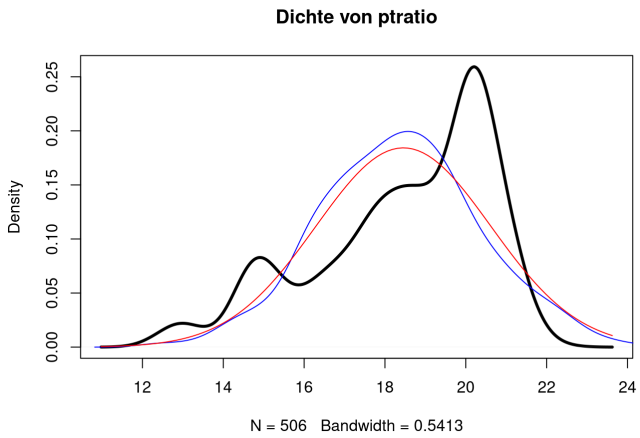
SWUE Projekt - Aufgabe 7

Tobias, Julia, Markus

Analyse und Berechnungen:



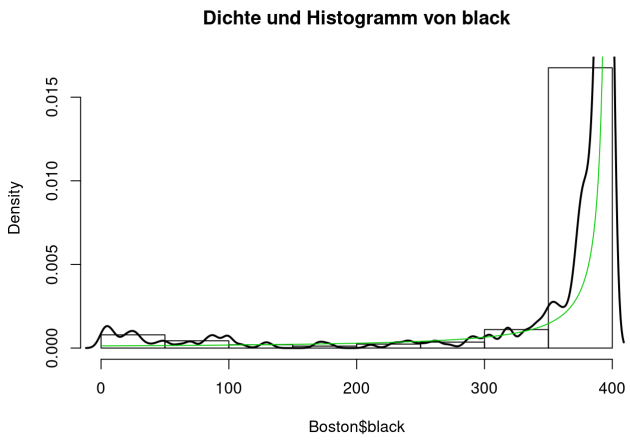
Boston$ptratio:

Schüler-Lehrer-Verhältnis nach Stadt.
Durch die density Funktion (schwarze Linie) kann man sehen und durch die Schiefe kann man berechnen, dass die Verteilungsform eher linksschief ist (Schiefe < 0).
Anfangs angenommen es ist normalverteilt, aber man sieht, dass das nicht so ist.
blaue Linie = rnorm density; rote Linie = dnorm
Die Verteilung hat 2 Peaks: bei 14 bis 15 und 19 bis 20.5.

5 (bzw. 6) Zahlen Zusammenfassung:

```
  summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```



Boston$black:

1000 (Bk - 0,63) ^ 2 wobei Bk der Schwarzanteil der Stadt ist.
Verteilung ist linksschief (Schiefe kleiner 0).

```
  skewness(Boston$black)
```

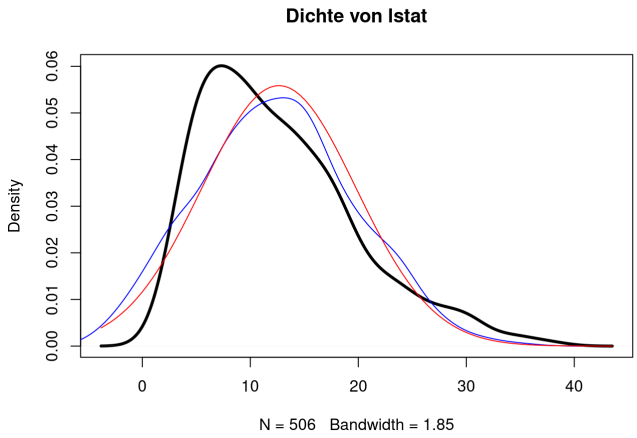
```
## [1] -2.881798
```

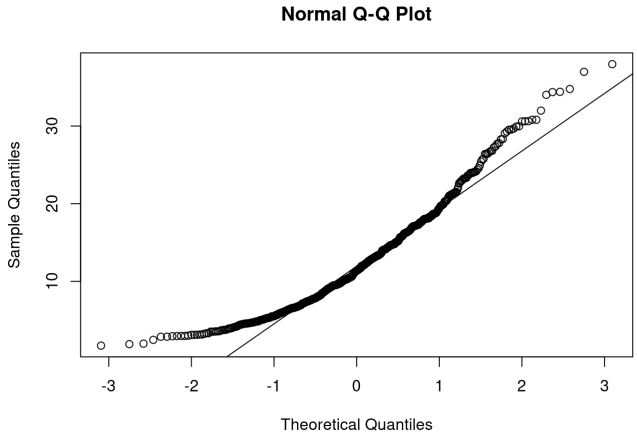
Annäherung an Paetro-Verteilung mit Intervall $[0, \infty]$ Die Paetro-Verteilung (grüne Linie) hat hier die meisten Vorkommen bei den höheren Werten.

Berechnung des Paramters $\hat{\xi}$:

```
  min(Boston$black)
```

```
## [1] 0.32
```





Boston$lstat:

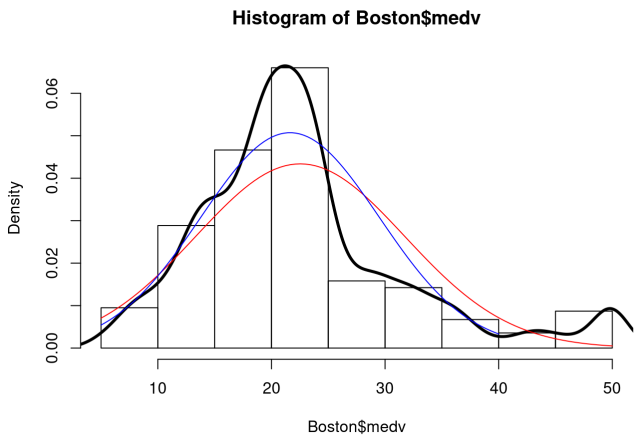
Prozentanteil der Bevölkerung mit niedriger Position in der sozialen Hierarchie. Dh. schlechte menschliche Lebensumstände (z.B.: wenig Bildung, keinen Schulabschluss, keine Ausbildung oder Studium, geringes Einkommen, Migrationshintergrund, …)
Annahme: ptratio ist normalverteilt
die blaue und rote Linie zeigen, dass die Annahme stimmt
blaue Linie=rnorm density
Parameter μ und σ:

```
  mean(Boston$lstat)
```

```
## [1] 12.65306
```

```
  sd(Boston$lstat)
```

```
## [1] 7.141062
```



Boston$medv

mittlerer Wert von Wohneigentum in $1000. Annäherung an Normalverteilung, aber rechtsschief (Schiefe > 0) und leptokurtisch (steilgipfelig) (Kurtosis > 3).

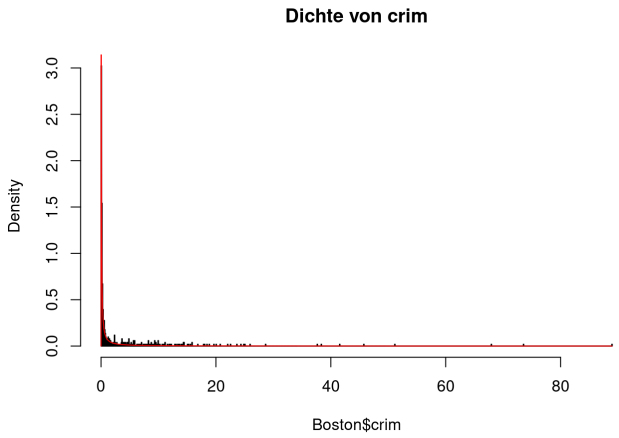
```
skewness(Boston$medv)
```

```
## [1] 1.104811
```

```
kurtosis(Boston$medv)
```

```
## [1] 4.468629
```

rote Linie = Normalverteilung, blaue Linie = Normalverteilung von medv Werten < 50 (Ausreißer größer/gleich 50 weggeschnitten)



Crim beschreibt die Verbrechensrate pro Einwohner einer Stadt. Die Größe lässt sich gut mit einer Pareto-Verteilung beschreiben. Die Parameter errechnen sich wie folgt:

$$\hat{\xi} = min_{1 \leq i \leq n}\, x_i = 0.00632$$

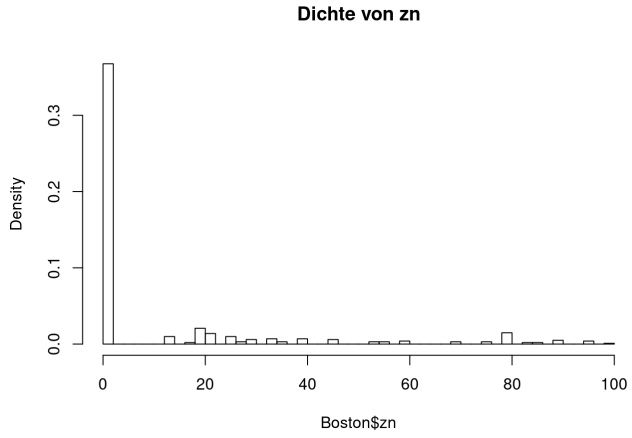
$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} log\left(\frac{x_i}{\hat{\xi}}\right)} = 0.00645905$$

Die Pareto-Verteilung ereignet sich gut für Datensätze die sich über mehrere Größenordnungen erstrecken. Das ist hier der Fall, da $\frac{max}{min} \approx 14000$ ist. Diese Paretto-Verteilund ist jedoch auf dem Interval $(0, \infty]$ definiert ist, und unsere Daten nur im Interval $[\hat{\xi}, 100]$ auftreten können könnte man denken, dass hier auch eine Exponentialverteilung mit $\tau = \frac{1}{\overline{X}} = 0.2767382$ zur Beschreibung verwendet werden kann. Jedoch fällt diese Kurve zu schnell ab und die Verteilungsfunktion erzeugt bereits bei $F(17)$ Werte jenseits von 99%. Daher wird zu Beschreibung der Verteilung eine angepasste Pareto-Verteilung verwendet. Dabei wird die Verteilungsfunktion so konstruiert, dass $F(100) = 1$ gilt. Die normale Verteilungsfunktion F_p hat den Wert $F_P(100) = 0.8953634$.

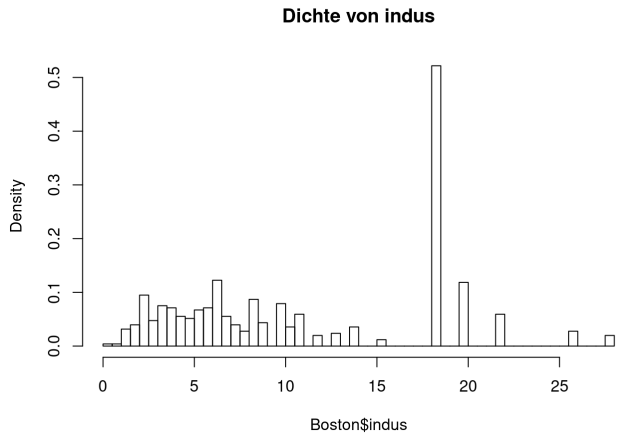
Nun konstruiert man die adjustierte Verteilungsfunktion $F = cF_p$, wobei c der Reziprokwert von $F_P(100)$ ist. Daher gilt die Verteilungsfunktion

- wenn $n < \hat{\xi}$: $F(n) = 0$
- wenn $n \in [\hat{\xi}, 100]$: $F(n) = 1.116865 F_P(n)$
- wenn $n > 100$: $F(n) = 1$

Die Dichtefunktion $F'(n)$ ist aufgrund der Linearität des Differenzialqoutienten $f(n) = 1.116865 f_P(n)$, wobei $f_P(n)$ die Dichtefunktion der Pareto-Verteilung mit den obigen Parametern ist.



Zn beschreibt den Anteil der Wohngrundstücke für Grundstücke mit mehr als 25.000 sq.ft. Es gibt viele 0-Werte (ca. 73%). Die restlichen Werte scheinen keiner konkreten Verteilung zu folgen. Am ehesten würden sich 2 weitere skalierte (Da diese maximal ~0.27 als Summe haben dürften) Binomialverteilungen (die Werte für zn sind nur $z \in \{0, 1, \ldots 99, 100\}$) mit den Mittelpunkten 20 und 80 eigenen, da es hier kleinere Spitzen in den Frequenzen gibt. Darüber hinaus könnte es auch eine Gleichverteilung im Intervall $[12.5; 100]$ sein.



Indus beschreibt den Anteil der Industriefläche pro Stadt. Wenn man sich das Datenset ansieht, kann man erkennen, dass Werte öfters vorkommen (zb. 18.10 kommt 132 Mal vor). Daher nehmen wir an, dass mehrere Vororte zu einem Industriegebiet zusammengefasst wurden. Für jeden Ort, der Teil eines Industriegebietes ist, wurde der Wert des gesamten Industriegebietes verwendet. Wenn diese Größe mit Hilfe einer Funktion beschrieben werden soll, dann würde sich eine skalierte Normalverteilung mit $\mu = 6$ und eine skalierte Exponential-Verteilung (oder ähnliches) ab ca. 18.

Der Anteil der Städte aus der Stichprobe, welche am Charles River liegen.

```
## [1] 0.06916996
```

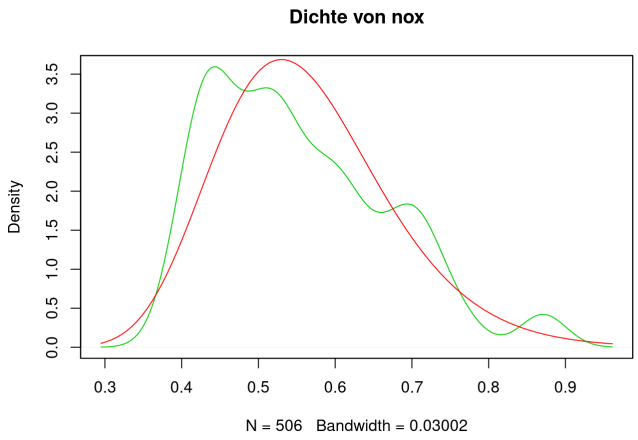
Der Anteil der Städte aus der Stichprobe, welche nicht am Charles River liegen.

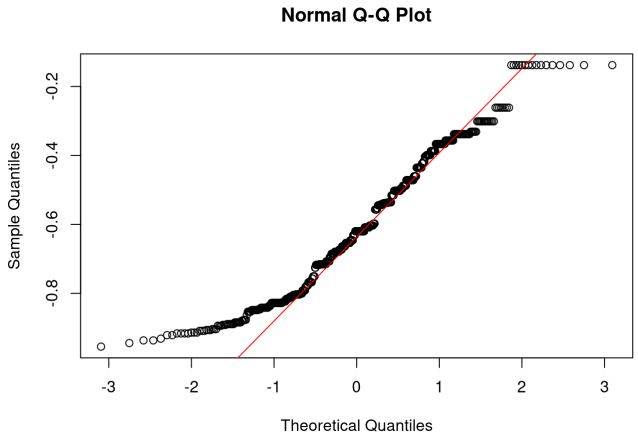
```
## [1] 0.93083
```

Chas gibt ab, ob der Vorort an den Charles River angrenzt. Hierbei handelt es sich um eine Bernoulli Verteilung. Die Wahrscheinlichkeit, dass das Bernoulli Experiment erfolgreich ist, lässt sich wie folgt berechnen:

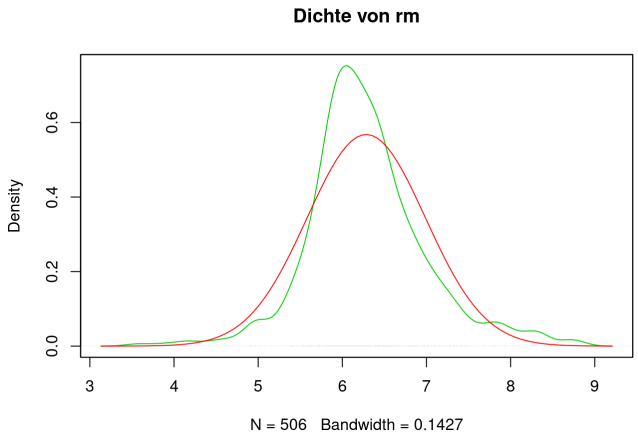
$$\hat{p} = \frac{\hat{p}_t}{n} = 0.06916996$$

Wobei $\hat{p}_t$ die Anzahl der erfolgreichen Experimente in der Stichprobe ist.





Nox gibt die Konzentration von Stickstoff-Oxiden an. Die Größe scheint nicht ganz normalverteilt zu sein, da die Dichtefunktion (siehe Plot links) nicht symmetrisch scheint - sie steigt stärker an, als sie abfällt. Daher wird versucht, die Größe mit einer logarithmischen Normalverteilung zu beschreiben. Die transformierte Zufallsvariable $Y = log(X)$ ist annähernd normalverteilt (siehe QQPlot). Jedoch ist das Ergebnis nur marginal besser, als durch eine normale Normalverteilung. Um die Zufallsvariabel Y analysieren zu können, wurde für jedes Element x_i aus der Stichprobe $y_i = log(x_1)$ gesetzt. Der so gewonnene Datensatz repräsentiert nun eine Stichprobe der Zufallsvariable Y.

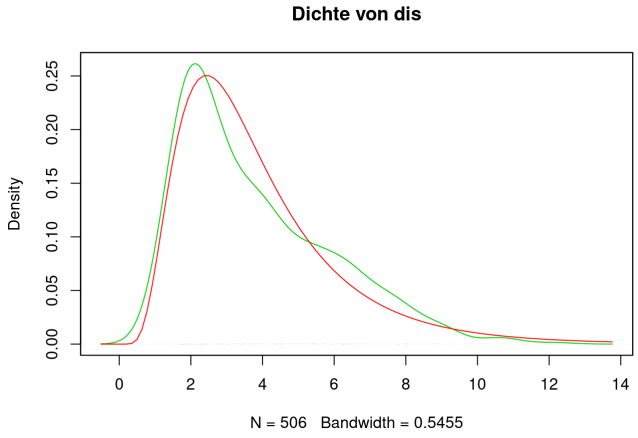


Der rm wert des Datensatzes boston beschreibt die durchschnittliche Nummer an Räumen von Wohnungen in Boston. Die Verteilung der Daten lässt sich gut mit einer Normalverteilung (siehe rote line) darstellen. Die Formel für die Normalverteilung ist:

$$\frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

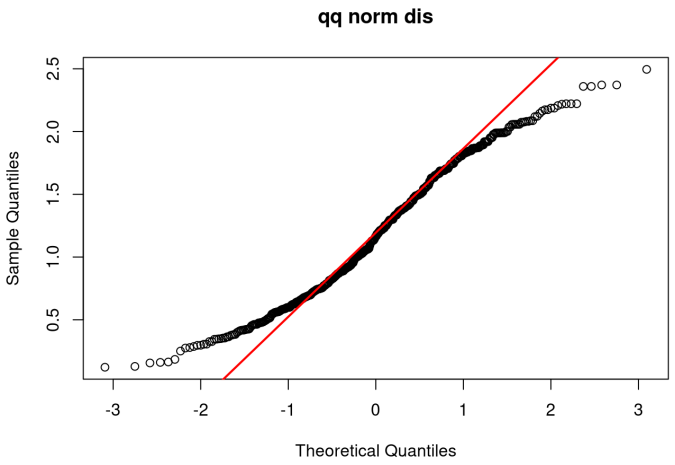
Wobei $\mu \in \mathbb{R}$ den Mittelwert und $\sigma > 0$ die Varianz darstellet. Bei dieser Verteilung sind die Werte: μ = 6.285 und σ = 0.702

Der age wert des Datensatzes Boston beschreibt den Anteil der Eigentumswohnungen, die vor 1940 gebaut wurden. Diese Werte stellen stellen keine konkrete Verteilung dar.

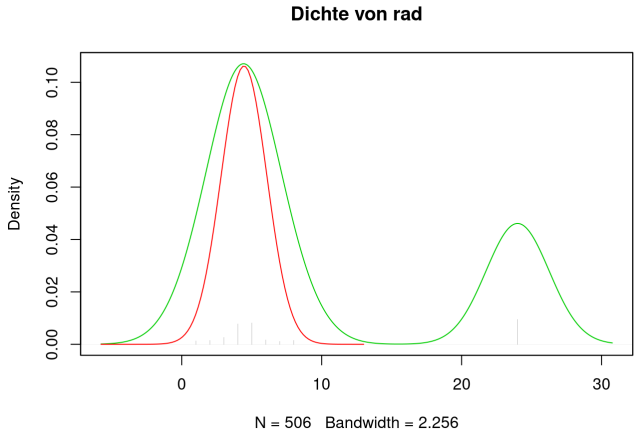


Der dis wert des Datensatzes Boston beschreibt den gewichteten Mittelwert der Entfernungen zu den fünf Bostoner Beschäftigungszentren. Die Verteilung der Daten lässt sich gut mit einer logarithmischen Normalverteilung beschreiben (siehe rote Line). Da die Werte auf der rechten Seite langsamer sinken als bei einer normalverteilung wird versucht die Kurve mittels einer logarithmischen Normalverteilung zu approximieren. Die Formel für die Logarithmische Normalverteilungist:

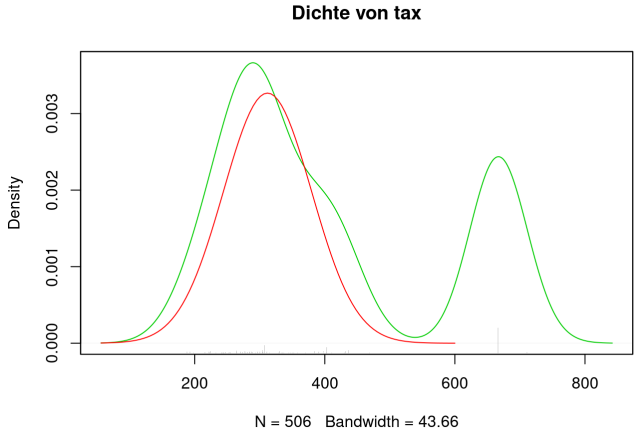
$$\frac{1}{\sigma x\sqrt{2\pi}} exp(-\frac{(ln(x)-\mu)^2}{2\sigma^2})$$



Die Transofmierte Zufallsvariable Y = ln(X) folgt ebenfalls einer Normalverteilung. (siehe qq norm dis) Die Werte dieser Funktion sind: μ =1.203 und σ = 0.559



Der rad wert des Datensatzes Boston beschreibt die Zugänglichkeit zu radialen Autobahnen. Die Kurve auf der linken Seite lässt sich annähernd mit einer Normalverteilung beschreiben. Die Kurve auf der rechten Seite ist nur ein Wert und stellt somit keine Verteilung dar. Die Werte der Normalverteilung sind: μ = 4.449 und σ = 1.633.



Der tax Wert des Datensatzes Boston beschreibt den Immobiliensteuersatz in voller Höhe pro 10.000 USD. Die Kurve auf der linken Seite lässt sich annähernd mit einer Normalverteilung beschreiben. Die Kurve auf der rechten Seite ist nur ein Wert und stellt somit keine Verteilung dar. Die Werte der Normalverteilung sind: μ = 311.927 und σ = 67.828.

Hypothese für Boston$lstat

Der Erwartungswert der Verteilung wird mit 11% angenommen, durch die aktuelle Flüchtlingskrise, wird eine Vergrößerung des Prozentanteils der Bevölkerung mit niedriger Position in der sozialen Hierarchie erwartet. Wird diese Erwartung bei einem Signifikantsniveau von 2.5% bestätigt?

Um die Erwartung zu bestätigen oder wiederlegen werden die folgenden Hypothesen aufgestellt:
$H_0 : \mu = 11, H_1 : \mu > 11$ (rechtsseitiger Test)
Wie oben festgestellt ist die Verteilung annähernd normalverteilt, und es handelt sich um eine große Stichprobe (506 Elemente). Daher kann man hier die Z-Statistik verwenden.
Die Erwartung kann bestätigt werden, wenn $Z > Z_\alpha$ bzw. wenn Z im kritischen Bereich liegt. kritischer Bereich: $[1.959964, +\infty]$
$Z = 28.2071466$
Z liegt in dieser kristischen Region, daher wird die Nullhypothese verworfen.

Hypothese für Boston$lstat (Boston$ptratio vorgegeben)

Ich teile den Datensatz Boston auf, sodass der 1. Teil ein Schüler-Lehrer Verhältnis pro Stadt von kleiner 19 haben und der 2. Teil ein Schüler-Lehrer Verhältnis pro Stadt größer gleich 19 (ptratio - nach Mittelwert aufgeteilt)

```
  mean(Boston$ptratio)
```

```
## [1] 18.45553
```

Dadurch erhält man im 1. Teil 249 Städte und im 2. Teil 257 Städte. Basierend auf der Stichprobe beträgt der durchschnittliche Prozentanteil der Bevölkerung mit niedriger Position in der sozialen Hierarchie im 1. Teil 9.5736948 und im 2. Teil 15.6365759. ($\hat{x}_1 und \hat{x}_2$) Nun wollen wir sehen ob bei einem gegebenen Signifikantsniveau von 1% der Durchschnitt des Prozentanteils der Bevölkerung mit niedriger Position in der sozialen Hierarchie des ersten Teils niedriger ist als der des 2. Teils ($\mu_1 und \mu_2$).

$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 < \mu_2$ (linksseitiger Test) Wie oben festgestellt ist Boston$lstat annähernd normalverteilt, es handelt sich um große Stichproben (506 Elemente) und die 2 Teile sind unabhängig voneinander. Daher kann man hier die Z-Statistik verwenden. (mit der Formel für unbekannte Standardabweichungen)

Die Nullhypothese kann verworfen werden, wenn $Z < Z_\alpha$ bzw. wenn Z im kritischen Bereich liegt. kritischer Bereich: $[-\infty, -2.3263479]$ Es wurde berechnet, dass Z = -26.9319105. Dieser Wert fällt in die rejection region und die Nullhypothese wird daher verworfen. Somit kann bestätigt werden, dass in Städte mit einem Schüler-Lehrer Verhältnis kleiner 19, der Durchschnitt des Prozentanteils der Bevölkerung mit niedriger Position in der sozialen Hierarchie auch kleiner ist als in Städten mit einem Schüler-Lehrer Verhältnis größer gleich 19.

Hypothese für Boston$rm

Marty Walsh, Der Bürgermeister von Boston, weiß, dass im Jahr 1970 die durchschnittliche Raumanzahl der Vororte bei 6.15 lag. Er möchte nun wissen, ob zur Zeit der Stichprobe, diese Behauptung noch gilt. Dabei soll $\alpha = 0.1$ verwendet werden.

Für die Verteilung von rm, gilt: $\hat{\mu} = 6.2846344$ und $s = 0.7026171$. $\mu = 6.15$, ist der angenommene Erwartungswert der Population (aus der Aufgabenstellung). Die Varianz der Population ist nicht gegeben. Da jedoch $n = 506 \geq 30$ ist, kann hier die Normalverteilung mit $\sigma = s$ verwendet werden. Daher kann eine Z-Statistik verwendet werden.

Die Hypothesen sind wie folgt:

- $H_0 : \mu = 5.8$
- $H_1 : \mu \neq 5.8$

Da es sich hier um einen beidseitigen Test handelt, bleibt die Nullhypothese bestehend, wenn $\hat{\mu}$ im 90%-Konfidenzintervall von $N(\mu, \sigma^2)$ liegt.

Dieses kann mit $\mu \pm z_{0.95} \frac{\sigma}{\sqrt{n}}$ berechnet werden.

Lower Bound:

```
## [1] 6.098623
```

Upper Bound:

```
## [1] 6.201377
```

Das Konfidenzintervall ist $[6.098623, 6.201377]$. Damit liegt $\hat{\mu}$ nicht innerhalb und die Nullhypothese wird zugunsten von H_1 verworfen.

Hypothese für Boston$dis

Der/Die Verantworliche des Bostoner Arbeitsmarkt legt großen Wert darauf, dass Beschäftigungszentren für die meisten Menschen leicht erreichbar sind. Er/Sie möchte sich sicher sein, dass der mittlere errechnete Abstand nicht abgenommen hat. Da zur Zeit nur wenig Budget verfügbar ist, soll eine schnelle Berechnung durchgeführt werden, die eine Entscheidungsgrundlage bietet, ob weitere Untersuchungen notwendig sind. Da die Mittelwerte der Verteilung bei größeren n normalverteilt sind, soll das aufgrund der Verteilung der Mittelwerte entschieden werden. Die damals gezogene Stichprobe (gleiches n) hatte $\overline{X}_o = 3.9$ und $\sigma_o^2 = 2.1^2$

Für die gezogene Stichprobe gilt $\hat{\mu} = 3.795043$. Sie können davon ausgehen, dass σ_o^2 und σ_n^2 gleuch sind.

Da es viele bei dieser Berechnung viele Unsicherheiten gibt, soll $\alpha = 0.2$ verwendet werden. Falls die Nullhypothese verworfen wird, folgen weitere Untersuchungen.

Muss der/die Verantwortliche weitere Untersuchungen einleiten?

Es werden folgende Hypothesen für die Verteilung der Mittelwerte verwendet:

- $H_0 : \mu = 3.9$
- $H_1 : \mu < 3.9$

Laut dem Zentralen Grenzwertsatz, der hier zum Einsatz kommen kann, weil $n = 506 \geq 30$ ist, sind die Mittelwerte einer solchen Stichproche $N(\mu, \sigma^2/\sqrt{n})$ verteilt.

```
# z-Wert
(mean(Boston$dis) - 3.9) / (2.1 / sqrt(506))
```

```
## [1] -1.124265
```

```
qnorm(0.2)
```

```
## [1] -0.8416212
```

Der kritische Wert ist -0.8416212. Da die Rejection Region sich auf der linken Seite des kritischen Wertes ist und der z-Wert kleiner ist, wird die Nullhypothese verworfen. Es sollten weitere Untersuchungen angestellt werden.

Hypothese für Boston$medv

Die Verwaltung von Bosten weiß, dass der Durchschnitt des durchschnittlichen Werts für Wohneigentum bei 22.5328063 * 1000 Doller liegt. Es wird angenommen, dass das Bevölkerungswachstum in Bosten in den nächsten Jahren weniger stark wächst als in den Jahren zuvor und daher der mittlere Wert sinkt. Da auf der offiziellen Website ein Wert Durchschnittswert angegeben werden muss, welcher möglichst lange gültig sein soll wird μ = 22 verwendet. Ist dies bei einem α Wert von 10% zulässig ?
Des Weiteren würde die Verwaltung gerne wissen, in welchem Bereich diese Angabe verwendet werden kann.
n = $8898 \geq$ 30; Daher kann die Z-Statistik verwendet werden.
$\alpha = 5\%$
$\overline{x} = 311.9268293$
$s = 67.8282883$
$H_0 : \mu = 22$
$H_1 : \mu > 22$

```
m <- mean(Boston$medv)
s <- sd(Boston$medv)
mu <- 22
z <- (m-mu)/(s/(sqrt(length(Boston$medv))))
p <- pnorm(z, lower.tail = FALSE)
show(p)
```

```
## [1] 0.09626223
```

```
show(p < 0.1)
```

```
## [1] TRUE
```

Da der p-Wert unter 0.1 liegt, akzeptieren wir die Nullhypothese H_0.

Kritischer Bereich:

```
z <- qnorm(0.05)
lower <- z*(s/sqrt(length(Boston$medv)))+mu
higher <- -z*(s/sqrt(length(Boston$medv)))+mu
```

Der kritische Bereich für $alpha = 0.05$ ist: $21.3274833 \leq \overline{x} \leq 22.6725167$. Das bedeutet, dass H_0 gültig ist, solange der wahre Mittelwert in diesem Bereich liegt.