Histogramm der Daten der roten Kurve (Dichtefunktion) sehen:

```r
hist(dauer, breaks = 20, probability = TRUE, main = "Histogramm der Tauchdauern",
     xlab = "Dauer (Sekunden)", ylab = "Dichte")
# Dichtefunktion angepasst
curve((tH^6 * x^5 / 120) * exp(-tH * x), col = "red", lwd = 2, add = TRUE)
```



2 Maximum-Likelihood-Schätzung (numerisch)

Sei $f_\theta(x) = 5\theta^2 x^4 e^{-(\theta x)^5}, x > 0$ die Dichte der gegebenen Verteilung zum Gewicht von 439 Neugeborenen ($i \in \{1, \cdots, 439\}$ und $n = 439$).

Allgemein

$$
\begin{aligned}
\ell(\theta) &= \log(f_\theta(x_1, \cdots, x_n)) \\
&= \log\left(\prod_{i=1}^{n} f_\theta(x_i)\right) \\
&= \sum_{i=1}^{n} (\log f_\theta(x_i)) \\
&= \sum_{i=1}^{n} \left(\log(5) + \log(\theta^5) + \log(x_i^4) + \log(exp(-(\theta x_i)^5))\right) \\
&= n\log(5) + 5n\log(\theta) + 4\sum_{i=1}^{n}\log(x_i) - \theta^5\sum_{i=1}^{n} x_i^5
\end{aligned}
$$

$$
\begin{aligned}
&\Rightarrow \ell'(\theta) = \frac{5n}{\theta} - 5\theta^4\sum_{i=1}^{n} x_i^5 \\
&\Rightarrow \ell''(\theta) = -\frac{5n}{\theta^2} - 20\theta^3\sum_{i=1}^{n} x_i^5
\end{aligned}
$$

$$
\Rightarrow \ell'(\theta) = 0 \Leftrightarrow \frac{5n}{\theta} = 5\theta^4\sum_{i=1}^{n} x_i^5 \Leftrightarrow \theta^5\sum_{i=1}^{n} x_i^5 = n \Leftrightarrow \theta^5 = \frac{n}{\sum_{i=1}^{n} x_i^5} \Rightarrow \hat{\theta} = \frac{1}{(\bar{x}^5)^{1/5}}
$$

Der allgemeine ML-Schätzer lautet $\hat{\theta} = \frac{1}{(\bar{x}^5)^{1/5}}$.

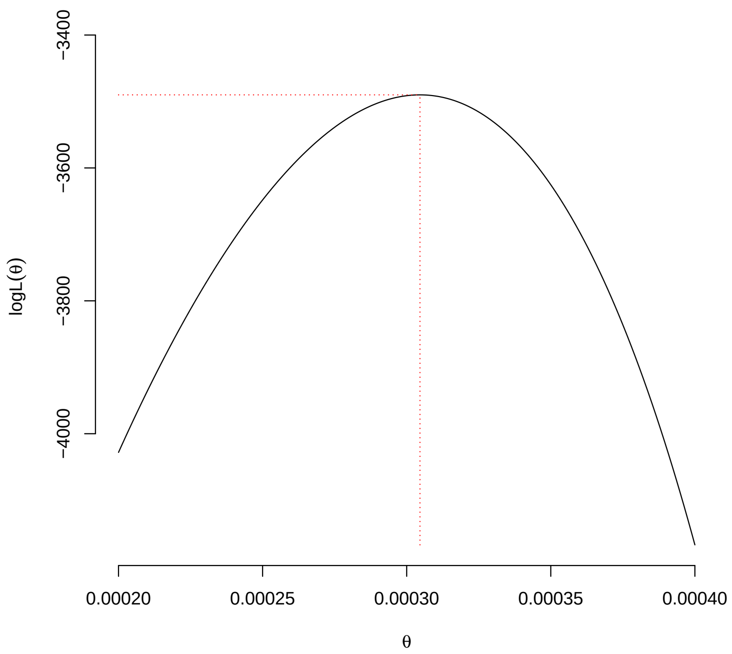
Konkret

Für die konkrete Berechnung werden $n = 439$ und $\hat{x}$ als arithmetisches Mittel der gegebenen Daten genutzt.

```r
gewicht <- read.csv("gewicht.csv")$x
n2 <- length(gewicht)
sum_gewicht <- sum(gewicht)

#Log-Likelihood Funktion
logL2 <- function(theta2){n2*log(5) + 5*n2*log(theta2) +
  4*sum(log(gewicht))-(theta2^5)*sum((gewicht)^5)}

theta2 <- seq(0.0002,0.0004,length=439)
logL2_theta2 <- rep(NA,439)
for (l in 1:439){
  logL2_theta2[l] <- logL2(theta2[l])
}
```

```r
tH2 <- 1/(mean(gewicht^5)^(1/5)); tH2 #mgl auch(n2 / sum(gewicht^5))^(1/5)

## [1] 0.0003046073
```

```r
minL <- min(logL2_theta2)
maxL <- max(logL2_theta2)

plot(theta2,logL2_theta2,type="l",xlab=expression(theta),
     ylab=expression(logL(theta)),bty="n", ylim = c(minL,maxL+100))
segments(x0=tH2,y0=minL,x1=tH2,y1=logL2(tH2),col="red",lty="dotted")
segments(x0=min(theta2),y0=logL2(tH2),x1=tH2,y1=logL2(tH2),col="red",lty="dotted")
```



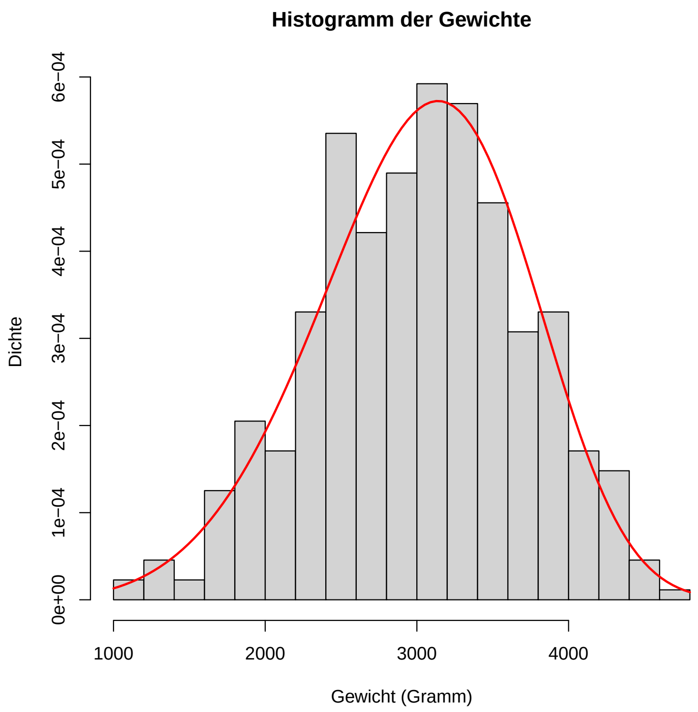
Der ML-Schätzer für die gegebenen Daten liegt rechnerisch und grafisch bei ca. $\hat{\theta} = 0.000305$.

Newton-Raphson Verfahren

```r
Theta <- rep(NA,5); Theta[1] <- 0.0002 # Startwert für die Suche
for (i in 2:6){
  loglike1 <- 5*n2/Theta[i-1]-5*(Theta[i-1])^4*sum(gewicht^5) # 1. Ableitung
  loglike2 <- -5*n2/(Theta[i-1])^2-20*(Theta[i-1])^3*sum(gewicht^5) # 2. Ableitung
  Theta[i]<-Theta[i-1]-loglike1/loglike2
}

Theta

## [1] 0.0002000000 0.0003179995 0.0003051930 0.0003046085 0.0003046073
## [6] 0.0003046073

all.equal(Theta[6], tH2)

## [1] TRUE
```

Mit dem Newton-Raphson Verfahren (in R) ergeben sich nach mindestens 4 Iterationen die selben Werte für den Schätzer $\hat{\theta}$, die auch zuvor exakt und grafisch bestimmt wurden.
Mithilfe des ML-Schätzers kann nun eine angepasste Dichtefunktion erstellt werden, die die gegebenen Daten gut abbildet. Das kann man im folgenden Histogramm der Daten der roten Kurve (Dichtefunktion) sehen:

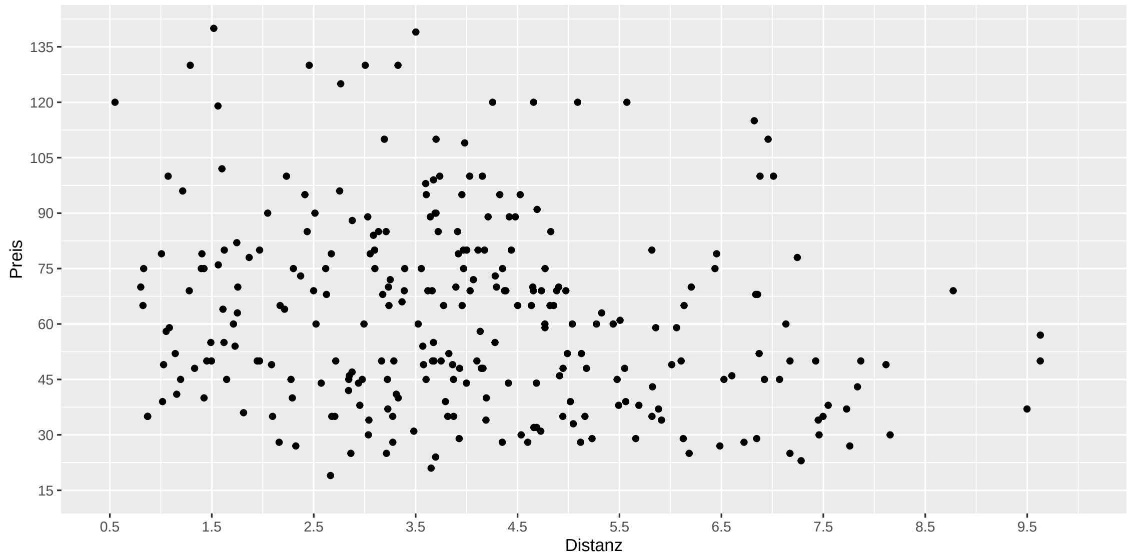
```r
hist(gewicht, breaks = 20, probability = TRUE, main = "Histogramm der Gewichte",
     xlab = "Gewicht (Gramm)", ylab = "Dichte")
# Dichtefunktion angepasst
curve((5*tH2^5 * x^4 * exp(-(tH2*x)^5)), col = "red", lwd = 2, add = TRUE)
```



3 Lineare Regression

a)

```
library(ggplot2)
airbnb <- read.csv("airbnb.csv"); attach(airbnb)

ggplot(airbnb, aes(x = Distanz, y = Preis)) + geom_point() +
  scale_x_continuous(breaks = seq(0.5, 9.5, by = 1), limits = c(0.5, 10)) +
  scale_y_continuous(breaks = seq(15, 140, by = 15), limits = c(15, 140))
```



b)

```
modell <- lm(Preis~Distanz)$coeff; modell

## (Intercept)     Distanz
##   73.384785   -2.813475

b_0 <- as.numeric(modell[1])
b_1 <- as.numeric(modell[2])
```

Das Regressionsmodell lautet:

$$Y_i = 73.385 - 2.814x_i + \epsilon_i.$$

Er ist also ein mittlerer Preis von ca 73.39 Euro zu erwarten bei einer Entfernung von 500 m und für jeden weiteren *km* von der Innenstadt entfernt, wird sich dieser Preis um ca. 2.81 Euro pro km verringen.

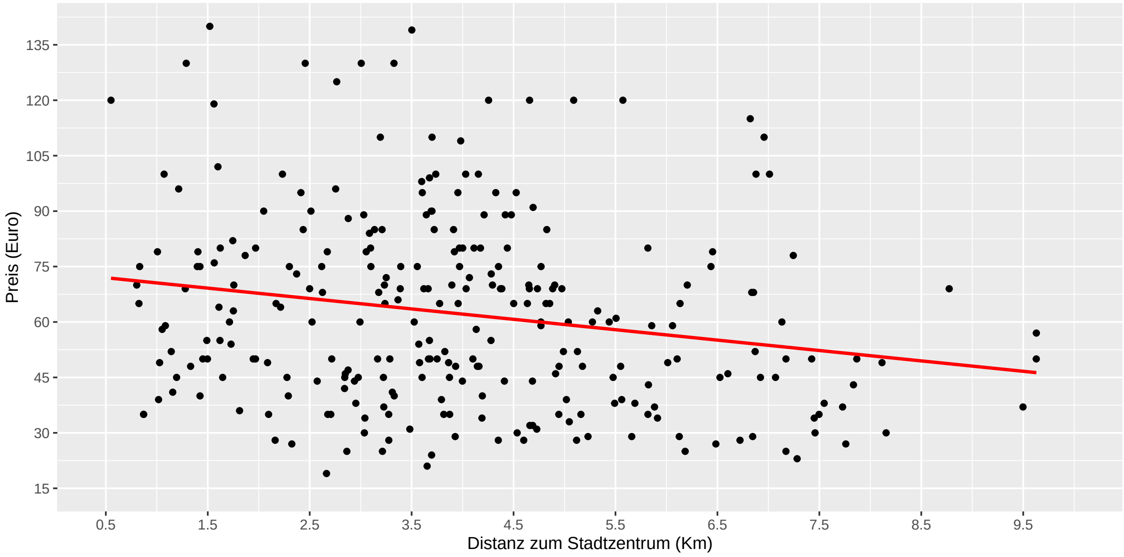
```
priceChange <- b_1 * 3
priceChange

## [1] -8.440424
```

Der mittlere Preis würde sich z.B. um 8.44 Euro verringen, wenn man 3 km zusätzliche Entfernung in Kauf nimmt.

c)

```r
ggplot(airbnb, aes(x = Distanz, y = Preis)) + geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Distanz zum Stadtzentrum (Km)", y = "Preis (Euro)") +
  scale_x_continuous(breaks = seq(0.5, 9.5, by = 1), limits = c(0.5, 10)) +
  scale_y_continuous(breaks = seq(15, 140, by = 15), limits = c(15, 140))
```

    ## `geom_smooth()` using formula = 'y ~ x'



d)

```r
price <- 60
dist_60 <- (price-b_0)/b_1
dist_60

## [1] 4.757386
```

Gemäß des Modells wäre die Distanz zum Stadtzentrum etwa 4.76 km, um einen mittleren Preis von 60 Euro zu erhalten.

4 Breite von Konfizendintervallen

a)

Die minimale Stichprobengröße n, die für $X_1, \cdots, X_n \overset{iid}{\sim} Bern(p)$ ein 90%-KI mit $max.B = 0.05$ garantiert, lautet:

$$n \geq \frac{z^2_{1-(\alpha/2)}}{B^2} = \frac{1.6449^2}{0.05^2} = 1082.278 \approx \underline{1083}.$$

b)

Die minimale Stichprobengröße n, die für $X_1, \cdots, X_n \overset{iid}{\sim} Bern(p)$ ein 95%-KI mit $max.B = 0.05$ und $\hat{p} \leq 0.2$ garantiert, lautet:

$$n \geq \frac{4z^2_{1-(\alpha/2)}\hat{p}(1-\hat{p})}{B^2} = \frac{4z^2_{0.975} \cdot 0.2 \cdot 0.8}{0.05^2} = \frac{4 \cdot 1.96^2 \cdot 0.2 \cdot 0.8}{0.05^2} = 983.45 \approx \underline{984}.$$

c)

Die minimale Stichprobengröße n, die für $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ ein 90%-KI mit $max.B = 0.05$ und $\sigma = 1$ garantiert, lautet:

$$B = 2z_{1-\alpha/2}\sqrt{\frac{Var(x)}{n}}$$
$$\Leftrightarrow B = 2z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$
$$\Leftrightarrow B = 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$\Leftrightarrow n = \left(\frac{2z_{1-\alpha/2}\sigma}{B}\right)^2$$

$$\Rightarrow n \geq \left(\frac{2z_{1-\alpha/2}\sigma}{B}\right)^2 = \left(\frac{2 \cdot 1.6449 \cdot 1}{0.05}\right)^2 = 4329.114 \approx \underline{4330}.$$

Wenn σ bzw. σ^2 unbekannt ist, gäbe es zwei unbekannte Variablen in der Rechnung und man müsste eine andere Stichprobenstandardabweichung nutzen. Dies würde dazu führen, dass eine t-Verteilung (also Student) anstatt der vorgegebenen Normalverteilung genutzt werden muss.