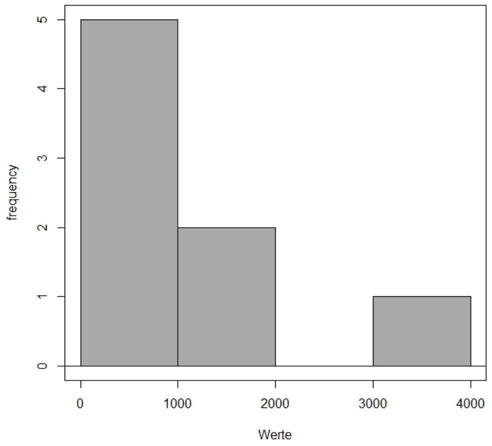
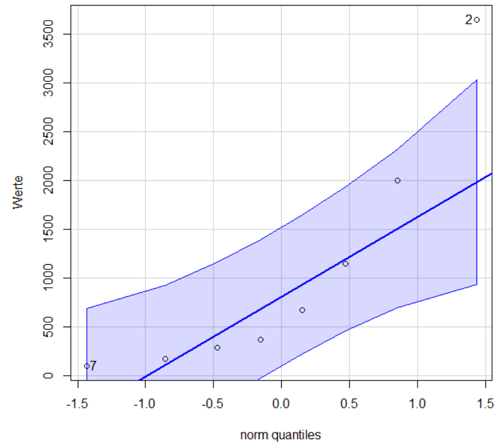
Statistik – Transformation von Daten

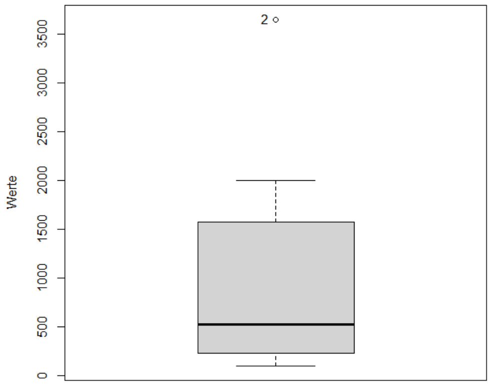
Gedankenexperiment

- Uns liegt ein Datensatz vor: 675 – 3650 – 175 – 1150 – 290 – 2000 – 100 – 375

- Histogramm, QQ-Diagramm und Box Plot deuten auf Daten hin, die nicht der Normalverteilung folgen







Gedankenexperiment

```
Anderson-Darling normality test

data:  Werte
A = 0.71578, p-value = 0.03661
```

- $p < \alpha = 0{,}05$
- Es gilt die Alternativhypothese: Die Daten sind nicht normalverteilt
- Der anschließende Test auf Normalverteilung zeigt: es handelt sich um nicht-normalverteilte Daten

Gedankenexperiment

- Diese Ergebnisse, gepaart mit einer geringen Stichprobengröße, führen dazu, dass verschiedene Tests (z.B. t-Tests, Bartlett, ANOVA, u.a.) für weitere Untersuchungen nicht herangezogen werden dürfen, da sie auf Normalverteilungen aufgebaut sind

- In Konsequenz muss man in nicht-parametrische Testverfahren wechseln

Gedankenexperiment

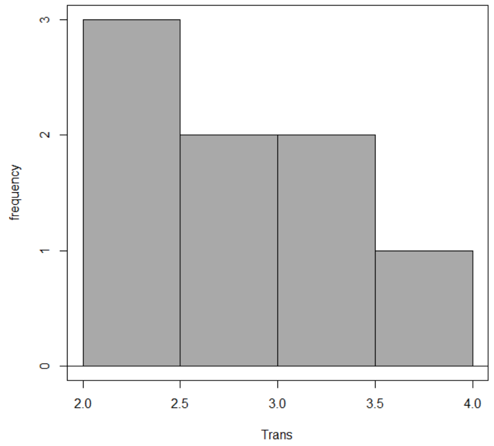
Was passiert, wenn man alle Datenpunkte logarithmiert?

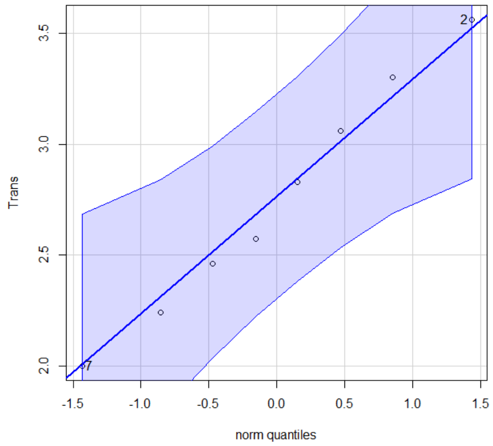
- Ursprüngliche Daten: 675 – 3650 – 175 – 1150 – 290 – 2000 – 100 – 375
- Logarithmierte Daten: 2,83 – 3,56 – 2,24- 3,06 – 2,46 – 3,30 – 2,00 – 2,57

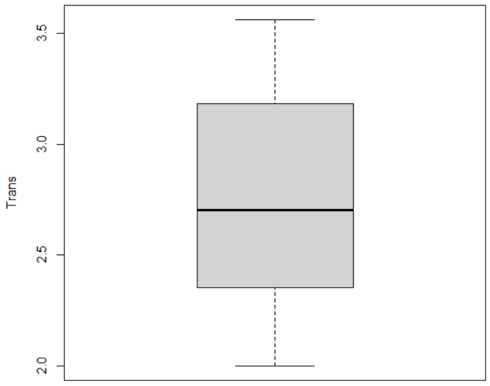
Gedankenexperiment

- Das Histogramm sieht nicht wirklich viel besser aus (nur wenige Daten)

- QQ-Diagramm und Box Plot lassen durchaus den Schluss auf Normalverteilung zu







Gedankenexperiment

```
Anderson-Darling normality test

data:  Trans
A = 0.13266, p-value = 0.9634
```

- $p > \alpha = 0{,}05$
- Wir bleiben bei der Nullhypothese: Die Daten sind normalverteilt
- Der anschließende Test auf Normalverteilung zeigt: es handelt sich um normalverteilte Daten

Gedankenexperiment

- Für die logarithmierten Daten sind nun alle Testverfahren nutzbar, die die Normalverteilung voraussetzen

- Weitere Gründe für Transformationen
  - Stabilisierung der Varianz von Datenreihen
  - Linearisierung

Zulässigkeit

- Bei der Transformation von Daten handelt es sich nicht um eine Manipulation im negativen Sinne, sondern um ein in der Mathematik und Statistik übliches Verfahren

- Solange alle Daten eines Datensatzes in gleicher Weise transformiert werden, alle weiteren Daten, die für Vergleiche u.a. herangezogen werden, ebenfalls transformiert werden und die Reihenfolge der Daten im Datensatz nicht verändert wird, ist eine Transformation zulässig

Typische Transformationen

- **Reziproke Transformation**

$g(x) = x^{-1}$ Anpassung an die Normalverteilung bei stark schiefen Verteilungen

- **Wurzeltransformation**

$g(x) = \sqrt{x + konst}$ für Poisson-Verteilungen

- **Logarithmische Transformation**

$g(x) = \ln(x + konst)$ für Lognormal-Verteilungen

Typische Transformationen

- **Arcus-Sinus-Transformation**

$$g(x) = \sqrt{n + konst_1} * \arcsin \sqrt{\frac{x+konst_2}{n+konst_3}}$$

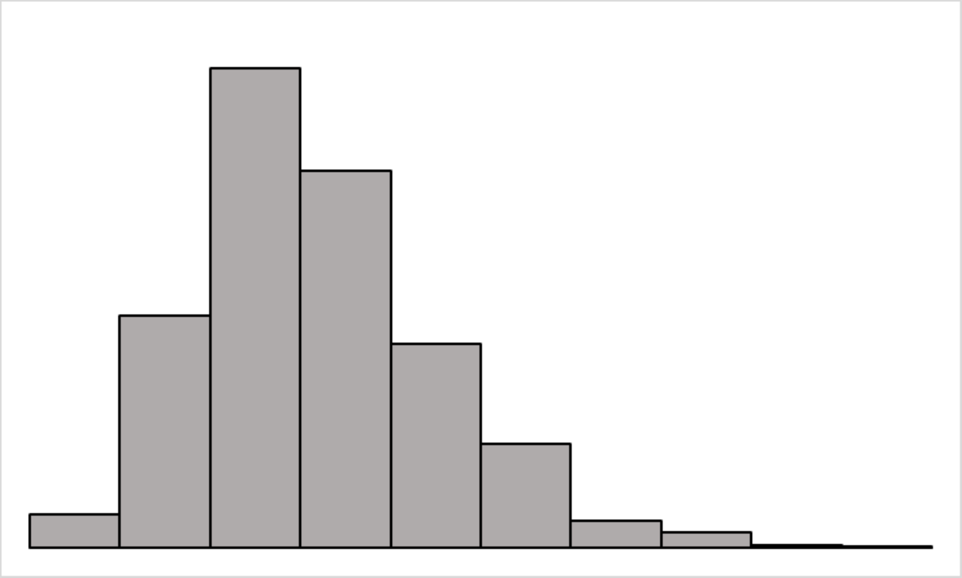
für binomiale Verteilungen

- **Fishersche Z-Transformation**

$g(x) = \operatorname{arctanh}(x)$ für Korrelationsrechnungen

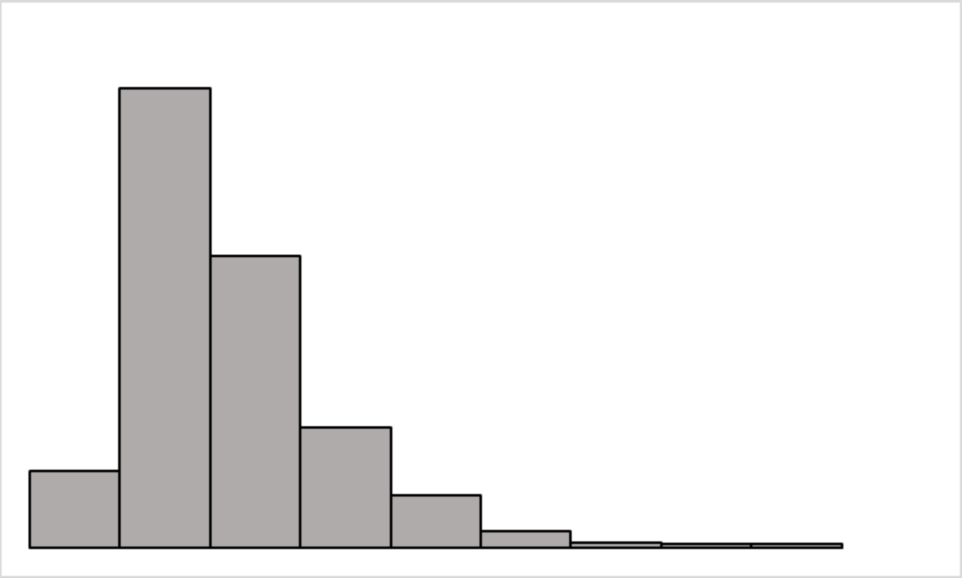
Transformationen für schiefe Verteilungen

- Schwach rechtsschiefe Verteilungen
  - Wurzeltransformation
  - $g(x) = \sqrt{x + konst}$
  - Sollten negative Werte vorhanden sein, können diese über die Konstante positiv gestellt werden



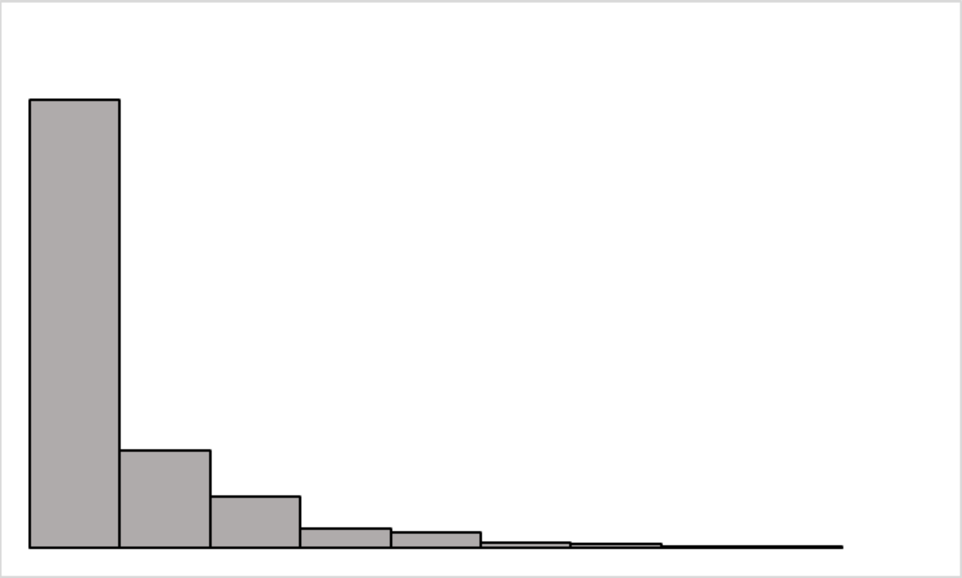
Transformationen für schiefe Verteilungen

- Rechtsschiefe Verteilungen
  - Logarithmische Transformation
  - $g(x) = \ln(x + konst)$
  - Sollten negative Werte vorhanden sein, können diese über die Konstante positiv gestellt werden



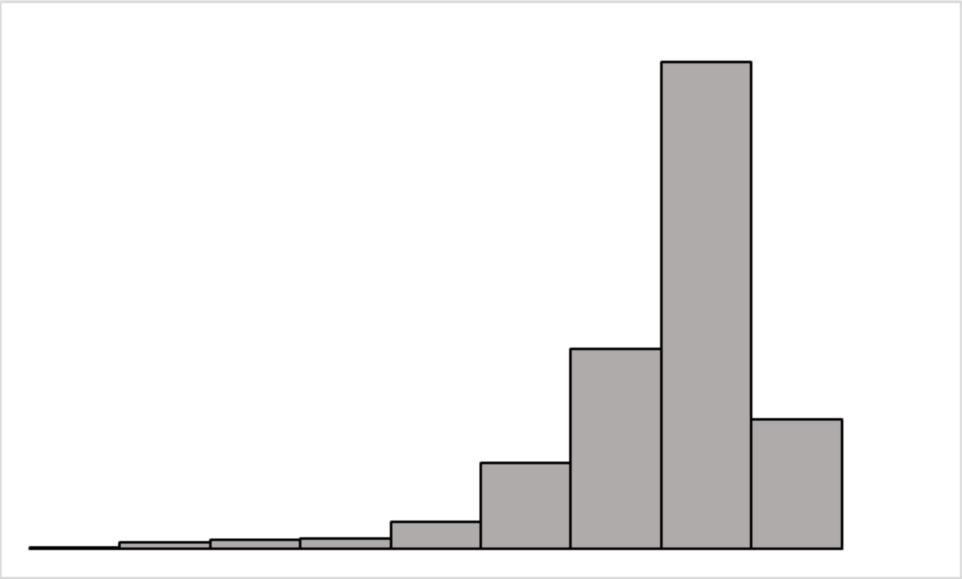
Transformationen für schiefe Verteilungen

- Stark rechtsschiefe Verteilungen
  - Reziproke Transformation
  - $g(x) = x^{-1}$
  - Sollten die Null im Datensatz vorkommen, muss vorher eine Verschiebung der Daten vorgenommen werden



Transformationen für schiefe Verteilungen

- Linksschiefe Verteilungen
  - Für linksschiefe Verteilungen werden dieselben Transformationen wie für Rechtsschiefe genutzt
  - Die Daten müssen vorher gespiegelt werden
  - Bsp.: $g(x) = \ln(Max + 1\ - x)$



Box-Cox-Transformation

- $g(x) = \begin{cases} \frac{(x+konst)^{\lambda}-1}{\lambda} & falls\ \lambda \neq 0 \\ \ln(x + konst) & falls\ \lambda = 0 \end{cases}$

- Je nach gewähltem λ ergeben sich einige der schon vorgestellten Transformationen

- Die Box Cox-Transformation dient in erster Linie der Varianzstabilisierung, wirkt dadurch aber auch normalisierend

- Der optimale λ-Wert kann in R bestimmt werden

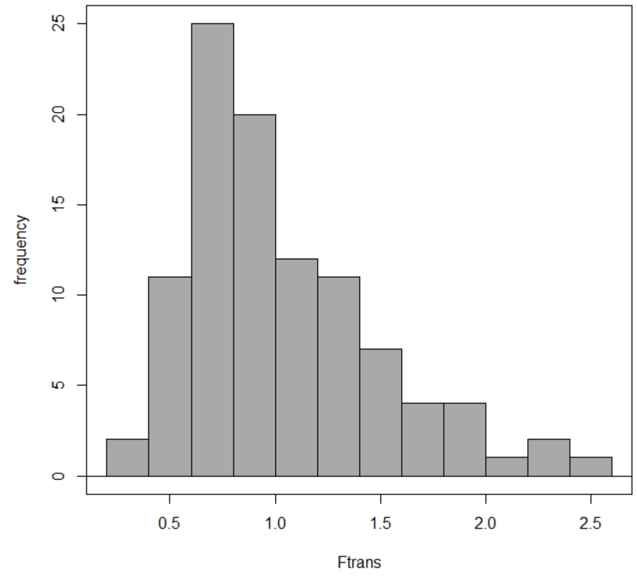
Box-Cox-Transformation

- Bei der Berechnung der Box Cox-Transformation schlägt der *RCommander* gerundete λ–Werte vor, die zu einfachen Transformationsformeln führen:
- $\lambda = 0{,}5 \qquad g(x) = 2 * (\sqrt{x + konst} - 1)$
- $\lambda = 0 \qquad g(x) = ln(x + konst) \qquad ln \triangleq log_e$
- $\lambda = -0{,}5 \qquad g(x) = \frac{1}{\sqrt{x+konst}}$
- $\lambda = -1 \qquad g(x) = \frac{1}{x+konst}$
- $konst$ ist so zu wählen, dass keine Wurzeln aus negativen Zahlen gezogen werden bzw. nicht durch 0 geteilt wird

Box-Cox-Transformation

- Beispiel F-Verteilung (df(Zähler)=30; df(Nenner)=25)

```
Anderson-Darling normality test

data:  Ftrans
A = 2.3292, p-value = 0.000006119
```



Box-Cox-Transformation

- Beispiel F-Verteilung (df(Zähler)=30; df(Nenner)=25)

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    -0.0366           0       -0.4356       0.3623

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                             LRT df    pval
LR test, lambda = (0) 0.03236961  1 0.85722

Likelihood ratio test that no transformation is needed
                           LRT df          pval
LR test, lambda = (1) 24.75681  1 0.00000065039
```

Box-Cox-Transformation

$\lambda = 0$ Transformation: $g(x) = \ln(x)$

- Nach der Transformation

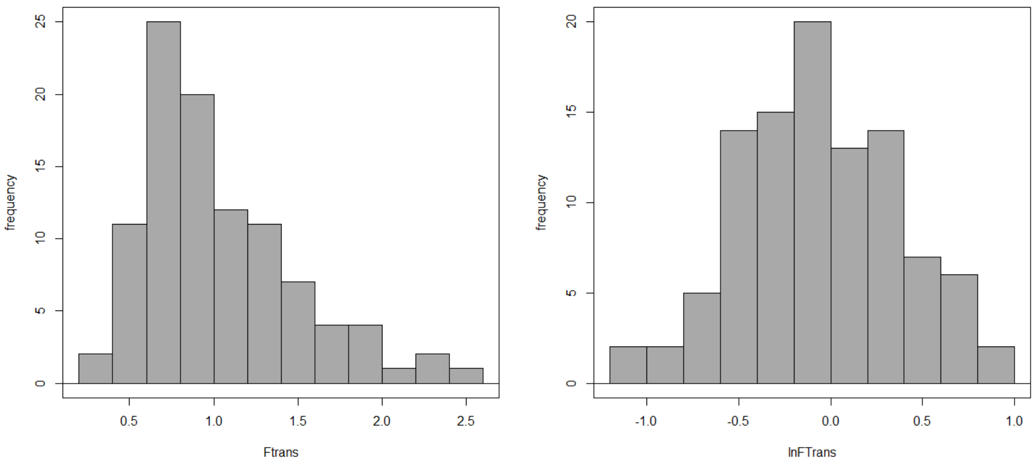
```
Anderson-Darling normality test

data:  lnFTrans
A = 0.31337, p-value = 0.5419

          mean        sd        skewness   kurtosis   n
Ftrans    1.03617808 0.4652247 1.0377115  0.7733478 100
lnFTrans -0.05842563 0.4358612 0.0413334 -0.3357070 100
```

Box-Cox-Transformation

- Vergleich Vorher – Nachher

	mean	sd	skewness	kurtosis	n
Ftrans	1.03617808	0.4652247	1.0377115	0.7733478	100
lnFTrans	-0.05842563	0.4358612	0.0413334	-0.3357070	100

Skewness:
Werte > 0: rechtsschief/linkssteil
Werte < 0: rechtssteil/linksschief



Yeo-Johnson-Transformation

- Im Gegensatz zu Box-Cox-Transformation zielt die Johnson-Transformation direkt auf die Normalisierung von Daten

- Ähnlich wie die Box-Cox-Transformation enthält die Johnson-Transformation eine Reihe von Funktionen und wird auch über einen λ-Wert angepasst

Yeo-Johnson-Transformation

$$g(x) = \begin{cases} ((x+1)^{\lambda} - 1)/\lambda & f\text{ür } \lambda \neq 0, x \geq 0 \\ \ln(x+1) & f\text{ür } \lambda = 0, x \geq 0 \\ -((-x+1)^{(2-\lambda)} - 1)/(2-\lambda) & f\text{ür } \lambda \neq 2, x < 0 \\ -\ln(-x+1) & f\text{ür } \lambda = 2, x < 0 \end{cases}$$

- Der optimale λ-Wert kann in R bestimmt werden

Yeo-Johnson-Transformation

- Beispiel F-Verteilung (df(Zähler)=30; df(Nenner)=25)

```
yjPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    -1.0744          -1     -1.9385      -0.2102

 Likelihood ratio test that transformation parameter is equal to 0
                          LRT df     pval
LR test, lambda = (0) 6.10714  1 0.013464
```

$$\lambda = -1$$

Transformation:

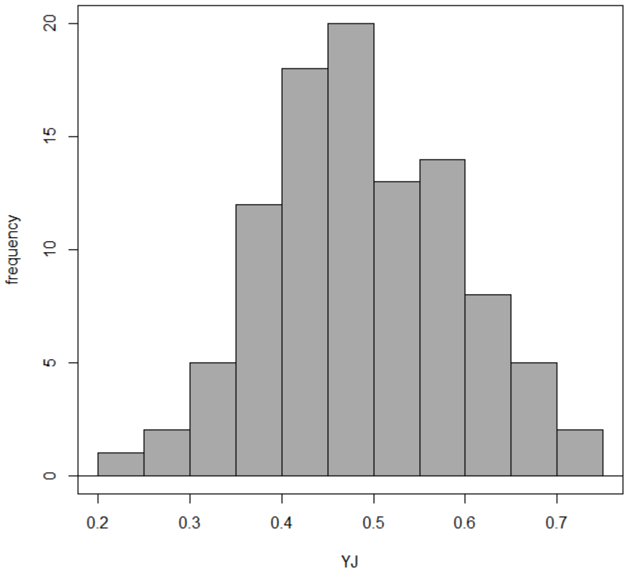
$\lambda \neq 0, x \geq 0$ $\qquad$ $g(x) = ((x+1)^{\lambda} - 1)/\lambda$

$$g(x) = -(x+1)^{-1} + 1$$

Yeo-Johnson-Transformation

- Beispiel F-Verteilung (df(Zähler)=30; df(Nenner)=25)

```
Anderson-Darling normality test

data:  YJ
A = 0.38381, p-value = 0.3892
```

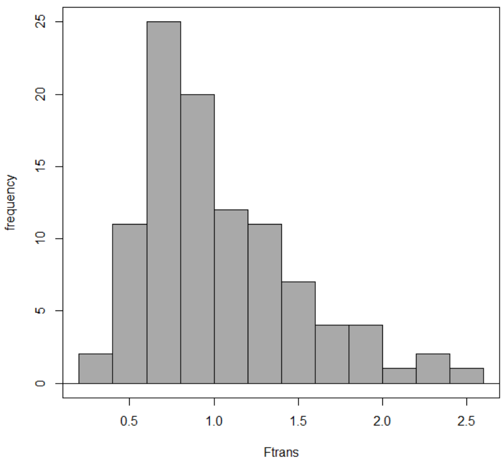


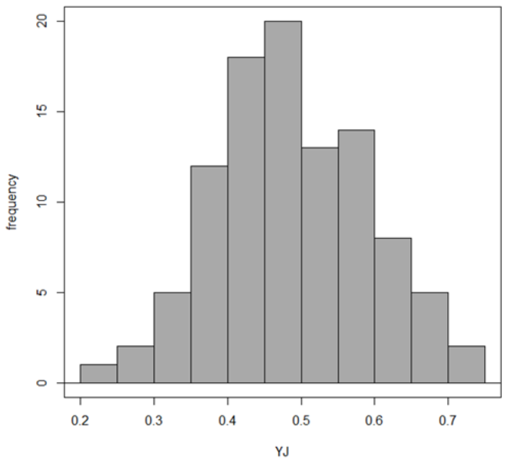
Yeo-Johnson-Transformation

- Vergleich Vorher – Nachher

	mean	sd	skewness	kurtosis	n
Ftrans	1.0361781	0.4652247	1.03771148	0.7733478	100
YJ	0.4859591	0.1047720	0.08128996	-0.5162111	100

Skewness:
Werte > 0: rechtsschief/linkssteil
Werte < 0: rechtssteil/linksschief





Fazit

- Transformationen sind ein zulässiges Mittel um Daten in eine erwünschte Form zu bringen
- Transformiert man auf Normalverteilung, ist im Anschluss die Normalverteilung zu prüfen
- Manche Datensätze lassen sich nicht in eine gewünschte Form transformieren, in diesem Fall muss man nicht-parametrische Tests nutzen
- Neben den eigentlichen Daten sind auch alle weiteren Informationen zu transformieren (Vergleichsdatensätze, Vorgabewerte, u.a.)

Fazit

- In machen Fällen ist die Interpretation der transformierten Daten schwierig

- Für diese Fälle wird i.a. empfohlen, auf eine Transformation zu verzichten, gegebenenfalls kann man dann eher geringe Abweichungen z.B. von der Normalverteilung tolerieren oder direkt auf nicht-parametrische Verfahren zugreifen