Anhang 2. Projekt DLBDBSC01

Code

2023-04-12

Deskriptive Statistik Ermitteln der empirischen Verteilungsfunktion: Test auf Normalverteilung der Variablen “SpendingProPerson”

Hide

```
#Mittelwerte und Standardabweichungen berechnen
Mittelwert_Spending<-mean(Arbeitstabelle02$SpendingProPerson)
Standardabweichung_Spending<-sd(Arbeitstabelle02$SpendingProPerson)
Mittelwert_Spending
```

```
[1] 4281.219
```

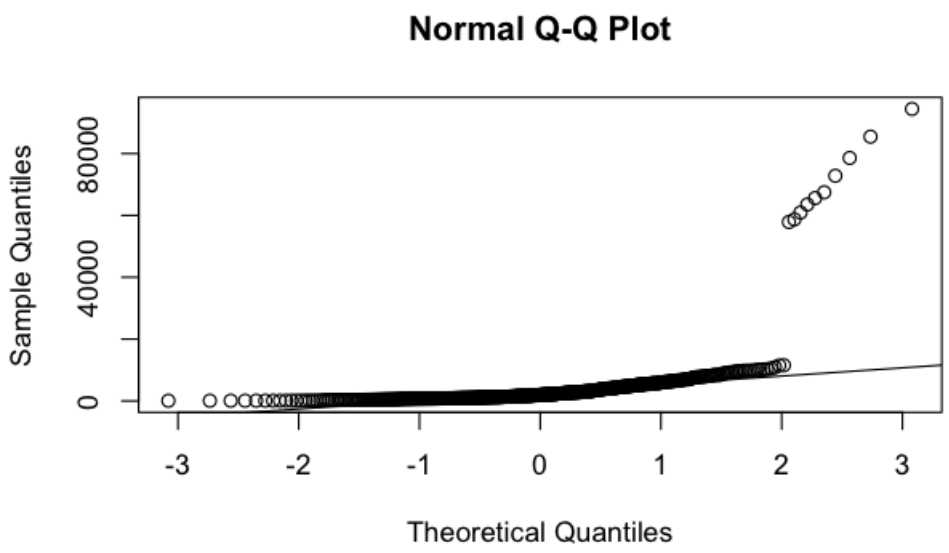
Hide

```
Standardabweichung_Spending
```

```
[1] 10131.34
```

Hide

```
#Q-Q Plot erstellen
qqnorm(Arbeitstabelle02$SpendingProPerson)
qqline(Arbeitstabelle02$SpendingProPerson)
```



Ermitteln der empirischen Verteilungsfunktion: Test auf Normalverteilung der Variablen “InternetAnteil”

Hide

```
Mittelwert_Internet<-mean(Arbeitstabelle02$InternetAnteil)
Standardabweichung_Internet<-sd(Arbeitstabelle02$InternetAnteil)
Mittelwert_Internet
```

```
[1] 73.49567
```

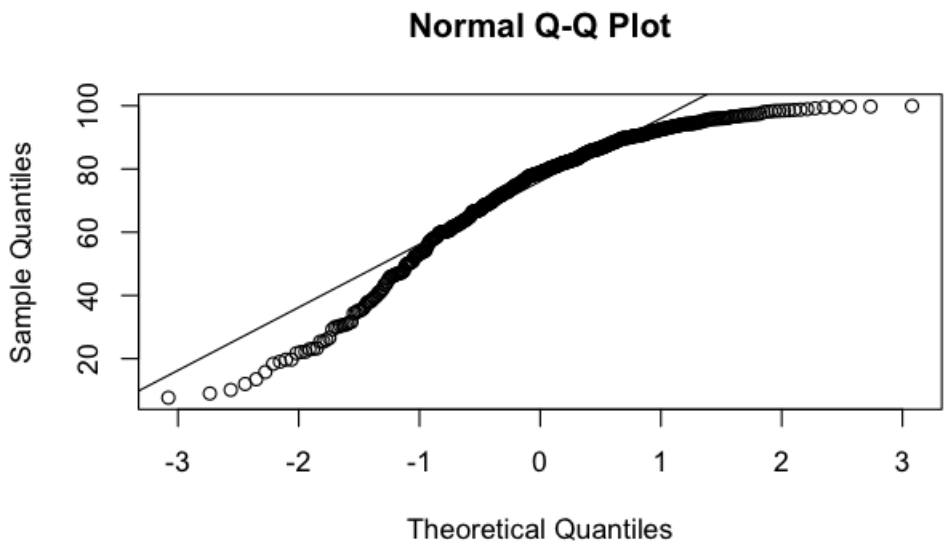
Hide

```
Standardabweichung_Internet
```

```
[1] 20.21313
```

Hide

```
#Q-Q Plot erstellen
qqnorm(Arbeitstabelle02$InternetAnteil)
qqline(Arbeitstabelle02$InternetAnteil)
```



Hide

```
# Erstellen eines Datensatzes mit zwei Variablen
set.seed(123)
data <- data.frame(
 SpendingVar = rnorm(100, mean = Mittelwert_Spending , sd = Standardabweichung_Spending),
  InternetVar = rnorm(100, mean = Mittelwert_Internet, sd = Standardabweichung_Internet))
# Shapiro-Wilk-Test für jede Variable
shapiro.test(data$SpendingVar)
```

```
    Shapiro-Wilk normality test

data:  data$SpendingVar
W = 0.99388, p-value = 0.9349
```

Hide

```
shapiro.test(data$InternetVar)
```

```
    Shapiro-Wilk normality test

data:  data$InternetVar
W = 0.97289, p-value = 0.03691
```

Schiefe und Kurtosis der Variable “InternetAnteil” berechnen, um sie in Log zu transformieren (wegen nicht-normalverteilte Variable)

Hide

```
# Das Paket "e1071" laden
# install.packages("e1071")
library(e1071)
# Berechnen der Schiefe und Kurtosis der Variable "InternetAnteil"
Schiefe_InternetAnteil<-skewness(Arbeitstabelle02$InternetAnteil)
Kurtosis_InternetAnteil<-kurtosis(Arbeitstabelle02$InternetAnteil)
Schiefe_InternetAnteil
```

```
[1] -1.06216
```

Hide

```
Kurtosis_InternetAnteil
```

```
[1] 0.5743557
```

Wenn Schiefe oder Kurtosis negativ sind, keine Log-Transformation möglich Wurzel- Transformation testen (für rechtsschief verteilte Variable), anschließend auf Normalverteilung testen

Hide

```
# Erstellen einer Wurzel-transformierten Variable
Arbeitstabelle02$sqrt_InternetAnteil <- sqrt(Arbeitstabelle02$InternetAnteil)
# Shapiro-Wilk Test auf Normalverteilung für die neue Variable
# Berechnen der Kenngrößen für den Shapiro-Wilk Test
Mittelwert_sqrt_Internet<- mean(Arbeitstabelle02$sqrt_InternetAnteil)
Standardabweichung_sqrt_Internet<-sd(Arbeitstabelle02$sqrt_InternetAnteil)
Mittelwert_sqrt_Internet
```

```
[1] 8.465729
```

Hide

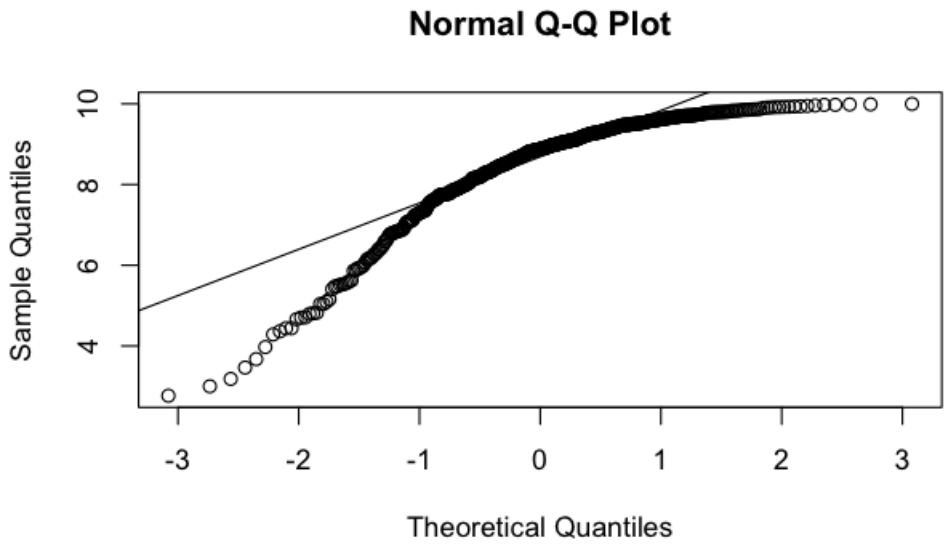
```
# Erstellen eines Beispiel-Datensatzes mit 100 Zufallszahlen
set.seed(123)
data01 <- rnorm(100, mean = Mittelwert_sqrt_Internet, sd = Standardabweichung_sqrt_Internet)
# Shapiro-Wilk-Test
shapiro.test(data01)
```

```
    Shapiro-Wilk normality test

data:  data01
W = 0.99388, p-value = 0.9349
```

Hide

```
# Q-Q Plot der transformierten Variablen
qqnorm(Arbeitstabelle02$sqrt_InternetAnteil)
qqline(Arbeitstabelle02$sqrt_InternetAnteil)
```



Kontrolle der neu erstellten Variablen

Hide

```
class(Arbeitstabelle02$sqrt_InternetAnteil)
```

```
[1] "numeric"
```

Problematik: Variable “SpendingProPerson” wurde als normalverteilt bewertet, obwohl die Standardabweichung höher als Mittelwert ist, beim Q-Q Plot ist keine einheitliche Korrelation der Quantile zu beobachten, daher wird die Transformation überprüft.

Hide

```
# install.packages("e1071")
library(e1071)
Schiefe_Spending <- skewness(Arbeitstabelle02$SpendingProPerson)
Schiefe_Spending
```

```
[1] 6.351827
```

Hide

```
Kurtosis_Spending <- kurtosis(Arbeitstabelle02$SpendingProPerson)
Kurtosis_Spending
```

```
[1] 42.65986
```

Log-Transformation der Variable “SpendingProPerson” möglich

Hide

```
Arbeitstabelle02$log_SpendingProPerson <- log(Arbeitstabelle02$SpendingProPerson)
Arbeitstabelle02
```

1
2
3
4
6
8
10
12
18
19

1-10 of 482 rows | 1-1 of 10 columns Previous **1** 2 Next

Test auf Normalverteilung

Hide

```
#Mittelwerte und Standardabweichungen berechnen
Mittelwert_log_Spending<-mean(Arbeitstabelle02$log_SpendingProPerson)
Standardabweichung_log_Spending<-sd(Arbeitstabelle02$log_SpendingProPerson)
Mittelwert_log_Spending
```

```
[1] 7.577889
```

Hide

```
Standardabweichung_log_Spending
```

```
[1] 1.173994
```

Hide

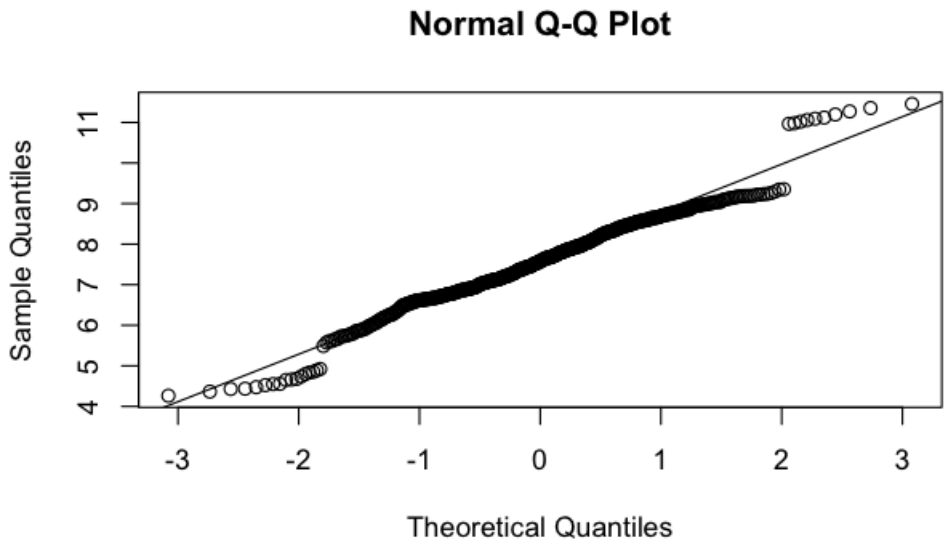
```
# Erstellen eines Beispiel-Datensatzes mit 100 Zufallszahlen
set.seed(123)
data02 <- rnorm(100, mean = Mittelwert_log_Spending, sd = Standardabweichung_log_Spending)
# Shapiro-Wilk Test auf Normalverteilung
shapiro.test(data02)
```

```
	Shapiro-Wilk normality test

data:  data02
W = 0.99388, p-value = 0.9349
```

Hide

```
#Q-Q Plot erstellen
qqnorm(Arbeitstabelle02$log_SpendingProPerson)
qqline(Arbeitstabelle02$log_SpendingProPerson)
```



Korrelationskoeffizient nach Pearson für normalverteile Variablen berechnen

Hide

```
# Korrelationskoeffizienten berechnen
Korr_Spending_Internet_Pearson<- cor(Arbeitstabelle02$log_SpendingProPerson, Arbeitstabelle02$sqrt_InternetAnteil) # nach Spearman (für nor
malverteilte Daten, monoton, N>10)
Korr_Spending_Internet
```

```
[1] 0.07818057
```

Inferenzstatistik Zusammenhang überprüfen zwischen Spending pro Einwohner und Internet-Anteil (ist keine Aufgabestellung)

Hide

```
cor.test(Arbeitstabelle02$log_SpendingProPerson, Arbeitstabelle02$sqrt_InternetAnteil)
```

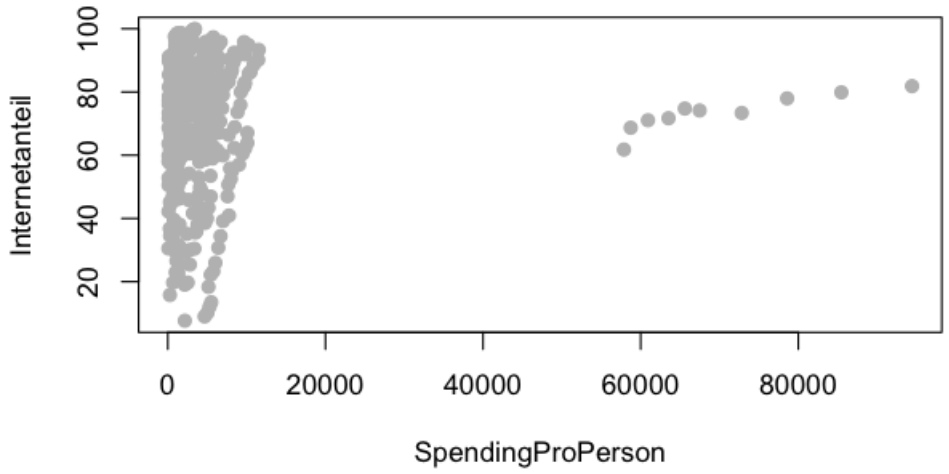
```
	Pearson's product-moment correlation

data:  Arbeitstabelle02$log_SpendingProPerson and Arbeitstabelle02$sqrt_InternetAnteil
t = 1.257, df = 480, p-value = 0.2094
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03219768  0.14584969
sample estimates:
       cor
0.05728146
```

Inferenzstatistik: Zweifaktorielle Varianzanalyse Fragestellung: Korrelieren die Anteile der Breitbandzugänge mit Haushaltsausgaben (Variable “InternetAnteil” und “SpendingProPerson”)?

Hide

```
x <- Arbeitstabelle02$InternetAnteil
y <- Arbeitstabelle02$SpendingProPerson
plot(y,x, pch = 19, xlab = "SpendingProPerson", ylab = "Internetanteil", col = "grey")
```



Hide

```
# Korrelationskoeffizienten berechnen
Korr_Spending_Internet<- cor(Arbeitstabelle02$SpendingProPerson, Arbeitstabelle02$InternetAnteil, method = "spearman") # nach Spearman (für nicht normalverteilte Daten, monoton, N>10)
Korr_Spending_Internet
```

```
[1] 0.07818057
```

Frage: Korrelieren die Anteile der Breitbandzugänge mit Haushaltsausgaben durch die Zeit? Partielle Korrelation berechnen

Hide

```
# Das Paket "ppcor" installieren
# install.packages(ppcor)
# Das Paket "ppcor" laden
library(ppcor)
# eine MAtrix aus nummerischen Variablen erstellen
PartielleKorrelation_Time <- as.matrix(Arbeitstabelle02[c("sqrt_InternetAnteil", "log_SpendingProPerson", "ZeitNum")])
# Partielle Korrelation berechnen
pcor(PartielleKorrelation_Time, method="pearson")
```

```
$estimate
                     sqrt_InternetAnteil
sqrt_InternetAnteil            1.00000000
log_SpendingProPerson         -0.02877769
ZeitNum                        0.63207746
                     log_SpendingProPerson   ZeitNum
sqrt_InternetAnteil             -0.02877769 0.6320775
log_SpendingProPerson            1.00000000 0.1152856
ZeitNum                          0.11528559 1.0000000

$p.value
                     sqrt_InternetAnteil
sqrt_InternetAnteil          0.000000e+00
log_SpendingProPerson        5.289353e-01
ZeitNum                      5.122694e-55
                     log_SpendingProPerson      ZeitNum
sqrt_InternetAnteil              0.52893526 5.122694e-55
log_SpendingProPerson            0.00000000 1.139736e-02
ZeitNum                          0.01139736 0.000000e+00

$statistic
                     sqrt_InternetAnteil
sqrt_InternetAnteil             0.0000000
log_SpendingProPerson          -0.6300914
ZeitNum                        17.8521070
                     log_SpendingProPerson   ZeitNum
sqrt_InternetAnteil             -0.6300914 17.852107
log_SpendingProPerson            0.0000000  2.540085
ZeitNum                          2.5400847  0.000000

$n
[1] 482

$gp
[1] 1

$method
[1] "pearson"
```

Fragestellung: Korrelieren die Anteile der Breitbandzugänge mit Haushaltsausgaben durch die Location?

Hide

```
# Das Paket "ppcor" laden
library(ppcor)
# eine Matrix aus nummerischen Variablen erstellen
PartielleKorrelation_Location <- as.matrix(Arbeitstabelle02[c("sqrt_InternetAnteil", "log_SpendingProPerson", "LocationNum")])
# Partielle Korrelation berechnen
pcor(PartielleKorrelation_Location, method="pearson")
```

```
$estimate
                     sqrt_InternetAnteil log_SpendingProPerson LocationNum
sqrt_InternetAnteil           1.00000000            0.06993367 -0.13640104
log_SpendingProPerson         0.06993367            1.00000000  0.09871434
LocationNum                  -0.13640104            0.09871434  1.00000000

$p.value
                     sqrt_InternetAnteil log_SpendingProPerson LocationNum
sqrt_InternetAnteil         0.000000000            0.12560879 0.002719833
log_SpendingProPerson       0.125608795            0.00000000 0.030415973
LocationNum                 0.002719833            0.03041597 0.000000000

$statistic
                     sqrt_InternetAnteil log_SpendingProPerson LocationNum
sqrt_InternetAnteil            0.000000            1.534330   -3.013447
log_SpendingProPerson          1.534330            0.000000    2.171073
LocationNum                   -3.013447            2.171073    0.000000

$n
[1] 482

$gp
[1] 1

$method
[1] "pearson"
```

Ergebnis: Die Variablen SpendingProPerson und InternetAnteil korrelieren nicht Die Variablen SpendingProPerson und InternetAnteil unter Berücksichtigung der Zeit korrelieren nicht Die Variablen SpendingProPerson und InternetAnteil unter Berücksichtigung der Location korrelieren nicht Die Variablen SpendingProPerson und InternetAnteil sind nicht zusammenhängend