HTW Chur

Hochschule für Technik und Wirtschaft
University of Applied Sciences

Normalverteilung

Quantitative Forschungsmethoden
Prof. Dr. Franz Kronthaler

Lernziele

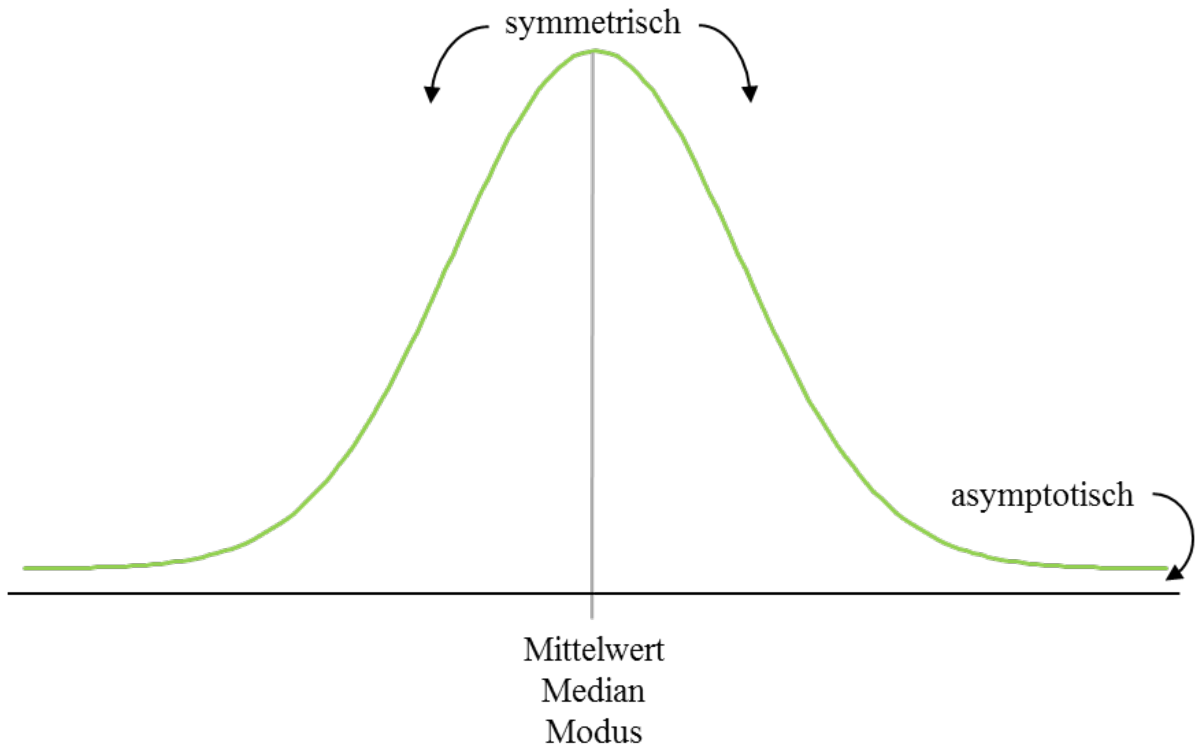
- verstehen wie eine normalverteilte Variable geformt ist
- lernen wie grafisch auf Normalverteilung geprüft werden kann
- lernen wie numerisch auf Normalverteilung getestet werden kann

Einleitung

- eine Beurteilung, ob die Daten normalverteilt sind, ist eine Voraussetzung für viele statistische Testverfahren
- normalverteilte Daten sind eine Voraussetzung für parametrische Testverfahren
- es gibt zwei Methoden der Beurteilung, ob eine Variable normalverteilt ist
  - grafische Verfahren
  - statistische Testverfahren
- statistische Testverfahren haben den Vorteil einer objektiven Beurteilung, aber
  - sie sind nicht sensibel genug wenn die Stichprobe klein ist
  - sie reagieren übersensibel wenn die Stichprobe gross ist
- grafische Verfahren haben einen Vorteil wenn statistische Testverfahren übersensibel oder zu wenig sensibel reagieren, aber
  - grafische Verfahren fehlt es etwas an Objektivität
  - grafische Verfahren erfordern mehr Erfahrung bei der Interpretation

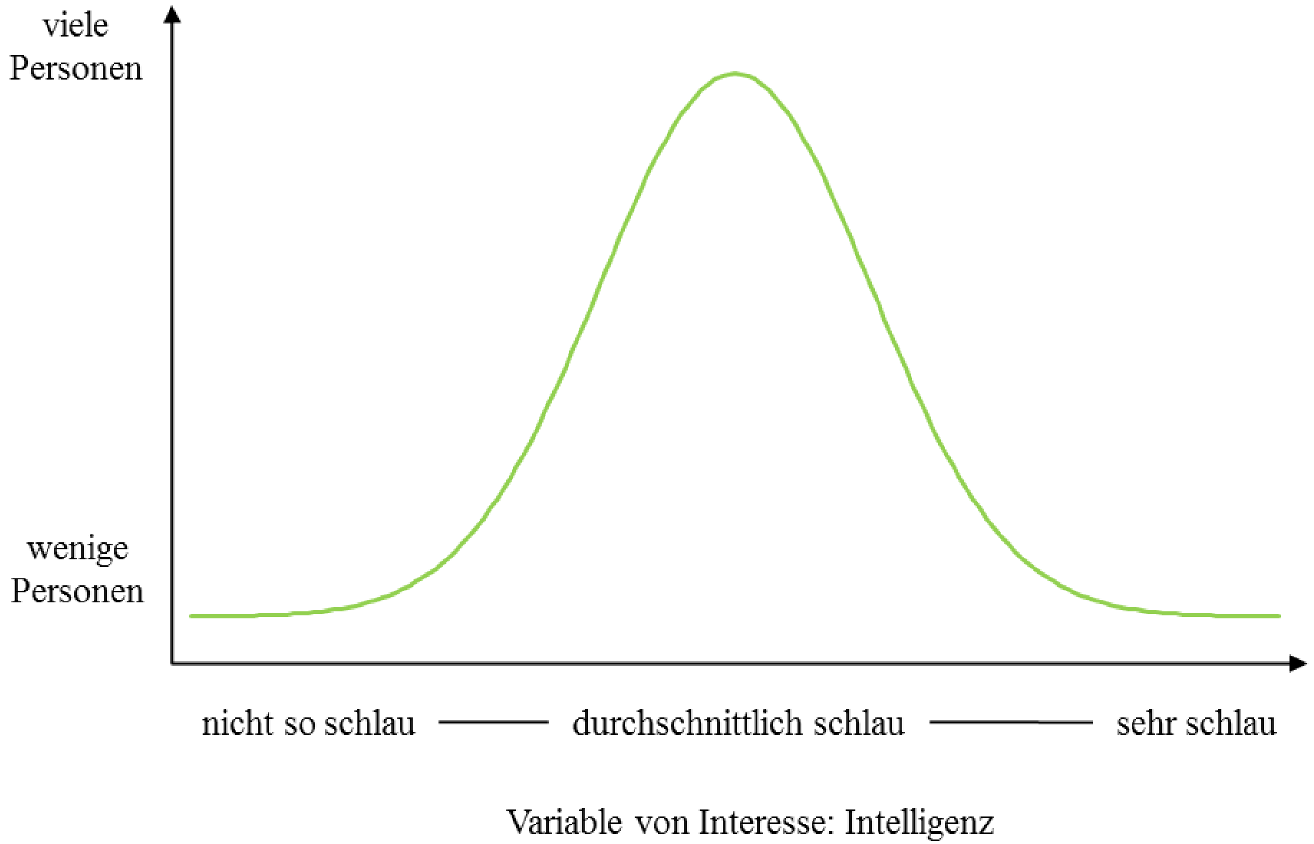
Normalverteilungskurve

- die Normalverteilungskurve (Gaußsche Glockenkurve) ist eine visuelle Darstellung von Werten mit drei Eigenschaften



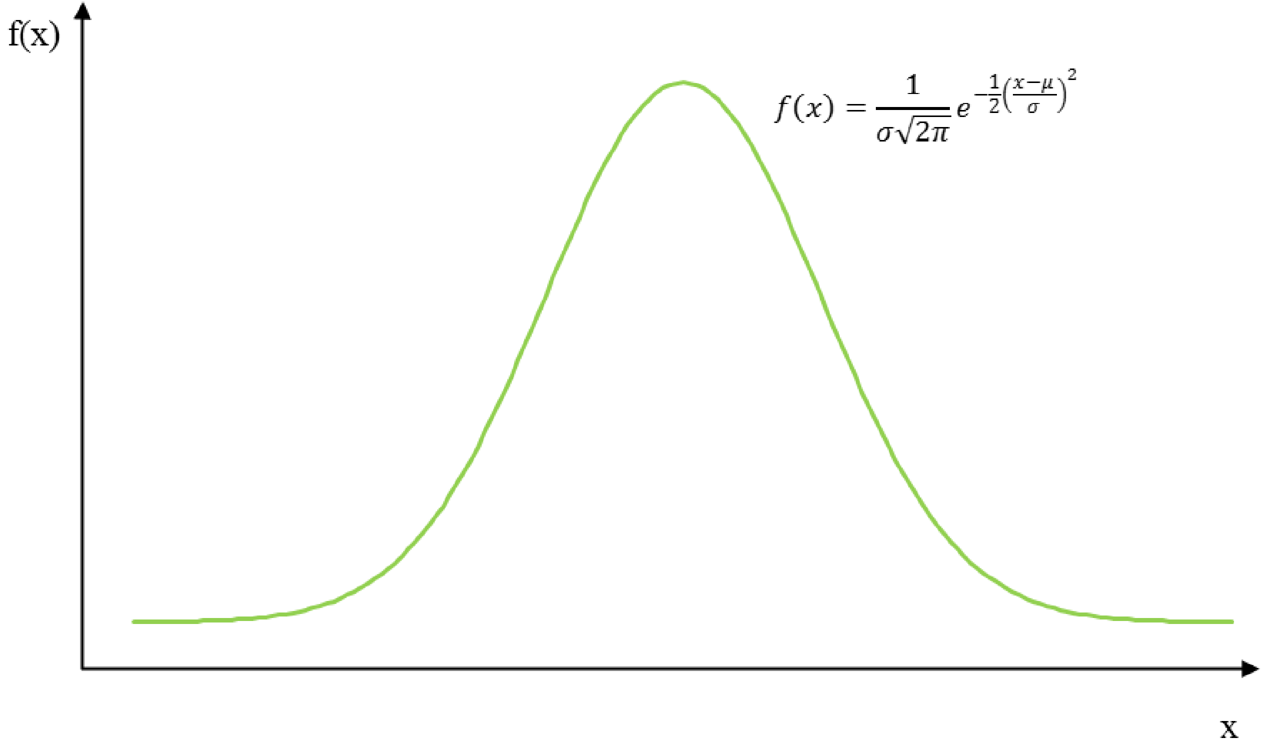
Normalverteilungskurve

- ein Beispiel für eine typischerweise normalverteilte Variable



Normalverteilungskurve

- die Form einer normalverteilten Variable ist bestimmt durch
  - den Mittelwert
  - die Standardabweichung



Normalverteilungskurve

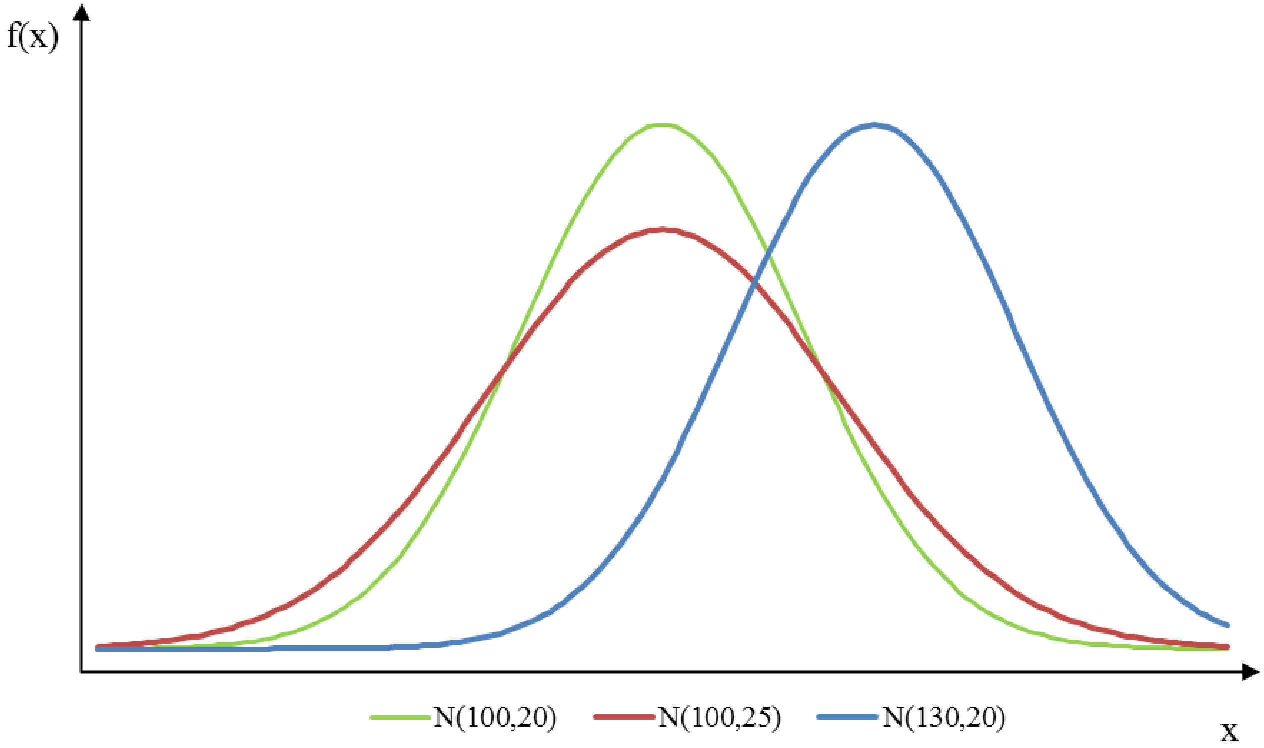
- die Funktion der Normalverteilungskurve

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  - $f(x)$ ist der Funktionswert
  - μ ist der Mittelwert
  - σ ist die Standardabweichung
  - π ist die Zahl pi mit 3.141…
  - e ist Eulersche Zahl mit 2.718…

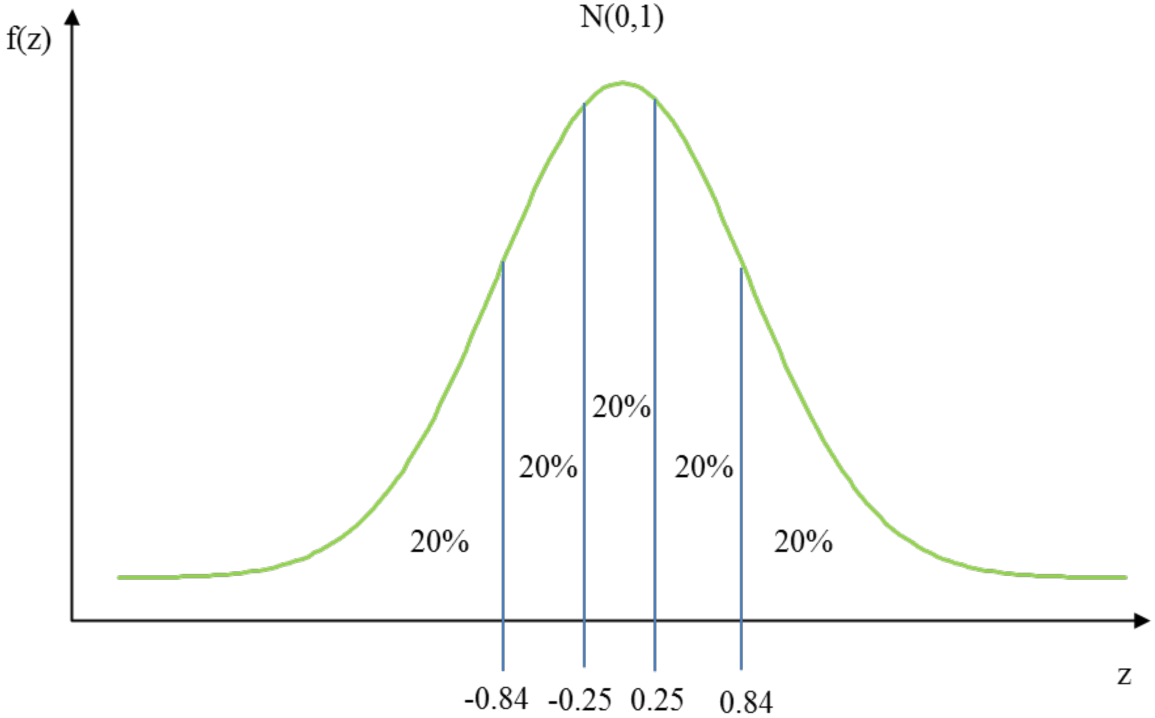
Normalverteilungskurve

- eine Vergrösserung des Mittelwertes führt zu einer Verschiebung nach rechts und umgekehrt
- eine Vergrösserung der Standardabweichung führt zu einer Verbreiterung und umgekehrt



Normalverteilungskurve

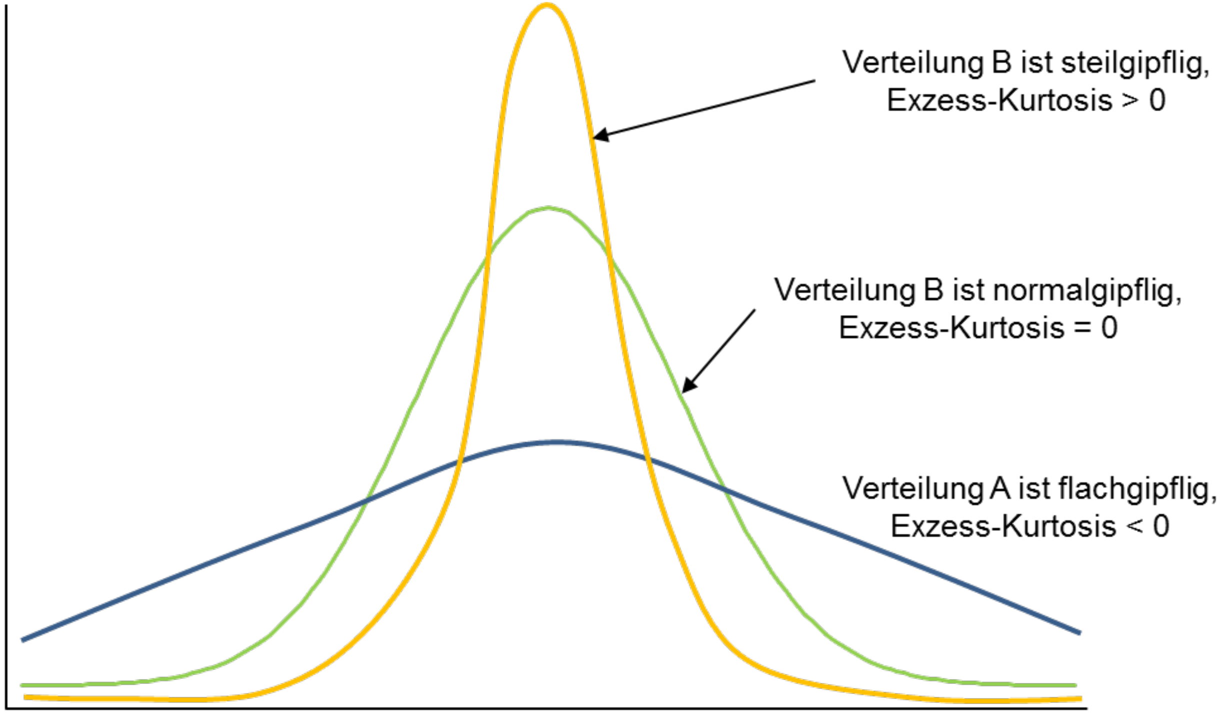
- die Normalverteilungskurve lässt sich in Quantile einteilen
- die folgende Abbildung zeigt die 20%-Quantile (Quintile) für die Standardnormal-verteilungskurve



- d. h., für jede normalverteilte Variable lassen sich die theoretischen Quantilswerte berechnen

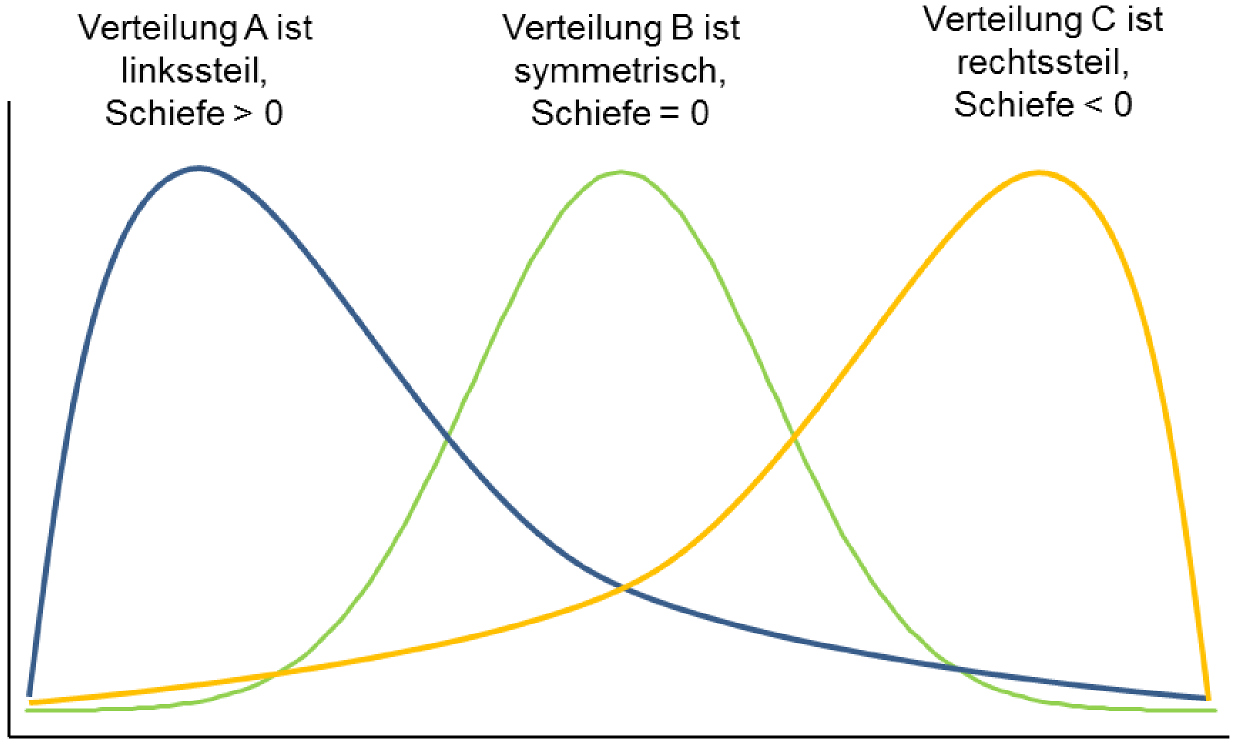
Normalverteilungskurve

- per Definition hat die Normalverteilung
  - eine Exzess-Kurtosis von 0 beziehungsweise eine Kurtosis von 3



Normalverteilungskurve

– eine Schiefe von 0



Testen auf Normalverteilung – Beispiel “marathon_1000.Rdata”

- um zu prüfen, ob eine Variable normalverteilt ist, wird der folgende Datensatz genutzt
  - der Datensatz ist benannt mit marathon_1000.Rdata
  - der Datensatz stammt aus dem Lehrbuch von Norusis (2008), ursprünglich umfasst der Datensatz 28’764 Personen, welche den Marathon von Chicago von 2001 beendet haben
  - der Datensatz enthält die folgenden Variablen (siehe marathon_1000.xlsx)
    - person: Person
    - age: Alter
    - sex: Geschlecht
    - hours: Stunden
    - name: Name

lade den Datensatz “marathon_1000.Rdata”

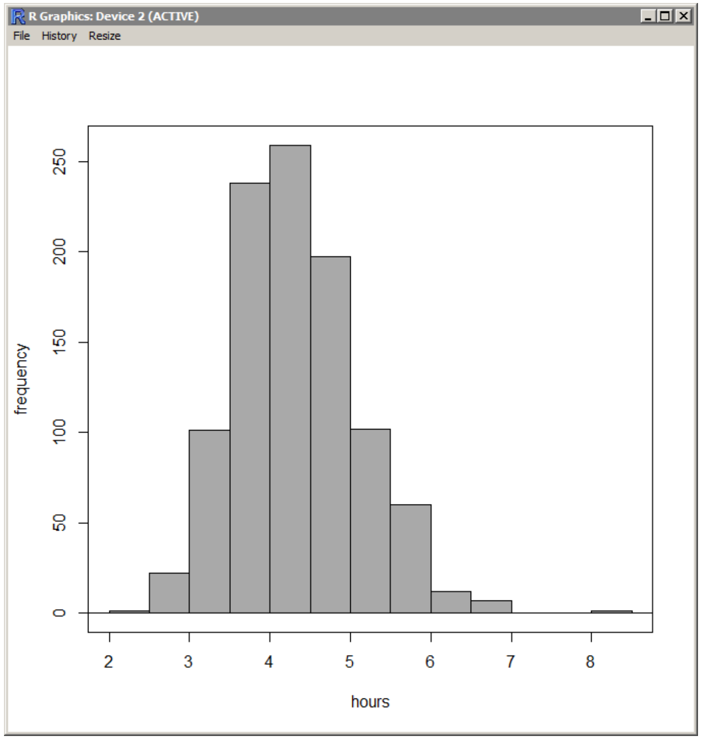
Testen auf Normalverteilung – grafisch

- das Prüfen der Variable mit Hilfe des Histogramms
  - zeichne das Histogramm
  - vergleiche das Histogramm mit der Normalverteilungskurve

```
# nutze die Benutzeroberflache des R Commanders oder die folgenden Befehle
attach(marathon_1000)
hist(hours, scale="frequency", breaks="Sturges", col="darkgray")
detach(marathon_1000)
```

Testen auf Normalverteilung – grafisch

- Abbildung: Histogramm der Variable hours (gebrauchte Zeit für den Marathon)



Testen auf Normalverteilung – grafisch

- das Prüfen der Variable mit der Hilfe des Quantil-Quantil-Plots (Q-Q-Plot)
  - der Q-Q-Plot vergleicht die theoretischen Quantilswerte einer normalverteilten Variable mit den tatsächlich beobachteten Werten
  - wenn die Punkte auf der Linie liegen ist die Variable normalverteilt
  - wenn die Punkte nicht auf der Linie liegen ist die Variablen nicht normalverteilt
  - Gründe, warum das Muster der Daten von der Linie abweicht

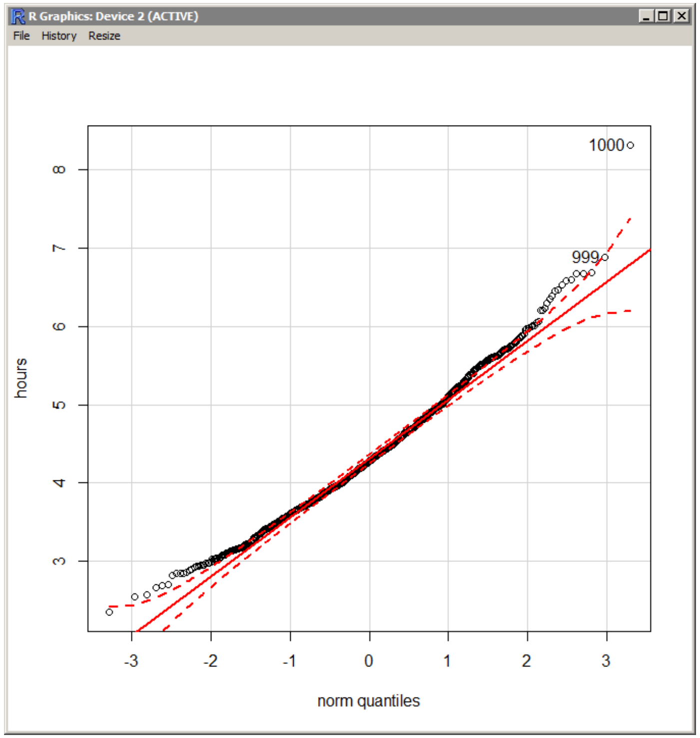
Beschreibung des Musters	mögliche Interpretation
alle bis auf wenige Punkte sind auf der Linie	es gibt Ausreisser in den Daten
das linke Ende ist unterhalb der Linie; das rechte Ende ist oberhalb der Linie	die Verteilung ist steilgipflig
das linke Ende ist oberhalb der Linie; das rechte Ende ist unterhalb der Linie	die Verteilung ist flachgipflig
kurviges Muster in der die Steigung von links nach rechts ansteigt	die Verteilung ist linkssteil
kurviges Muster in der die Steigung von links nach rechts abfällt	die Verteilung ist rechtssteil

Testen auf Normalverteilung – grafisch

```
# nutze die Benutzeroberflache des R Commanders oder die folgenden Befehle
attach(marathon_1000)
qqPlot(hours, dist="norm", id.method="y", id.n=2, labels=rownames(marathon_1000))
detach(marathon_1000)
```

Testen auf Normalverteilung – grafisch

- Abbildung: Q-Q-Plot für die Variable hours (gebrauchte Zeit für den Marathon)



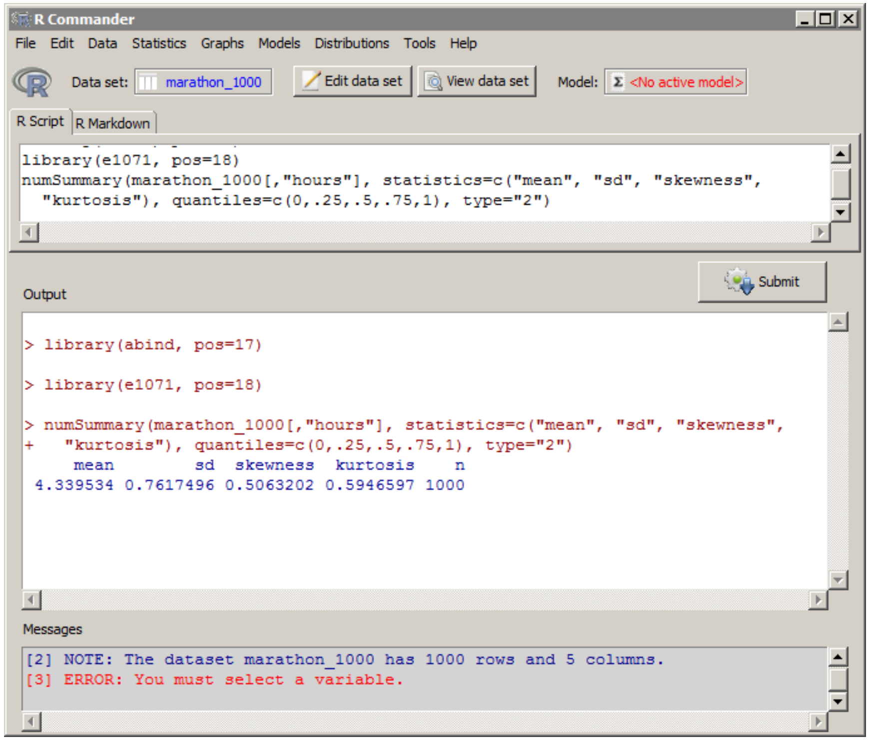
Testen auf Normalverteilung – Schiefe und Kurtosis

- das Prüfen einer Variable mit Hilfe der Schiefe und der Exzess-Kurtosis
  - berechne die Exzess-Kurtosis und die Schiefe und prüfe ob die Variable
    - linkssteil, Schiefe > 0
    - rechtssteil, Schiefe < 0
    - steilgipflig, Exzess-Kurtosis > 0
    - flachgipflig, Excess-Kurtosis < 0
  - in der Empirie ist keine Variable perfekt normal verteilt, d. h. wir finden immer Abweichungen von 0

  # am Besten nutze die Benutzeroberflache des R Commanders mit dem Kommando numerical statistics

Testen auf Normalverteilung – Schiefe und Kurtosis

- Abbildung: Schiefe und Exzess-Kurtosis für die Variable hours (gebrauchte Zeit für den Marathon)



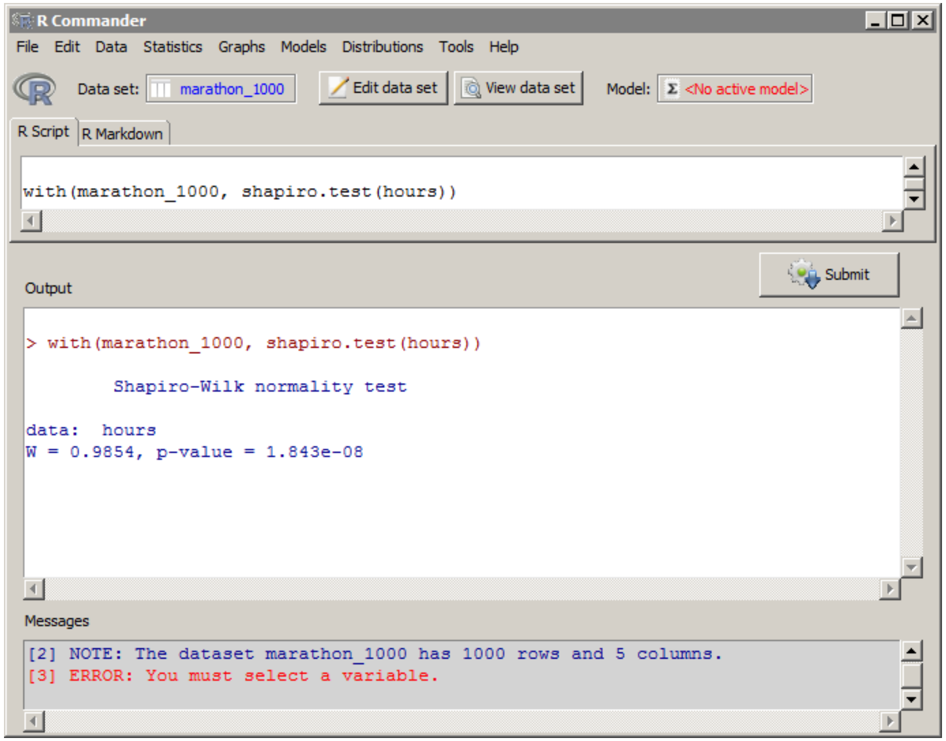
Testen auf Normalverteilung – Shapiro-Wilk-Test

- das Testen einer Variable mit der Hilfe des Shapiro-Wilk-Tests
  - der Shapiro-Wilk-Test nutzt die Nullhypothese, die Stichprobe stammt von einer normalverteilten Variable
  - einfach:
    - H_0: Die Variable ist normalverteilt.
    - H_A: Die Variable ist nicht normalverteilt.
    - das typischerweise verwendete Signifikanzniveau ist 5% (α=0.05)
      - ❖ wenn p kleiner 0.05 wird H_0 abgelehnt (wir sind zu 95% sicher, dass die Daten in der Grundgesamtheit nicht normalverteilt sind)
      - ❖ wenn p grösser 0.05 wird H_0 nicht abgelehnt (wir sind zu 95% sicher, dass die Variable in der Grundgesamtheit normalverteilt ist)

```
# nutze die Benutzeroberflache des R Commanders oder die folgenden Befehle
shapiro.test(marathon_1000&hours)
```

Testen auf Normalverteilung – Shapiro-Wilk-Test

- Abbildung: Shapiro-Wilk-Test für die Variable hours (gebrauchte Zeit für den Marathon)



Testen auf Normalverteilung

- für die Entscheidung sollten die Ergebnisse zusammengefasst werden, für unser Beispiel könnte die Interpretation folgendermassen aussehen
  - das Histogramm folgt weitgehend einer Normalverteilung, ist aber etwas linkssteil
  - die meisten Punkte des Q-Q-Plots sind auf der Linie beziehungsweise innerhalb des Konfidenzintervalls, mit einigen Ausnahmen
  - die Schiefe ist leicht grösser als 0, etwas linkssteil
  - die Exzess-Kurtosis ist leicht grösser als 0, etwas steilgipflig
  - der Shapiro-Wilk-Test verwirft H_0, d. h. die Variable ist nicht normalverteilt

→ zusammenfassend können wir schliessen, dass die Variable nicht normalverteilt ist

- aber erinnern Sie sich, im Fall einer grossen Stichprobe reagiert der Shapiro-Wilk-Test übersensibel

Anwendungen

- Eine zufällige Stichprobe von 100 Beobachtungen wurde aus dem Datensatz marathon_1000.Rdata entnommen, benannt marathon_100.Rdata. Nutze den neuen Datensatz.
  - Teste für die Variable hours erneut, ob die Variable normalverteilt ist.
  - Teste für die Variable age, ob die Variable normalverteilt ist.