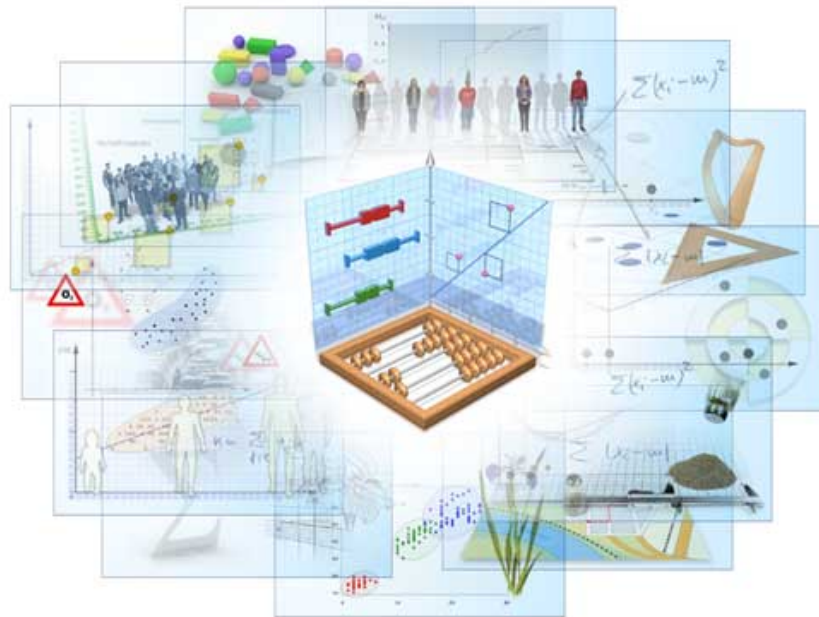


Hinweis:

Diese Druckversion der Lerneinheit stellt aufgrund der Beschaffenheit des Mediums eine im Funktionsumfang stark eingeschränkte Variante des Lernmaterials dar. Um alle Funktionen, insbesondere Verlinkungen, zusätzliche Dateien, Animationen und Interaktionen, nutzen zu können, benötigen Sie die On- oder Offlineversion.
Die Inhalte sind urheberrechtlich geschützt.
©2024 Berliner Hochschule für Technik (BHT)

STB - Statistik in Beispielen



Lernziele und Überblick

Mit dieser Lerneinheit wollen wir Sie neugierig auf das machen, was Ihnen im Laufe dieses Kurses begegnen wird.

Da der Mensch wesentlich weniger Anstrengung, Kraft und Energie benötigt, wenn sich seine Gesichtszüge zu einem Schmunzeln oder Lachen entspannen, haben wir uns Mühe gegeben, den für viele sicher sehr trockenen Inhalt in buntem, nicht immer ganz ernstem Gewand zu präsentieren.



Lernziele

In der Statistik haben wir es vor allem mit Daten und mit Methoden zu tun, die uns eine systematische Betrachtung der Daten erleichtern sollen. Davon gleich mehr.

Beim Studium der folgenden Seiten sollen Sie versuchen, sich in die einzelnen Beispiele und Fragestellungen einzudenken und einzufühlen. Lassen Sie Ihrer Fantasie freien Lauf, finden Sie heraus, ob Sie aus eigener Erfahrung oder Beobachtung ähnliche Situationen kennen.

Falls Sie sich vielleicht noch nicht ganz sicher im Umgang mit dem Blättern und Navigieren auf unseren Seiten fühlen, ist es eine gute Idee, das während des Studiums dieser Lerneinheit zu üben.

Die Lernziele finden Sie in der Regel knapper formuliert zu Beginn jeder Lerneinheit.



Gliederung der Lerneinheit

1. Einführung in Beispielen
2. Hinweise zur Arbeitsweise und zum didaktischen Konzept
3. Übungen mit der Statistiksoftware **R**



Zeitbedarf und Umfang

Von uns geschätzter Zeitbedarf für das Erarbeiten der theoretischen und praktischen Elemente der Lerneinheit.

Für diese Lerneinheit benötigen Sie etwa 60 Minuten. Zusätzlich sollten Sie zur Installation der Statistiksoftware **R** in etwa 60 Minuten veranschlagen.

1 Einleitung



Am Anfang jeder statistischen Analyse steht die inhaltliche Fragestellung. Ohne diese Fragestellung ergibt die Statistik keinen Sinn.

- Werden die Studierenden immer fauler?
- Sind Lebensmittel in den letzten Jahren teurer geworden?
- Was ist das durchschnittliche Monatseinkommen einer Berufsanfängerin mit dem Abschluss Wirtschaftsingenieur?
- Wie stark variiert die Auslastung des Servers eines Unternehmens von Monat zu Monat?
- Wie zuverlässig öffnen sich Airbags im neuen Auto?
- Verfügen Absolventen aus den südlichen Bundesländern über besonders gute Fremdsprachenkenntnisse?
- Sind MathematikerInnen besonders musikalisch?

Diese Menge an Fragen soll fürs erste genügen. Sie werden nach erfolgreichem Studium des Kursmaterials gelernt haben, auf welcher Datenbasis und mit welchen Methoden Sie Antworten darauf geben können.

Vielleicht stellen Sie auch fest, dass Sie mit den erworbenen Kenntnissen noch nicht zufrieden sind. Dann war der Kurs wirklich erfolgreich. Das Methodenspektrum der modernen Statistik ist sehr umfangreich und viele Verfahren werden ständig weiterentwickelt.

Bevor Sie sich in unseren Lerneinheiten Schritt für Schritt die ersten Grundbegriffe, Kennzahlen zur Beschreibung von Daten und Methoden zur Analyse von Zusammenhängen aneignen, haben Sie jetzt die Gelegenheit, statistische Fragestellungen und Daten kennen zu lernen.

Sie finden in dieser Lerneinheit wichtige Hinweise für Ihre Arbeitsweise, zum didaktischen Konzept und der Statistiksoftware **R** - einer Übungssoftware für statistische Datenanalysen.

Wenn Sie die Titelgraphiken der Lerneinheiten etwas genauer betrachten, werden Sie feststellen, dass darin einiges an Inhalt transportiert wird. Mit der folgenden Abbildung wollen wir Sie beispielsweise darauf aufmerksam machen, wie man aus Zusammenhängen lernen kann.

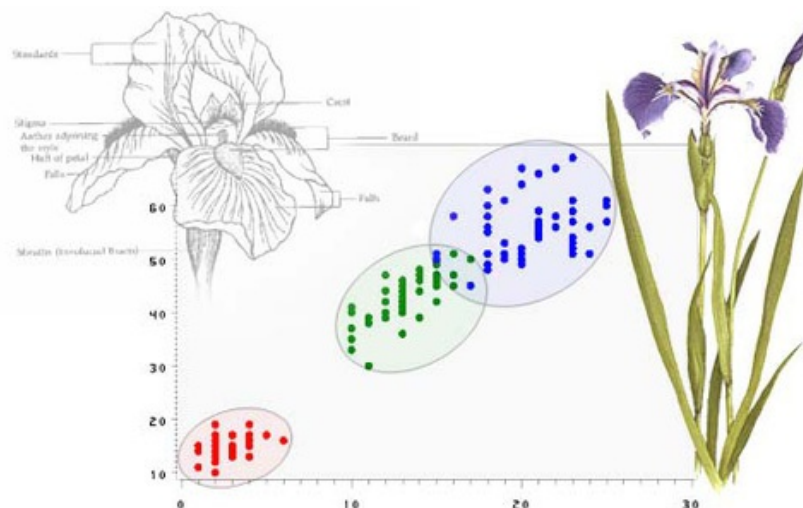


Abb: Titelgrafik der Lerneinheit STB

Die in der Abbildung dargestellten Punkte geben die Wertepaare der Breite und Länge von Blütenkelchen verschiedener Iris-Arten wieder. Im Unterschied zu den bisher betrachteten Fragestellungen werden hierbei gleichzeitig mehrere Eigenschaften der Objekte untersucht. Multivariate Betrachtungen machen Statistik richtig interessant. Sie kommen in allen Bereichen von Wirtschaft, Technik und Forschung zum Einsatz.

2 Datenquellen

In unserem Alltag sammeln wir ständig Informationen. Eine der an diesem Projekt beteiligten Studentinnen heißt Claudia. Einige ihrer Eigenschaften haben wir aufgeschrieben. Mehr wollte uns Claudia für diesen Zweck nicht verraten.



Beispiel

Wir beobachten Claudia

Von Claudia erfahren wir folgende Eigenschaften:

Alter = 29,
Geschlecht = weiblich,
Länge = 173,
Zufriedenheit = 2



Hinweis

Wenn Sie jetzt neugierig geworden sind, sollten Sie gleich initiativ werden. Verabreden Sie mit Ihren Mitstudentinnen und -studenten eine kleine Datensammlung von Eigenschaften, die Sie untereinander austauschen können. Dann haben Sie gleich einen ersten Datensatz zum Üben. Ihre Kursbetreuung hilft Ihnen sicher gern, die Sammlung im Lernraum zu organisieren.

3 Datensammlung



Beobachten wir dieselben Eigenschaften an mehreren Personen, erhalten wir eine Datensammlung, mit deren Hilfe wir versuchen können, die Gruppe der Beobachteten zu beschreiben. Wir können die erhobenen Daten in Form einer Liste oder Tabelle aufschreiben und erhalten so eine Datenmatrix:

Index	Name	Größe (cm)	Gewicht (kg)
1	Magda	158	48
2	Anna	160	59
3	Roland	163	102
4	Swetlana	165	57
5	Alexander	165	80
6	Tamara	168	53
7	Iwan	168	58
8	Eva	168	58
9	Karoline	169	66
10	Nikolaj	170	87
11	Alexandra	171	70
12	Ingrid	171	79
13	Oksana	172	68
14	Jörg	173	73
15	Volker	174	76
16	Stefan	174	63
17	Heike	174	83
18	Karpo	177	65
19	Vladimir	177	77
20	Stanislaw	178	72
21	Felix	178	85
22	Andrej	186	67
23	Walerij	190	80
24	Jost	191	95
25	Stefan	192	90
26	Witalij	194	72

Tab.: Beispielhafte Datenmatrix befragter Personen

In dieser Datenmatrix sind alle erfassten Informationen enthalten. Machen Sie sich ein Bild, wie groß, wie schwer sind die befragten Studentinnen und Studenten.

Was fällt Ihnen an den Daten auf?

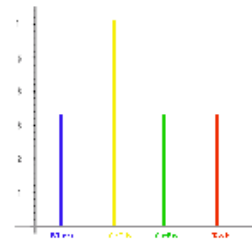
Sicher ist, dass wir auf diese Weise keine Übersicht über alle Studierende einer Hochschule oder der Mitarbeiter und Mitarbeiterinnen eines Unternehmens erhalten könnten. Das sind in der Regel einfach zu viele! So lange Tabellen könnten wir keinesfalls überblicken.

4 Datenanalyse – fast ein Kinderspiel

Bei sehr geringem Datenumfang genügt es, die Werte aufzulisten. Sobald aber eine Liste mehr als zehn oder gar mehr als zwanzig Einträge besitzt, können wir uns aus der Datenliste allein nur noch schlecht ein Bild der gesammelten Informationen machen. Geht es uns um eine einzelne Eigenschaft, genügt es stattdessen aufzuschreiben, wie oft die verschiedenen Möglichkeiten in den Daten vorkommen.



9 mal rot,
16 mal gelb,
9 mal grün,
9 mal blau,



Das hätte auch mit mehr als den hier gezeigten 43 Bauklötzen funktioniert. Zur vollständigen Beschreibung der Farben genügen die Häufigkeiten.

5 Ordnung muss sein

Bei Eigenschaften wie der Körpergröße kann uns eine Aufstellung der Größe nach helfen, Übersicht zu gewinnen.

Vorher



Nachher



Wer ist am kleinsten, wer am größten, wer steht in der Mitte? Das können wir jetzt mühelos herausfinden. Das funktionierte auch, wenn wir anstatt der 13 Freiwilligen unseres Teams die halbe Hochschule versammelt hätten.

6 Statistiksoftware R und R-Studio

Zur Datenanalyse verwendet man in der Praxis spezielle Statistiksoftware. Teilweise kommen auch einfachere Tabellenkalkulationsprogramme wie EXCEL oder CALC zur Anwendung.

Statistik lernt man durch Anwenden, behaupten wir. Damit Sie diese Behauptung überprüfen können – wir sind uns ziemlich sicher, dass sie stimmt – haben wir für Sie eine Möglichkeit gefunden, wie Sie schnell und unkompliziert auch größere Mengen von Daten handhaben und bearbeiten können – die Statistiksoftware **R**.



R ist eine Interpretersprache zur statistischen Datenanalyse sowie zur Erstellung professioneller Grafiken. Als Open Source Software steht **R** kostenfrei zur Verfügung und läuft auf diversen Plattformen. Mit seiner Vielfalt an wichtigen statistischen Funktionen und Zusatzpaketen sowie der Möglichkeit, Datensätze zu importieren und zu exportieren, bietet **R** viele Möglichkeiten im Bereich der Datenbearbeitung. Die vielfältigen graphischen Darstellungsvarianten ermöglichen dabei anschauliche Visualisierungen der statistischen Daten.


Informationen zum Download und Installationshinweise finden Sie auf der Webseite der R-Projekts.

<http://www.r-project.org/>



Mit der Software **R-Studio** steht mittlerweile eine integrierte Entwicklungsumgebung für **R** zur Verfügung. Sie ist kostenfrei als Open Source aber auch in einer kommerziellen Version verfügbar. Wir empfehlen Ihnen die Nutzung von R-Studio. Weitere Informationen hierzu finden Sie unter:

<http://www.rstudio.com/ide/>

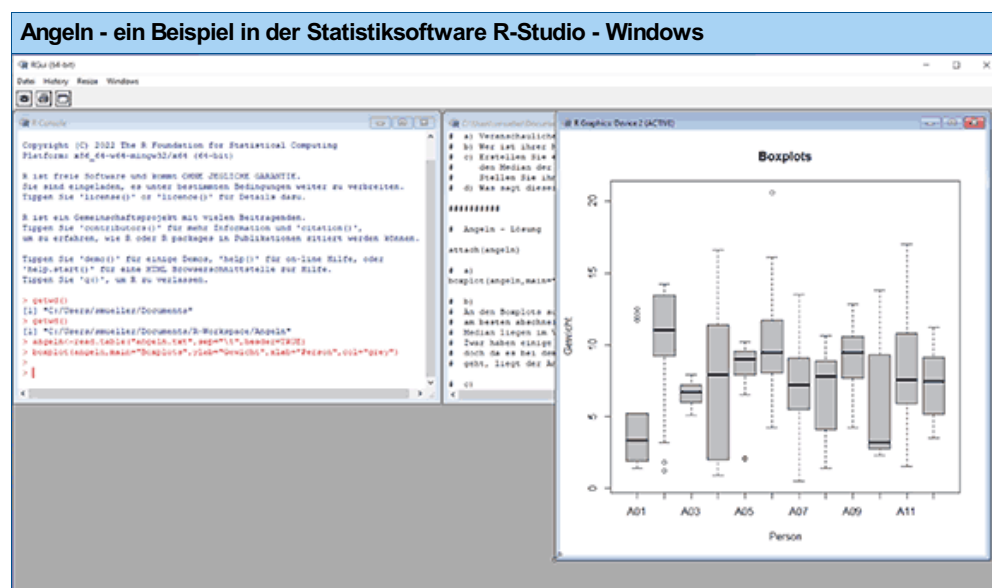
Sie finden am Ende jeder Lerneinheit spezielle Übungsaufgaben für die Statistiksoftware **R**, die Sie einfach speichern und dann über das Programm öffnen können. Die Dateien erkennen Sie am Symbol  vor dem Dateinamen.

Wer fängt die dicksten Fische?

Wie Sie umfangreiche Datensätze schnell auswerten können, zeigt Ihnen die kurze Simulation einer Übungsaufgabe bei der es um den Angelsport geht. Die vorliegenden Resultate eines Wettangels werden mit Hilfe der Funktionen der Statistiksoftware **R-Studio** aufbereitet.



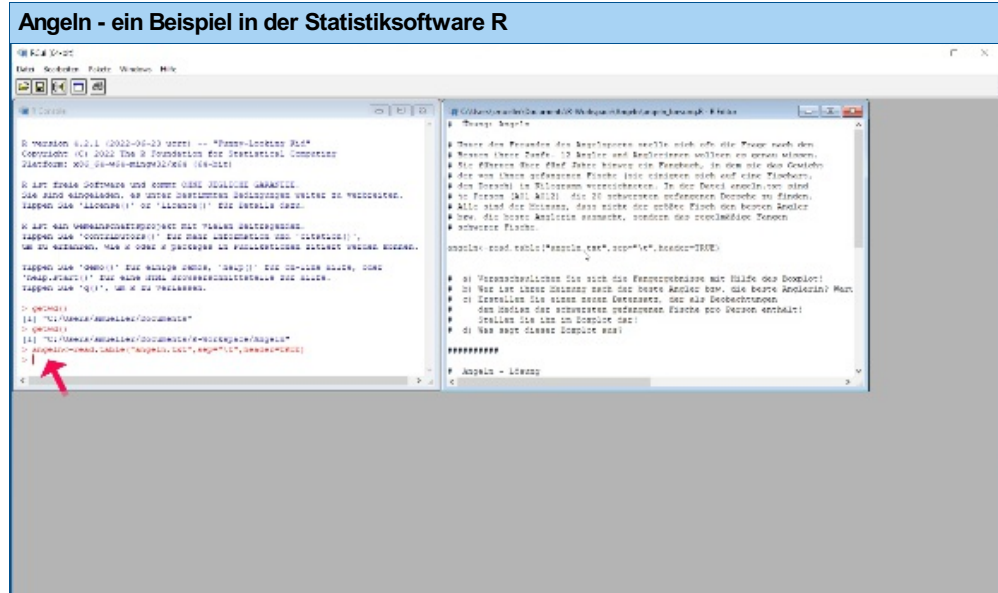
Film



Haben Sie auf R-Studio verzichtet und nur die Software **R** installiert, dann sehen sie im folgendem Video wie das Beispiel Angeln ausgeführt wird.



Film



Mit den Beispieldateien können Sie in der folgenden Übung ihre **R**-Installation testen. Sie sollten sich bald in die Bedienung von **R** einarbeiten, um vom „learning by doing“ profitieren zu können.



Statistiksoftware R

Übung STB-01

Installation der Software R oder R-Studio

Wir schlagen vor, dass Sie sich nun die Statistiksoftware **R** auf Ihrem Rechner installieren. Hinweise zum Download und zur Installation finden Sie unter:

<http://www.r-project.org/> oder für R-Studio unter:

<http://www.rstudio.com/ide/>

Wenn Sie die Installation erfolgreich abgeschlossen haben, rufen Sie sich die folgende Einführung in **R** als PDF-Datei auf und erkunden Sie die beschriebenen Funktionen - somit sind Sie für die kommenden Lerneinheiten vorbereitet.

Einführung in die Sprache **R** von F. Müller [0.3 MB]

Sie können zusätzlich das Beispiel „Angeln“ mit den unten stehenden Quelldateien selber durchführen. Beachten Sie, dass die Datei **angeln.txt** den Datensatz beinhaltet. Speichern Sie die Datei im gleichen Ordner wie die R-Datei und stellen Sie das Arbeitsverzeichnis von **R** auf diesen Ordner ein.

Angeln (angeln_aufgabe.pdf)

Angeln (angeln_loesung.R)

angeln.txt (2 KB)

Bearbeitungszeit: 90 Minuten

Tutorials und Portable Version

Auf der Webseite der Universität Wien finden Sie weitere hilfreiche Informationen zu **R**.

<https://statistik.boku.ac.at/>

Anleitungen

Eine umfangreiche und gut aufbereitete Seite mit Beispielen und Anleitungen finden Sie auf der englischsprachigen Seite Quick-R:

<http://www.statmethods.net/>

7 Statistik in digitalen Zeiten – Praxisbeispiele

Wir wollen mit den folgenden Praxisbeispielen demonstrieren, dass statistische Datenanalysen durch die Digitalisierung immer selbstverständlicher verwendet werden. Dazu bedarf es keiner Gurus oder nerdiger Freaks. Für den Anfang genügen solide Grundlagenkenntnisse, ein ausreichendes Verständnis der Fragestellung und die Fähigkeit, ein wenig mit Daten, Computern und Programmen umgehen zu können.

Wir haben für Sie zwei Beispiele ausgewählt: eine öffentlich zugängliche Datenbank des Umweltbundesamts (UBA) und einen relativ großen Datensatz aus der Immobilienwirtschaft.

Die Datensammlung aus dem Luftqualitätsmessnetz des Umweltbundesamts kann Ihnen dabei helfen, eigene Fragestellungen zu entwickeln und diese dann mit Hilfe der abrufbaren Daten zu beantworten.

Im zweiten Beispiel verarbeiten wir Angebotsdaten von Mietwohnungen und wir zeigen Ihnen, wie Sie mit Hilfe von **R** die Datenaufbereitung und einfache statistische Berechnungen durchführen können.

Wir haben bereits darauf hingewiesen, dass es für Sie sehr hilfreich sein kann, sich mit der Statistiksoftware **R** oder mit **R-Studio** zu befassen. Diese Investition in zusätzliches Lernen kann sogar Spaß machen und zahlt sich auf lange Sicht garantiert aus.

Abschließend möchten wir Ihre Aufmerksamkeit noch auf ein Statistikportal mit hohem Suchtfaktor lenken - auf die Initiative Gapminder (www.gapminder.org) . Hier erfahren Sie unter anderem wie viel vernünftiger wir entscheiden und analysieren könnten, wenn wir statistische Fakten statt falscher Vermutungen und Vorurteilen als Grundlage unserer Einschätzungen nutzen würden.

7.1 Luftdaten des Umweltbundesamt

EU-weit gültige Regulierungen legen Grenzwerte für die zulässige Konzentration mehrerer potentiell gesundheitsschädlicher Stoffe fest. Dazu gehören unter anderem Stickoxide NO_2 , SO_2 Ozon O_3 und Feinstaub. Schadstoffe können auf verschiedenen Wegen in die Umwelt gelangen. Quellen sind unter anderem Heizungen, Fahrzeuge mit Verbrennungsmotor, Industrieanlagen und die Landwirtschaft.

Wie diese Grenzwerte festgelegt werden, erläutern wir hier nicht weiter. Soviel sei gesagt: es werden auf gesundheitswissenschaftlichen Erkenntnissen und politischen Entscheidungen beruhende, vorsichtige aber pragmatische Werte festgelegt.

Eine nur kurzfristige Überschreitung der Werte bedeutet in der Regel keine akute Gefährdung. Je mehr aber aktuelle Werte den Grenzwerten nahe rücken, desto größer wird das Gefahrenpotential. Um die Bevölkerung ausreichend vor Gefahren durch Luftverschmutzung zu schützen, muss dann engmaschiger gemessen werden. Das Luftdatenportal des UBA stellt in übersichtlicher Form Daten und grafische Darstellungen zur Verfügung.

<https://www.umweltbundesamt.de/daten/luft/luftdaten>

Im folgenden Beispiel betrachten wir unter Schritt 1 die Feinstaubbelastung Ende November 2021. Der Kartenausschnitt zeigt den westlichen Teil Berlins. Wir wählen unter Schritt 2 auch zwei Stationen (126 und 032) aus und schauen uns im Schritt 3 das Diagramm an.

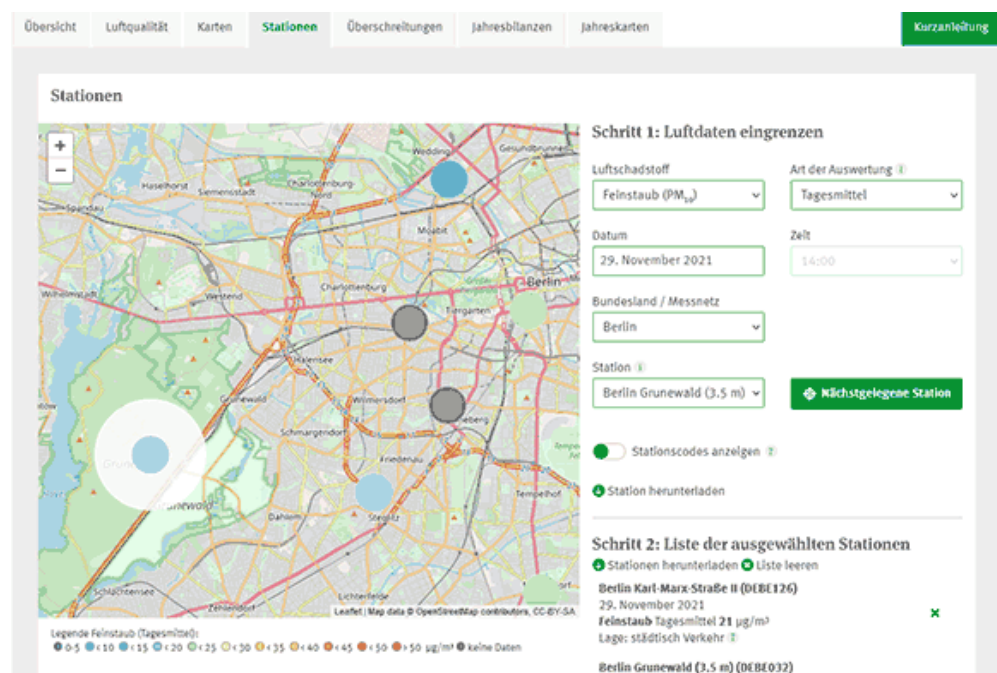


Abb.: Screenshot Luftdaten des Umweltbundesamt (1)

Scrollen Sie auf der Seite etwas nach unten, in Schritt 3 werden detaillierte Messdaten der ausgewählten Messstationen dargestellt.

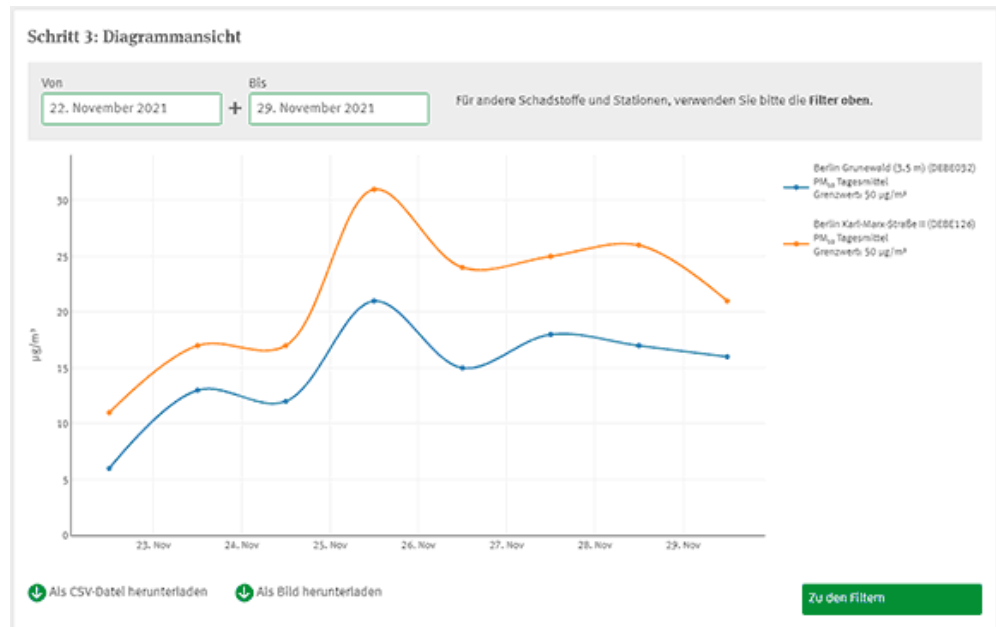


Abb.: Screenshot Luftdaten des Umweltbundesamt (3)

Die beiden Kurven zeigen den Verlauf der Feinstaubkonzentrations-Mittelwerte in der Woche vom 23. Bis zum 29. November. In der Legende auf der rechten Seite werden zusätzlich auch die Grenzwerte angegeben. Die zugehörigen Daten können Sie auch aus einer CSV-Datei ablesen, die Sie bspw. in Excel oder Calc öffnen können.

Stationen_2021-11-22-2021-11-29.csv [3 KB]

Wieviel Statistik können Sie eigentlich jetzt schon ohne den Kurs? Einen Überblick über **alle** Messstationen zu erhalten wird schwierig, aber einzelne Stationen können Sie sich schon anschauen. Also wie wäre es mit „Berlin Mitte“ oder „Grunewald“? Suchen Sie sich eine oder mehrere Stationen im Portal aus, lassen Sie sich aktuelle Werte und das Diagramm anzeigen. Was erkennen Sie? Versuchen Sie mal die folgenden Fragen zu beantworten:

- Wie oft wurden Grenzwerte überschritten?
- Wie hoch war die Feinstaubkonzentration ungefähr im Mittel?
- Von welchen Gegenden ging die höchste Gefahr aus?


Sie können die Antworten aus den Diagrammen schätzen und interpretieren oder die exakten Daten aus der CSV-Datei entnehmen. Je mehr Daten Sie haben, desto schwieriger wird allerdings die Suche und Auswertung. Im nächsten Beispiel werden Sie erfahren, wie Daten aus CSV-Dateien in Statistiksoftware eingelesen und bearbeitet werden. Das, was Sie jetzt ohne spezielle Kenntnisse mit ganz normaler natürlicher Intelligenz und Neugier herausfinden konnten, wird dann systematisch und professionell erzeugt.

7.2 Mietwohnungen in Potsdam

Quelle: OpenStreetMap



Bei dem zweiten Beispiel, das wir Ihnen vorstellen möchten, handelt es sich um Daten zu Mietwohnungen in Potsdam aus den Jahren 2014 bis 2021. Dieser Datensatz wurde uns von der Firma ImmoScout24 für Lehrzwecke zur Verfügung gestellt.

 ImmoScout24 ist ein Unternehmen der Immobilienwirtschaft, das Angebote zu Wohnraum Interessenten zugänglich macht. Es stellt ein Suchportal zur Verfügung, auf dem Anbieter Wohnungen einstellen können und das Wohnungssuchende nach verschiedenen Kriterien durchsuchen können. Durch die gesammelten Angebote über Jahre für viele Städte verfügt ImmoScout über sehr viele Informationen sowohl über den aktuellen Stand als auch über die Entwicklung des Wohnungsmarkts. ImmoScout und andere Unternehmen haben ein großes Interesse an Absolventen und Absolventinnen die in der Lage sind, diese Art von Daten systematisch mit modernen statistischen Methoden auszuwerten.

Der Datensatz „Potsdammieten“ enthält Informationen zu knapp 20.000 Wohnungen mit jeweils 30 Merkmalen. Würden die Daten in eine Tabelle eingefügt werden, wären weit über 500.000 Zellen nötig. Wir haben es also mit einem recht großen Datensatz zu tun. Zur Auswertung dieser Datenmenge ist zwingend Softwareunterstützung notwendig.

Der Datensatz darf von uns zu Lehrzwecken genutzt werden, eine Weitergabe oder Veröffentlichung der digitalen Daten ist nicht gewünscht. Informationen und Daten die eine Identifizierung der Wohnungen möglich machen könnten, wurden entfernt. So sind bspw. weder Fotos noch Grundrisse der Wohnungen verfügbar und auch die Angabe der Etage in der sich die Wohnungen befinden fehlen. Für Lehr- und Übungszwecke ist der auf realen Informationen basierende Datensatz dennoch sehr gut geeignet. Er ermöglicht die Untersuchung einer Vielzahl an Fragestellungen, bspw.:

- Wie haben sich die Quadratmeterpreise für bestimmte Wohnungsgrößen entwickelt?
- Wie hat sich die Zahl der angebotenen Wohnungen verändert?
- Welche Bezirke haben die stärksten Veränderungen erfahren?
- Welche Wohnungstypen sind am zahlreichsten vorhanden?

Die Beantwortung dieser Fragen wird nicht Bestandteil dieses Studienmoduls sein, aber wir wollen Ihnen an einem Beispiel das Vorgehen mit Hilfe von **R** zeigen.



Beispiel

Wohnungsdaten aufbereiten und darstellen mit R

Unsere Fragestellung lautet: Welche unterschiedlichen Quadratmeterpreise haben Wohnungen mit 3 und mehr Zimmern in Potsdam Babelsberg (PLZ 14482) im Zeitraum von März bis September 2015?

Zugegeben, dies ist eine sehr spezielle Fragestellung. Sie gibt uns aber die Möglichkeit Ihnen einige der Funktionalitäten von **R** vorzustellen.

Haben Sie bereits **R** oder **R-Studio** installiert? Dann können Sie anhand der folgenden Beschreibung das Beispiel schon rechnen.

Schritt 1: Daten einlesen

Die Wohnungsdaten befinden sich in der Datei  `Potsdammieten.csv` [4.3 MB]. Damit R-Studio die Datei findet, muss sie im Arbeitsverzeichnis von R gespeichert sein. Das Arbeitsverzeichnis wird in R-Studio angezeigt. Sie könnten es mit dem Befehl `setwd` auch ändern. Ist aber jetzt nicht nötig.

Mit der Anweisung in Zeile 001 wird der Datensatz eingelesen und erhält die Bezeichnung **allW** - ist nicht sehr einfallsreich und steht für „*alle Wohnungen*“. Sie können die Namen der Bezeichner auch selber wählen.

Mit der Anweisung in Zeile 002 lassen wir uns die **Dimension** des Datensatzes anzeigen. Die Ausgabe in Zeile 003 zeigt an, dass der Datensatz 19868 Zeilen und 30 Spalten hat - wenn wir uns es als Tabelle vorstellen.

Dimension anzeigen

```
001 > allW<-read.csv2("Potsdammieten.csv",as.is=TRUE)
002 > dim(allW)
003 [1] 19868 30
004
```

Schauen wir uns jetzt die 30 Spalten mit den Merkmalen der Wohnungen an. Dazu verwenden wir den Befehl `colnames` und wenden ihn auf unsere eingelesenen Daten (`allW`) an.

Spaltennamen anzeigen

```
001 > colnames(allW)
002 [1] "lfdNr" "letzter_aktiver_monat"
003 [3] "lon" "lat"
004 [5] "gkz" "IS24_Stadt_Kreis"
005 [7] "IS24_Bezirk_Gemeinde" "plz"
006 [9] "ort" "strasse"
007 [11] "hausnr" "Immobilientyp"
008 [13] "Objektkategorie2" "Objektzustand"
009 [15] "Ausstattungsqualitaet" "mietekalt"
010 [17] "mietewarm" "nebenkosten"
011 [19] "heizkosten" "heizkosten_in_wm_enthalten"
012 [21] "baujahr" "letzte_modernisierung"
013 [23] "wohnflaeche" "zimmeranzahl"
014 [25] "parkplatz" "parkplatzpreis"
015 [27] "Heizungsart" "energieausweistyp"
016 [29] "ev_kennwert" "ev_wwenthalten"
017
```

Die Nummer in der eckigen Klammer an Anfang jeder der Zeile gibt lediglich die Nummer des rechts daneben stehenden Spaltennamen an. Also die Spalte [9] ist "ort". Nun aber zurück zu unserer Fragestellung. Die Spalte 8 enthält die Postleitzahl "plz", und wir **selektieren** nun alle Einträge mit der Postleitzahl von Babelsberg "14482" aus `allW` und nennen diese `allWB`. Das B steht für - Sie ahnen es schon "Babelsberg". Dann lassen wir uns die **Dimension** unserer neuen Auswahl `allWB` anzeigen und sehen in Zeile 004 - es sind nur noch 2580 Zeilen, mit 30 Spalten.

Einfache Anweisung

```
001 > sel <- (allW$plz=="14482")
002 > allWB <- allW[sel,]
003 > dim(allWB)
004 [1] 2580 30
005
```

Mehrere Anweisungen

Nun könnten wir so fortfahren und die Daten immer weiter reduzieren. Es geht etwas einfacher indem wir mehrere Anweisungen kombinieren.

```
001 > sel2015 <- (allWB$zimmeranzahl >=3) &
002 + ((allWB$letzter_aktiver_monat>=201503)&(allWB$letzter_aktiver_monat<=201509) )
003 >
```

In Zeile 001 sprechen wir die Spalte 24 an - dieses Mal aber über den Spaltennamen "zimmeranzahl" und wählen aus unserer Auswahl `allWB` alle Einträge aus die größer oder gleich 3 sind. In Zeile 002 schränken wir zusätzlich noch den Zeitraum aus unserer Fragestellung ein - März bis einschließlich September 2015. Dem Ergebnis geben wir den Bezeichnernamen `sel2015`. Dieses hat 2580 Elemente – eines für jede Zeile von `allWB` – die angeben, ob die

jeweilige Zeile von `allWB` ausgewählt wird oder nicht.

Für die geplante Auswertung brauchen wir nur die lt. „sel2015“ auszuwählenden Zeilen. Wir benötigen auch nicht alle 30 Merkmale, sondern nur die Spalten 1, 2, 16, 23 und 24. Deshalb erstellen wir den neuen Arbeitsdatensatz `allWB2015`, der nur die benötigten Zeilen und Spalten enthält, und lassen uns die verbliebenen fünf Spaltennamen mit `colnames(allWB2015)` anzeigen.

Auswahl einschränken

```
001 > allWB2015 <- allWB[sel2015,c(1,2,16,23,24)]
002 > colnames(allWB2015)
003 [1] "lfdNr" "letzter_aktiver_monat" "mietekalt"
004 [4] "wohnflaeche" "zimmeranzahl"
005 >
```

Laut unserer Fragestellung wollen wir Quadratmeterpreise vergleichen, aber es gibt keine Spalte die diesen Wert enthält. Wir können diesen aber aus den vorhandenen Werten "wohnflaeche" und "mietekalt" errechnen und unseren bisherigen Daten (`allWB2015`) hinzufügen als Spalte mit dem Namen "qmPreis".

Spalte hinzufügen

```
001 > attach(allWB2015) # setzt den Suchpfad auf allWB2015
002 > allWB2015$qmPreis<-mietekalt/wohnflaeche
003 > detach() # rückgängig gemacht
004 >
```

Jetzt sollten wir uns mal ansehen, wie der Datensatz `allWB2015` aussieht und ob der Quadratmeterpreis (`qmPreis`) jetzt verfügbar ist. Dazu verwenden wir den Befehl `print` mit dem Parameter `head`. Dieser bewirkt, dass nur die ersten sechs Zeilen ausgegeben werden. Damit können Sie das Ergebnis prüfen und ggf. anpassen bevor Sie den vollständigen Datensatz anzeigen lassen.

`print(head)` - zeigt die ersten sechs Zeilen.

```
001 > print(head(allWB2015),digits=3)
002      lfdNr letzter_aktiver_monat mietekalt wohnflaeche zimmeranzahl qmPreis
003 3415   1179           201507      389         58           3      6.70
004 5310   2384           201506      457         59           3      7.70
005 8259   1161           201508      560         86           3      6.51
006 8471   1204           201508      568         66           3      8.61
007 8937   1535           201509      590         67           3      8.87
008 9511   2029           201509      610         72           4      8.57
009 >
```

Wenn sie möchten, können Sie sich nun die komplette Liste anzeigen lassen. Dazu lassen sie den Parameter `head` einfach weg.

Liste ausgeben

```
001 > print(allWB2015,digits=3)
```

Wir verzichten hier auf die vollständige Darstellung der Liste aber Sie wissen ja jetzt schon, mit welchem Befehl Sie die Dimension von `allWB2015` anzeigen lassen. Versuchen Sie es!

Als nächstes wollen wir die Anzahl der Wohnungen je Zimmerzahl anzeigen lassen. Dazu geben wir die Spalte "zimmeranzahl" aus `allWB2015` als Tabelle (`table`) aus. Die Striche zwischen den Werten müssen Sie sich denken.

Tabelle ausgeben

```
001 > table(allWB2015$zimmeranzahl)
002
003  3 3.5  4  5  6  7
004 68  3 23  7  4  4
```

Wir sehen in der Tabelle, dass 3- und 4-Zimmerwohnungen am häufigsten angeboten wurden. Kommen wir damit zur letzten Frage: Welche Quadratmeterpreise haben die einzelnen Wohnungen je Zimmeranzahl in unserer Auswahl `allWB2015`? Da wir schon so gut vorgearbeitet haben, erhalten wir die Antwort mit nur einer Zeile in Form eines Boxplots.

Tabelle ausgeben

```
001 > boxplot(qmPreis~zimmeranzahl, data=allWB2015, horizontal=TRUE )
```

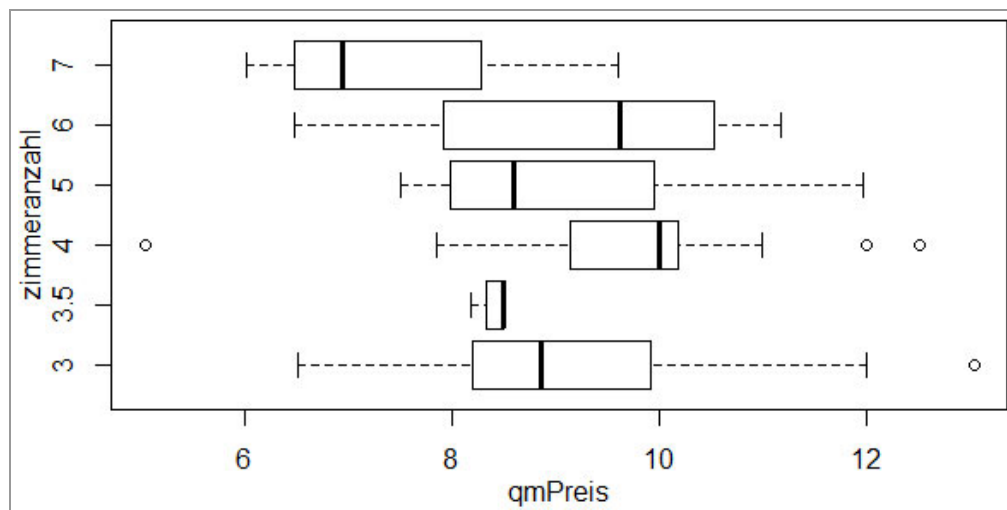


Abb.: Boxplot Mieten
Babelsberg

Die Abbildung zeigt Ihnen die Verteilung der Quadratmeterpreise aufgliedert nach der Zimmeranzahl. Beachten Sie, dass die Boxplots je nach Zimmeranzahl unterschiedlich viele Angebote abbilden - die genaue Anzahl an angegebenen Wohnungen konnten wir der Tabelle oben entnehmen. Die Interpretation der Ergebnisse überlassen wir jetzt Ihnen. Die statistische Zuverlässigkeit der Ergebnisse bei den Wohnungen mit vielen Zimmer ist eher gering, weil es eben wenige davon gibt.



Hinweis

Boxplots werden Sie in den folgenden Lerneinheiten noch kennen und schätzen lernen. Das Prinzip ist ganz einfach: Innerhalb der Boxen liegen 50 % der Werte, der Strich in der Mitte trennt die Daten in die Hälfte der eher kleineren und der eher größeren Werte. Außerhalb liegen jeweils 25 %. Extreme sind mit einem Kreis (o) gekennzeichnet.

Damit haben wir nun unsere Fragestellung beantworten können und wir hoffen Sie konnten das Beispiel ohne viel Mühe mit **R** nachrechnen. Zu Beginn ist sicherlich noch einiges ungewohnt aber wir hoffen, dass es Ihnen Spaß macht.

7.3 Wissen und Vorurteile - Gapminder

Was haben wir für Vorurteile? Diese beruhen oft auf Wissenslücken lassen sich zum Glück mit statistischen Daten korrigieren. Die Plattform **Gapminder** gibt mit vielen Beispielen Anreize, um über die Statistik spielerisch Antworten auf Fragen unserer Zeit zu finden.

Die Plattform ist nur auf Englisch verfügbar aber die Inhalte sind so beeindruckend, dass es sich auf jeden Fall lohnt, die Englischkenntnisse dabei zu trainieren. Stöbern Sie durch die Webseiten, wählen Sie aus einer Vielzahl an Kategorien und schätzen Sie bei den Antworten auf die gestellten Fragen. Sie werden überrascht sein und im besten Fall viel Motivation finden, sich mit Statistik zu beschäftigen.

<https://www.gapminder.org>

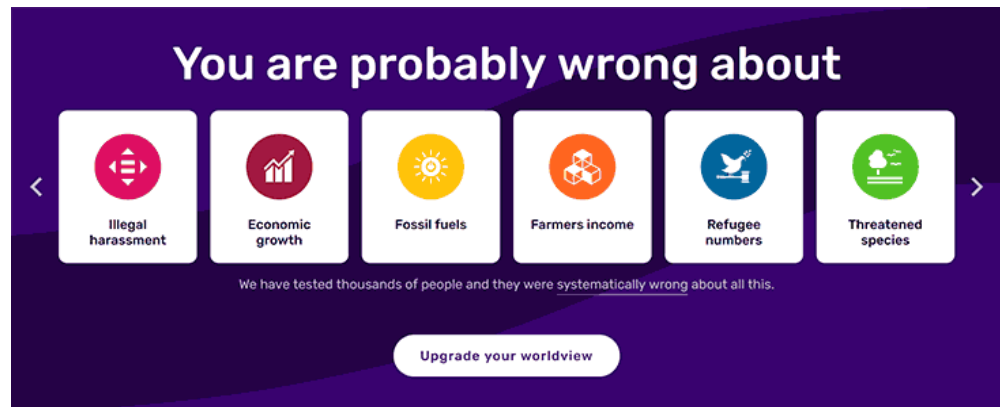


Abb.: Screenshot
Gapminder.org

8 Hinweise zum Bearbeiten der Lerneinheiten

Statistik lernt man durch Anwenden. Eigentlich sind die erforderlichen Mathematikkenntnisse minimal, und dennoch behaupten viele, für Statistik und Mathematik völlig unbegabt zu sein.

Wir glauben das nicht und stellen in unserem Lernmodul zu Wirtschaftsmathematik und Statistik viele Hilfen und Anreize zur Verfügung, die Abneigung gegen formales und mathematisches Argumentieren zu überwinden.

Das Lernmaterial führt Sie in der Regel über einführende Beispiele zu Definitionen und Hinweisen. Übungen mit ausführlichen Lösungshinweisen ermöglichen es Ihnen, den Stoff Schritt für Schritt nachzuvollziehen. Viele der Aufgaben sind recht einfach, Fragen werden auch wiederholt gestellt. Sie sollen Ihnen Sicherheit geben.

Als Ergänzung steht Ihnen die Software **R** zur Verfügung, wo Sie Ergebnisse ohne mühsame Handrechnungen erhalten können. Sie haben jederzeit die Möglichkeit, eigene Beispiele und Datensätze einzubinden.

Das Modul ist in wöchentlich zu bearbeitende Lerneinheiten strukturiert. Die Benennung der Lernziele am Anfang jeder Lerneinheit sowie die Zusammenfassung am Ende ermöglichen ein effizientes individuelles Lernen über einen Zeitraum von drei bis vier Monaten. Die interaktive Wissensüberprüfung am Ende jeder Lerneinheit gibt Ihnen ein schnelles Feedback über Ihren Wissensstand. Die Übungen und die Aufgaben der Wissensprüfung liegen auf dem Niveau der Klausur, die am Ende des Kurses geschrieben wird.

Die Kommunikation zwischen den Studierenden und den Mentoren erfolgt über die Foren im Lernmanagementsystem, per E-Mail oder anderen Verbindungswegen wie bspw. Webkonferenzen. Während der Präsenzphasen an der eingeschriebenen Hochschule steht die gemeinsame Bearbeitung von Übungsaufgaben im Team im Mittelpunkt.

9 Didaktisches Konzept

Mit Hilfe der folgenden vierzehn Lerneinheiten können Sie sich ein solides Grundwissen über Begriffe und Methoden der beschreibenden Statistik aneignen.

Damit Sie sich gut zurechtfinden, folgen hier einige der Regeln, nach denen Form und Inhalte konzipiert wurden, sowie Tipps, wie Sie unser Material am besten nutzen können.

Die Anschaulichkeit und das Vermitteln eigener Erfahrungen mit der Anwendung der Begriffe und Methoden waren unser wichtigstes Anliegen.

Außerdem wurde die Statistiksoftware **R** integriert, damit Sie alles Gelernte auf reale Beispiele aus der Praxis anwenden können.

Alle Lerneinheiten haben eine einheitliche innere Gliederung. Überprüfen Sie ruhig im Lauf des Kurses, ob Sie diese jeweils wiedererkennen.

Wozu?



Audio

In der einleitenden Motivation wird erläutert, wozu die Inhalte benötigt werden, wie sie sich von Bekanntem absetzen etc. Teilweise werden Ihnen die Texte von Sprecherinnen angeboten. Wir zeigen Ihnen damit, dass man Statistik auch aussprechen kann.

Was ist zu tun?



Beispiel

Soweit möglich, werden die neuen Sachverhalte an einem einfachen Beispiel erläutert, auch wenn das formale Vorgehen noch nicht bekannt ist. Versuchen Sie das Prinzip aus dem Beispiel abzuleiten.

Wie geht es genau?



Definition

Definitionen, Formeln, eine exakte Darstellung halten die im Beispiel gemachten Erkenntnisse fest.

Was ist zu beachten?



Hinweis

In Anmerkungen und Hinweisen werden Besonderheiten und bestimmte mathematische Eigenschaften im Zusammenhang mit den neu definierten Begriffen oder Verfahren erklärt.

Dezimalpunkt statt Dezimalkomma

Im Studienmodul WMS wird für Dezimalzahlen der Punkt und nicht das Komma als Trennzeichen verwendet. Dies vereinfacht für Sie die Nutzung der Statistiksoftware **R** und gilt zudem als vorherrschende internationale Konvention.

Eine Ausnahme bilden die Dateien für das Programm Excel. Ob hier Dezimalzahlen mit Punkt oder Komma angezeigt werden, hängt von den Ländereinstellungen Ihres Computers ab. In der Regel wird es hier wohl das Komma sein.

Das ist es!



Zusammenfassung

Wissensüberprüfung und Zusammenfassung geben Ihnen die Möglichkeit Ihr objektives und gefühltes Verständnis zu kontrollieren.

Was bieten wir Ihnen als Hilfen an?



Multiple Choice

Salopp formuliert: Training und Unterhaltung.

Training?



Berechnen



Statistiksoftware R

Wenn Sie alle Beispielrechnungen und möglichst viele der Übungen sorgfältig durcharbeiten, werden Ihnen die zunächst ungewohnten Dinge in Fleisch und Blut übergehen.

Mit der Statistiksoftware **R** können sie vorbereitete Übungen oder auch eigene Fragestellungen bearbeiten, statistische Maßzahlen ermitteln oder graphische Darstellungen erzeugen. Sie werden sehen, dass Sie sich mehr und mehr auf die Ergebnisse konzentrieren können, die Berechnungen erledigen sich fast wie von selbst.

Für einige der Übungen stehen auch Musterlösungen bereit, die mit der Software Microsoft Excel erstellt wurden. Bitte haben Sie Verständnis dafür, dass wir keinen Support für die Programme **R** und Excel bieten können.

Unterhaltung?



Rolloverbild



Animation

Die einzelnen Lerneinheiten sind reich an Grafiken, Animationen und kleinen Illustrationen. Diese sollen Ihnen die Möglichkeit zum assoziativen Begreifen geben.

Wir haben uns damit viel Mühe gegeben und hoffen, dass Sie sich ausreichend Zeit nehmen, Ihre Kreativität immer wieder anregen zu lassen.

Zusammenfassung

- ✅ Datenquellen und Datensammlungen bilden die Grundlage für die Datenanalyse.
- ✅ Einfache statistische Fragen können durch Bestimmung von Häufigkeiten beantwortet werden.
- ✅ Durch Sortieren von Werten werden Ordnung und Überblick gewonnen.
- ✅ Das Programm **R** bietet Übungsmöglichkeiten, die Sie nutzen sollten.
- ✅ Sie haben sich mit unseren Vorstellungen zum Bearbeiten der Lerneinheiten vertraut gemacht.
- ✅ Sie kennen das didaktische Konzept unserer Lerneinheiten und wissen, welche Bausteine und Hilfen Ihnen zur Verfügung stehen.
- ✅ Die erfolgreiche Installation von **R** haben Sie durch die Übung „Angler“ überprüft.

Wenn Sie alle Aussagen abhaken können, steht einem Start in die Lerneinheit „*GST - Grundbegriffe der Statistik*“ nichts mehr im Wege.

Wir wünschen Ihnen viel Erfolg, Neugier und Spaß!
