Tut 2

```
faithful <- faithful
waiting <- faithful$waiting
eruptions <- faithful$eruptions
```

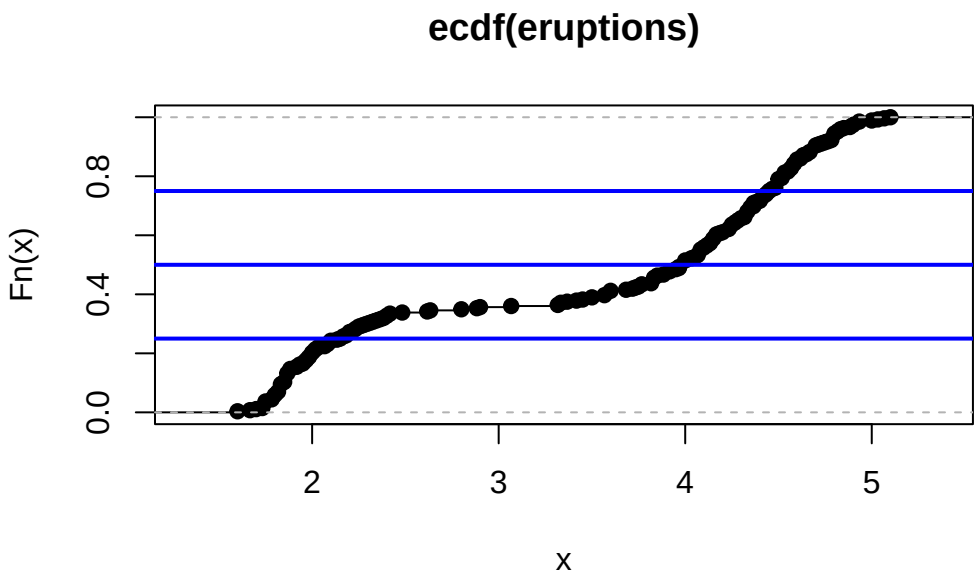
Aufgabe 1: Explorative Datenanalyse und ECDF

a)

Erstellen Sie mit plot(ecdf(eruptionen)) die empirische Verteilungsfunktion der insgesamt 272 Beobachtungen. Identifizieren Sie (grob) anhand des Plots das 25%-, 50%- und 75%-Quantil. Überprüfen Sie anschließend Ihre Werte mit dem summary(…) Befehl.

```
plot(ecdf(eruptions))

abline(h = 0.25, col = "blue", lwd = 2) # 25% Quantil
abline(h = 0.5, col = "blue", lwd = 2) # 50% Quantil
abline(h = 0.75, col = "blue", lwd = 2) # 75% Quantil
```




```
summary(eruptions)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.600   2.163   4.000   3.488   4.454   5.100
```

1. **Quartil:**Liegt bei einer Eruptionsdauer von etwa 2,2
2. **Quartil:**Liegt bei einer Eruptionsdauer von etwa 4,0
3. **Quartil:**Liegt bei einer Eruptionsdauer von etwa 4,5

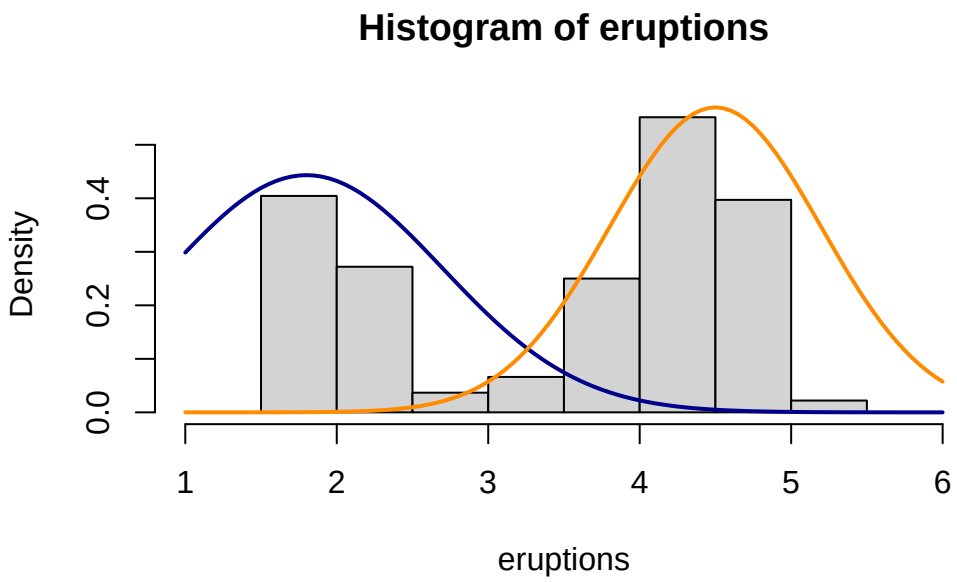
b)

Betrachten Sie nun ein Histogramm der Daten:

```
hist(eruptions, breaks = seq(1.5, 5.5, 0.5), prob = TRUE)
hist(eruptions, breaks = seq(1.5, 5.5, length = 8 + 1), prob = TRUE)
```

Das Histogramm hat eine bimodale Verteilung und könnte durch zwei Gammaverteilungen oder Normalverteilungen beschrieben werden.

```
hist(eruptions, breaks = seq(1.5, 5.5, 0.5), prob = TRUE,
xlim = c(1, 6))
curve(dnorm(x, mean = 1.8, sd = 0.9),
col = "darkblue", lwd = 2, add = TRUE)
curve(dnorm(x, mean = 4.5, sd = 0.7),
col = "darkorange", lwd = 2, add = TRUE)
```



c)

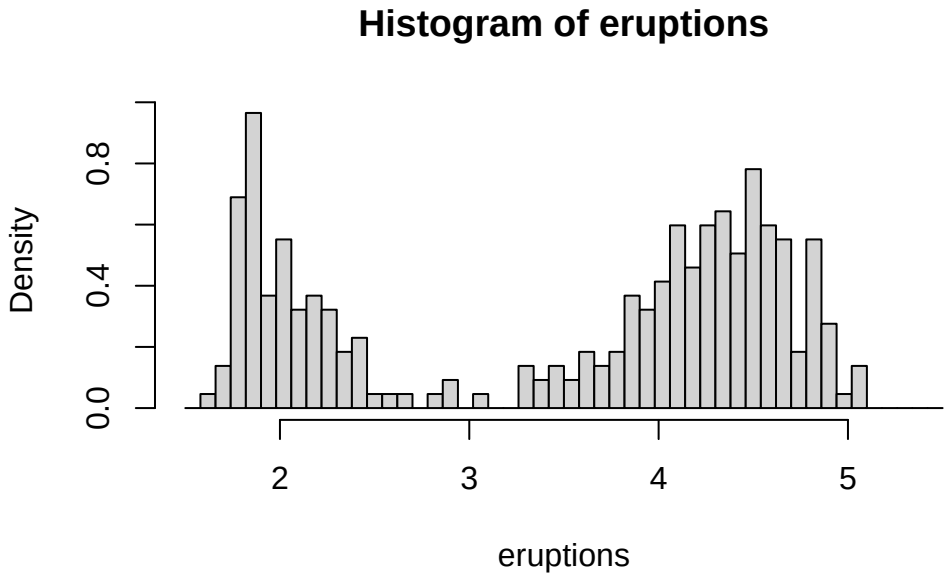
Erstellen Sie Histogramme mit verschiedenen Klassenzahlen (z.B., 2, 10, 50, 100) und erinnern Sie sich daran, inwiefern dies den Bias-Varianz-Trade-off illustriert. Wie viele Klassen würden Sie für die Eruptionsdauer wählen?

```
hist(eruptions, breaks = seq(1.5, 5.5, length = 3), prob = TRUE)
```

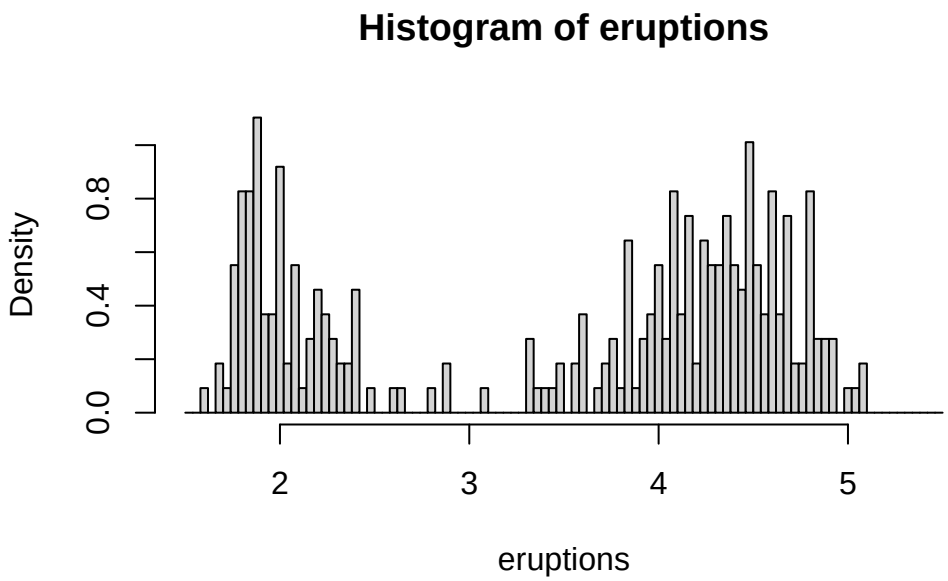
```
hist(eruptions, breaks = seq(1.5, 5.5, length = 11), prob = TRUE)
```

```
hist(eruptions, breaks = seq(1.5, 5.5, length = 21), prob = TRUE)
```

```
hist(eruptions, breaks = seq(1.5, 5.5, length = 51), prob = TRUE)
```



```r
hist(eruptions, breaks = seq(1.5, 5.5, length = 101), prob = TRUE)
```



Wenige Klassen: Histogramm ist nicht wirklich an die Daten angepasst Hoher Bias (Form nicht erkennbar), niedrige Varianz

Viele Klassen: Histogramm ist stark an die Daten angepasst Niedriger Bias, hohe Varianz

10 Klassen scheinen eine gute Balance zu finden

Aufgabe 2: Histogramme mit dynamischen Klassengrenzen

Diese Aufgabe befasst sich mit der Modifikation des Histogramms, bei dem die Klassengrenzen dynamisch sind (siehe Slide 22 der 2. Vorlesung). Bestimmen Sie für die Eruptionsdauer aus Aufgabe 1 den Wert des entsprechenden Schätzers f(x) an der Stelle x = 3.75 für b = 0.5.

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} \mathbb{I}\left(x_i \in \left[x - \frac{b}{2}, x + \frac{b}{2}\right]\right)$$

```r
x <- 3.75
b <- 0.5
fxhat <- 1/(length(eruptions) * b) *
sum(x-b/2 <= eruptions & eruptions <= x+b/2)
fxhat
```

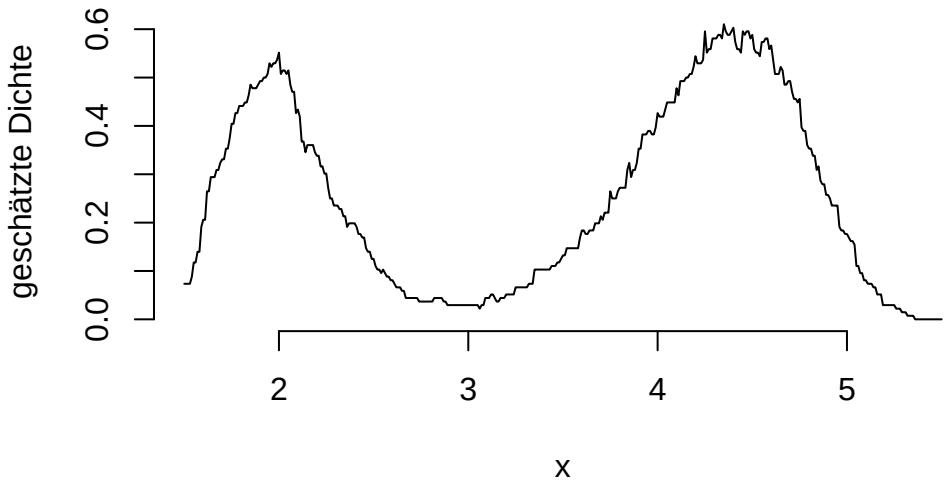
```
[1] 0.2647059
```

Die Plausibilität dieses Wertes kann anhand eines Histogramms aus der vorherigen Aufgabe überprüft werden.

Zusatzaufgabe (falls Sie gerne programmieren):

Schreiben Sie eine for-Schleife, welche die Berechnung für x =1.5,1.51,1.52,...,5.48,5.49,5.5 durchführt. Plotten Sie anschließend die so erhaltenen Werte gegen x, um die gesamte geschätzte Funktion zu erhalten. Sie könnten anschließend auch noch den Wert von b vari ieren.

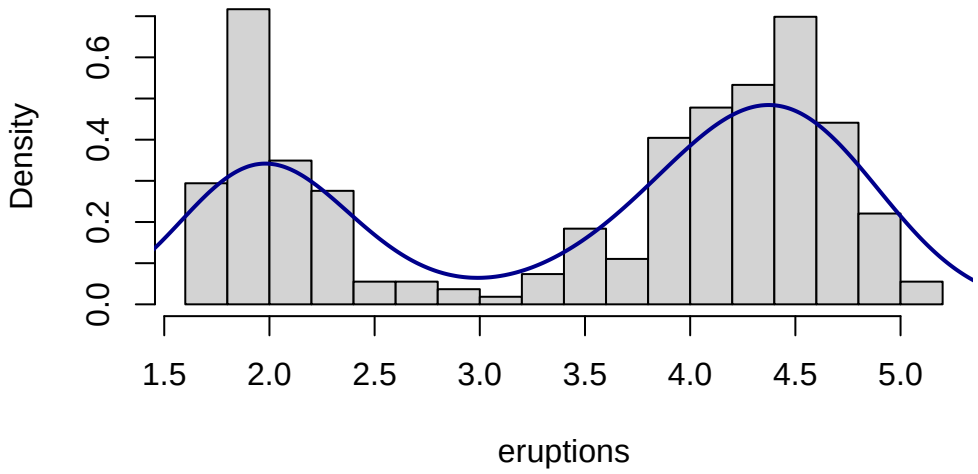
```
b <- 0.5
x <- seq(1.5, 5.5, by = 0.01)
fxhat <- rep(NA, length(x))
for(k in 1:length(x)){
fxhat[k] <- 1/(length(eruptions)*b)*
sum(x[k]-b/2 <= eruptions & eruptions <= x[k]+b/2)
}
plot(x, fxhat, type = "l", ylab = "geschätzte Dichte", bty = "n")
```



Aufgabe 3: Kerndichteschätzer

Mit den folgenden Befehlen erstellen Sie ein Histogramm der 272 Beobachtungen und fügen anschließend dem gleichen Plot einen Kerndichteschätzer hinzu:

```
hist(eruptions, breaks = 15, prob = TRUE, main = "")
lines(density(eruptions), col = "darkblue", lwd = 2)
```



a)

Welche beiden wesentlichen Vorteile hat der Kerndichteschätzer gegenüber dem His togramm?

1. ist nicht konstant auf von uns festgelegten Intervallen (wie beim Histogramm)

2. Stetigkeit (sofern nicht der Rechteckkern verwendet wird)

b)

Nutzen Sie die Hilfefunktion (?density), um herauszufinden, welche Kernfunktion und welche Bandweite von density(…) per default zur Kerndichteschätzung genutzt wer den.

1. Kern: Gaußkern

2. Bandweite: “nrd0”-Schätzer nach Silverman Definiert die Breite des Einflussbereichs jedes Datenpunkts