Statistische Zeitreihenanalyse: Brownsche Bewegung, Aktienkurse und Temperaturdaten

Daniel C. Wagner, Dr. Rudi Schäfer

März 2015

1 Einleitung

In diesem Versuch sollen grundlegende Methoden der statistischen Analyse und Beschreibung von Zeitreihen vermittelt und angewandt werden. Wir werden dabei drei sehr unterschiedliche Systeme betrachten: die Bewegung von Pollen auf einer Wasseroberfläche, die zeitliche Entwicklung von Aktienkursen, sowie Temperaturzeitreihen.

Abbildung 1 zeigt exemplarisch Zeitreihen der drei genannten Systeme. Auf den ersten Blick erscheinen alle drei Zeitreihen als ziemlich erratisch oder zufällig. Diese qualitative Beobachtung wollen wir in diesem Versuch präzisieren.

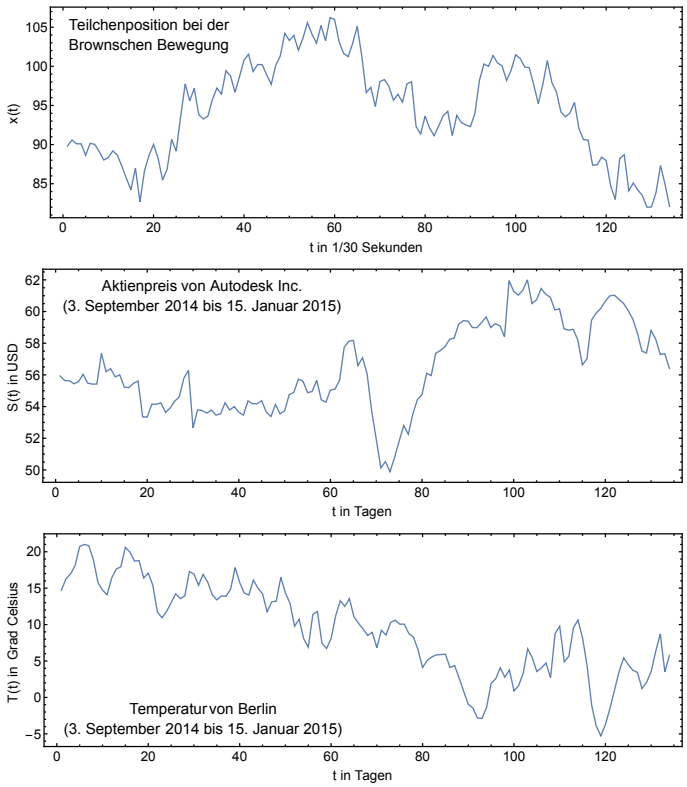

Abbildung 1: Beipiele von Zeitreihen für die Brownsche Bewegung (oben), einen Aktienkurs (Mitte), sowie einen Temperaturverlauf (unten).

1.1 Brownsche Bewegung

Der schottische Botaniker Robert Brown beobachtete im Jahre 1827 unter dem Mikroskop, wie Pollen auf einem Wassertropfen unregelmäßig zuckende Bewegungen machten. Diese Wärmebewegung von Teilchen in Flüssigkeiten und Gasen ist heutzutage nach ihm benannt. Allerdings wurde sie erstmals 1785 von Jan Ingenhousz, einem niederländischen Arzt und Botaniker, beschrieben. Dieser untersuchte die Bewegung von Holzkohlestaub auf Alkohol. Die Zeitreihen, die man aus solchen Experimenten erhält, sind die in regelmäßigen Abständen aufgezeichneten Positionen der Kohle- bzw. Pollenteilchen (Kolloide).

Obwohl es sich bei einem solchen System um ein rein deterministisches handelt, erscheint die Bewegung völlig erratisch. Sie läßt sich durch einen Zufallsprozess beschreiben [Wiener, Einstein]. Warum ist dem so? Es gibt pro Sekunde eine große Zahl von Stößen zwischen dem Kolloid und den Flüssigkeitsmolekülen (Größenordnung 10^{21}). Bei jedem dieser Stöße erfährt das Kolloid eine kleine Impulsänderung. Dies führt effektiv zu einer zufälligen Bewegung des Kolloids. Für die Statistik der Ortsänderungen ergibt sich nach dem zentralen Grenzwertsatz eine Gaußverteilung. Wir wollen in diesem Versuchsteil ergründen, wie gut empirische Zeitreihen der Kolloidbewegung durch einen Zufallsprozess mit unabhängigen, gaußverteilten Inkrementen beschrieben werden.

1.2 Aktienpreise

Wie in Abbildung 1 zu sehen, sieht auch die zeitliche Entwicklung von Aktienkursen ähnlich erratisch aus wie die Brownsche Bewegung von Kolloiden auf einer Flüssigkeit. Und tatsächlich hat die Modellierung von Aktienkursen als Zufallsprozess eine lange Tradition, die auf die Doktorarbeit des französischen Mathematikers Louis Bachelier aus dem Jahre 1900 zurückgeht. Weshalb aber macht eine solche stochastische Beschreibung für Aktienkurse Sinn, wo doch auf lange Sicht der Aktienkurs das wirtschaftliche Wachstum eines Unternehmens widerspiegeln sollte und die Aktienhändler, die den Preis letztlich bestimmen, auch keineswegs zufällig handeln? Ersteres führt sicher zu einem deterministischen Anteil, einer sogenannten Drift. Die Argumentation für den stochastischen Anteil verläuft analog zur Brownschen Bewegung: Obwohl die einzelnen Aktienhändler klare Absichten und Strategien verfolgen, macht die große Vielzahl der Handelsinteraktionen (Größenordnung 10^4 pro Tag) eine stochastische Modellierung sinnvoll.

Bachelier nahm für die Änderungen der Aktienpreise in einem festen Zeitintervall eine Gaußverteilung an. Dies kann im Modell jedoch einerseits schnell zu negativen Preisen führen, und spiegelt andererseits nicht den auf längeren Zeithorizonten zu beobachtenden exponentiellen Verlauf der Aktienkurse wieder. Um diesen Aspekten gerecht zu werden, nimmt man eine Gaußverteilung für die relativen Preisänderungen (Renditen, englisch: *returns*) an. Empirisch ist jedoch seit den 1920er Jahren bekannt, dass Renditeverteilungen in der Regel nicht gaußisch sind, sondern ein stärkeres Gewicht in den Flügeln der Verteilung steckt. Das bedeutet, dass sehr große relative Preisänderungen viel wahrscheinlicher sind, als man von einer Gaußverteilung erwarten würde. Die einfache Argumentation über den zentralen Grenzwertsatz scheint also bei Aktienkursen nicht zu funktionieren. Die Gründe dafür sind vielfältig und wir werden ihnen im Rahmen dieses Versuchsteils auf den Grund gehen.

1.3 Temperaturen

Das Wetter gilt gemeinhin als Paradebeispiel für chaotische Systeme. Gerne wird etwa der Flügelschlag eines Schmetterlings in China angeführt, der womöglich so große Auswirkungen haben könne, dass sich das Wetter in Europa änderte. Am besten wird die Komplexität der Wetterphänomene jedoch deutlich, wenn man bedenkt, wie außerordentlich schwierig es ist, Vorhersagen zu machen. Obwohl die Wetterdienste über die weltweit leistungsstärksten Großrechner verfügen und ein immenser Aufwand in die Messung empirischer Daten, sowie in die Modellbildung fließt, können zuverlässige Vorhersagen oft nur für wenige Tage gemacht werden.

Wir werden uns hier auf Temperaturdaten beschränken. Auch diese zeigen auf den ersten Blick ein recht erratisches Verhalten, siehe Abbildung 1. Jedoch wird es sicher auch starke räumliche

und zeitliche Korrelationen im Temperaturverlauf geben, sowie eine starke Abhängigkeit von der Tages- und Jahreszeit. In diesem Versuchsteil wollen wir diese systematischen Aspekte, sowie den stochastischen Anteil näher ergründen.

1.4 Vorbereitung, Literatur

Machen Sie sich mit den Grundbegriffen der Statistik vertraut. Orientieren Sie sich dazu an den folgenden Leitfragen:

- Was versteht man unter Verteilung?
- Was ist eine Wahrscheinlichkeitsdichte?
- Wie schätzt man eine Wahrscheinlichkeitsdichte aus empirischen Daten?
- Wie ist die Gaußverteilung definiert?
- Was besagt der zentrale Grenzwertsatz?
- Welche Voraussetzungen müssen für den zentralen Grenzwertsatz gelten?
- Wie misst man statistische Abhängigkeiten zwischen empirischen Zeitreihen? (Korrelationskoeffizient, Copula)
- Was versteht man unter einer Autokorrelationsfunktion?
- Was ist ein Markov-Prozess?
- Wie testet man die Markov-Eigenschaft von empirischen Zeitreihen?

Für einen schnellen Einstieg in diese grundlegenden Fragen eignet sich ein Blick in Wikipedia, sowie die folgende Literatur:

1. Ulrich Krengel (2005) Einführung in die Wahrscheinlichkeitstheorie und Statistik, Vieweg+Teubner Verlag.
2. Rudi Schäfer (2015) Introduction to copulas: Studying statistical dependencies, Semesterapparat. Siehe hierzu auch den Infotext auf der Homepage des Praktikums.

2 Daten und Auswertung

Wir werden Sie Schritt für Schritt durch die Auswertung Ihrer Ergebnisse führen. Die Verwendung von Wolfram Mathematica 10, dessen Befehle in dieser Anleitung in `Schreibmaschinenschrift` gedruckt werden, ist dabei obligatorisch. Es ist oftmals eine Vereinfachung, `Map` bzw. `/@` zu nutzen. Zudem werden sogenannte *pure functions* (also `Function` bzw. `#` und `&`) als bekannt vorausgesetzt. Machen Sie bei der Bearbeitung der Aufgaben ausführlichen Gebrauch von der Mathematica-Hilfe, in der sämtliche Funktionen an vielen Beispielen erläutert werden.

2.1 Brownsche Bewegung

Zeichnen Sie die Trajektorien der Partikel aus dem Ihnen zur Verfügung gestellten Video auf und exportieren Sie die Ergebnisse. Die statistische Auswertung dieser Daten erfolgt anschließend in Mathematica. Gehen Sie wie folgt vor:

1. Starten Sie das Programm ImageJ und öffnen Sie das Video zur Brownschen Bewegung: `File` ⇒ `Open` ⇒ Datei auswählen
2. `Convert to Grayscale` aktivieren ⇒ `OK`
3. `Plugins` ⇒ `Mosaic` ⇒ `Particle Tracker 2D/3D` ⇒ No 3D data ⇒ `OK` ⇒ Warten!
4. Kalibrierung gemäß Abbildung 2 ⇒ (`Preview Detected` ⇒) `OK`
   (Mehr Informationen unter `http://mosaic.mpi-cbg.de/ParticleTracker/`.)

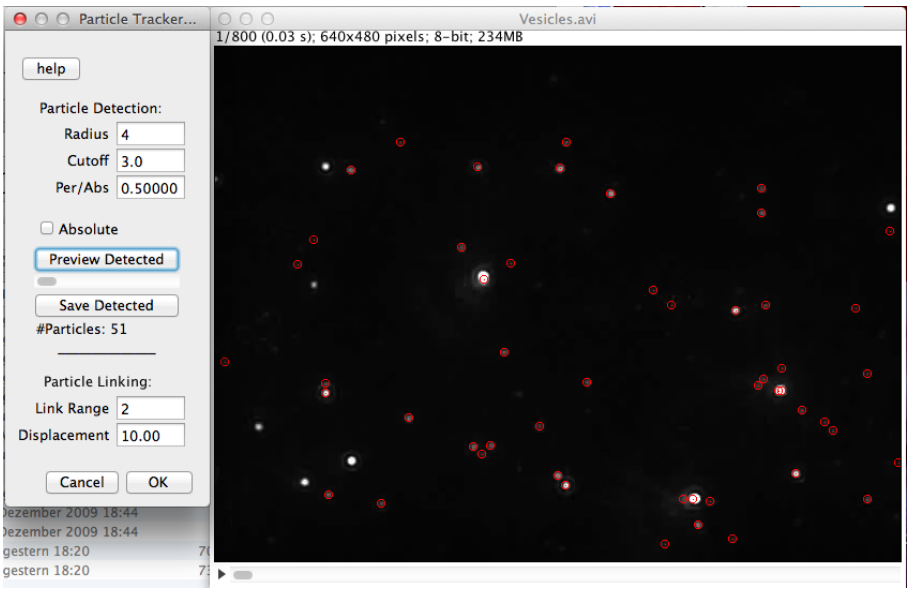

Abbildung 2: Paramter für den *particle tracker* für `Vesicles.avi`.

5. Im sich öffnenden Fenster `Results` auf `All Trajectories to Table` klicken und sie über `File` ⇒ `Save As` als Tabelle abspeichern.
6. Lesen Sie die detektierten Trajektorien in Mathematica ein. Nutzen Sie die Funktion `Import` und importieren Sie die Daten mit den Optionen `"Data"`, sowie `"HeaderLines" -> 1`. Um nun zwischen den einzelnen Trajektorien separieren zu können, empfehlen wir die Funktion `Gather`. Alternativ, jedoch viel langsamer, funktioniert dies auch mit `Select`.

7. Im Folgenden filtern Sie ungewünschte Trajektorien heraus:

   - Berechnen Sie die Anzahl der Zeitschritte für sämtliche Trajektorien und verwerfen Sie jene, die kürzer als 50 Zeitschritte sind. Benutzen Sie dafür die Funktion `Delete` bzw. `Extract` in Verbindung mit `Position`.
     Tipp: Wenn Sie aus einer Liste `a` die Positionen aller Werte, die kleiner als `x` sind, ermitteln möchten, dann lautet ein möglicher Mathematica-Befehl für diese Aufgabe: `Position[a, _?(# < x &)]`.
   - Berechnen Sie für alle M übrig gebliebenen Trajektorien die Differenzen (`Differences`) $\Delta x_n = x_{n+1} - x_n$ für $n \in [1,\, L_i - 1]$, wenn L_i die Anzahl der Datenpunkte der Trajektorie i ($i \in [1,\, M]$) ist. Wiederholen Sie dies für die zweite Dimension.
   - Verwerfen Sie alle Trajektorien, deren Zeitreihen Δx_n und/oder Δy_n eine zu große Wölbung (`Kurtosis`) aufweisen. Welche Grenze haben Sie hier gewählt und warum?
   - Die Anzahl der so gefilterten Trajektorien sei fortan N.

8. Stellen Sie exemplarisch die längste Trajektorie und ihre zugehörigen Zeitreihen Δx_n und Δy_n dar. Berechnen Sie für alle N Zeitreihen die zentralen Momente, also den Mittelwert, die Standardabweichung, die Schiefe (`Skewness`) und die Wölbung.

9. Normieren Sie nun jeweils die beiden Zeitreihen Δx_n und Δy_n jeder der N Trajektorien. Ziehen Sie dafür von jedem Wert den Mittelwert der gesamten Zeitreihe ab und teilen Sie das Ergebnis durch die Standardabweichung. Fassen Sie sodann alle N Zeitreihen Δx_n mittels `Flatten` in eine Liste, die wir nun Γ_x nennen wollen, zusammen. Wiederholen Sie dies für die zweite Dimension.

10. Wie groß sind die zentralen Momente von Γ_x und Γ_y und warum verhalten Sie sich so? Stellen Sie die Wahrscheinlichkeitsverteilungen von Γ_x und Γ_y zusammen mit einer Normalverteilung dar. Welche Werte für den Mittelwert und die Standardabweichung der Normalverteilung müssen eingestellt werden? Was fällt beim Vergleich mit den empirischen Ergebnissen auf?
    Tipp: Bei der Verwendung von `Histogram[a, Automatic, "PDF"]` wird das Histogramm einer Liste `a` bereits so normiert, wie es für eine Wahrscheinlichkeitsverteilung vonnöten ist. Nutzen Sie in diesem Kontext auch eine logarithmische Darstellung, für die die Option `ScalingFunctions -> "Log"` sorgt.

11. Generieren Sie ein Streudiagramm aus den Werten von Γ_x und Γ_y.

12. Die Funktion `qrank[x_] := (Ordering@Ordering@x - 0.5)/Length@x` soll nun auf Γ_x und Γ_y angewandt werden, bevor Sie ein weiteres Streudiagramm erstellen. Was macht diese Funktion und was fällt auf? Erstellen Sie zudem ein `Histogram3D` aus diesen Daten. Wie nennt man diese Darstellung?

13. Schreiben Sie eine Autokorrelationsfunktion unter der Verwendung von `Correlation` und `Drop`. Wie sieht sie aus, wenn sie für die längste Trajektorie auf Δx bzw. Δy und $(\Delta x)^2$ bzw. $(\Delta y)^2$ angewandt wird, und warum?

14. Plotten Sie nun ein Streudiagramm, auf dem für die längste Trajektorie die Liste Δx ohne ihren letzten Wert gegen die Liste Δx ohne ihren ersten Wert aufgetragen ist. Wiederholen Sie dies für Δy. Wenden Sie auf diese verkürzten Listen auch die obige Funktion `qrank` an und erstellen Sie hier ebenfalls ein `Histogram3D`.

2.2 Aktiendaten

Die nun folgenden Aufgaben zur Auswertung der Börsendaten ähneln in besonderem Maße jenen zur Brownschen Bewegung. Diskutieren Sie daher jeweils die Gemeinsamkeiten und Unterschiede der Ergebnisse in bezug auf den ersten Teil.

1. Die benötigten Aktiendaten können direkt mit Mathematica abgerufen werden. Mit dem Befehl `FinancialData["FB", "Jan. 1, 2014"]` erhalten Sie alle Tagesschlusspreise der Aktie des Unternehmens Facebook seit dem 1. Januar 2014.
   Rufen Sie auf diese Weise die Tagesschlusspreise von $K = 9$ Aktien des Aktienindex S&P 500 für jeweils zehn Jahre ab. Stellen Sie sicher, dass für jede Aktie dieselbe Anzahl an Datenpunkten abgerufen wird. Warum ist dies zum Beispiel bei der Aktie FB nicht der Fall?

2. Stellen Sie die Preiszeitreihe einer dieser Aktien mittels `DateListPlot` dar.

3. Berechnen Sie für den gesamten Zeitraum jeder Aktie k ($k \in [1, K]$) die Renditen (Returns) $R_k(t) := (S(t + \tau) - S(t))/S(t)$ mit $\tau = 1$ Tag.

4. Stellen Sie exemplarisch die Renditen jener Aktie dar, deren Preiszeitreihe Sie bereits geplottet haben. Berechnen Sie außerdem die zentralen Momente der Renditen aller zehn Aktien.

5. Normieren Sie nun die Returnzeitreihen $R_k(t)$ jeder der zehn Aktien. Ziehen Sie dafür von jedem Wert den Mittelwert der gesamten Zeitreihe ab und teilen Sie das Ergebnis durch die Standardabweichung.

6. Stellen Sie jeweils die Wahrscheinlichkeitsverteilungen aller $R_k(t)$ zusammen mit einer Normalverteilung dar. Was fällt hier beim Vergleich mit den empirischen Ergebnissen auf?

7. Generieren Sie ein Streudiagramm aus $R_m(t)$ und $R_n(t)$ für von Ihnen gewählte Aktien m und n mit $m \neq n$.

8. Wenden Sie nun `qrank` auf diese Returnzeitreihen $R_m(t)$ und $R_n(t)$ an, bevor Sie ein weiteres Streudiagramm erstellen. Was fällt auf? Erstellen Sie zudem ein `Histogram3D` aus diesen Daten.

9. Wie sieht die Autokorrelationsfunktion aus, wenn sie auf $R_m(t)$ und $R_m^2(t)$ angewandt wird, und warum? Wie nennt man das Phänomen, das hier zutage tritt, und wie kann man es noch visualisieren?

10. Plotten Sie nun ein Streudiagramm, auf dem $R_m(t + \tau)$ (mit $\tau = 1$ Tag) gegen $R_m(t)$ aufgetragen ist (analog zum ersten Teil). Wenden Sie auf diese verkürzten Listen auch die obige Funktion `qrank` an und erstellen Sie hier ebenfalls ein `Histogram3D`.

2.3 Temperaturdaten

Auch dieser Teil zur Auswertung der Temperaturdaten umfasst ähnliche Gesichtspunkte wie in den bisherigen beiden Abschnitten. Heben Sie deshalb auch hier oder in einem separaten Kapitel sämtliche Gemeinsamkeiten und Unterschiede zur Brownschen Bewegung und zu den Aktiendaten hervor.

In diesem Aufgabenteil ist es besonders wichtig, dass Sie keine ältere Version als Mathematica 10 verwenden, da einige der hier verwendeten Funktionen erst ab dieser Version zur Verfügung stehen.

1. Die Temperaturdaten können ebenfalls direkt mit Mathematica abgerufen werden. Über `TimeSeriesResample[WeatherData["Berlin", "Temperature", {{2007, 1, 1}, {2007, 12, 31}, "Day"}]]` erhalten Sie sie zum Beispiel für die Stadt Berlin im Jahr 2007 in Form eines `TimeSeries`-Objekts. Wofür dient die Funktion `TimeSeriesResample` im angegebenen Befehl?

Rufen Sie auf diese Weise die täglichen Wetterdaten für neun Städte ab. Jeder Kontinent soll mindestens einmal vorkommen und zwei der Städte müssen in Deutschland liegen. Verwenden Sie die englische Schreibweise dieser Orte. Das Zeitintervall ist zehn Jahre. Wählen Sie als Enddatum aber nicht das heutige Datum, sondern eine Woche davor. Andernfalls kann es passieren, dass nicht alle Zeitreihen dieselbe Anzahl an Einträgen haben.
Tipp: Wenn Sie auf die Temperaturwerte des `TimeSeries`-Objekts zugreifen möchten, dann wählen Sie das Element `ts["Values"]`, wenn `ts` der Bezeichner des `TimeSeries`-Objekts ist. Es ist ferner empfehlenswert, die Einheiten mithilfe der Funktion `QuantityMagnitude` zu entfernen, da Mathematica sonst Schwierigkeiten mit der (noch folgenden) Normierung der Zeitreihen hat.

2. Wir betrachten nun die absoluten Temperaturen:

   - Stellen Sie eine beliebige Temperaturzeitreihe in einem `DateListPlot` dar.
   - Berechnen Sie für jeden Ort die zentralen Momente der Temperaturzeitreihen. Was fällt hier auf?
   - Normieren Sie jede der Temperaturzeitreihen. Ziehen Sie dafür von jedem Wert den Mittelwert der gesamten Zeitreihe ab und teilen Sie das Ergebnis durch die Standardabweichung. Stellen Sie dann die Wahrscheinlichkeitsverteilungen zusammen mit einer Normalverteilung dar. Was beobachten Sie?
   - Erstellen Sie ein Streudiagramm der normierten Temperaturen zweier Orte, die in Deutschland liegen. Wenden Sie dann auch hier `qrank` an und erstellen Sie ebenfalls ein `Histogram3D`. Recherchieren Sie, wie die analytische Form, die dieses Histogramm beschreibt, lautet. Erstellen Sie auch ein Streudiagramm für einen Ort, der in Deutschland liegt, und einen anderen, der sich auf der Südhemisphäre befindet.
   - Wie sieht die Autokorrelationsfunktion einer normierten Temperaturzeitreihe aus? Welche Besonderheit ist hier offensichtlich?

3. Jetzt geht es um die Temperaturdifferenzen aufeinanderfolgender Tage:

   - Stellen Sie eine beliebige Temperaturdifferenzenzeitreihe dar.
   - Berechnen Sie für jeden Ort die zentralen Momente der Temperaturdifferenzenzeitreihen. Was fällt an dieser Stelle auf?
   - Normieren Sie jede der Temperaturdifferenzenzeitreihen. Ziehen Sie dafür von jedem Wert den Mittelwert der gesamten Zeitreihe ab und teilen Sie das Ergebnis durch die Standardabweichung. Stellen Sie die Wahrscheinlichkeitsverteilungen zusammen mit einer Normalverteilung dar. Was beobachten Sie diesmal?
   - Erstellen Sie ein Streudiagramm der normierten Temperaturdifferenzen zweier Orte, die in Deutschland liegen. Wenden Sie dann auch hier `qrank` an und erstellen Sie ebenfalls ein `Histogram3D`. Erstellen Sie auch ein Streudiagramm für einen Ort, der in Deutschland liegt, und einen anderen, der sich auf der Südhemisphäre befindet.
   - Wie sieht die Autokorrelationsfunktion einer normierten Temperaturdifferenzenzeitreihe aus und wie deuten Sie dieses Verhalten? Ermitteln Sie, wann die Autokorrelationsfunktion ihr absolutes Minimum erreicht.

3 Fragen zur Selbstkontrolle

Beantworten Sie die folgenden Fragen nachdem Sie sich auf den Versuch mit der empfohlenen Literatur vorbereitet haben, um herauszufinden, ob Sie bei manchen Themen noch Nachholbedarf haben.

1. Erklären Sie mit eigenen Worten, wovon dieser Versuch handelt.

   Mit welchen Aspekten befasst sich die Auswertung?

   Welche Unterschiede zwischen den drei Versuchsteilen erwarten Sie?

2. Worin unterscheiden sich die Mathematica-Operatoren `/@` und `@`?

   Nennen Sie jeweils ein sinnvolles Anwendungsbeispiel.

   Wann benötigen Sie in diesem Zusammenhang die sogenannten *pure functions*?

3. Was ist eine PDF (*probability density function*) und was beschreibt die CDF (*cumulative distribution function*)? Welcher Zusammenhang besteht zwischen ihnen?

4. Was besagt der zentrale Grenzwertsatz?

5. Wie wird die Gaußverteilung noch genannt und weshalb?

   Wie lautet die mathematische Beschreibung ihrer PDF?

   Wie sieht sie in einer logarithmischen Darstellung aus?

6. Wie sind die zentralen Momente einer Zufallsvariablen definiert und was beschreiben sie?

   Welche Synonyme sind dafür geläufig? Erklären Sie diese anschaulich.

   Was ergibt sich im Falle gaußverteilter Zufallsvariablen?

7. Was beschreibt der Pearsonsche Korrelationskoeffizient?

   Wie ist er definiert?

   Warum nennt man ihn auch den linearen Korrelationskoeffizienten?

   Sind unkorrelierte Größen zwangsläufig auch voneinander unabhängig?

   Was ist eine Autokorrelationsfunktion und wie ist sie definiert?

8. Was ist ein Streudiagramm?

   Was versteht man unter einem gerankten Streudiagramm?

   Wie nennt man seine zweidimensionale Verteilung?

   * Stellen Sie einen Zusammenhang zum linearen Korrelationskoeffizienten her.

   * Stellen Sie einen Zusammenhang zur Autokorrelationsfunktion her.