Statistik in der Testanalyse und Implementierung mit *Mathematica*

Bachelorarbeit
bei Prof. Dr. Wolfram Koepf

vorgelegt von Katrin Heumann
am Fachbereich Mathematik / Informatik der
Universität Kassel
UNIKASSEL
VERSITÄT

Die Implementierung wurde mit Unterstützung des
Deutschen Zentrums für Luft- und Raumfahrt
speziell für die Wissenstests im Astronautentraining
der Europäischen Raumfahrt Agentur erstellt.
Betreuer vor Ort: Dr. Rüdiger Seine

Köln, April 2006

Diese Bachelorarbeit ist durch die Kooperationsbereitschaft des Prüfungsausschusses *Computational Mathematics* der Universität Kassel, Prof. Dr. Wolfram Koepf, und dem Leiter der Astronautentrainingseinheit des Deutschen Zentrums für Luft- und Raumfahrt, Klaus Wasserberg, entstanden. Mein besonderer Dank gilt diesen beiden Herren.

Der Leiter des *COLUMBUS*-Trainings der Europäische Raumfahrt Agentur, Dr. Rüdiger Seine, gab viele Anregungen zu dieser Bachelorarbeit, vielen Dank!

Inhaltsverzeichnis

0 Einleitung

Die Testanalyse ist ein interessantes, aber im Allgemeinen wenig bekanntes Gebiet. Hier geht es um die Frage, wie weit man einen Test[1] in Bezug auf seine Qualifikation für den individuellen Einsatz analysieren kann. Desweiteren ist es Ziel der Testanalyse, bisher verwendete Tests so zu verändern, dass sie den an sie gestellten Anforderungen genügen, eine Art individuelle Anpassung des Tests auf seinen spezifischen Prüfzweck hin. Schließlich geht es in der Testanalyse auch darum, *Testergebnisse* und *Testanalyseergenisse* auf ihre jeweilige Aussage hin zu interpretieren.

Testergebnisse wurden schon seit geraumer Zeit interpretiert. Jedoch hängt es stark von der Haltung des Prüfers ab, ob schlechte Testergebnisse nur auf Faulheit und Unwissen der Prüflinge geschoben werden, oder ob er bereit ist, den eigenen Test selbstkritisch zu überprüfen. Es gilt für den Tester zu hinterfragen, ob der Test in dieser Form ein faires und geeignetes Mittel ist. Eine Möglichkeit der Objektivierung von Tests wurde in der Vergangenheit häufig im Einsatz eines zweiten Prüfers mit entsprechender Kontrollfunktion gesehen. Hierbei hängt das Ergebnis jedoch weiterhin von einem Menschen und seiner Subjektivität ab. Dieses grundsätzlich menschliche Phänomen kann durch einen zweiten Prüfer nur relativiert, niemals jedoch ausgeschaltet werden. Es gilt Kriterien zu schaffen die von Menschen unabhängig sind und an Hand derer Tests auf ihre „Güte“ untersucht und bewertet werden können.

Testanalyseergebnisse sind normierte Messwerte über einen Test. Während Testergebnisse widerspiegeln, welcher Proband besonders gut abgeschnitten hat und wieviele den Test insgesamt bestanden oder nicht bestanden haben, geben die Testanalyseergebnisse einen tieferen Einblick in den Test an sich. Historisch gewachsen, in Zusammenarbeit vieler Psychologen und Statistiker, hat man heute einen umfangreichen Katalog an mathematischen Prüfverfahren für einen Test. Das gesamte Forschungsgebiet der Testanalyse ist in mehrere Teilgebiete untergliedert. Zum Beispiel sind die Bereiche *Testdiagnostik* und *Testkonstruktion* eine amerikanische Domäne. Die deutschen Forscher RIEGER, KRAEPELIN und Mitarbeiter des WUNDT'schen Instituts haben wiederum viele neue Wege in der *Testmethodik* gefunden[2]. Diese entstandenen Prüfverfahren beziehen sich auf die sogenannten Gütekriterien des Tests, sie spiegeln seine Qualität wieder. Die Objektivität dieser Verfahren liegt in einer einheitlichen Anwendung, die für fast jede Testform möglich ist, und in einer normierten Ausgabe. Das bedeutet, ein mathematisches Messverfahren mit dem Namen „Schwierigkeitsindex einer Aufgabe“[3] liefert als normierte Ausgabe für jede einzelne Testaufgabe einen Wert zwischen 0 und 1, unabhängig vom Ort, Personen oder Zeitpunkt.

Ein einzelnes dieser Prüfverfahren verfügt nicht über genügend Aussagekraft zur Qualität des Tests. Erst das Zusammenspiel mehrerer Prüfverfahren führt zu einer ganz-

[1] Die Definition von „Test“ ist in Kapitel 2.1 auf S. 32 nachzulesen.

[2] Vertiefende und geschichtliche Informationen dazu sind in [LiRa] zu finden.

[3] siehe Formel (2.16) auf Seite 40

heitlichen Betrachtung des Tests und damit zu einem optimalen Analyseergebnis. Die Erkenntnisse der Testanalyse werden zum Beispiel bei der Entwicklung und Weiterentwicklung komplexer Testverfahren in Wissenschaft und Forschung eingesetzt, da es hier auf aussagekräftige, genaue Ergebnisse ankommt. Durch unzureichende Tests verfälschte oder nicht genügend differenzierte Ergebnisse kann es zu weitreichenden Schwierigkeiten kommen. So widmet sich diese Bachelorarbeit der Entwicklung eines zur Anwendung bestimmten Programms bezüglich der Testanalyse. Die Anwendung des Programms bezieht sich auf einen Bereich, der in höchstem Maße auf Genauigkeit und Zuverlässigkeit angewiesen ist – die Ausbildung von Astronauten für ihren Einsatz im Weltall.

Die Bachelorarbeit ist durch eine Zusammenarbeit der Universität Kassel und dem Deutschen Zentrum für Luft- und Raumfahrt (DLR) entstanden. Das DLR arbeitet im Bereich der bemannten Raumfahrt eng mit der Europäischen Raumfahrtagentur (ESA) zusammen. Das Ziel der Kooperation mit der Uni Kassel ist die Entwicklung eines Programms zur Testanalyse aus dem Bereich des Astronautentrainings. Der DLR- und ESA-Standort Köln-Porz ist die Trainingsstätte des europäischen Astronautencorps, genannt EAC (European Astronaut Centre). In Köln werden die Astronauten auf ihre Missionen vorbereitet, wie zum Beispiel die aktuell anstehende Langzeitmission des Deutschen ESA-Astronauten, Thomas Reiter, zur Internationalen Raumstation, ISS. Training bedeutet für einen Astronauten zum einen das Training am Simulator für ganze Raumeinheiten wie das europäische Raumlabor *COLUMBUS*, welches im nächsten Jahr der ISS angegliedert werden soll. Zum anderen geht es im Astronautentraining auch um ganz spezielle Experimente, die der Astronaut in der Schwerelosigkeit für verschiedene Auftraggeber durchführen soll. Einzelne Prozeduren und manuelle Abläufe müssen gut vorbeitet sein, um unter den veränderten Bedingungen des Weltraums reibunglos abzulaufen. Ein dritter Punkt ist das körperliche Training des Astronauten.

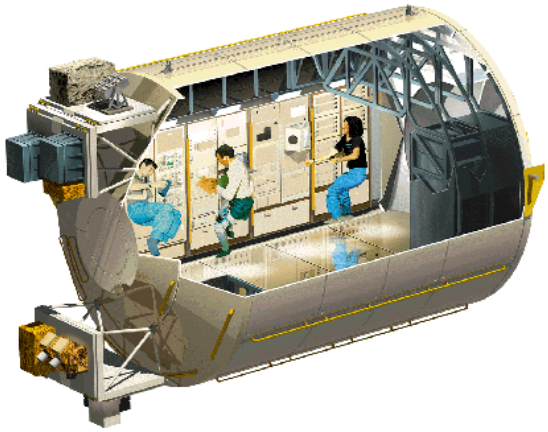


Abbildung 1: Das europäische Raumlabor *COLUMBUS*

Das Raumlabor *COLUMBUS* besteht aus verschiedenen Systemen. Darunter zum Bei-

spiel das *Thermal Control System* (TCS) oder *Data Management System* (DMS). In Abständen finden im EAC immer wieder Trainingswochen statt, während derer die Teilnehmergruppe – bestehend aus Astronauten, Bodencrew, etc. – im Gruppenunterricht auf die einzelnen Systeme geschult wird. Der grundlegende Kurs heißt *COLUMBUS-User-Level-Training*, später gibt es als Aufbau auch das *Operator-* und *Specialist-Training*. Am Ende jeder Trainingswoche wird das erlernte Wissen der Teilnehmer in einem Wissenstest überprüft. Der Unterrichtsstoff wird für diesen Test in sechs inhaltliche Bereiche gegliedert. Der Test wird in Printform ausgegeben und hat den Aufbau eines Multiple-Choice-Tests[4]. Fragen mit mehreren Antwortmöglichkeiten werden angekreuzt, es können eine bis mehrere Antworten richtig sein, bei einzelnen Fragen sind Wörter einzufüllen oder Bereiche in Graphiken anzustreichen. Ziel der Trainingswoche ist, dass jeder Teilnehmer möglichst viel verstehen und lernen soll. Der abschließende Test wird nicht mit Noten bewertet, sondern nur nach bestanden und nichtbestanden kategorisiert. Anschließend wird mit der Gruppe jede Testfrage besprochen. Da es für einen erfolgreich absolvierten Test ein Zertifikat gibt, ist es für die Teilnehmer wichtig, den Test zu bestehen. Für Teilnehmer, die keinen Erfolg hatten, wird eine Wiederholungsmöglichkeit angeboten.

Ziel dieser Bachelorarbeit ist die Entwicklung einer technisch „automatisierten“ Testanalyse für die sogenannten *COLUMBUS*-Tests. Der Leiter der *COLUMBUS*-Schulung, Dr. Rüdiger Seine, gab den Anstoß ein „*tool*“[5] zu entwickeln, welches die Testanalyse auf die speziellen Bedingungen seines Tests abgestimmt zuverlässig durchführt. Die Handhabung ist individuell auf die Gebrauchswünsche des Leiters ausgelegt. Als Eingabe dienen Tabellen mit Werten über das Abschneiden eines jeden Probanden pro Testaufgabe. Dabei sind diese Daten so anonymisiert, dass ein Rückschluss auf das Abschneiden einzelner Personen vermieden wird. Die Testanalyse zielt auf die Betrachtung der Qualität der Testfragen ab und nicht auf eine Betrachtung des Abschneidens der Probanden. Die Ausgabe liefert Hinweise auf einzelne Testfragen, die nach Anwendung verschiedener Prüfverfahren als verbesserungsbedürftig ausgewiesen werden. Dr. Seines Aufgabe ist es im Anschluss, die Ursachen für das ungenügende Abschneiden dieser Fragen herauszufinden. Als mögliche Ursachen kommen zum Beispiel Schulungsdefizite oder unklare Aufgabenstellungen in Betracht. Das entstandene Testanalyse-Werkzeug wird seit April 2006 für die Arbeit im EAC eingesetzt.

Nach dem in Kapitel 1 zunächst die relevanten Definitionen und Formeln der Statistik dargestellt werden, wird in Kapitel 2 auf die Methoden der Testanalyse genauer eingegangen. Hierbei ist der Fokus auf die Verfahren gerichtet, deren Voraussetzungen von den ESA-Tests erfüllt werden. Eine umfangreiche Darstellung kann im Rahmen dieser Arbeit aus Platzgründen nicht geleistet werden. Hierzu sei auf die entsprechende Literatur im Index, insbesondere [LiRa], verwiesen. In Kapitel 3 dieser Arbeit werden Details der Implementierung vorgestellt. Der gesamte Algorithmus ist mit dem Com-

[4] mehr dazu auf S.11

[5] [Wiki, tool] Ein Dienstprogramm, oder auch Hilfsprogramm (engl. *utility* oder *tool*), das für den Benutzer allgemeine oder spezielle Aufgaben ausführt.

puteralgebrasystem **`Mathematica`** umgesetzt worden, der komplette Quellcode und ein Beispieldatensatz befindet sich im Anhang und auf der beigelegten CD. Im abschließenden 4. Kapitel werden mögliche Verbesserungen des Testanlyse*tools* diskutiert. Es wird reflektiert, in wie weit das Programm zuverlässig und aussagekräftig arbeitet und ob die formulierte Zielsetzung erreicht wurde.

1 Grundlagen der Statistik

In diesem Kapitel werden wir uns mit den grundlegenden Formeln der Statistik beschäftigen. Dabei werden Begriffe und Formeln mit der mathematischen Notation einer „Definition“ eingeführt. Wir gehen nur auf die Formeln ein, die für die Implemetierung benötigt wurden und zum Verständnis beitragen. Begleitend zeigen wir die Anwendung der Formeln mit unserem Beispieldatensatz **Tabelle**.

```
In[1]:= Tabelle := Import["D : Beispiel.txt", "Table"];
        MatrixForm[Tabelle]
```

Out[1]=

per	i01	i02	i03	i04	i05	i06	i07	i08	i09	i10	i11	i12	i13	i14
p01	1	a	0	1	1	1	1	1	1	1	1	1	1	1
p02	1	a	1	1	1	1	1	0	1	1	1	1	1	1
p03	1	a	1	1	1	1	1	1	1	1	1	1	1	0
p04	1	a	1	1	1	1	1	1	0	1	1	1	1	0
p05	1	a	1	1	1	1	1	1	1	1	1	0	1	0
p06	1	0	0	1	1	1	1	1	1	0	1	0	1	1
p07	1	1	1	1	1	1	1	1	1	0	1	0	1	0
p08	1	1	a	1	1	1	1	1	1	0	1	0	1	0
p09	1	1	a	0	1	0	1	0	1	0	0	0	0	0

Diese **Tabelle** ist der Beipieldatensatz, den wir immer wieder betrachten werden. In der ersten Spalte stehen individuelle Personenkürzel. Man sieht, dass 9 Probanden aufgeführt sind. Die erste Zeile beinhaltet eine eindeutige Kennzeichnung der *Testitems*.

item Das englische Wort *item* beschreibt in der Testanalyse ein Untersuchungskriterium. Für unseren speziellen Fall kommt das Wort *item* dem Begriff *Antwortmöglichkeit* gleich. Wir gehen von Multiple-Choice-Fragebögen[6] aus, bei denen jede Frage mehrere Antwortmöglichkeiten besitzt. Jede dieser möglichen Antworten betrachten wir in der Testanalyse einzeln. Eine Frage mit vier Antwortmöglichkeiten teilen wir also für die Beispieltabelle in vier *items* auf, wir vergeben vier Spalten und jeder Spalte einen eindeutigen Namen.

Die eingetragenen Werte haben folgende Bedeutung:

- Der Wert „1“ steht für ein richtig beantwortetes *item*.
- Der Wert „0“ steht für ein falsch beantwortetes *item*.
- Der Wert „a“ steht für ein nicht beantwortetes *item*. Das ist der Fall, wenn einer Person eine bestimmt Frage erst gar nicht vorgelegt wurde. Wir alle kennen das Schema von Professoren, eine Klausur in Gruppe A und Gruppe B einzuteilen. Je nach Sitzordnung im Raum werden abwechselnd Aufgabenblätter der Gruppe A und B ausgegeben um ein Abschreiben zwischen den Studenten zu verhindern.

[6] [Wiki, Multiple Choice] Das Multiple-Choice-Verfahren (kurz: *MC*) (deutsch: *mehrfache Auswahl* im Sinne von mehreren Möglichkeiten, die zur Auswahl stehen) wird bei Tests bzw. Prüfungen verwendet. Hierbei werden zu einer Frage verschiedene Antwortmöglichkeiten vorgegeben, aus denen der Prüfling eine oder mehrere auswählen muss, die er für richtig hält.

Zu dem Wert „a“ sei außerdem folgendes gesagt: Wir müssen am Anfang der Testanalyse entscheiden, ob solche *items* mit in die Betrachtung kommen, die nicht von allen Personen beantwortet wurden. Es kann ja der Fall sein, wie in unserem Beipiel, dass Person **p01** bis **p05** das *item* **i02** nicht in ihrem Test hatten und die Personen **p08** und **p09** das *item* **i03** nicht in ihrem Test hatten. Nun wäre ein Vergleich aller Personen falsch, da die Personen **p06** und **p07** mehr Punkte erreichen konnten als die anderen Personen. Also werden wir alle Spalten entfernen, in denen mindestens ein „a“ vorkommt[7].

```
In[2]:= Tabelle
```

Out[2]=

per	i01	i04	i05	i06	i07	i08	i09	i10	i11	i12	i13	i14
p01	1	1	1	1	1	1	1	1	1	1	1	1
p02	1	1	1	1	1	0	1	1	1	1	1	1
p03	1	1	1	1	1	1	1	1	1	1	1	0
p04	1	1	1	1	1	1	0	1	1	1	1	0
p05	1	1	1	1	1	1	1	1	1	0	1	0
p06	1	1	1	1	1	1	1	0	1	0	1	1
p07	1	1	1	1	1	1	1	0	1	0	1	0
p08	1	1	1	1	1	1	1	0	1	0	1	0
p09	1	0	1	0	1	0	1	0	0	0	0	0

Zum allgemeinen Verständnis legen wir nun ein paar Begriffe fest: *N* steht für die Anzahl der Probanden, *n* für die Anzahl der *items*. Mit *Mathematica* bestimmen wir die Zeilen- und Spaltenanzahl und ziehen jeweils eins davon ab, wegen der Zelle links oben in **Tabelle**.

```
In[3]:= N = First[Dimensions[Tabelle]] - 1;
Out[3]= 9

In[4]:= n = Last[Dimensions[Tabelle]] - 1;
Out[4]= 12
```

Um einen Begriff für die individuell erreichte Punktezahl der Probanden zu haben, führen wir die folgende Definition ein. Diese Definition benötigen wir nur, um in der Testanalyse einen „Namen“ für die Gesamtpunktzahl zu haben. Die tatsächliche Gesamtpunktzahl, die die Probanden im Test erreicht haben, wird anders berechnet[8]. Der Begriff des *Rohwertes* taucht **nur** in der Theorie auf, dieser *Rohwert* entspricht aber **nicht** der tatsächlich erreichten Punktzahl des Probanden.

Definition 1 (Rohwert) *Die erreichte Gesamtpunktzahl unserer Probanden wird Rohwert genannt. Der Rohwert X_i einer Person i wird folgendermaßen bestimmt:*

$$X_i \quad = \quad \textit{Anzahl der richtigen items von Proband i.}$$

[7] Details zu dieser Reduktion sind in Kapitel 3.1.2 aufgeführt.
[8] siehe dazu Seite 39

Wir legen uns hier auf folgende Unterschiede in der Notation fest:

$$\begin{aligned} X_i &= \textit{Rohwert von Person } i \\ X_{(i)} &= \textit{i-ter Rohwert in einer ordinalskalierten Menge} \\ &\quad \textit{(aufsteigende Reihenfolge der Werte).} \end{aligned}$$

Wir arbeiten generell mit *absoluten* Häufigkeiten wie hier in der Definition für den Rohwert gezeigt. Daneben gibt es genauso die *relative* Häufigkeit. Relative Häufigkeiten werden meist als Prozent-Anteile ausgedrückt. So bedeuten 10 von 12 möglichen Punkten in der relativen Schreibweise der Rohwertermittlung gerade $\frac{10}{12} = 83.\overline{33}\%$. Wir arbeiten aber mit dem Absolutwert und dieser ist gleich 10.

1.1 Lagemaße

Einen ersten Überblick über eine Datenmenge erhalten wir mit *Lagemaßen*. Diese geben Auskunft über das Maß der zentralen Tendenz, den „Schwerpunkt“. Die Frage ist, in wie fern *Lagemaße* Eigenschaften der Daten herausstellen können. Wir betrachten zunächst einige grundlegende Definitionen.

Definition 2 (Arithmetisches Mittel) *Das arithmetische Mittel, auch als Mittelwert oder Durchschnitt bekannt, errechnet sich aus den Summen der Rohwerte und der Anzahl der Probanden:*

$$\overline{X} := \frac{\sum_{i=1}^{N} X_i}{N}. \tag{1.1}$$

Das arithmetische Mittel ist nicht zu verwechseln mit dem geometrischen oder harmonischen Mittel[9].

Definition 3 (Median) *Der Median ist der Wert, der eine Menge Merkmalswerte in zwei gleich große Teile teilt. Voraussetzung für die Bestimmung des Medians ist eine ordinalskalierte Wertemenge:*

$$X_{1/2} := \begin{cases} X_{(\frac{N+1}{2})}, \textit{ falls N eine ungerade Zahl ist,} \\ \\ \frac{X_{(\frac{N}{2})} + X_{(\frac{N}{2}+1)}}{2}, \textit{ falls N eine gerade Zahl ist.} \end{cases} \tag{1.2}$$

Durch diese Festlegung wird erreicht, dass mindestens 50% aller X_p kleiner oder gleich $X_{1/2}$ und mindestens 50% aller X_p größer oder gleich $X_{1/2}$ sind.

[9] [Henze, S.38]

Definition 4 (p-Quantil) *Als Verallgemeinerung des Medians heißt für eine Zahl p mit* $0 < p < 1$

$$X_p := \begin{cases} X_{([N\cdot p+1])}, \textit{ falls } N \cdot p \notin \mathbb{N}, \\ \\ \frac{1}{2} \cdot (X_{(N\cdot p)} + X_{(N\cdot p+1)}), \textit{ falls } N \cdot p \in \mathbb{N}, \end{cases} \tag{1.3}$$

das p-Quantil von $X_{(1)}, \ldots, X_{(N)}$.
Dabei bezeichnet allgemein der Ausdruck $[y] := max\{k \in \mathbb{Z} : k \leq y\}$ *die größte ganze Zahl, welche kleiner oder gleich einer reellen Zahl y ist,*
also z. B. $[1, 2] = 1, [-0, 3] = -1, [5] = 5$.

Die obige Festlegung bewirkt, dass mindestens $p \cdot 100\%$ aller Werte kleiner oder gleich X_p und mindestens $(1 - p) \cdot 100\%$ aller Werte größer oder gleich X_p sind. Das p-Quantil X_p teilt also unsere Werte „im Verhältnis p zu $(1 - p)$“ auf.

Den Median nennt man deshalb auch das $0,5$-Quantil. Daneben gibt es noch einige Namen für häufig verwendete Quantile. $X_{0,25}$ und $X_{0,75}$ nennt man das *untere* bzw. *obere Quartil* und $X_{j\cdot 0,1}$ das j-te *Dezil* $(j = 1, \ldots, 9)$[10].

Zu den hier genannten Mittelwerten und dem Median gibt es noch den sogenannten **Modus**, oder Modalwert. Das ist der Wert, der in der gegebenen Menge von Merkmalswerten am häufigsten auftritt. Wenn wir die Situation haben, dass mehrere Werte gleich häufig vorkommen, unterscheiden wir zwei Fälle:

- Befindet sich zwischen zwei gleich häufig auftretenden Werten mindestens ein weiterer, seltener vorkommender Wert, so sprechen wir von einer *bimodalen* Verteilung[11].
- Ist jedoch zwischen den beiden Werten mit der höchsten Frequenz kein weiterer Wert, so handelt es sich um eine Verteilung mit nur einem Modalwert, die allerdings *breitgipflig* ist. Der Modus entspricht hier der Grenze bzw. Mitte zwischen diesen beiden Werten.

Betrachtet man alle drei *Lagemaße* zusammen, ergeben sich verschiedene Möglichkeiten der Verteilung. Wir sehen uns exemplarisch die *symmetrische und asymmetrische Verteilung* an:

[10] [Henze, S.32]

[11] Neben *bimodalen* Verteilungen gibt es ebenso *unimodale* Verteilungen. Hierbei handelt es sich um eine Art Verteilung, bei der eindeutig nur **ein** Wert mit höchster Häufigkeit auftritt.

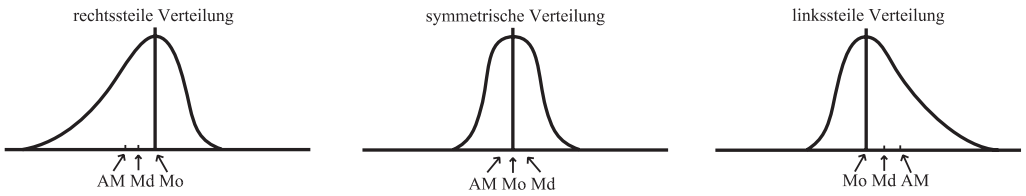


Abbildung 2: Die unterschiedlichen Verteilungen

Nun interessieren wir uns für die *Lagemaße* und Verteilung unserer Beispielrohwerte, die bereits ordinalskaliert vorliegen:

```
In[5]:= X = {4, 9, 9, 10, 10, 10, 11, 11, 12};
```

Wir berechnen Arithmetisches Mittel (AM), Median (Md) und Modus (Mo) und zum Vergleich betrachten wir auch die **Mathematica** Befehle:

```
In[6]:= AM = Sum[X[[i]], {i, 1, N}] / N
Out[6]= 9.55556

In[7]:= Mean[X]
Out[7]= 9.55556

In[8]:= Md = X[[(N + 1) / 2]]
Out[8]= 10

In[9]:= Median[X]
Out[9]= 10

In[10]:= Mo = Mode[X]
Out[10]= 10
```

Nun betrachten wir das *untere Quartil*:

```
In[11]:= [N * 0.25 + 1]
Out[11]= [3.25]
```

Da $3.25 \notin \mathbb{N} \Rightarrow$ das $0,25$-Quantil ist also das dritte Element aus X ($[3,25] = 3$):

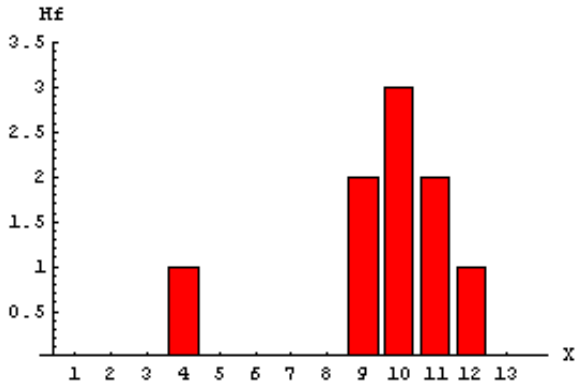
```
In[12]:= X[[IntegerPart[3.25]]]
Out[12]= 9
```

Der entsprechende **Mathematica** Befehl lautet:

```
In[13]:= Quantile[X, 0.25]
Out[13]= 9
```

Zur Übersicht der Rohwertverteilung nehmen wir ein Balkendiagramm. Bei diesem repräsentiert die X-Achse die auftretenden Rohwerte, auf der Y-Achse sind die entsprechenden Häufigkeiten abzulesen:

```
In[14]:= Needs["Statistics`DataManipulation`"];
         Needs["Graphics`Graphics`"];
         BarChart[Frequencies[X], AxesLabel->"X", Hf];
```



Out[14]=

Für unser Beispiel stellt sich heraus, dass Median und Modus gleich zehn sind, der Mittelwert liegt etwas niedriger. Wir haben keinen großen Unterschied zwischen den drei *Lagemaßen*. Bei der Verteilung sieht man allerdings den Rohwert 4 als *Ausreißer*. Er „verzieht“ das Bild der Verteilung zu einer rechtssteilen Verteilung. Typisch dafür ist der Wert des Arithmetischen Mittels, welcher den kleinsten der drei Lagewerte darstellt.

Bezogen auf die eingangs gestellte Frage, in wie weit wir anhand von *Lagemaßen* etwas über die Eigenschaften einer Datenmenge sagen können, lässt sich folgendes festhalten: Sollten wir eine *bimodale* Verteilung vorfinden, so ist die Berechnung des Mittelwertes evtl. wenig sinnvoll. Liegen Arithmetisches Mittel und Median bzw. Modus sehr weit auseinander, weist das auf einen *Außreißer* hin. Generell bekommen wir durch die Betrachtung der *Lagemaße* nicht genügend Informationen über die gesamte Datenmenge, betrachten wir also ein weitere Maße.

1.2 Streuungsmaße

Die Kenntnis allein über Lagemaße, wie zum Beispiel den Mittelwert, hat nicht viel Aussagekraft. Wir bekommen keine Information über die *Streuung* der Werte. Die Beispielwerte 9, 10, 11 und 0, 10, 20 haben das gleiche arithmetische Mittel, die Streuung ist jedoch im zweiten Fall viel größer.

Als beispielhafte Anekdote dient ein Gedicht von Professor Dr. P.H. List[12]

Ein Mensch, der von Statistik hört,
denkt dabei nur an Mittelwert.
Er glaubt nicht dran und ist dagegen,
ein Beispiel soll es gleich belegen:

[12][Henze, S.33], vgl. Krafft O. (1977): Statistische Experimente: Ihre Planung und Analyse. Zeitschrift für angewandte Mathematik und Mechanik 57, T17-T23.

Ein Jäger auf der Entenjagd,
hat einen ersten Schuss gewagt.
Der Schuss, zu hastig aus dem Rohr,
lag eine gute Handbreit vor.

Der zweite Schuss mit lautem Krach
lag eine gute Handbreit nach.
Der Jäger spricht ganz unbeschwert
voll Glauben an den Mittelwert:
Statistisch ist die Ente tot.

Doch wär' er klug und nähme Schrot
- dies sei gesagt, ihn zu bekehren -
er würde seine Chance mehren:
Der Schuss geht ab, die Ente stürzt,
weil Streuung ihr das Leben kürzt.

Betrachten wir also die wichtigsten Formeln zur Streuung. Das heißt, wir suchen ein Maß für die „Abweichung vom Mittel" der gegebenen Wertemenge.

Definition 5 (Varianz) *Das klassische Streuungsmaß ist die Varianz σ^2 von $X_1, \ldots, X_N$:*

$$\sigma^2 := \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}. \tag{1.4}$$

Definition 6 (Standardabweichung) *Die Wurzel aus der Varianz heißt Standardabweichung σ von $X_1, \ldots, X_N$:*

$$\sigma := \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}}. \tag{1.5}$$

Wir berechnen als nächstes mit **Mathematica** die Varianz und die Standardabweichung unserer Rohwerte. Der Vergleich mit den eingebauten **Mathematica** Funktionen bringt unterschiedliche Ergebnisse. Der Grund für diese Differenzen liegt in der Definition der Varianz. Wir haben die Varianz „allgemein" definiert, wie man es in der beschreibenden Statistik tut. Das Computeralgebrasystem **Mathematica** benutzt hingegen die Definition zur „erwartungstreuen" Varianz aus der schließenden Statistik. Der Unterschied zwischen „allgemeiner" und „erwartungstreuer" Varianz ist, das im erwartungstreuen Fall $N - 1$ anstatt N in den Nenner des Bruches geschrieben wird. In diesem Fall taucht zum Beispiel eine Definitionslücke für $N = 1$ auf, da dann der Nenner gleich Null ist. Für alle Berechnungen dieser Bachelorarbeit und im Quellcode der Implementierung verwenden wir ausschließlich die allgemeine Definition der Varianz.

```
In[15]:= XQuer = N[Sum[X[[p]], p, 1, N] / N]
Out[15]= 9.55556
```

```
In[16]:= N[Mean[X]]
Out[16]= 9.55556

In[17]:= sigma2 = Sum[(X[[p]] - XQuer)^2, p, 1, N] / N
Out[17]= 4.69136

In[18]:= Variance[RWe] / / N
Out[18]= 5.27778

In[19]:= sigma = Sqrt[sigma2]
Out[19]= 2.16595

In[20]:= N[StandardDeviation[X]]
Out[20]= 2.29734
```

Weitere Streuungsmaße sind:

Definition 7 (Quartilsabstand)

$$\begin{aligned} \textit{Quartilsabstand} &= \textit{oberes} - \textit{unteres Quartil} \\ &= X_{0,75} - X_{0,25} \end{aligned} \tag{1.6}$$

Definition 8 (Stichprobenspannweite)

$$X_{(N)} - X_{(1)} = \max X_p - \min X_p, \quad \forall \quad 1 \le p \le N. \tag{1.7}$$

Man darf sich an dieser Stelle nicht von dem Wort Stichprobe verwirren lassen. Es geht nicht um eine *Stichprobe* aus unseren Rohwerten X, sondern um den vollen Umfang der Beispielrohwerte. Das Wort *Stichprobenspannweite* kommt daher, dass in der Statistik häufig von der *Stichprobe* die Rede ist. Man geht von großen Datensätzen, der sogenannten Grundgesamtheit, aus und untersucht nur einzelne Teilmengen. Als **Grundgesamtheit** (oder Population) bezeichnet man alle potentiell untersuchbaren Einheiten oder „Elemente“, die ein gemeinsames Merkmal oder eine gemeinsame Merkmalskombination aufweisen. So sprechen wir beispielsweise von der Population aller Menschen, die jemals einen *COLUMBUS-User-Level-Test* bei der ESA mitgeschrieben haben. Theoretisch können Grundgesamtheiten unbegrenzten Umfang aufweisen. Allgemein stellt eine *Stichprobe* eine Teilmenge aller Untersuchungsobjekte dar, die die untersuchungsrelevanten Eigenschaften der Grundgesamtheit möglichst genau abbilden. In diesem Fall ist eine *Stichprobe* ein „Miniaturbild“ der Grundgesamtheit. *Stichproben* lassen sich noch genauer differenzieren. Es gibt verschiedene Methoden, die *Stichprobe* zu bilden. Ein solches Kriterium für unseren Beispieldatenstz **`Tabelle`** ist das Merkmal „alle Elemente aus der Grundgesamtheit, die ihren *COLUMBUS-User-Level-Test* am 02.02.2002 oder 03.03.2003 geschrieben haben“. Das heißt, an diesen Daten fand, zusammengenommen, ein Test mit $N = 9$ Personen und $n = 12$ items statt. Wir errinnern uns an die Situation auf Seite 11. Im Beispieldatensatz **`Tabelle`** tauchte bei *item* **`i02`** und **`i03`** der Wert „a“ auf, woraufhin wir diese *item*-Spalten löschten, um

eine vergleichbare Analyse zu bekommen. Mit dem Wort *Stichprobe* bezeichen wir die Menge aller Personen, die in die anstehenden Berechnungen zur Testanalyse eingehen wird[13].

Betrachten wir doch exemplarisch den Quartilsabstand und die Stichprobenspannweite unserer Beispielrohwerte:

```
In[21]:= X

Out[21]= {4, 9, 9, 10, 10, 10, 11, 11, 12}


In[22]:= Needs["Statistics`DescriptiveStatistics`"];
         InterquartileRange[X]

Out[22]= 2


In[23]:= range = Max[X] - Min[X]

Out[23]= 8
```

1.3 Statistik-Plots

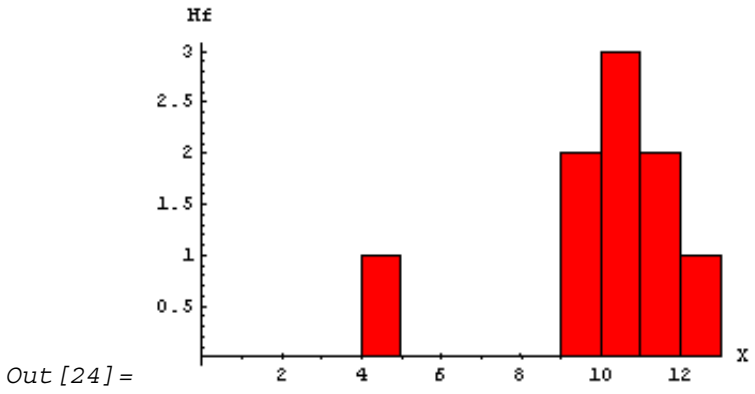
Neben dem vorgestellten Balkendiagramm, gibt es in der Statistik außerdem die häufig genutzten Histogramme und Box-Plots.

Histogramm Ein Histogramm ist die graphische Darstellung der Häufigkeitsverteilung von Rohwerten. Man geht dabei von den nach Größe geordneten Daten aus und teilt den gesamten Bereich der Stichprobe in k Klassen auf. Diese müssen nicht notwendig gleich breit sein. Über jeder Klasse wird ein Rechteck errichtet, dessen Fläche proportional zur klassenspezifischen Häufigkeit ist. Ist die Fläche des Rechtecks gleich der absoluten Häufigkeit, wird das Histogramm *absolut* genannt. Wenn die relativen Häufigkeiten verwendet werden, wird das Histogramm entsprechend *relativ* oder auch *normiert* genannt[14].

Sehen wir uns das Histogramm zu unseren Rohwerten X an:

```
In[24]:= Needs["Statistics`NormalDistribution`"];
         Histogram[X, AxesLabel-> {"X", Hf}];
```

[13] [Bortz, S.86]
[14] [Wiki, Histogramm]



Box-Plot Der Box-Plot, auch *Kisten-Diagramm* oder *Box-Whisker-Plot* genannt, dient dem schnellen visuellen Vergleich zwischen verschiedenen Stichproben. Auch für einzelne Betrachtungen ist er durchaus attraktiv, da die „Kerninformationen“ schnell erkennbar sind. Er benutzt Quantile zur graphischen Darstellung von Lage und Streuung, und er hebt potentielle Ausreißer hervor. Zur Anfertigung des Box-Plots wird eine *Kiste* vom *unteren* zum *oberen Quartil* gezeichnet und beim Median unterteilt. Die Kiste umfasst somit 50% der Daten. Der Endpunkt des nach oben aufgesetzten Stabes, auch *oberer Whisker* genannt, ist die größte Beobachtung, die kleiner als das *obere Quartil* plus das $1,5$-fache des Quartilsabstands, also kleiner als $X_{0,75} + 1,5 \cdot (X_{0,75} - X_{0,25})$ ist (sog. *größte normale*[15] *Beobachtung*). In gleicher Weise ist der Endpunkt des nach unten angebrachten Stabes, entsprechend auch *unterer Whisker* genannt, die kleinste Beobachtung, die größer als $X_{0,25} - 1,5 \cdot (X_{0,75} - X_{0,25})$ ist (sog. *kleinste normale Beobachtung*). Extrem große Beobachtungen und somit mögliche „Ausreißer nach oben“ sind konventionsgemäß jene, die oberhalb der Grenze $X_{0,75} + 1,5 \cdot (X_{0,75} - X_{0,25})$ liegen, sie werden jeweils durch einen Punkt gekennzeichnet. Analog behandelt man extrem kleine Beobachtungen als potentielle „Ausreißer nach unten“[16].

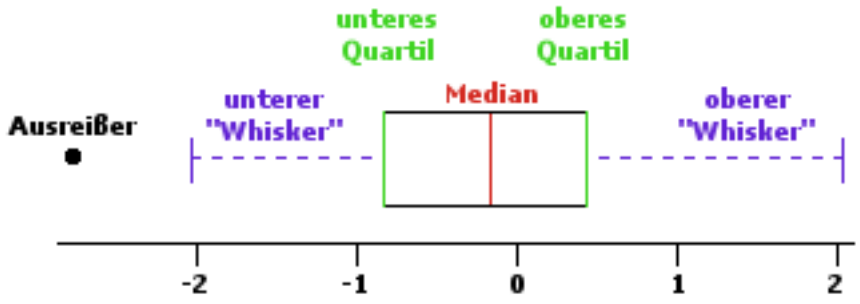


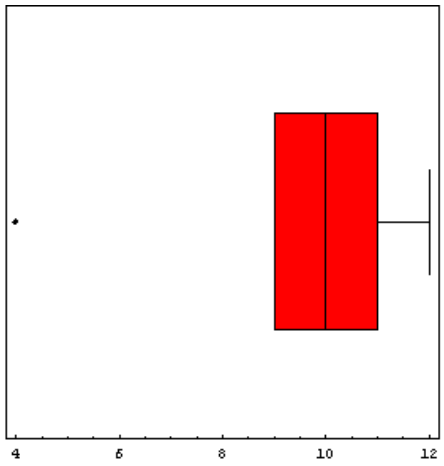
Abbildung 3: Der Box-Plot

[15] Zu dem Begriff „normal“ kommen wir in Kapitel 1.4.
[16] [Henze, S.35]

Betrachten wir den Box-Plot zu unseren Rohwerten X:

```
In[25]:= Needs["Statistics`StatisticsPlots`"];
         BoxWhiskerPlot[X];
```

Out[25]=



Die „Kerninformationen“ Median, *unteres* und *oberes Quartil*, Quartilsabstand und Ausreißer sind deutlich zu erkennen. Es fällt als Besonderheit auf, dass das *untere Quartil* und die *kleinste normale Beobachtung* zusammenfallen.

1.4 Normalverteilung

In diesem Abschnitt wollen wir uns unter Anderem mit der Bedeutung der Standardabweichung beschäftigen. Wir betrachten die Verteilung der Rohwerte unserer Beispielprobanden. Gehen wir davon aus, dass eine Verteilung *unimodal* und *symmetrisch* ist und zudem einen *„glockenförmigen Verlauf“* aufweist. Eine solche Verteilung wird als **Normalverteilung** bezeichnet. Sie ist die wichtigste Verteilung der Statistik[17].

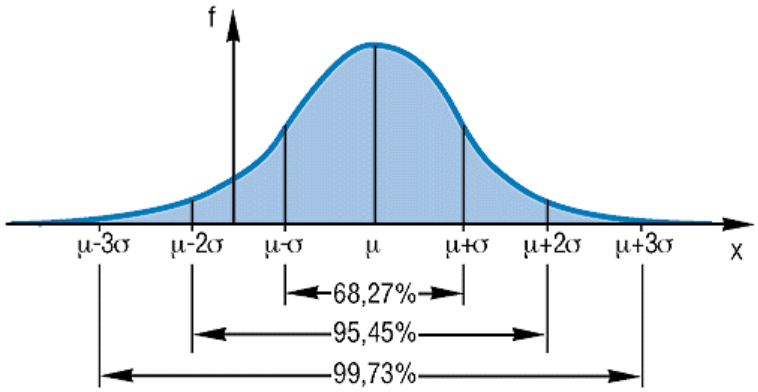


Abbildung 4: Die Normalverteilung

[17] laut [Bortz, S.73]

Betrachten wir die Eigenschaften, die die Normalverteilung definieren[18]:

- Die Normalverteilung wird beschrieben durch die Dichtefunktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \tag{1.8}$$

  wobei $\pi \approx 3,14$ (die Kreiszahl Pi) und $e \approx 2,72$ (die Eulersche Zahl).

- Die Verteilung hat einen glockenförmigen Verlauf. Dies brachte der Funktion f auch den Namen „Gaußsche Glockenkurve“ ein.

- Die Funktion ist stetig und ihre Funktionswerte sind positiv:
  $f(x_i) \geq 0 \quad \forall \; -\infty \leq x_i \leq +\infty$

- Die Funktionswerte nähern sich asymptotisch dem Wert Null an:
  $\lim_{x\to\pm\infty} f(x_i) = 0$

- Das Maximum der Funktion wird erreicht bei $x_i = \mu$:
  $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$

- Die Funktion besitzt Wendepunkte bei $x_i = \pm\sigma$.

- Für f gilt die Beziehung $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

- Die Funktion ist unimodal und symmetrisch. Es gilt: $AM = Md = Mo$

- Für Normalverteilungen gilt, dass zwischen den Werten $\overline{X} - \sigma$ und $\overline{X} + \sigma$ ca. $2/3$ aller Fälle $(68,27\%)$ liegen. Erweitern wir diesen Bereich auf $\overline{X} \pm 2\sigma$, befinden sich in ihm ca. 95% $(95,45\%)$ aller Fälle.

- Ein Vorteil für die *praktische* Arbeit ist, dass die Normalverteilung nur von zwei Parametern abhängt. Dem Mittelwert, in diesem Zusammenhang auch *Erwartungswert* μ genannt, und der Standardabweichung σ. Diese beiden Parameter legen die Funktion f eindeutig fest. Wir symbolisieren diese Abhängigkeit, indem wir kurz schreiben $NV(\mu;\sigma)$.

[18] [Litz, S.255], [Bortz, S.73], [Henze, S.200]

Zur Veranschaulichung, wie sich die Normalverteilung anhand ihrer Parameter μ und σ verändert, hier ein Beispiel-Plot:

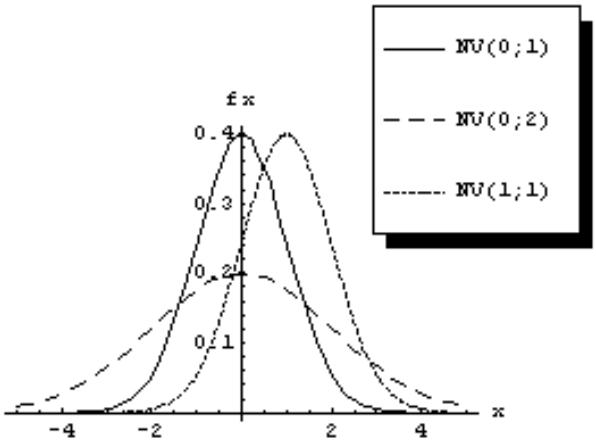


Abbildung 5: Beispiele für $NV(\mu;\sigma)$

Die Beispiele zeigen zum einen die horizontale Verschiebung, wenn bei gleicher Standardabweichung ein anderer Erwartungswert auftaucht (vgl. `NV(0;1)` und `NV(1;1)`). Zum anderen erkennt man eine „Abflachung“ des Funktionsverlaufs mit steigender Standardabweichung (vgl. `NV(0;2)`).

Die Normalverteilung als empirische Verteilung: Wir haben bisher die Normalverteilung als eine rein theoretische Verteilung mit bestimmten mathematischen Eigenschaften kennengelernt. Ihre Bedeutung ist jedoch zum Teil darauf zurückzuführen, dass sich einige human- und sozialwissenschaftlich relevanten Merkmale zumindest angenähert normalverteilen. Das Modell der Normalverteilung wurde erstmalig im 19. Jahrhundert von dem Belgier Adolph Quetelet[19] auf menschliche Eigenschaften angewandt. Quetelet war es aufgefallen, dass sich eine Reihe von Messungen, wie zum Beispiel die Körpergröße, das Körpergewicht, Testleistungen usw. angenähert normalverteilten, was ihn zu dem Schluss veranlasste, dass die Normalverteilung psychologischer, biologischer und antropologischer Merkmale einem Naturgesetz entspricht. Er ging davon aus, dass die Natur eine ideale, normative Ausprägung aller Merkmale anstrebe, dass jedoch die individuelle Ausprägung eines Merkmals von einer großen Zahl voneinander unabhängiger Faktoren abhänge, sodass die entgültige Merkmalsausprägung sowohl von der „idealen Norm“ als auch von Zufallseinflüssen determiniert wird. Das Ergebnis dieser beiden Wirkmechanismen sei die Normalverteilung.

Dieser Ansatz über die Normalverteilung hat inzwischen an Bedeutung verloren. Vor allem wird der Gedanke, dass sich in der Normalverteilung ein Naturgesetz abbilde, heute eindeutig abgelehnt. Empirische Merkmalsverteilungen können zwar angenähert normalverteilt sein, es existieren jedoch auch andere empirische Verteilungen, die mit

[19] vgl. Boring, 1950

der Normalverteilung nicht die geringste Ähnlichkeit haben[20].

Große Bedeutung kommt der Normalverteilung auf Grund der Aussage des **Zentralen Grenzwertsatzes** (ZGWS) zu: In Kapitel 1.3 haben wir bezüglich des Histogrammes erfahren, dass die Fläche der Rechtecke proportional zur klassenspezifischen Häufigkeit ist. Stellen wir uns nun ein „idealisiertes Histogramm“ mit unendlich feiner Klasseneinteilung auf der X-Achse und glockenförmigen Verlauf vor. Der ZGWS besagt, dass beim Grenzübergang $n \to \infty$ für ein gegebenes Intervall $[a, b]$ der X-Achse mit der Klassenanzahl n die Fläche des Histogrammes gegen die Fläche unter der Gaußschen Glockenkuve in den selben Grenzen, also gegen das Integral $\int_a^b f(x)\,\mathrm{d}x$, konvergiert[21].

Genauso leitet sich die Bedeutsamkeit der Normalverteilung aber auch aus anderen Eigenschaften ab. Sie ist in Verbindung zu bringen als Verteilungsmodell für statistische Kennwerte, eine mathematische Basisverteilung und in der statistischen Fehlertheorie[22].

Für die Bedeutung der Normalverteilung im Falle der Testanalyse geht es uns hauptsächlich um eine Aussage über die Zugehörigkeit von Rohwerten zum Bereich $\bar{X} - \sigma$ bis $\bar{X} + \sigma$. Wir wollen prüfen, ob Wissenstests eine brauchbare Struktur aufweisen. Als Kriterium dafür bietet sich natürlich die Kontrolle der erreichten Rohwerte auf eine Normalverteilung an. Wir wollen zum Beispiel feststellen, wieviele – und welche – Probanden nicht im normalverteilten Bereich abgeschnitten haben. Waren das Personen, die durch Krankheit einen Teil des Unterrichts verpasst haben, oder waren es gerade die „Fleißigsten und Tüchtigsten“ des Unterrichts gewesen? Klärung dieser und ähnlicher Fragen kann eine Hilfe sein, wenn es darum geht, einen bestehenden Test zu optimieren.

Bei unserem Beispieldatensatz `Tabelle` haben wir mit dem Balkendiagramm auf S.16 und dem Histogramm auf S.20 eine rechtssteile Verteilung festgestellt. Der Verlauf ist nicht „gaußglockenförmig“. Prüfen wir also, in wie weit unsere Verteilung einer Normalverteilung angenähert ist.

1.5 Chi-Quadrat-Test

Häufig, besonders bei kleinen Analysenstichproben, beobachtet man eine unregelmäßige Häufigkeitsverteilung. Es ist nicht klar zu erkennen, wie gut die Anpassung der Häufigkeitsverteilung an die Normalverteilung gegeben ist. In diesem Fall teilt man zunächst die Testpunkte in eine Anzahl „K“ von Klassen ein und benutzt dann den χ^2-Test (Chi-Quadrat-Test). Die Anzahl der Klassen sollte zwischen 10 und 15 liegen, in jeder Klasse sollte mindestens eine Frequenz von 5 Häufigkeiten liegen. Damit ist zu erkennen, dass wir einen Stichprobenumfang von ca $10 \cdot 5 = 50$ Probanden benötigen,

[20] Nachzulesen in der Studie von *Miccerie*, 1989

[21] [Henze, S.200]

[22] [Bortz, S.76]

um den χ^2-Test anwenden zu können[23].

Definition 9 (Chi-Quadrat-Test) *Für die „Güte der Anpassung“ an die Normalverteilung stellt der χ^2-Test ein geeignetes Messinstrument dar.*

$$\chi^2 = \sum_{k=1}^{K} \frac{(fo_k - fe_k)^2}{fe_k}, \text{ wobei } fe_k = \frac{hN}{\sigma_X} \cdot y_k \tag{1.9}$$

In dieser Formel bedeuten:

fo_k = *beobachtete Frequenz in Klasse k*

fe_k = *aufgrund der Normalverteilungshypothese erwartete Frequenz in Klasse k*

h = *Intervallbreite der Klassen*

y_k = *Ordinatenhöhe des Standardwertes* $z = \frac{(X_k - \bar{X})}{\sigma_X}$, *die der Klassenmitte* X_k *entspricht. Das heißt, man setzt z in die Funktion (1.8) auf S.22 ein.*

σ_X = *Standardabweichung der Rohwerte X*

N = *Anzahl der Probanden*

Das errechnete χ^2 wird im Anschluss mit den Werten der $(1 - \alpha)$-Quantile-Tabelle verglichen.

$m \setminus \alpha$	0,995	0,990	0,950	0,900	0,500	0,100	0,050	0,010	0,005
1	$0,0^439$	$0,0^316$	$0,0^239$	0,016	0,455	2,71	3,84	6,63	7,9
2	0,010	0,020	0,103	0,211	1,39	4,61	5,99	9,21	10,6
3	0,072	0,115	0,352	0,584	2,37	6,25	7,81	11,3	12,8
4	0,207	0,297	0,711	1,06	3,36	7,78	9,49	13,3	14,9
5	0,412	0,554	1,15	1,61	4,35	9,24	11,1	15,1	16,7
6	0,676	0,872	1,64	2,20	5,35	10,6	12,6	16,8	18,5
7	0,989	1,24	2,17	2,83	6,35	12,0	14,1	18,5	20,3
8	1,34	1,65	2,73	3,49	7,34	13,4	15,5	20,1	22,0
9	1,73	2,09	3,33	4,17	8,34	14,7	16,9	21,7	23,6
10	2,16	2,56	3,94	4,87	9,34	16,0	18,3	23,2	25,2

Tabelle 1: Quantile von χ^2-Verteilungen (Freiheitsgrad m, Signifikanzniveau α) aus [Ziez, S.102]

Liegt der errechnete χ^2-Wert unter dem entsprechenden Tabelleneintrag mit dem passenden m und gewähltem α, dann können wir von einer ausreichenden Anpassung

[23][LiRa, S.149]

unserer Rohwerte an die Normalverteilung ausgehen. Die Freiheitsgrade m bestimmt man in unserem Fall der Klasseneinteilung mit der Formel[24]

$$m = (\text{Anzahl der Klassen mit Häufigkeit } \geq 5) - 3. \tag{1.10}$$

Der Chi-Quadrat-Test für die Güte der Anpassung an eine Normalverteilung bei einer unregelmäßigen Rohwertverteilung ist nur bei kleinen Stichprobenumfängen (N unter 200) sinnvoll[25]. In Kapitel 3.1.2 wird der „Bauplan“ des Chi-Quadrat-Algorithmus vorgestellt.

1.6 Korrelation

Die Korrelation ist eine Beziehung zwischen zwei oder mehr statistischen Variablen. Wenn eine hohe Korrelation vorhanden ist, ist noch nicht gesagt, ob eine Größe die andere kausal[26] beeinflusst, oder ob beide von einer dritten Größe kausal abhängen, oder ob sich überhaupt ein Kausalzusammenhang folgern lässt.

Es gibt positive und negative Korrelation. Ein Beispiel für positive Korrelation (je mehr, desto mehr) ist: *Je mehr Sonnentage im Jahr, desto besser läuft der Verkauf von Sonnencreme.* Ein Beispiel für negative Korrelation (je mehr, desto weniger) ist: *Je mehr Verkauf von Regenschirmen, desto weniger Verkauf von Sonnencreme.*

Häufig benutzt man zurecht die Korrelation, um einen Hinweis darauf zu bekommen, ob zwei statistische Größen ursächlich miteinander zusammenhängen. Das funktioniert immer dann besonders gut, wenn beide Größen durch eine „Je… , desto… “-Beziehung miteinander zusammenhängen und eine der Größen nur von der anderen abhängt. Die Korrelation beschreibt nicht unbedingt eine Ursache-Wirkungs-Beziehung in die eine oder andere Richtung. So darf man über die Tatsache, dass man die Feuerwehr oft bei Bränden findet nicht folgern, dass die Feuerwehr die Ursache für Brände ist. Die direkte Kausalität kann auch gänzlich fehlen. So kann es durchaus eine Korrelation zwischen dem Rückgang der Störche in Nordhessen und einem Rückgang der Anzahl der Neugeborenen geben, diese Ereignisse haben aber nichts miteinander zu tun – weder bringen Störche Kinder, noch umgekehrt!

Definition 10 („Bravais-Pearson-Korrelation“) *Die ersten Anwendungen des Korrelationskoeffizienten stammen von Francis Galton und Karl Pearson, die mit diesem Zusammenhangsmaß die Beziehung von Körperbaumaßen zwischen Eltern- und Kindergenerationen untersuchten. Die klassische und noch heute verwendete Korrelationsrechnung begann schließlich* 1846*. Der Name dieser ursprünglichen Korrelationsformel lautet „Bravais-Pearson-Korrelation“, nach den beiden Hauptbegründern der*

[24] [Bortz, S.165]

[25] Ein gut erklärtes Beispiel zur Anwendung dieser Rechenformal kann in [LiRa, S.150] nachgelesen werden.

[26] ursächlich

Korrelationsuntersuchung.

$$r = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} \tag{1.11}$$

Die Bezeichnung r kommt von dem Begriff Regression, der sinngemäß für die Untersuchung von Merkmalen und ihren Bezug zueinander steht[27]*. Die Kovarianz* $cov(X,Y)$ *zweier Merkmale* X *und* Y*, ist ein Maß, das Informationen über die Enge des Zusammenhangs der Merkmale liefert.*

Wir werden im Folgenden zeigen, dass für die Berechnung des Korrelationskoeffizienten die Formel (1.11) noch verändert bzw. vereinfacht werden kann. Zunächst machen wir aber einige Vorbetrachtungen zum besseren Verständnis.

Die Berechnung der Kovarianz beruht auf der Formel

$$cov(X,Y) = \frac{\sum_{i=1}^{N} X_i \cdot Y_i - \frac{\sum_{i=1}^{N} X_i \cdot \sum_{i=1}^{N} Y_i}{N}}{N}. \tag{1.12}$$

Gebräuchlicher ist aber eine andere Formal für die Kovarianz, da an Hand dieser das Verhältnis der beiden zu untersuchenden Merkmale sichtbar wird. Wir zeigen die Umformungsschritte zwischen beiden Kovarianzformeln[28].

Zu zeigen:

$$cov(X,Y) = \frac{\sum_{i=1}^{N} X_i \cdot Y_i - \frac{\sum_{i=1}^{N} X_i \cdot \sum_{i=1}^{N} Y_i}{N}}{N} = \frac{\sum_{i=1}^{N} (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{N} \tag{1.13}$$

[27] mehr dazu in [Bortz, Kap.6.1]
[28] [Bortz, S.189]

Beweis:

$$\begin{aligned}
\frac{\sum_{i=1}^{N}(X_i-\bar{X})\cdot(Y_i-\bar{Y})}{N} &= \frac{\sum_{i=1}^{N}X_i\cdot Y_i-\sum_{i=1}^{N}X_i\cdot\bar{Y}-\sum_{i=1}^{N}\bar{X}\cdot Y_i+\sum_{i=1}^{N}\bar{X}\cdot\bar{Y}}{N}\\
&\text{An dieser Stelle benutzen wir die Umformung}\\
&N\cdot\bar{X}=N\cdot\frac{\sum_{i=1}^{N}X_i}{N}=\sum_{i=1}^{N}X_i\\
&\text{und setzen sie im nächsten Schritt ein:}\\
&= \frac{\sum_{i=1}^{N}X_i\cdot Y_i}{N}-\frac{\bar{Y}\cdot N\bar{X}}{N}-\frac{\bar{X}\cdot N\bar{Y}}{N}+\frac{N\cdot\bar{X}\bar{Y}}{N}\\
&= \frac{\sum_{i=1}^{N}X_i\cdot Y_i}{N}-\frac{\frac{\sum_{i=1}^{N}Y_i}{N}\cdot N\frac{\sum_{i=1}^{N}X_i}{N}}{N}-\frac{\frac{\sum_{i=1}^{N}X_i}{N}\cdot N\frac{\sum_{i=1}^{N}Y_i}{N}}{N}\\
&\quad+\frac{N\cdot\frac{\sum_{i=1}^{N}X_i}{N}\cdot\frac{\sum_{i=1}^{N}Y_i}{N}}{N}\\
&= \frac{\sum_{i=1}^{N}X_i\cdot Y_i}{N}-\frac{\sum_{i=1}^{N}X_i\cdot\sum_{i=1}^{N}Y_i}{N^2}-\frac{\sum_{i=1}^{N}Y_i\cdot\sum_{i=1}^{N}X_i}{N^2}\\
&\quad+\frac{\sum_{i=1}^{N}Y_i\sum_{i=1}^{N}X_i}{N^2}\\
&= \frac{\sum_{i=1}^{N}X_i\cdot Y_i}{N}-\frac{\sum_{i=1}^{N}X_i\cdot\sum_{i=1}^{N}Y_i}{N^2}\\
&= \frac{\sum_{i=1}^{N}X_i\cdot Y_i-\frac{\sum_{i=1}^{N}X_i\cdot\sum_{i=1}^{N}Y_i}{N}}{N}
\end{aligned}$$

□

Die Standardabweichung kennen wir aus Formel (1.5) auf S.17. Auch hier wollen wir uns kurz eine Umformung ansehen, die uns im Anschluss das Verstehen des Korrelationskoeffizienten erleichtern wird[29].

Zu zeigen:

$$\sigma_X=\sqrt{\frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{N}}=\sqrt{\frac{\sum_{i=1}^{N}X_i^2}{N}-\bar{X}^2} \tag{1.14}$$

[29][Bortz, S.43]

Beweis:

$$\begin{aligned}
\sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}} &= \sqrt{\frac{\sum_{i=1}^{N}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)}{N}} \\
&= \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - 2\overline{X}\sum_{i=1}^{N} X_i + N\overline{X}^2}{N}} \\
&= \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - 2\overline{X} \cdot N\overline{X} + N\overline{X}^2}{N}} \\
&= \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - N\overline{X}^2}{N}} \\
&= \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N} - \overline{X}^2}
\end{aligned}$$

□

Nun betrachten wir die Umformung der Bravais-Pearson-Korrelationsformel (1.11). Dadurch wird der Korrelationskoeffizient rechnerisch einfacher und weniger anfällig für Rundungsfehler[30].

Zu zeigen:

$$r = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{N \cdot \sum_{i=1}^{N} X_i Y_i - (\sum_{i=1}^{N} X_i) \cdot (\sum_{i=1}^{N} Y_i)}{\sqrt{\left[N\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2\right] \cdot \left[N\sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2\right]}} \qquad (1.15)$$

[30][Bortz, S.205]

Beweis:

$$
\begin{aligned}
\frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} &= \frac{\frac{\sum_{i=1}^{N} X_i \cdot Y_i - \frac{\sum_{i=1}^{N} X_i \cdot \sum_{i=1}^{N} Y_i}{N}}{N}}{\sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N} - \overline{X}^2} \cdot \sqrt{\frac{\sum_{i=1}^{N} Y_i^2}{N} - \overline{Y}^2}} \\
&= \frac{N(\sum_{i=1}^{N} X_i \cdot Y_i - \frac{\sum_{i=1}^{N} X_i \cdot \sum_{i=1}^{N} Y_i}{N})}{N^2 \sqrt{(\frac{\sum_{i=1}^{N} X_i^2}{N} - \overline{X}^2) \cdot (\frac{\sum_{i=1}^{N} Y_i^2}{N} - \overline{Y}^2)}} \\
&= \frac{N \sum_{i=1}^{N} X_i \cdot Y_i - \sum_{i=1}^{N} X_i \cdot \sum_{i=1}^{N} Y_i}{\sqrt{N^2(\frac{\sum_{i=1}^{N} X_i^2}{N} - (\frac{\sum_{i=1}^{N} X_i}{N})^2) \cdot N^2(\frac{\sum_{i=1}^{N} Y_i^2}{N} - (\frac{\sum_{i=1}^{N} Y_i}{N})^2)}} \\
&= \frac{N \sum_{i=1}^{N} X_i \cdot Y_i - \sum_{i=1}^{N} X_i \cdot \sum_{i=1}^{N} Y_i}{\sqrt{(N \sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2) \cdot (N \sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2)}}
\end{aligned}
$$

□

Die Werte einer Korrelationsberechnung liegen zwischen 1 und −1. Wobei 1 für positive, −1 für negative Korrelation und 0 für keinen Zusammenhang zwischen den Merkmalen steht. Generell kann man sagen, je näher die positiven Werte des Korrelationskoeffizienten an 1 herankommen, desto größer ist der Merkmalszusamenhang und „linear ansteigender“ die Kurve der Merkmale, je näher die Werte an −1 herankommen, desto geringer ist der Merkmalszusammenhang und „linear abfallender“ die Kurve. Je näher der Korrelationkoeffizient an 0 herankommt, desto verteilter bzw. verbreiteter sind die Merkmalswerte, ein Merkmalszusammenhang ist nicht vorhanden[31].

[31] [Wiki, Korrelation]

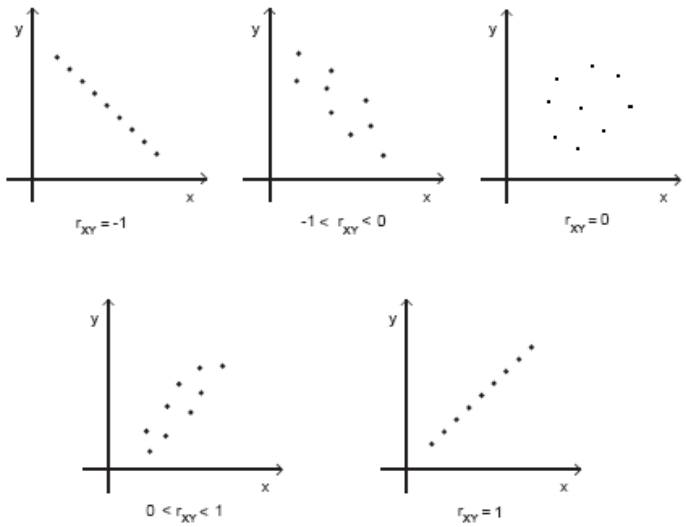


Abbildung 6: Unterschiedliche Möglichkeiten der Korrelation zweier Merkmale *x* und *y* [MaOn, S.2].

Wir betrachten *item i*04 aus dem Beispieldatensatz **`Tabelle`**. Unsere Merkmale für die Korrelation sind zum einen die Rohwerte der Probanden im gesamten Test, zum anderen die erreichte Punktzahl jedes Probanden bei *i*04. **`Mathematica`** rechnet bei dem Befehl „**`Correlation[]`**“ mit der Bravais-Pearson-Korrelationsformel, wie die *Mathematica*-Hilfe erläutert.

```
In[26]:= X = {12, 11, 11, 10, 10, 10, 9, 9, 4};
         Y = {1, 1, 1, 1, 1, 1, 1, 1, 0};
         Needs["Statistics`MultiDescriptiveStatistics`"];
         Correlation[X, Y] // N
Out[26]= 0.906845
```

Das *item i*04 weist eine Korrelation von $0,9$ auf. Ein Korrelationskoeffizient in dieser Höhe stellt uns zufrieden, da er positiv ist und nahe am Maximumswert 1. Wir haben erfahren, dass der Zusammenhang zwischen dem Abschneiden eines jeden Probanden bei dem **einzelnen** *item i*04 und dem Abschneiden im **gesamten** Test in einer „*je mehr, desto mehr*“-Beziehung gegeben ist.

2 Methoden der Testanalyse

2.1 Test

Das Wort „Test“ stammt aus dem englischen Sprachgebrauch und bedeutet soviel wie Probe. Dieser Begriff ist inzwischen in den deutschen Sprachschatz eingegangen, wobei sich für den Plural die englische Form „Tests“ durchgesetzt hat[32]. Das Wort „Test“ hat eine *mehrfache* Bedeutung. Man versteht darunter:

a) Ein Verfahren zur Untersuchung eines Persönlichkeitsmerkmals.

b) Den Vorgang der Durchführung der Untersuchung.

c) Die Gesamtheit der zur Durchführung notwendigen Materialien.

d) Jede Untersuchung, sofern sie Stichprobencharakter[33] hat.

e) Gewisse mathematisch-statistische Prüfverfahren (z.B. χ^2-Test, S.24).

f) Kurze, außerplanmäßige „Zettelarbeiten“ im Schulunterricht[34].

Unter diesen Bedeutungen ist die erste für uns am wichtigsten; sie soll in der folgenden Definition in ihren in diesem Zusammenhang wesentlichen Punkten festgelegt werden:

Definition 11 (Test) *Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch*[35] *abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung*[36].

Zu der Testanalyse gehören verschiedene Prüfverfahren. Um aber erst einmal einheitlich zu definieren, auf **was** geprüft werden soll, schauen wir uns die „Gütekriterien“ eines Tests im nächsten Abschnitt an. Anhand dieser wird später die Signifikanz der Ergebnisse und damit die Qualität des Tests beurteilt. Zunächst aber noch zwei wichtige Erklärungen, damit wir entsprechende Begriffe aus der Statistik richtig benutzen können.

[32] [LiRa, S.1]

[33] (vgl. S.18)

[34] Punkt f) ist umstritten, da es Psychologen gibt, die Lehrern nicht genügend Kompetenz zuschreiben, einen Test sachgerecht zu konzipieren bzw. auszuwerten.

[35] [Wiki, empirisch] Die **Empirie** (v.griech.: die Erfahrung) ist im eigentlichen Sinne nur wissenschaftlich, d.h. auf methodischem Weg gewonnenen Erfahrung.

[36] Die Definition lehnt sich unter Hervorhebung der wesentlichen Merkmale eng an WARREN (1934).

Signifikanz In der Statistik heißen Unterschiede **signifikant**, wenn sie durch Zufall nur mit einer bestimmten geringen Wahrscheinlichkeit zustande kommen. Die Überprüfung der statistischen Signifikanz geschieht mit Hilfe einer Nullhypothese, die verworfen wird, wenn das zufällige Zustandekommen des Unterschieds sehr unwahrscheinlich ist. Das Quantil der zu überprüfenden Unwahrscheinlichkeit wird vorher festgelegt und mit α bezeichnet, beispielsweise $\alpha = 0,05$ für 5% Irrtumswahrscheinlicheit. Je geringer diese, desto höher die Informationsqualität. Diesen Wert α bezeichnet man auch als Signifikanzniveau. Das Signifikanzniveau wird als umso höher bezeichnet, je kleiner α ist[37].

Nullhypothese Die Nullhypothese ist eine Annahme, die wir treffen und anschließend überprüfen. Formulieren wir etwa die Nullhypothese „Das Medikament *XY* hat bei bestimmten Personen schädliche Nebenwirkungen“, dann sollte das Risiko, diese Hypothese fälschlicherweise abzulehnen, sehr gering gewählt werden (z. B. mit einer Signifikanz von $\alpha = 0,01$). Bei der Nullhypothese „Heute Nachmittag scheint die Sonne“, könnte ein Risiko von $\alpha = 5\%$ noch tragbar sein, wenn die Konsequenz einer Ablehnung dieser richtigen Hypothese darin besteht, dass man bei Sonnenschein einen Regenschirm spazieren trägt. In den meisten Untersuchungen, so auch bei uns, gibt man sich mit einem Signifikanzniveau von 5% zufrieden[38].

2.2 Gütekriterien eines Tests

Ein guter Test soll als *Hauptgütekriterien* folgende drei Eigenschaften erfüllen:

- er soll reliabel,
- er soll valide,
- er soll objektiv sein.

Es gibt außerdem vier *Nebengütekriterien*: Normiert, vergleichbar, ökonomisch, nützlich. In dieser Arbeit werden wir nur auf die Hauptgütekriterien eingehen[39].

2.2.1 Reliabilität

Unter der *Reliabilität* oder Zuverlässigkeit eines Tests versteht man den Grad der Genauigkeit, mit dem er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal misst,

[37] [Wiki, Statistische Signifikanz], [Bortz, S.113]
[38] [Litz, S.313]
[39] [LiRa, S.7]

gleichgültig, ob er dieses Merkmal auch zu messen beansprucht (welche Frage ein Problem der Validität ist). Ein Test wäre demnach vollkommen reliabel, wenn die mit seiner Hilfe erzielten Ergebnisse den Probanden genau, d.h. fehlerfrei beschreiben bzw. auf der Testskala lokalisieren. Diese Genauigkeit betrifft lediglich den beobachteten Messwert und nicht auch seinen Interpretationswert, also nicht die Frage, ob er auch das misst, was er messen soll. Der Grad der Reliabilität wird durch den *Reliabilitätskoeffizienten*[40] bestimmt, der angibt, in welchem Maße unter gleichen Bedingungen gewonnenene Messwerte über ein und denselben Probanden übereinstimmen, in welchem Maße also das Testergebnis reproduzierbar ist[41].

2.2.2 Validität

Die *Validität* oder Gültigkeit eines Tests gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal, das er messen soll, tatsächlich auch misst. Ein Test wäre demnach vollkommen valide, wenn seine Ergebnisse einen unmittelbaren und fehlerfreien Rückschluss auf den Ausprägungsgrad des zu erfassenden Persönlichkeitsmerkmals zulassen, wenn also der individuelle Testpunktwert eines Probanden diesen auf der Messskala eindeutig lokalisiert[42].

2.2.3 Objektivität

Unter *Objektivität* eines Tests verstehen wir den Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind. Ein Test wäre demnach vollkommen objektiv, wenn verschiedene Untersucher bei denselben Probanden zu gleichen Ergebnissen gelangen. Man spricht deshalb auch von „interpersoneller Übereinstimmung“ der Untersucher. Als Maß für die Objektivität verschiedener Untersucher könnte der durchschnittliche Korrelationskoeffizient zwischen den durch verschiedene Untersucher an einer Stichprobe von Probanden erhobene Testbefund gelten[43].

Dieses Kriterium findet im Rahmen der ESA keine Verwendung. Die Wissenstests von Herrn Dr. Seine wurden bisher ausschließlich von ihm persönlich durchgeführt. Helfer treten nur bei der Unterrichtsgebung auf. Außerdem hat es keinen Sinn, einen reinen Wissenstests – es geht ja nicht um Zeit – von den selben Personen zweimal schreiben zu lassen. Ein Austauschen der Fragen würde keine hohe Vergleichbarkeit zwischen den Tests zulassen, die Fragen könnten als **zu** unterschiedlich schwer empfunden werden.

[40] mehr dazu in Kapitel 2.4
[41] [LiRa, S.9]
[42] [LiRa, S.10]
[43] [LiRa, S.7]

2.3 Testform

Bevor wir uns im Detail ansehen, wie die Gütekriterien anhand ihrer Koeffizienten bestimmt werden können, stellen wir die Informationen über unsere Teststruktur zusammen.

2.3.1 Teststruktur

Die heutige Form des *COLUMBUS*-Tests wurde von den ESA-Mitarbeitern selbst entwickelt. Es ist wichtig, die „Eckdaten“ des zu analysierenden Tests zu kennen, da für die Prüfung der Gütekriterien viele Möglichkeiten in Bezug auf unterschiedliche Teststrukturen und Aufgabentypen vorhanden sind[44].

- Vorab zum Aufbau der ESA-Tests: Die *COLUMBUS* Tests werden aus einem *Pool* an Fragen zusammengestellt. Über jedes der sechs Wissensgebiete in *COLUMBUS* liegen eine Vielzahl Fragen vor. Der gesamte Test besteht in der Regel aus $6 \cdot 6 = 36$ Fragen. Durch dieses Gleichgewicht wird versucht, jedes der sechs Themen gleichbedeutend zu behandeln.
  Die *Testlänge* entspricht der Anzahl der Testaufgaben (36 Stück).

- Nach dem Allgemeinheitsgrad ihrer Anwendbarkeit unterscheidet man zwischen standardisierten (geeichten) und nichtstandardisierten (informellen[45]) Tests.
  Standardisierte Tests müssen wissenschaftlich entwickelt, hinsichtlich der wichtigsten Gütekriterien untersucht und unter Standardbedingungen durchführbar und normiert sein. Im anderen Fall handelt es sich um *nichtstandardisierte* Tests, wie sie Psychologen und Lehrer gewissermaßen für den Hausgebrauch benutzen und auswerten.
  Bei unserer Testform handelt es sich also um einen **nichtstandardisierten** Test.

- Ein wichtiger Unterscheidungspunkt bei verschiedenen Testformen ist die Zeitbemessung.
  Schnelligkeitstests (Speed-Tests) enthalten leichte oder mittelschwere Aufgaben[46] in einer solchen Anzahl, dass kein Proband in einer bestimmten vorgegebenen Testzeit alle Aufgaben beantworten kann. Es kommt hier im wesentlichen auf die Schnelligkeit (speed) bei der Lösung der Aufgaben an.
  Niveautests (Power-Tests) enthalten Aufgaben, die in ihrem Schwierigkeitsgrad kontinuierlich so weit ansteigen, dass die letzen Aufgaben trotz fehlender oder

[44] [LiRa, S.14 ff]

[45] Der englische Ausdruck „informal test“ lässt sich nicht voll zutreffend durch einen deutschen Ausdruck ersetzen. Für informelle Schultests wird auch der Begriff „Standardarbeiten“ oder „teacher-made Tests“ benutzt.

[46] Auf die „Schwierigkeit einer Aufgabe“ gehen auf S.40 genauer ein.

großzügiger Zeitbemessung kaum von einem Probanden richtig beantwortet werden können. Hier kommt es darauf an, „Denkkraft“ und „geistiges Niveau“ (power) zu beweisen, Schnelligkeit spielt keine Rolle.
Der ESA-Test hat zwar ein Zeitlimit von 60 **min Testzeit**[47], die Teilnehmer dürfen aber im Prinzip so lange schreiben, bis sie freiwillig abgeben. Aus Erfahrung zeigt sich, dass diese Zeitgrenze meist nicht überschritten wird. Von einem kontinuierlichen Ansteigen des Aufgabeschwierigkeitsgrades kann man nur begrenzt sprechen. Das hat zwei Gründe: Zum einen liegt den Probanden von Anfang an der gesamte Test vor, sie können also die Aufgaben in beliebiger Reihenfolge bearbeiten. Und zum anderen ist der Test in sechs Themenblöcke untergliedert. Je einer für jedes der sechs *COLUMBUS* Systeme. Man kann nun den Schwierigkeitsanstieg besser in jedem der sechs Themen für sich betrachten, anstatt über den gesamten Test. Da wir anhand dieser Bedingungen keine eindeutige Zuordnung der ESA-Tests zu den Niveautests machen können, verwenden wir den Namen „Wissenstest“. Dieser Name meint ein Testverfahren, dessen Hauptziel es ist, Wissen abzufragen. Die Schwierigkeit zwischen den Aufgaben kann dabei variieren, ein Zeitlimit existiert zwar, ist aber zu vernachlässigen.

- Nach dem Interpretationsbezug gibt es *direkte* (psychometrische) und *indirekte* (projektive) Testverfahren (z. B. Fragebogen zum Ankreuzen im Gegensatz zu Charaktertests in Gespächsform).
  Der Umfang, in dem das Testergebnis durch das subjektive Urteil des Auswerters mitbestimmt ist, veranlasst zu einer Gruppierung in *objektive* und *nichtobjektive* Tests. Diese Einteilung deckt sich praktisch mit der in direkte und indirekte Tests insofern, als direkte Tests zugleich objektiv und indirekte Tests im Allgemeinen nicht objektiv sind.
  Da bei unseren Tests die Auswertung anhand eines eindeutigen Schemas gemacht wird, welches für alle Probanden gleich ist, gibt es keine Unterschiede der Testergebnisse, egal welche Person die Korrektur übernimmt. Es handelt sich also um einen **direkten** und **objektiven** Test.

- Je nach Art der Testdurchführung und der dabei benötigten Materialien spricht man von *Befragungstests* (in Interviewform), *Papier- und Stifttests* (dazu gehört die Mehrzahl der gebräuchlichen Tests), und *Materialbearbeitungstests*.
  Ganz klar erkennbar liegt in unserem Fall der **Papier- und Stifttest** vor.

- Eine weitere, grundlegende Einteilung ist die in *Individual-* und *Gruppentests*.
  Diese Klassifizierung sortiert den ESA-Test, da er in der Gruppe aller Kursteilnehmer durchgeführt wird, ganz klar als **Gruppentest**. Allerdings muss man beachten, dass bei Nichtbestehen jeder Proband die Chance hat, den Test zu wiederholen[48]. Sollte nur eine Person diese Wiederholung nutzen, hätten wir dann

[47] Die Testzeit ist die zur Beantwortung der Testaufgaben vorgegebene Zeit.

[48] Die Bewertung funktioniert so: Für jeden der sechs Themenbereiche gibt es eine Mindestpunktzahl. Welcher Proband eine dieser Punktegrenzen nicht erreicht, wiederholt nur diesen Teil des Tests.

den Fall des Individualtests.

- Nach der Art der Bezugsgröße und dem Zweck wird zwischen *normorientierten* und *kriterienorientierten* (lehrzielorientierten) Tests unterschieden. Bei normorientierten Tests wird das individuelle Testergebnis zum Populationsmittelwert in Beziehung gesetzt und ein Normwert bestimmt. Bei kriterienorientierten Tests wird das Ergebnis auf die Gesamtzahl der Aufgaben bezogen. Normorientierte Tests sollen Probanden möglichst gut differenzieren, kriterienorientierte Tests sollen prüfen, ob ein Kriterium (Lehrziel) erreicht worden ist oder nicht.
  In unserem Fall wird der Mindestpunktwert, der zu erreichen ist, um jeden Themenbereich zu bestehen, mit $2/3$ der möglichen Punkte für dieses Thema festgelegt. Da jeder Bereich 6 Fragen hat, sind also pro Themenblock mindestens $6 \cdot \frac{2}{3} = 4$ Punkte zu holen. Jede Aufgabe gibt höchstens einen und mindestens Null Punkte[49]. Für unsere Tests liegt ganz klar der Anspruch einer **kriterienorientierten** Struktur vor. Ziel ist schließlich die Überprüfung, ob die Probanden möglichst viel Wissen aus der Trainingswoche „mitgenommen“ haben, ein interpersoneller Vergleich ist für diesen Zweck uninteressant.

- Bei der Struktur des gesamten Tests kann man zwischen *einfachen* und *komplexen* Formen unterscheiden. Einfache Tests ergeben als Testresultat für jeden Probanden nur einen einzigen Punktwert. Komplexe Tests ergeben mehrere Punktwerte, die entweder zu einem Gesamtpunktwert vereinigt oder unabhängig nebeneinander bestehen bleiben.
  Wir haben bei dem Punkt *Gruppentest* schon gesehen, dass die Korrektur sich auf sechs einzelne Punktwerte konzentriert. Danach wird das Bestehen und Nichtbestehen entschieden. Die Struktur der ESA-Tests ist daher in komplexer Form.

2.3.2 Aufgabentypen

Nach der formalen Struktur der Aufgaben gibt es *hochstrukturierte* und *niedrigstrukturierte* Tests. Hochstrukturierte Test setzen sich zusammen aus Aufgaben, die nur eine einzige richtige Lösung zulassen, sie sind voll objektiv auswertbar. Mit abnehmendem Strukturgrad wird die Zahl der möglichen richtigen oder teilrichtigen Antworten größer bis zu einer Stufe, auf der man von richtigen oder falschen Antworten nicht mehr sprechen kann. Oft deckt sich diese Betrachtung recht gut mit der Einteilung in direkte und indirekte sowie objektive und nichtobjektive Tests. Wir hatten bereits festgestellt, dass die ESA-Tests der Form „direkt und objektiv“ entsprechen. Sehen wir uns also die Aufgabentypen an, um zu entscheiden, ob der Test hoch- oder niedrigstrukturiert ist[50].

[49] Auf die genaue Bewertung pro Aufgabe gehen wir in Kapitel 2.3.2 ein, weil wir uns dort auch mit den verschiedenen Aufgabentypen beschäftigen.

[50] [LiRa, S.17 ff, 33, 48]

Grundsätzlich erlaubt die Art der Aufgabenbeantwortung die Einteilung in Tests mit *freier* und in solche mit *gebundener* Aufgabenbeantwortung. Muss der Proband selbst eine Antwort auf die gestellte Testaufgabe finden, so spricht man von einem Test mit freier Aufgabenbeantwortung. Sind dagegen mehrere Antworten in der Aufgabe bereits vorgegeben, aus denen der Proband die richtige(n) auswählen soll, so liegt ein Test mit gebundener Aufgabenbeantwortung vor. Zu den Tests mit freier Aufgabenbeantwortung gehören Schlüsselwortergänzungstests, Aufsatztests, Lückentests und alle nichtobjektiven Tests. Zu denen mit gebundener Aufgabenbeantwortung zählen Richtig-Falsch-Antwort-Tests, Mehrfach-Wahl-Tests, Aufgaben-Zuordnungstests und Neu-Anordnungstests.

Die gebundene Aufgabenbeantwortung Gebundene Aufgabenbeantwortung bedeutet, dass dem Probanden mehrere Möglichkeiten, die ihrerseits festgelegt sind, für die Beantwortung vorgeschlagen werden. Er ist an diese „gebunden“. Sie steht im Gegensatz zu der freien Aufgabenbeantwortung, bei welcher der Proband Form und Inhalt der Antwort nach Ermessen wählen kann.

Wir betrachten nun die Aufgabentypen, die in den ESA-Tests vorkommen:

(Die ersten drei Typen gehören zur gebundenen Aufgabenbeantwortung, Typ d) zur freien.)

a) Der einfachste, wenn auch nicht am häufigsten verwendete Aufgabentyp ist der *Richtig-Falsch-Antworttyp* (RF-Aufgabe) bei dem der Proband zwischen einer richtigen und einer falschen, einer „Ja“- oder „Nein“-, einer „Plus“- oder „Minus“, einer „Stimmt“- oder „Stimmt nicht“-Antwort zu wählen hat.

b) Der am meisten gebrauchte Aufgabentyp ist die *Mehrfach-Wahl-Aufgabe* (MW-Aufgabe) oder „Multiple-Choice“-Aufgabe. Der Proband hat hierbei aus mehreren zur Wahl gestellten Antwortmöglichkeiten diejenige(n) zu kennzeichnen, die er für richtig hält. MW-Aufgaben enthalten meist nur eine einzige richtige Antwort, die als *Schlüsselantwort* oder *Bestantwort* bezeichnet wird. Wenn die Aufgaben jedoch mehrere (z. B. stets zwei) oder abwechselnd verschieden viele (z. B. Aufg. 3 eine, Aufg. 4 drei) Bestantworten enthalten, bezeichnet man sie als *Mehrfach-Antwort-Aufgaben* (MA-Aufgabe). Es darf natürlich nicht passieren, dass die Bestantwort immer an der gleichen Position der Antwortmöglichkeiten steht. So würde sich ein Schema durch die Aufgaben ziehen. Das will man natürlich vermeiden, da es die Probanden zum „schematischen“ Ankreuzen animieren könnte, der Anspruch des Wissenstests[51] geht verloren.

c) Eine besonders bei Wissens- und Kenntnisprüfungen benutzte Aufgabenform ist die *Zuordnungsaufgabe* (ZO-Aufgabe). Hier müssen die beiden Elemente einer

[51] Das wichtigste Ziel der *COLUMBUS*-Schulung ist der Wissenszuwachs bei den Teilnehmern. Der Test soll die Teilnehmer zum Lernen, Verstehen und Behalten des neuen Lernstoffs animieren.

Aufgabe – Problem und Lösung, Frage und Antwort – zusammengefügt, einander zugeordnet werden.

d) Von den Typen mit freier Aufgabenbeantwortung kommt der *Ergänzungs-Aufgabe* (EG-Aufgabe) – auch Schlüsselwortergänzungs- oder Offenantwort-Aufgabe genannt – nach den MW-Aufgaben die größte Bedeutung zu. Die EG-Aufgabe ist die einzige Form der freien Aufgabenbeantwortung, die sogar in standardisierten Tests zulässig ist. Es kommt bei ihr darauf an, eine Aufgabe durch ein Wort (Schlüsselwort) oder eine kurze Darstellung (z. B. ein Symbol, eine Skizze) zu beantworten.

Zusammenfassend kann man sagen, dass die Fragen des *COLUMBUS*-Tests zum einen *Tatsachenwissen* und zum anderen *praktische* und *theoretische Intelligenz* kombinieren. Die erfolgreiche Beantwortung der Testfragen beruht auf Gedächtnisleistung, der Anwendung von Wissen und Kenntnissen und der Einsicht in komplexe Zusammenhänge.

Die Aufgabenbewertung: Wir haben im Test 6 Fragen pro Themenbereich, das macht $6 \cdot 6 = 36$ Fragen insgesamt. Pro Frage können zwischen 0 und 1 Punkt erreicht werden. Und zwar betrachtet man die Anzahl der Antwortmöglichkeiten (*items*) pro Frage. Jedes *item* kann auf die richtige oder falsche Art und Weise beantwortet worden sein. Der maximale Punkwert, die 1, einer Frage wird in soviele *k*-tel unterteilt, wie es *items* für diese Frage gibt, die als „richtig“ hätten markiert weren müssen. Dann gibt es bei richtiger *item*-Markierung je $\frac{1}{k}$, bei falscher *item*-Markierung je $\frac{-1}{k}$ Punkte. Jedoch können nicht weniger als Null Punkte pro Frage erreicht werden.

Beispiel 2.1 (Aufgabenbewertung)
Angenommen wir betrachten eine Frage, welche 5 Antwortmöglichkeiten hat. Die maximale Punktzahl 1 wird erreicht, wenn das 1., 2. und 5. *item* angekreuzt sind.

Fall 1: Proband *p*05 hat *item* $1, 2, 5$ für richtig gehalten und diese angekreuzt. Er erhält $\frac{3}{3} = 1$ Punkt, da er sich bei jedem der 3 relevanten *items* richtig verhalten hat.

Fall 2: Proband *p*09 hat *item* $1, 2$ für richtig gehalten und diese angekreuzt. Er erhält $\frac{2}{3}$ Punkte, da er sich bei zwei der drei relevanten *items* richtig verhalten hat.

Fall 3: Proband *p*10 hat *item* $1, 2, 3$ für richtig gehalten und diese angekreuzt. Er erhält $\frac{1}{3}$ Punkte, da er sich bei *item* $1, 2$ richtig verhalten hat und bei *item* 3 falsch ($\frac{2}{3} - \frac{1}{3} = \frac{1}{3}$).

Fall 4: Proband *p*14 hat *item* 3 für richtig gehalten und dieses angekreuzt. Er erhält 0 Punkte ($\frac{0}{3} - \frac{1}{3} = \frac{0}{3} = 0$).

Ein spezielles Kriterium für die Testaufgaben ist die *Aufgabenschwierigkeit*. Es gibt mehrere Formeln, um diese zu bestimmen. Dabei kommt es darauf an, ob man zum Beispiel zwischen den Aufgabentypen differenzieren möchte (RF-Fragen, MW-Fragen)

oder den Probanden eine gewisse „Raterei“ unterstellt. Da wir das **nicht** tun[52], verwenden wir die folgende Formel zur Bestimmung der Aufgabenschwierigkeit.

Definition 12 (Aufgabenschwierigkeit) *Die Aufgabenschwierigkeit P_j für ein item j berechnet sich durch*

$$P_j = 100 \cdot \frac{N_{Rj}}{N}, \tag{2.16}$$

wobei N die Anzahl der Probanden ist und N_{Rj} die Anzahl der Probanden, die item j richtig beantwortet haben.
Ein hoher Schwierigkeitsindex von knapp 100 *deutet auf eine „leichte“ Aufgabe hin, ein niedriger auf eine „schwere“.*

Betrachten wir die Aufgabenschwierigkeit unseres Beispieldatensatzes `Tabelle`:

```
In[27]:= P[j_] := 100 * (NRe[[j]] / N);
```

Der Name `NRe` steht für eine Liste mit den N_{Rj}-Werten für jedes *item*. Mit der Funktion `P[j_]` entscheiden wir nun für jedes *item*, wie „schwierig“ es ist und färben den Schwierigkeitsindex entsprechend ein. Wir listen in `PTabelle` alle *items* mit ihrer Aufgabenschwierigkeit `P[j_]` auf.

```
In[28]:= PTabelle
```

Out[28]=

item	P
i01	100.
i04	88.88
i05	100.
i06	88.88
i07	100.
i08	77.77
i09	88.88
i10	55.55
i11	88.88
i12	44.44
i13	88.88
i14	33.33

Die Farbeinteilung hat folgende Bedeutung: Grün steht für eine „gute“, das heißt angemessene Schwierigkeit ($P_j \geq 40$), blau signalisiert „keine Aussage“, weil bei diesem *item* alle Probanden richtig geantwortet haben, womit eine Berechnung von $P_j = 100 \cdot \frac{N}{N} = 100$ ergibt und rot steht für eine „zu hohe“ Schwierigkeit des *items* ($P_j < 40$).

Wünschenswert für einen gut konzipierten Testteil ist ein ansteigender Schwierigkeitsverlauf für die sechs Fragen pro Thema des ESA-Tests. Der Proband kann sich zunächst mit sogenannten „Aufwärmfragen“ beschäftigen und hat psychologisch einen

[52] Wir gehen fest davon aus, dass die Teilnehmer im Kurs soviel Wissen aufnehmen konnten, dass sie ein „Raten“ bei den Testantworten nicht nötig haben.

motivierenden, weil nicht zu schwierigen Einstieg in den Themenblock. Gegen Ende kann die Schwierigkeit der Fragen ansteigen, es ist gut zu beobachten, ab welchem Level die jeweiligen Probanden Schwierikeiten bekommen.

Für die Testanalyse sollten wir anhand dieser Schwierigkeitsbetrachtung *item* $i12$ und $i14$ im Auge behalten. Ergibt zum Beispiel später die Validitätskontrolle auch noch einen kritischen Wert für eines dieser *items*, hätten wir schon ein zweites Argument, dieses *item* aus dem Test zu nehmen oder zu verändern. Solche Identifizierungen sind typische Ergebnisse in der Testanalyse.

2.4 Reliabilitätskoeffizient

Grundsätzlich kommen zur Bestimmung der Testreliabilität vier verschieden Methoden in Frage[53]:

1. Die *Testwiederholungsmethode*. Ein Test wird bei einer Stichprobe von Probanden nach einem angemessenen Zeitabstand wiederholt. Die Rohwertpaare aus Test und Testwiederholung werden korreliert. Der Korrelationskoeffizient ist eine Schätzung der Reliabilität.

   Diese Methode eignet sich nicht für die ESA-Tests, da eine Wiederholung von denselben Testfragen bei einem Wissenstest keinen Sinn macht.

2. Die *Paralleltestmethode*. Einer Stichprobe werden zwei Parallelformen eines Tests sofort hintereinander oder in einem gewissen zeitlichen Abstand nacheinander in Zufallsfolge dargeboten. Die Rohwertpaare werden korreliert. Auch hier ist der Korrelationskoeffizient eine Schätzung der Reliabilität.

   In unserem Fall ist es kaum möglich zwei gleichwertige Tests zusammenzustellen. Die Fragen können nicht objektiv nach ihrer Schwierigkeit klassifiziert werden. So könnte der Fall auftreten, dass dem einen Probanden eine spezielle Frage leichter fällt als einem anderen. Es ist also nicht sinnvoll zwei vergleichbare Paralleltests aufzustellen, die erreichten Rohwerte der Personen sind nicht uneingeschränkt vergleichbar.

3a. Die *Testhalbierungsmethode*. Der Test wird bei einer Stichprobe von Probanden einmal durchgeführt. Dann wird der Test in zwei äquivalente Aufgabengruppen aufgeteilt, und die beiden Hälften werden separat ausgewertet. Die Rohwerte der einen Hälfte werden mit den Rohwerten der anderen Hälfte korreliert. Aus dem Korrelationskoeffizienten berechnet man mit speziellen Formeln Schätzwerte für die Reliabilität.

   Eine Aufteilung der Fragen ist für uns nicht sinnvoll. Teilt man jeden der sechs Themenblöcke in der Hälfte ergibt sich ein Problem der Aufgabenschwierigkeit,

[53] [LiRa, S.180]

weil in den Blöcken meist ansteigende Schwierigkeit herscht. Zu der Aufgabenschwierigkeit kommt außerdem noch der Punkt des individuellen Aufgabentyps. Eine RF-Frage ist in der Regel „leichter“ als eine MA-Aufgabe. Genauso ist es problematisch, dass nur so wenige Fragen aufzuteilen sind (36 Stück). Bei dieser geringen Anzahl fallen Extreme bei einzelnen Aufgaben stärker auf.

3b. Die *Konsistenzanalyse*[54]. Der Test wird bei einer Stichprobe von Probanden einmal durchgeführt. Aus den Aufgabenkennwerten[55] erhält man mittels spezieller Rechenformeln wiederum Schätzwerte für die Reliabilität. Diese Methode 3b ist eine ins Extrem gesteigerte Variante der Methode 3a. Wir teilen die Testfragen nicht nur in zwei, sondern in „mehr“ Teile ein (siehe Kapitel 2.4.1).

Die ESA-Tests eignen sich gut für die Reliabilitätsbestimmung auf diesem Wege, betrachten wir eine genauere Beschreibung der Methode 3b.

2.4.1 Methode der Konsistenzanalyse

Diese Methode der Reliabilitätsbestimmung ist die „jüngste“ unter den oben genannten. Sie wurde in den späten 30er Jahren entwickelt und ist dem Halbierungsverfahren (Methode 3a) in vieler Hinsicht überlegen. Die Methode der inneren Konsistenz geht davon aus, dass man einen Test nicht nur in zwei vergleichbare Hälften, sondern in drei, vier oder noch mehr äquivalente Teile, im Extremfall in ebenso viele Teile untergliedern kann, wie Aufgaben bzw. *items* vorhanden sind. Man kann den inneren Kontext der einzelnen Aufgaben bzw. *items* auf diese Weise ebenso gut überprüfen, wie dies für zwei Testhälften möglich ist. In der Literatur [LiRa, S.191 ff] werden zwölf Formeln angegeben, mit denen man die Reliabilität bzw. die innere Konsistenz eines aufgeteilten Tests berechnen kann.

Für diese Formeln gibt es unterschiedliche Voraussetzungen. Die Struktur der ESA-Tests passt zu den Formeln *„α-Koeffizient von CRONBACH (1951)“* und *„KUDER-RICHARDSON-Formula 20 (KR 20)“*.

Definition 13 (α-Koeffizient von CRONBACH (1951)) *Die Definition des CRONBACH α's beruht auf folgender Voraussetzung: Alle c Testteile sind gleich lang und parallel, außerdem messen sie das gleiche Merkmal (dieses Merkmal könnte zum Beispiel „das erlernte Wissen über das COLUMBUS Labor“ sein).*

$$\alpha = \frac{c}{c-1}\left[1 - \frac{\sum_{j=1}^{n}\sigma_j^2}{\sigma_X^2}\right] \tag{2.17}$$

[54][Wiki, Konsistenz] Das Wort **Konsistenz** bedeutet Bestand, Zusammenhalt, Geschlossenheit und In-sich-Ruhen.

[55]Der „Aufgabenkennwert“ eines *items* entspricht der Anzahl der Probanden, die dieses *item* richtig beantwortet haben (salopp gesagt, „der Rohwert des *items*“).

In dieser Formel bedeuten:

$$\begin{aligned} n &= \textit{Anzahl der items} \\ c &= \textit{Anzahl der Teile, in die der Test aufgeteilt wurde} \\ \sigma_X^2 &= \textit{Varianz der Test-Rohwerte X} \\ \sigma_j^2 &= \textit{Varianz der Aufgabenstreuung (siehe Formel (2.18))} \end{aligned}$$

So wie es eine Streuung der Punktwerte eines Tests gibt, so gibt es auch eine Streuung der Punktwerte der einzelnen Aufgaben, die *Aufgabenstreuung*.

Definition 14 (Aufgabenstreuung) *Die Aufgabenstreuung s_j wird aus der Aufgabenschwierigkeit P_j[56] ermittelt:*

$$s_j := \sqrt{p_j \cdot q_j}, \tag{2.18}$$

wobei $p_j = \frac{P_j}{100}$ und $q_j = 1 - p_j$ ist
(s_j wird auch als „Standardabweichung" der Punktwerte einer Aufgabe angesehen).

Die α-Formel ist zur Reliabilitätsschätzung bei weitestgehend homogenen Tests[57] sehr universell anwendbar. Sie kann genutzt werden

a) bei einem Test mit c äquivalenten Teilen gleicher Länge,

b) bei einem Test mit Aufgaben, bei denen es für Lösungen unterschiedlich viele intervallskalierte Punkte gibt (z.B. Persönlichkeitsfragebogen),

c) im Grenzfall, wo der Test in soviele Teile aufgeteilt wird, wie er *items* enthält.

Im letzten Fall geht die Formel von *CRONBACH* in die *KR 20* von *KUDER-RICHARDSON* über.

Definition 15 (KUDER-RICHARDSON-Formula 20) *Die KR 20 ist eine Spezialisierung von CRONBACHs α. Der Reliabilitätskoeffizient r_{rel} ist nach der KR 20 folgendermaßen definiert. Insbesondere bei der dritten Schreibweise erkennt man die Äquivalenz zu Formel (2.17).*

$$\begin{aligned} r_{rel} &= \frac{n}{n-1}\left[\frac{\sigma_x^2 - n \cdot \overline{p \cdot q}}{\sigma_x^2}\right] \\ &= \frac{n}{n-1}\left[\frac{\sigma_X^2 - \sum_{j=1}^{n} p_j q_j}{\sigma_X^2}\right] \\ &= \frac{n}{n-1}\left[1 - \frac{\sum_{j=1}^{n} p_j q_j}{\sigma_X^2}\right] \end{aligned} \tag{2.19}$$

[56] vgl. Formel (2.16) auf S.40

[57] [LiRa, S.36] Die Aufgaben eines Tests können mehr oder weniger *homogen* und im Grenzfall sogar *heterogen* sein. Homogenität bedeutet inhaltliche Einheitlichkeit bei vollkommen erhaltener formaler Unabhängigkeit der einzelnen Aufgaben voneinander.

In dieser Formel bedeuten:

$$\begin{aligned} n &= \textit{Anzahl der Test-items (Aufgaben)} \\ \overline{pq} &:= \frac{\sum_{j=1}^{n} p_j q_j}{n} \textit{, Mittelwert aller } p_j \cdot q_j \\ \sigma_X^2 &= \textit{Varianz der Test-Rohwerte X} \end{aligned}$$

Wir haben mit der *KR 20* also eine Formel, die den Reliabilitätskoeffizienten unseres Tests bestimmt. Berechnen wir diesen doch einmal mit **Mathematica**. Die Varianz **sigma2** haben wir bereits im Beispiel auf S.18 bestimmt.

```
In[29]:= p[i_] := NRe[[i]] / N;
         q[i_] := 1 - p[i];
         Rel = (n / (n - 1)) * (1 - (Sum[p[i] * q[i], i, 1, n] / sigma2))
Out[29]= 0.769
```

Wie im nächsten Abschnitt zu lesen ist, sind wir mit einem Reliabilitätskoeffizienten von $0,769$ sehr zufrieden. Das Gütekrieterium Reliabilität braucht uns bei dem Beispieldatensatz **Tabelle** also keine Sorgen zu machen.

2.4.2 Bewertung und Interpretation

Für den Wert des Reliabilitätskoeffizienten sind anerkannte Normen festgelegt[58]:

- Für die Beurteilung individueller Differenzen gilt der Mindestwert $r_{rel} = 0,7$.
- Standardisierte Tests sollten eine Re- oder Paralleltestreliabilität von $r_{rel} \geq 0,8$ aufweisen.
- Für die Beurteilung von Gruppendifferenzen sind Tests mit einer Reliabilität von $r_{rel} \geq 0,5$ verwendbar.

Für uns trifft Punkt eins zu, da wir individuelle Differenzen bei den *items* herausfiltern wollen. Die Reliabilität der ESA-Tests sollte also $0,7$ oder besser sein. Mit einem solchen Wert kann man den Test weiterhin benutzen. Er misst das gewünschte Kriterium genau genug. Unser Kriterium ist „das Verständnis über die gelernte Information zum *COLUMBUS* Labor“. In der Implementierung bewerten wir den Reliabilitätskoeffizienten mit folgender Skala: Ein $r_{rel} \geq 0,7$ ist „gut“, $0,7 > r_{rel} \geq 0,6$ ist zu überprüfen und $r_{rel} < 0,6$ ist zu ersezten. Das heißt, es wird anhand der Güte des Reliabilitätskoeffizienten entschieden, wie **genau** die *items* des gesamten Tests messen, ob das Lernziel erreicht wurde.

[58] [LiRa, S.269]

2.5 Validitätskoeffizient

Die in Kapitel 2.2.2 auf S.34 gegebene Definition der Validität lässt das eigentliche Problem noch unberührt: Valide ist ein Test, wenn er dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das bzw. die er messen oder vorhersagen soll, tatsächlich misst oder vorhersagt. Wie – so muss die erste Frage lauten – will man nun beurteilen, ob ein Test wirklich das misst, was er messen soll, ob er tauglich und angemessen ist zur Erfassung eines bestimmten Persönlichkeitsmerkmals?

Das Vorgehen bei der Überprüfung der Validität wird als *Validierung* bezeichnet. Erste korrelationsstatistische Ansätze zur Frage der Validitätskontrolle wurden Anfang des 20. Jahrhunderts geschaffen. Leistungsmerkmale wurden untereinander korreliert, von denen das eine als psychologisch hinreichend gekennzeichnet galt und später zum sogenannten Validitätskriterium wurde. Dieser Ansatz zielt auf den heute so genannten, und im Folgenden unter Punkt c) erklärten Aspekt ab, die *kriterienbezogene Validität*[59].

Zur Differenzierung verschiedener Aspekte bei der Validitätsbetrachtung gibt es grundsätzlich drei Unterscheidungen[60]:

a) **Inhaltliche Validität** (content validity). Der Test bzw. seine Elemente sind so beschaffen, dass sie das zu erfassende Persönlichkeitsmerkmal oder die in Frage stehende Verhaltensweise repräsentieren. Mit anderen Worten: Der Test selbst stellt das optimale Kriterium für das Persönlichkeitsmerkmal oder die Verhaltensweise dar. Beispielsweise würde ein Schulkenntnistest dann inhaltlich valide sein, wenn seine Aufgaben inhaltlich eine repräsentative Auswahl aus dem Unterrichtsstoff darstellen. Ebenso ist evident, dass eine Schreibprobe inhaltlich valide ist für die Erfassung der Schnelligkeit und Genauigkeit, mit der eine Sekretärin schreiben kann. Inhaltliche Validität wird einem Test in der Regel durch ein *Rating* von Experten als „Konsens von Kundigen“ zugebilligt.

b) **Konstruktvalidität**. Auf Grund theoretischer – sachlogischer und begrifflicher – Erwägungen und anhand daran anschließender empirischer Untersuchungen wird entschieden, ob ein Test ein bestimmtes Konstrukt zu erfassen vermag. Die Konstruktvalidität zielt direkt auf die psychologische Analyse der einem Test zugrunde liegenden Eigenschaften und Fähigkeiten ab, also auf Beschreibungsmerkmale, die nicht in eindeutiger Weise operational erfassbar, sondern theoretischen Charaker haben, wobei freilich eine empirische Basis gegeben ist. Ein Test etwa, der den individuellen Ausprägungsgrad von Angst messen soll, hätte dann eine hinreichende Konstruktvalidität, wenn nachgewiesen wurde, dass das vom Test erfasste Merkmal in genügender Übereinstimmung mit dem theoretischen Konstrukt „Angst“ steht.

[59] [LiRa, S.220]
[60] [LiRa, S.10]

c) Die **kriterienbezogene** Validität (criterion validity). Während sich im Fall der beiden oben genannten Validitätsaspekte im Allgemeinen keine Maßzahl für den Grad der Validität eines Tests ermitteln und angeben lässt, ist es bei der kriterienbezogenen Validität dadurch möglich, dass man die Testergebnisse einer Stichprobe von Probanden mit einem sogenannten *Außenkriterium* korreliert, das vom Test unabhängig erhoben wird. Man kann entweder direkt von der Testleistung auf die Kriterienleistung schließen, oder man kann das Kriterium als einen ausreichend validen Repräsentanten für das Persönlichkeitsmerkmal, das es zu erfassen gilt, ansehen und so direkt Aussagen über dieses Merkmal treffen.

Für unsere Testform ziehen wir eine Kombination aus inhaltlicher und kriterienbezogener Validität heran. Zum einen erfüllen unsere Testfragen das Kriterium, gute Repräsentanten für den Unterricht zu sein, zum anderen können wir uns aber auch ein Kriterium aus dem Test heraus definieren und mit diesem und den Rohwerten ähnlich zu Methode c) eine Korrelation anstreben. Ein „klassisches“ Außenkriterium für uns wäre zum Beispiel: Astronaut XY hat den *COLUMBUS*-Test fehlerfrei abgeschlossen und er hat sich auf der ISS in der Arbeit mit dem *COLUMBUS*-Raumlabor immer souverän und überlegen verhalten. Doch da das Labor noch in Bremen steht, gibt es leider (noch) kein Außenkriterium. Im `Mathematica` Beispiel am Ende dieses Kapitels sehen wir jedoch eine gute Alternative mit einem *inneren* Kriterium zu arbeiten.

Statistische Methoden zur Ermittlung eines Validitätskennwertes:

a) Die Extremgruppen-Methode
Die Extremgruppenmethode besteht darin, dass man zwei Gruppen von Probanden miteinander vergleicht, von denen die einen das fragliche Persönlichkeitsmerkmal in extrem hohem Grade und die anderen es in extrem niedrigem Grade besitzen: Eine Gruppe habe zum Beispiel das Abitur mit „sehr gut“ bestanden, die andere Gruppe habe es nicht bestanden. In diesem und ähnlichen Fällen des Extremgruppenvergleichs ist das Merkmal in der Stichprobe alternativ und nicht normal verteilt. Für die ESA-Tests ist die Extremgruppen-Methode nicht anwendbar, da wir kein sinnvolles Aufteilungskriterium für die zwei Probandengruppen haben. Wir möchten die Ergebnisse **aller** Teilnehmer eines Tests und **aller** *items* in die Validierung einbringen, da wir ohnehin schon mit kleinen Stichproben arbeiten[61]. Eine Gruppeneinteilung nach „besserer“ und „schlechterer“ Hälfte der Testteilnehmer entspricht aber nicht der Anforderung für die Extremgruppen-Methode, zwei Gruppen mit einem **extrem** unterschiedlich ausgeprägtem Merkmal zu untersuchen.

b) Die Mischgruppen-Methode
Diese Methode gilt für dichotome[62] Tests und dichotome Kriterien. Bei diesem Ansatz geht man von zwei Populationen A und B aus, bei denen das Kriterium unterschied-

[61] Viele statistische und testanalytische Verfahren liefern erst ab Stichproben von 100 bis 200 Probanden verlässliche Ergebnisse.

[62] [DuFremd, dichotom] (gr.: „zweigeteilt“) in Begriffspaare eingeteilt

lich ausgeprägt ist. Es wird die Wahrscheinlichkeit eines positiven Testergebnisses geschätzt[63]. Der Vorteil dieser Methode liegt darin, dass man das Kriterium bei den Probanden nicht messen muss. Das ist zum Beispiel in der Medizin ein Vorteil, wenn man die Kriterien *Herzinfarkt* oder *Selbstmordneigung* untersuchen möchte. Der Nachteil an diesem Verfahren ist aber die große Ungenauigkeit bei der Schätzung. Für unsere Tests kommt die Mischgruppen-Methode nicht in Frage, da wir Messergebnisse in Form der Rohwerte vorliegen haben und uns nicht unnötig mit Schätzfehlern belasten wollen.

Die dritte Methode ist die Repräsentativgruppen-Methode, die im folgenden Kapitel genauer erklärt wird. Sie trifft für die ESA-Tests zu.

2.5.1 Repräsentativgruppen-Methode

Bei dieser Methode werden nicht zwei speziell ausgewählte und hinsichtlich des Validitätskriteriums extrem unterschiedlichen Gruppen miteinander verglichen, sondern der Test wird bei der kompletten Stichprobe durchgeführt. Es gibt in diesem Fall sechs verschiedene Ansätze, den Validitätskoeffizienten zu berechnen[64]. Für die ESA-Tests verwenden wir die folgende Methode, die der Bravais-Pearson-Korrelationsformel[65] entspricht.

Definition 16 (Maßkorrelation) *Liegt im Falle einer Repräsentativgruppenvalidierung das Kriterium quantitativ abgestuft vor, benutzt man die Maßkorrelation zur Berechnung des Validitätskoeffizienten r_{val}. Man bezeichnet konventionsgemäß die Testrohwerte mit X und die Kriteriumrohwerte mit Y.*

$$r_{val} = \frac{N \cdot \sum_{i=1}^{N} X_i Y_i - (\sum_{i=1}^{N} X_i) \cdot (\sum_{i=1}^{N} Y_i)}{\sqrt{\left[N \sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2\right] \cdot \left[N \sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2\right]}} \tag{2.20}$$

Die Testrohwerte X sind in unserem Fall die Rohwerte der Probanden. Als Kriteriumwerte Y wählen wir die Bewertung der einzelnen *items* - null oder eins. Damit ist das Kriterium quantitativ abgestuft und wir können den Validitätskoeffizienten für jedes *item* separat untersuchen. Für unseren Beispieldatensatz `Tabelle` bekommen wir folgende Ergebnisse mit `Mathematica`. Dabei gehen wir *item*weise vor, berechnen für jedes *item* den Zähler und Nenner des Maßkorrelationskoeffizienten getrennt[66] und schreiben schließlich den jeweiligen Wert der Maßkorrelation für die *items* in eine Tabelle[67].

[63] zur Vertiefung: [LiRa, Kap.11.6.2]

[64] zum Weiterlesen: [LiRa, S.245]

[65] (1.11), S.27

[66] Wenn alle Kriteriumrohwerte gleich sind, steht in der Maßkorrelation $\frac{0}{0}$, daraus macht `Mathematica` eine Fehlermeldung. Bei getrennter Zähler-und Nennerbetrachtung definieren wir einfach „von Hand“, dass $\frac{0}{0} = 0$ ist.

[67] Die einzelnen Berechnungsschritte werden in Kapitel 3.1.2 genauer erklärt.

```
In[30]:= ValTab
```

item	v
i01	0
i04	0.9
i05	0
i06	0.9
i07	0
i08	0.5
i09	-0.07
i10	0.64
i11	0.9
i12	0.59
i13	0.9
i14	0.47

```
Out[30]=
```

Die Bedeutung der Farbeneinteilung ist folgende: Grün steht für „gut“ ($r_{val} \geq 0,5$), orange steht für „zu prüfen“ ($0,5 > r_{val} \geq -0,1$) und rot steht für „sollte ersetzt werden“ ($r_{val} < -0,1$).

2.5.2 Bewertung und Interpretation

Anders als bei dem Reliabilitätskoeffizienten lässt sich für den Validitätskoeffizienten keinerlei starre Norm angeben. Man orientiert sich an Richtlinien, die aber zum Teil größeren Schwankungen unterliegen können. Als statistisch nicht überaus gut, aber für die Praxis ausreichend wird ein Wert von $0,5$ oder höher angesehen. An dieser Stelle sei noch einmal darauf hingewiesen, dass bei unserer geringen Anzahl von Probanden pro Test ohnehin die Korrektheit mancher statistischer Untersuchungen vermindert sein kann. Zuverlässige Ergebnisse erhält man in der Regel ab Stichprobengrößen von 100 Probanden.

Betrachten wir ein paar allgemeine Richtlinien über die erforderliche Höhe des Validitätskoeffizienten[68]:

- Ein Test muss in einem solchen Umfang valide sein, dass seine Anwendung eine bessere Voraussage ermöglicht als seine Unterlassung!
  Diese „Selbstverständlichkeit“ sollte sich der Tester immer wieder vor Augen führen. Besonders dann, wenn ein Validitätskoeffizient keine besondere Höhe aufweißt, die Probanden aber unter diversen Außenkriterien zeigen, dass ein Lernerfolg stattgefunden hat.

- Höhere Validität muss von einem Test oder seinen *items* dann gefordert werden, wenn von dem Testresultat eine wichtige Entscheidung für die Probanden abhängt, das heißt, wenn sich ein Fehlurteil für sie sehr nachteilig auswirken würde.
  Für die ESA-Tests ist ein eventuelles Fehlurteil und die daraus resultierende

[68] [LiRa, S.270]

Nachprüfung nicht so dramatisch wie zum Beispiel in einer Abiturklausur, durch welche zum Teil „Weichen für die Zukunft“ gestellt werden. Trotzdem sollte auch beim *COLUMBUS*training nach den Ursachen eines niedrigen Validitätskoeffizienten geforscht werden.

- Der Validitätskoeffizient ist auch danach zu beurteilen, ob das durch den Test geprüfte Persönlichkeitsmerkmal ohne Verwendung von Testverfahren leicht oder schwer zu erfassen ist.
  Dieser Punkt ist für uns zur Zeit dadurch zu beantworten, dass ohne den praktischen Einsatz der Astronauten im *COLUMBUS*labor nur eine Rücksprache mit den Probanden Aufschlüsse über einen Lernerfolg geben kann oder das Training am Simulator. Die Überprüfung des Kriteriums „Wissenszuwachs“ ist daher ohne Verwendung eines Testverfahrens schwer möglich.

Betrachten wir anhand der gegebenen Farbeinteilung die Validitätskoeffizienten für unser Beispiel `Tabelle`, so fallen *item* *i*09 und *i*14 als „zu prüfen“ auf. Die *items* *i*01, *i*05 und *i*07 weisen den Wert Null auf, da sie von **allen** Probanden richtig beantwortet wurden. Erinnern wir uns an die Ergebnisse der Schwierigkeitsanalyse auf S.40, so wird deutlich, dass *i*14 nun schon zum zweiten Mal negativ auffällt. Das wäre ein Fall in der Testanalyse, wo man dieses *item* auf Fehler zu untersuchen hätte. Es sollte geprüft werden ob ein Druck- oder Schreibfehler im *item* vorliegt, im Unterricht das betreffende Thema vergessen oder ungenügend erklärt wurde, man könnte prüfen ob die Aussage womöglich umständlich – und damit vielleicht verwirrend – aufgeschrieben wurde oder ob sonst irgendwelche Auffälligkeiten damit in Verbindung zu bringen sind.

3 Details zur Implementierung

Die Umsetzung der Testanalyse, abgestimmt auf die Bedingungen der ESA-Tests, wurde vollständig mit der **Mathematica** Version 5.2 durchgeführt. Zum einen haben wir die individuellen Eckdaten dieses Tests in Kapitel 2.3.1 auf S.35 betrachtet. Zum anderen zählen zu den Bedingungen, nach denen das Testanalyse*tool* gestaltet wurde, auch die Bedienungswünsche von Herrn Dr. Seine.

Folgende Eigenschaften in der Bedienung sind wichtig:

- Möglichst einfache und schnelle Handhabung bezüglich Öffnen, Starten und Speichern der Dateien.
- Leicht überschaubare und einfach strukturierte Benutzeroberfläche (das Gegenteil dazu wären Seiten voller Quellcode).
- Einfache und für jeden Benutzter verständliche Ausgabe (zum Beispiel: farblich an einer Skala eingestufte Werte oder Antwortsätze anstatt nur eine Ausgabe von „Zahlenkolonnen“).
- Kein großer Aufwand bei der Erklärung und Einarbeitung neuer Mitarbeiter, die das Testanalyse*tool* verwenden sollen.

3.1 Aufbau

Der Aufbau der Implementierung ist zweigeteilt. Zum einen geschieht der Programmaufruf über ein **Mathematica** Notebook mit dem Namen „testanalyse.nb“. Auf der anderen Seite wird von diesem Notebook aus ein **Mathematica** Package geladen, dass das Testanalyse*tool* in Form einer GUI[69] startet. Dieses **Mathematica** Package heißt “tool.m“ und für seinen Aufruf wird mit **Needs[“GUIKit`“];** das **Mathematica** Package für GUIs geladen. Dieses gibt es ab der Version 5.1. Die Ausgabe des Testanalyse*tools* wird wieder in „testanalyse.nb“ ausgegeben und kann dort unter einem sinnvollen Namen bezüglich der eingelesenen und analysierten Daten abgespeichert werden. Es ist also wichtig, immer eine „Blanko-Version“ von „testanalyse.nb“ zu behalten, um immer wieder neue Testanalysen starten zu können.

3.1.1 GUI

Der englische Ausdruck „Graphical User Interface“, kurz GUI, bedeutet wörtlich übersetzt „grafische Benutzerschnittstelle“. Im deutschen Sprachgebrauch hat sich jedoch neben „GUI“ der Begriff „grafische Benutzeroberfläche“ durchgesetzt. Eine grafische

[69] mehr über GUIs im folgenden Abschnitt 3.1.1

Benutzeroberfläche ist eine Softwarekomponente, die einem Computerbenutzer die Interaktion mit der Maschine über grafische, metapherhafte Elemente[70] (Arbeitsplatz, Symbole, Papierkorb, Menü) unter Verwendung eines Zeigegerätes (wie einer Maus) erlaubt[71].

Wir sehen uns eine kleine GUI an, um uns mit der Grundstruktur vertraut zu machen.

```
In[31]:= Needs["GUIKit`"];
         GUIRun[Widget["Panel",Widget["Button","text"->"ButtonA"]]]
```

Out[31]=

Die Prozedur **GUIRun[]** startet den GUI-Modus, den wir zuvor mit dem Package **Needs[“GUIKit`“];** geladen haben. Der Befehl **Widget[]** legt einen Rahmen an, in den wir **Panel** legen. Dieses kann nun mit Input gefüllt werden. Wir legen ein erneutes **Widget** mit der Form **Button** an. Diesen **Button** beschriften wir mit Text.

Die GUI-Gebilde lassen sich nach Belieben erweitern. In der **Mathematica**-Hilfe gibt es viele Beispiele zu allen möglichen GUI-Optionen. Zum Beispiel kann man neben Buttons auch Häkchenkästchen (Check-Box), Schieberegler und Eingabefelder anlegen, man kann Felder beschriften, aufklappbare Menüs bauen, Rahmen definieren, Farben vergeben, Funktionen hinter Action-Felder legen und mit Hilfe von sogenannten **Wizards** Verlinkungen zwischen GUIs schaffen.

Im Testanalyse*tool* „tool.m“ finden wir eine einseitige GUI mit einer Liste von Check-Box-Feldern, die durch Mausklick mit Häkchen markiert werden können. Es gibt keine Links zu anderen Seiten, jeder Haken aktiviert eine Berechnung, der Output wird in das Notebook „testanalyse.nb“ geschrieben, von dem aus „tool.m“ gestartet wird.

Der Grund für die Sprachwahl der GUI in Englisch liegt in der Tatsache, dass viele Mitarbeiter des EAC-Teams nicht aus Deutschland kommen. Für einen flexiblen Einsatz des *tools* ist es deshalb nötig, die Sprache allgemeinverständlich zu wählen.

[70] [Wiki, Metapher] Die **Metapher** (griechisch „Übertragung“) ist eine rhetorische Figur, bei der ein Wort nicht in seiner wörtlichen Bedeutung, sondern in einer übertragenen Bedeutung gebraucht wird, und zwar so, dass zwischen der wörtlich bezeichneten Sache und der übertragen gemeinten eine Beziehung der Ähnlichkeit besteht. Beispiel: leeres Stroh dreschen - inhaltslos reden

[71] [Wiki, GUI]

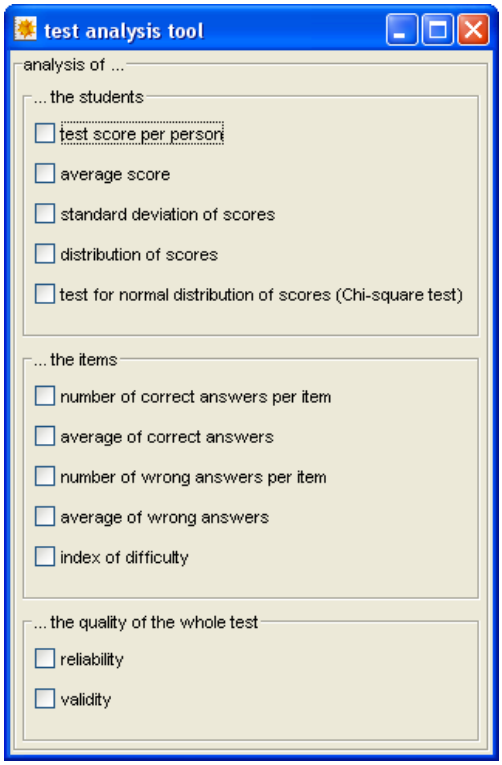


Abbildung 7: Die GUI zur Testanalyse.

Der Teil in „tool.m“ von Zeile 1 bis Zeile 112 legt die Struktur der GUI an[72]. In Zeile 113 bis 136 wird der Datensatz importiert und es werden alle *items* gelöscht, in denen mindestens ein „a“ auftaucht (das Problem haben wir in Kapitel 1 auf S.11 besprochen). Danach kommen die Rechenfunktionen, die bei Betätigung der Check-Box Häkchen ablaufen (Zeile 137 bis 502).

Wir werden im folgenden Kapitel die Implementierung der einzelnen Rechenfunktionen betrachten. Der Aufbau der GUI kann aus Platzgründen nicht detaillierter vorgestellt werden. Umfassende Informationen und Beispiele zum Bau und Aufbau von GUIs sind jedoch in der `Mathematica`-Hilfe ab der Version 5.1 nachzulesen.

3.1.2 Funktionen

Wir betrachten nun die Funktionen in der gegebenen Reihenfolge. Der Anwender kann beliebige Feldern markieren. Eine Berechnung findet nur bei markierten Feldern statt. Die Reihenfolge der Markierung spielt keine Rolle. Für die ersten zwei Punkte der folgenden Gliederung existieren keine Häkchenfelder in der GUI. Die Aktionen werden am Anfang jedes GUI-Aufrufs automatisch vollzogen.

[72] Vergleiche hierzu den Quellcode von „tool.m“ im Anhang auf Seite 77

Die Zeilenangabe zu jeder Funktion gibt die Stelle im Quellcode des Anhangs „tool.m“ an, wo diese Funtion implementiert ist. Ein Beispiel für jede Ausgabe der Funktionen ist in Kapitel 3.3 aufgeführt.

import *(Zeile 116-120)* Um den gewünschten Datensatz zu importieren, muss im Quellcode von „tool.m“ für jede Analyse der entsprechende Pfad zum Ordner der Datei eingegeben werden. Mit dem **Mathematica** Befehl „Input – Get file path...“ aus der Menüleiste kann der Pfad zum Ordner der Datei gesucht werden. Dieser ist im Quellcode in Zeile 116 einzusetzen:

```
path="D:\\Uni\\Bachelorarbeit";
```

Nun wird die „Arbeitsumgebung“ auf diesen Pfad gesetzt. Wir arbeiten mit sogenannten *relativen* Pfaden: `SetDirectory[path];`

Im nächsten Schritt folgt der Import der gewünschten Datei. Diese hat die Namensendung „txt“ und interpretiert Zellen anhand von Tabs[73].

```
Tabelle:=Import["BeispielTabelle.txt","TSV"];
```

Bei dem `Import` Befehl steht das Format `"TSV"` für ein Tabellenformat, welches importiert werden soll.

delete *items* **which include „a“** *(Zeile 124-135)* Die importierte Tabelle wird transponiert, da im Excel- bzw. txt-Format die *items* in Zeilen angeordnet waren. Das hat den Grund, dass eine Exceltabelle eine höhere maximale Zeilen- als Spaltenanzahl hat.

```
Tabelle=Transpose[Tabelle];
```

Im nächsten Schritt werden die Positionen von auftretenden „a’s“ bestimmt.

```
as=Position[Tabelle,"a"];
```

Dann löschen wir die Spalten der *items*, in denen ein „a“ vorkommt.

```
Del=Partition[Union[Table[as[[i]][[2]],
{i,Length[as]}]],1];
Tabelle=Transpose[Delete[Transpose[Tabelle],Del]];
```

Nun werden die Anzahl der Zeilen und Spalten festgestellt, wir bekommen $N+1$ (Anzahl der Probanden plus eins) und $n+1$ (Anzahl der *items* plus eins) und bezeichnen diese mit x und y um mit **Mathematica** nicht in Konflikte bezüglich des Befehls **N[]** zu kommen.

```
x=First[Dimensions[Tabelle]];
y=Last[Dimensions[Tabelle]];
```

Die Namen der Personen werden in einer Liste abgelegt.

```
PersID=Tabelle[[Range[2,x],1]];
```

[73] Im Tabellenprogramm Excel können Tabellen im Textformat unter „name.txt“ abgespeichert werden.

Anschließend werden Funktionen definiert, an Hand derer auf die Namen und Werte der *items* zugegriffen werden kann.

```
item[i_]:=Tabelle[[1,i]];
values[i_]:=Tabelle[[Range[2,x],i]];
```

Um die Laufzeit der GUI zu beschleunigen, wird schließlich eine Tabelle aus Listen erstellt, die nur noch die Nullen und Einsen enthält.

```
valueList=Table[values[i],{i,2,y}];
```

Nun kommen wir zu den Berechnungen zu jeder Option der GUI. Die Sichtbarkeit einiger Variablen und Funktionen ist in „tool.m“ *global*. Sie sind in den Zeilen 165 bis 212 angelegt. Für die Auflistung in diesem Kapitel werden sie jedoch den einzelnen Funktionen zugeordnet, was ausschließlich dem besseren Verständnis und der Lesbarkeit dient.

test score per person *(Zeile 216-226)* Der Rohwert der Personen wird anhand einer Zeilensumme bestimmt. Die Summe läuft nur bis $(y - 1)$, da wir an dieser Stelle aus Schnelligkeitsgründen über die statische Liste von Listen **`valuelist`** summieren, die nur noch Größe $(x - 1) \times (y - 1)$ $(= N \times n)$ hat, da die erste Spalte mit den Personen und die erste Zeile mit den *items* entfernt wurden.

```
RWPers[p_]:=Sum[valueList[[i,p]],{i,1,y-1}];
```

Die Rohwerte für alle Personen kommen in eine Liste.

```
RWe=Table[RWPers[p],{p,1,x-1}];
```

Aus dieser Liste und der `PersID` Liste wird für die Ausgabe der GUI eine Tabelle mit zwei Spalten gemacht.

average score *(Zeile 227-237)* Für den durchschnittlichen Rohwert – berechnet nach der Formel (1.1) von S.13 für das Arithmetische Mittel – bilden wir die Summe über die Rohwertliste und dividieren durch die Anzahl der Personen.

```
RWQuer=N[Sum[RWe[[p]],{p,1,x-1}]/(x-1)];
```

standard deviation of scores *(Zeile 238-248)* Die Berechnung der Standardabweichung erfolgt analog zu Formel (1.5) auf S.17.

```
RWSA=Sqrt[(Sum[(RWe[[p]]-RWQuer)^2,{p,1,x-1}])
/(x-1)];
```

distribution of scores *(Zeile 249-300)* Bezüglich der Rohwertverteilung gibt die GUI zwei Graphiken aus. Die eine ist „selbst gebaut“, die andere mit der eingebauten **`Mathematica`** Funktion **`BoxWhiskerPlot[]`** erstellt. Mit der ersten Graphik zeigen wir das Arithmetische Mittel, die Standardabweichung und die Häufigkeitsverteilung der Rohwerte in Form eines Balkendiagramms in einer gemeinsamen Ausgabe. Bei dem Box-Whisker-Plot werden der Median, das *obere* und *untere Quartil*, der obere und untere Whisker und eventuelle Ausreißer dargestellt[74]. Die Details der Implementierung können im Anhang eingesehen werden, eine ausführliche Darstellung des Quellcodes an dieser Stelle ist zu unübersichtlich. In Kapitel 3.3 sind beide Graphiken für den Beispieldatensatz **`Tabelle`** zu sehen.

[74] vergleiche hierzu Kapitel 1.3, S.20.

test for normal distribution of scores (Chi-square test) *(Zeile 301-386)* Der in Kapitel 1.5 vorgestellte Chi-Quadrat-Test prüft, wie gut die Anpassung der Rohwerte an eine Normalverteilung gegeben ist. Bei der Implementierung wird analog zu den genannten Schritten von Seite 24 vorgegangen:

Algorithmus: Chi-Quadrat-Test

Eingabe: Rohwertliste `RWe`, Anzahl der *items* $(y-1)$
Ausgabe: Informationen über die Anpassung an die Normalverteilung
Einteilung in 10 *bis* 15 *Rohwertklassen*

```
1  Klassen K = 0;
2  Intervallbreite der Klassen h = 0;
3  Rest R = 0;
4  Für j = {10, 11, 12, 13, 14, 15}
5      Wenn Mod[y − 1, j]== 0,
6          dann K = j, h = (y − 1)/ j,
7      ansonsten K = 10, R = Mod[y − 1, 10], h = IntegerPart[(y − 1)/10].
8  RWKlasse[1]:= Join[0, h + R];
9  RWKlasse[i_]:= Join[{RWKlasse[i − 1][[2]]+1}, {RWKlasse[i − 1][[2]]+h}];
10 RWKlassen = MatrixForm[Table[RWKlasse[k], {k, 1, K}]];
```

Klassenmitte bestimmen

```
11 KlM=Table[N[(RWKlassen[[1, k]][[1]]+RWKlassen[[1, k]][[2]])/2], {k, 1, K}];
12
```

Häufigkeiten den Klassen zuordnen

```
13 Für p = 1, ..., x − 1
14     Setze AnzKl[k_]:= 0;
15     Für k = 1, ..., K
16         Wenn RWe[[p]]>=RWKlasse[k][[1]] und RWe[[p]]<=RWKlasse[k][[2]],
17             dann AnzKl[k]+ = 1,
18         ansonsten AnzKl[k]]
19 Die Liste mit den Häufigkeiten nennen wir HF.
20
```

Die zwei Spalten von `RWKlassen` *jeweils einzeln*

```
21 A=Table[RWKlassen[[1, k]][[1]], {k, 1, K}];
22 B=Table[RWKlassen[[1, k]][[2]], {k, 1, K}];
23 F=Table[KlM[[k]]−RWQuer, {k, 1, K}];
24 z=Table[F[[k]]/RWSA, {k, 1, K}];
```

Die Normalverteilungsfunktion

```
25 f[x_]:= (1/Sqrt[2Pi])*\[ExponentialE] hoch ((x^2)/(−2));
26 Y=Table[N[f[z[[k]]]], {k, 1, K}];
27 fe=Table[((h ∗ (x − 1))/RWSA)*Y[[k]], {k, 1, K}];
```

Klassen zusammenfassen, deren Häufigkeitswert < 5 *ist*

```
28  hf=HF;
29  AB=Transpose[Join[{A}, {B}]];
30  Setze k = 1;
31  Tue bis (Length[hf]−1):
32      Wenn hf[[k]]< 5,
33          dann AB[[k, 2]]=AB[[k + 1, 2]];
34          AB=Delete[AB, k + 1];
35          hf[[k]]+ =hf[[k + 1]];
36          hf=Delete[hf, k + 1];
37          fe[[k]]+ =fe[[k + 1]];
38          fe=Delete[fe, k + 1];
39      ansonsten k = k + 1.
```

Das letzte Listenelement der `hf` *überprüfen*

```
40  Wenn Last[hf]<5,
41      dann k =Length[hf];
42      AB[[k − 1, 2]]=AB[[k, 2]];
43      AB=Delete[AB, k];
44      hf[[k − 1]]+ =hf[[k]];
45      hf=Delete[hf, k];
46      fe[[k − 1]]+ =fe[[k]];
47      fe=Delete[fe, k];
48
```

Anhand der evt. reduzierten Listen `hf` *und* `fe` *Chi-Quadrat nach Formel (1.9) bestimmen*

```
49  ChiQua=Sum[((hf[[j]]−fe[[j]])^2)/fe[[j]],{j, 1,Length[fe]}];
```

`Mathematica` *Funktion um die Quantile der* χ^2*-Verteilung zu bestimmen*

```
50  ChiQuantile=If[(df−3) > 0,
    Quantile[ChiSquareDistribution[df−3],0.95]];
51
```

Anpassung an die Normalverteilung prüfen

```
52  Wenn NumberQ[ChiQuantile],
53      dann
54      Wenn ChiQua<=ChiQuantile,
55          dann Print[“OK, the score is nearly normal distributed (level of
            significance = 5% ).“],
56      ansonsten Print[“The scores are not covered by the Normal Distribution
        (level of significance = 5% ).“].
57  ansonsten Print[“It’s not possible to calculate the Quantile of
    Chi-Square-Distribution, because the degree of freedom is too low (zero).“].
58
```

Rückgabe: `ChiQua`, Text

number of correct answers per item *(Zeile 387-394)* Die Anzahl der richtigen Antworten pro *item* errechnen sich aus der Spaltensumme über die Nullen und Einsen des jeweiligen *items*:

```
NR[i_ ]:=Sum[valueList[[i,p]],{p,1,x-1}];
```

Die komplette Liste aller NCs wird erstellt, um diese für die GUI zusammen mit den Namen der *items* in einer zweispaltigen Tabelle auszugeben:

```
NRe=Table[NR[i],{i,1,y-1}];
```

average of correct answers *(Zeile 395-405)* Der Durchschnitt über die Liste NRe:

```
NRQuer=N[Sum[NRe[[i]],{i,1,y-1}]/(y-1)];
```

Für die Berechnung der GUI Felder **number of wrong answers per item** und **average of wrong answers** verfährt man entsprechend ähnlich wie bei den letzten beiden Berechnungen.

index of difficulty *(Zeile 425-449)* Der Schwierigkeitsindex aus der Definition von Seite 40 wird hier genauso angelegt:

```
P[i_]:=100*(NRe[[i]]/(x-1));
```

Für die GUI Ausgabe wird wieder eine zweispaltige Tabelle erstellt mit den Namen der *items* und den dazugehörigen P[i]. Außerdem gibt es eine Graphik, die den Schwierigkeitsverlauf abbildet. Diese Graphik zeigt auf der X-Achse die *items* und auf der Y-Achse die Schwierigkeit. Geht man nach den sechs Themenblöcken von links nach rechts durch die Graphik, ist es wünschenswert für jedes Thema einen ansteigenden Schwierigkeitsverlauf[75] zu beobachten. Siehe auch hierzu das Beispiel in Kapitel 3.3.

reliability *(Zeile 450-473)* Die Reliabilität des Tests bestimmen wir anhand der *KR 20* von Seite 43. Dazu benötigen wir unter Anderem die Varianz der Rohwerte. Diese berechnen wir als Quadrat der Standardabweichung, wie in der Definition in Kapitel 1.2 erläutert:

```
RWVar=RWSA^2;
p[i_]:=NRe[[i]]/(x-1);
q[i_]:=1-p[i];
```

An dieser Stelle genügt uns ein Reliabilitätskoeffizient mit zweistelliger Genauigkeit. In **Mathematica** Notebooks verwenden wir gewöhnlich die Notation **N**[Rel, 2], um den Wert von Rel mit zweistelliger Genauigkeit anzugeben. Warum diese Funktion in dem **Mathematica** Package „tool.m“ nicht funktioniert, ist unklar. So geschieht diese Reduzierung auf zwei Nachkommastellen alternativ mit der folgenden Prozedur:

```
Rel=N[IntegerPart[((y-1)/((y-1)-1))*
(1-(Sum[p[i]*q[i],{i,1,y-1}]/RWVar))*100]/100];
```

[75] Das heißt, man wünscht ein abfallende Kurve in der Graphik, da P nah an 100 für „leicht“ steht und die Beschriftung der Y-Achse im Ursprung bei P „schwer“ beginnt.

validity *(Zeile 474-499)* Zur Bestimmung der Validität der *items* berechnen wir, wie schon in Kapitel 2.5.1 erwähnt, den Zähler und Nenner des Bruchs der Maßkorrelation getrennt.
Zähler:

```
rtc1[i_]:=(x-1)*
Sum[RWe[[p]]*valueList[[i, p]],{p,1,x-1}]-
(Sum[RWe[[p]],{p,1,x-1}]*
Sum[valueList[[i,p]],{p,1,x-1}]);
```

Nenner:

```
rtc2[i_]:=Sqrt[((x-1)*Sum[RWe[[p]]^2,{p,1,x-1}]-
Sum[RWe[[p]],{p,1,x-1}]^2)*
((x-1)*Sum[valueList[[i,p]]^2,{p,1,x-1}]-
Sum[valueList[[i,p]],{p,1,x-1}]^2)];
```

Im nächsten Schritt wird eine Liste mit den einzelnen Validitätskoeffizienten erstellt:

```
Validit=Table[If[rtc1[w]==0&&rtc2[w]==0,0,
N[IntegerPart[N[rtc1[w]/rtc2[w]]*100]/100]],
{w,1,y-1}];
```

Alle weiteren Einzelheiten der Implementierung sind dem Anhang zu entnehmen.

3.2 Probleme

Im Laufe der Entwicklungszeit des gesamten Algorithmus für das Testanalyse*tool* tauchten verschieden Schwierigkeiten auf.

Es gab zum Beispiel Probleme eingebaute **Mathematica** Packages in der GUI zu verwenden. Die Lösung dieser Schwierigkeit besteht in einer Deklaration des benötigten Packages zu Beginn der Implementierung. Schon im Aufbau der GUI Struktur lädt man die speziellen Packages und nicht erst später bei der Implementierung der Berechnungsfunktion.

Dieses Beispiel zeigt den Aufbau der Check-Box für die Rohwertverteilung, in der die Packages für den späteren Gebrauch bereits geladen werden[76]:

```
Widget["CheckBox", {"text" -> "distribution of scores",
Script[Needs["Statistics`DataManipulation`"];
Needs["Graphics`Graphics`"];
Needs["Statistics`StatisticsPlots`"];],
BindEvent["action",
Script[Distr[]]]}],
WidgetAlign[]
```

[76] *Zeile 025-032* des Anhangs

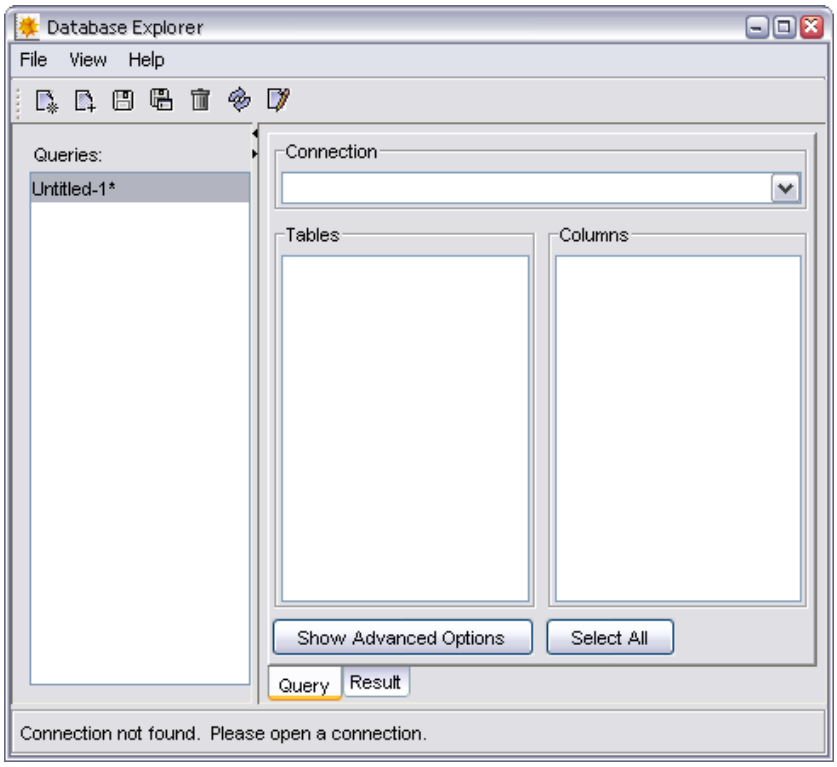
Ein Problem der Rechenzeit tritt auf, wenn der Zugriff für sämtliche Berechnungen über einzelne Matrixeinträge geschieht. Wesentlich schneller läuft hingegen der Zugriff auf statisch abgelegte Listen ab. Würde man zum Beispiel für den gesamten Testanalyse Algorithmus immer wieder die Rohwerte der Probanden als Zeilensumme aus der importierten **`Tabelle`** berechnen lassen, dauert das bei steigender *Item*anzahl erheblich länger, also der Zugriff auf die **einmalig** berechnete Liste `RWe`.

Ein ungelöstes Problem des Testanalysealgorithmus ist die Verbindung mit einer Datenbank. Es wäre wünschenswert, die Null / Eins-Informationen über das Abschneiden der Probanden bei den *items* **direkt** aus einer Datenbank nach **`Mathematica`** importieren zu können. Auf diesem Weg würde der „Umweg“ über eine (Excel-) Tabelle umgangen. In der **`Mathematica`**-Hilfe gibt es unter dem Stichwort „database“ Informationen zu den Verbindungsmöglichkeiten. Unter den Beispielen für GUIs ist sogar ein Datenbank Explorer angegeben, der mit den GUI Funktionen erstellt wurde. Doch reichen diese Informationen nicht aus, eine entsprechende Importfunktion für die GUI der Testanalyse zu erstellen.

Das Beipiel kann aufgerufen werden mit dem Befehl:

```
In[32]:= Needs["DatabaseLink`"];
         DatabaseExplorer[];
```

Out[32]=



3.3 Beispiel

Nachdem wir in Kapitel 3.1.2 die Implementierungsschritte der Check-Box Häkchen des Testanalyse*tools* betrachtet haben, wollen wir uns nun die jeweilgen Ausgaben ansehen. Bei einigen Ausgaben wird zu den errechneten Werten auch die verwendete Formel abgebildet. Das soll das Verständnis der Benutzer fördern, sie können einschätzen, auf welchen Berechnungen die Ausgabe beruht. Die Reihenfolge der Check-Boxen wird weiterhin beibehalten. Aus Formatierungsgründen ist bei einigen Ausgaben nicht die komplette Breite abgebildet.

Das ist der „Kopf“ des Notebooks „testanlyse.nb“. Die Zellen sind der Reihe nach auszuführen, dann startet die GUI, in der die Häkchen zu setzen sind.

- **to execute each cell click into it and push Shift+Enter**

```
Needs["GUIKit`"];
```

```
path = "D:\\Uni\\Bachelorarbeit";
SetDirectory[path];
```

```
Rechner := GUIRunModal["tool.m"]
```

```
Rechner
```

Wir sehen die zweispaltige Tabelle zur Ausgabe der Personen und ihrer Rohwerte.

Test score per person (X)

"pers"	X
"p01"	12
"p02"	11
"p03"	11
"p04"	10
"p05"	10
"p06"	10
"p07"	9
"p08"	9
"p09"	4

Die Ausgabe von Durchschnitt und Standardabweichung der Rohwerte:

Average score (X cross)

```
9.55`
```

$$\bar{X} = \frac{\sum_{p=1}^{N} X_p}{N}$$

N = number of persons

X_p = score of person p

Standard deviation of scores (sigma)

```
2.16`
```

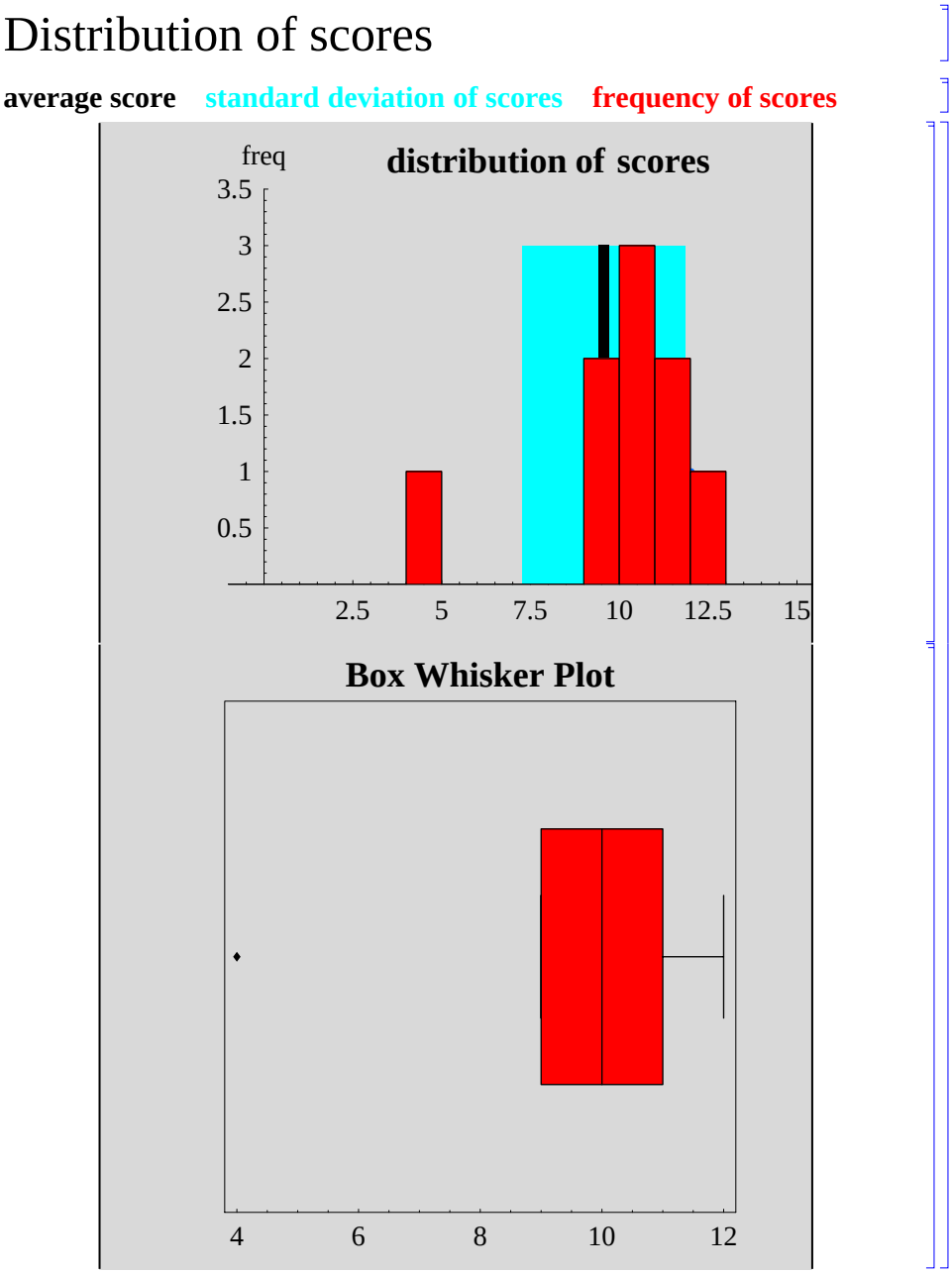
$$\sigma = \sqrt{\frac{\sum_{p=1}^{N} (X_p - \bar{X})^2}{N}}$$

N = number of persons

X_p = score of person p

$$\bar{X} = \frac{\sum_{p=1}^{N} X_p}{N}$$

Zwei Graphiken zur Veranschaulichung der Rohwertverteilung.



Beim Chi-Quadrat-Test wird zuerst die Aussage bezüglich der Normalverteilungsanpassung ausgegeben, im Anschluss der Wert von χ^2.

Test for normal distribution of scores (Chi-square test)

```
"It's not possible to calculate the
  Quantile of Chi-Square-Distribution,   because
  the degree of freedom is too low (zero)."
```

```
0.09`
```

$$\chi^2 = \sum_{c=1}^{C} \frac{(fo_c - fe_c)^2}{fe_c}$$

fo_c = noticed frequency in score catego
(should be ≥ 5)

$fe_c = \frac{hN}{\sigma} \cdot y_c$
expected frequency in score categ
(because of normal distribution h

h = spectrum of category

$y_c = \frac{e^{-\frac{z_c^2}{2}}}{\sqrt{2\pi}}$

$z_c = \frac{X_c - \bar{X}}{\sigma}$

X_c = average of score category c

C = divide into 10-15 categories

Die Anzahlen der richtigen und falschen Antworten pro *item* und der jeweilige Durchschnitt:

Number of correct answers per item (NC)

item	NC
"i01"	9
"i04"	8
"i05"	9
"i06"	8
"i07"	9
"i08"	7
"i09"	8
"i10"	5
"i11"	8
"i12"	4
"i13"	8
"i14"	3

Average of NC

```
7.16`
```

$$\bar{N}_C = \frac{\sum_{i=1}^{n} N_{Ci}}{n}$$

n = number of items

N_{Ci} = number of correct answers per it

Number of wrong answers per item (NW)

item	NW
"i01"	0
"i04"	1
"i05"	0
"i06"	1
"i07"	0
"i08"	2
"i09"	1
"i10"	4
"i11"	1
"i12"	5
"i13"	1
"i14"	6

Average of NW

```
1.83`
```

$$\bar{N}_W = \frac{\sum_{i=1}^{n} N_{Wi}}{n}$$

n = number of items

N_{Wi} = number of wrong answers per ite

Die Aufgabenschwierigkeit in Tabellenform und Graphik:

Index of difficulty per item (P)

good **no action** **replace**

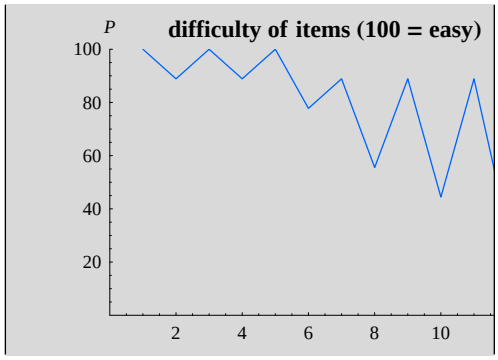
item	P
"i01"	100.`
"i04"	88.88`
"i05"	100.`
"i06"	88.88`
"i07"	100.`
"i08"	77.77`
"i09"	88.88`
"i10"	55.55`
"i11"	88.88`
"i12"	44.44`
"i13"	88.88`
"i14"	33.33`

$$P_i = 100 \cdot \frac{N_{Ci}}{N}$$

high (100) = very easy
low (0) = very difficult

N = number of persons
N_{Ci} = number of correct answer

Curve of difficulty



Bei der Graphik ist schön zu erkennen, wie die Schwierigkeit der *items* gegen Ende ansteigt.

Der Reliabilitätskoeffizient, berechnet mit der *KR 20* von Seite 43:

Reliability of test (r)

good **check** **replace**

0.76`

„*KUDER – RICHARDSON –*

$$r = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} p_i q_i}{\sigma_X^2}\right]$$

n = number of items

$p_i = \frac{N_{Ci}}{N}$

$q_i = 1 - p_i$

$\sigma_X^2 = \frac{\sum_{p=1}^{N} (X_p - \bar{X})^2}{N}$

N = number of persons

N_{Ci} = number of correct answers per ite

X_p = score of person p

$\bar{X} = \frac{\sum_{p=1}^{N} X_p}{N}$

Der Validitätskoeffizient, berechnet mit Formel (2.20) von Seite 47:

Validity per item (v)

good **check** **replace**

item	v
"i01"	0
"i04"	0.9`
"i05"	0
"i06"	0.9`
"i07"	0
"i08"	0.5`
"i09"	-0.07`
"i10"	0.64`
"i11"	0.9`
"i12"	0.59`
"i13"	0.9`
"i14"	0.47`

$$v_i = \frac{N\sum_{p=1}^{N} V_{p,i}X_p - \sum_{p}^{N}}{\sqrt{\left[N\sum_{p=1}^{N} V_{p,i}^2 - \left(\sum_{p=1}^{N} V_{p,i}\right)^2\right]}}$$

N = number of persons

$V_{p,i}$ = value (0 or 1) of person p (item i)

X_p = score of person p

4 Fazit

Im Rahmen dieser Bachelorarbeit wurden zunächst die für das Vorhaben notwendigen Grundlagen der Statistik erarbeitet und dargestellt. Hierbei wurden aufgrund der Komplexität besondere Schwerpunkte in den Teilbereichen *Normalverteilung* und *Chi-Quadrat-Test* gesetzt. Anschließend folgte die Sichtung und Klassifizierung von Methoden, Formeln und Prozeduren im Hinblick auf ihre individuelle Brauchbarkeit für dieses Projekt. Bei diesem Arbeitsschritt lag die eigenständige Leistung vor allem in der umfangreichen Materialsichtung und Reduktion auf die wesentlichen Elemente zur Bestimmung der Gütekriterien Reliabilität und Validität. Das dritte Hauptgütekriterium, die Objektivität, konnte jedoch keine Verwendung finden, da die wichtigen Voraussetzungen unterschiedlicher Testleiter und Re-Tests aus institutionellen Gründen nicht erfüllt werden konnten. Auf Basis der erarbeiteten Grundlagen wurde anschließend die gesamte Implementierung zur Testanalyse dargestellt, wobei der Fokus auf den Chi-Quadrat-Test und seine Kommentierung gesetzt ist. Hierbei erwies sich die Implementierung als umfangreich, aber zuverlässig und stabil. Zuverlässig meint, dass die implementierten Funktionen ohne Fehlermeldungen durchlaufen und immer zu einem Ergebnis kommen. Das Testanalyse*tool* ist stabil gegenüber großen Datensätzen, es kommt nicht zu Programmabstürzen oder Rechenzeitproblemen. Die Genauigkeit der Berechnungen entspricht der erwünschten Aussagekraft. Vor diesem Hintergrund kann man feststellen, dass die geforderte Testanalyse für das Astronautentraining ganzheitlich umgesetzt wurde. Mit dieser Bachelorarbeit liegt eine detaillierte Beschreibung für das *tool* vor.

Ein großer Vorteil des entstandenen *tools* ist die Transparenz der verwendeten Algorithmen. Da der Quellcode der GUI in „tool.m“ einsehbar ist, tritt nicht der „Black-Box-Effekt“ einer Berechnung auf, die nur Ein- und Ausgabe offenlegt, nicht jedoch die Berechnungsschritte. Das ist ein großer Vorteil gegenüber Programmen wie `Mathematica` und `SPSS`. Diese kommerzielle Software bietet jeweils umfangreiche Pakete mit zahlreichen Funktionen im Bereich Statistik, hält jedoch ihren Quellcode geschützt. Der Benutzer bekommt keinerlei Informationen mit welchen Algorithmen bestimmte Berechnungen ausgeführt werden. Um die besagte Transparenz zu erschaffen wurden für das Testanalyse*tool* alle Funktionen[77] „selbst geschrieben“. Zur Demonstration der identischen Rechenergebnisse wurde an mehreren Stellen der Vergleich zwischen eingebauten und selbstgeschriebenen Funktionen vorgeführt, wobei nur mit Ausnahme der Varianzberechnung immer gleiche Berechnungsergebnisse erzielt wurden[78]. Identische Ergebnisse lassen jedoch keinen Rückschluss auf eine identische Implementierung zu. Für den Benutzer des Testanalyse*tools* ergibt sich der Vorteil, dass

[77] Ausgenommen ist der Box-Whisker-Plot, da er im Rahmen dieses *tools* lediglich illustrierenden Charakter hat. Zur Berechnung der Quantile der Chi-Quadrat-Verteilung wurde aus Effektivitätsgründen auf die eingebaute Funktion zurückgegriffen.

[78] Der Grund für die Differenzen bei der Varianz und Standardabweichung liegt in der unterschiedlichen Definition der verwendeten Varianzformeln.

er die verwendeten Berechnungsschritte einsehen kann. Ein weiterer Vorteil ist die Ausgabeform der GUI Funktionen. Für umfangreiche Funktionen wird dem Rechenergebnis zusätzlich ein Bild angehängt, das die verwendete Formel darstellt. So kann sich der Benutzer bereits ohne Einsicht in den Quellcode einen Einblick verschaffen welche Formeln zur Berechnung benutzt werden. Das *tool* ist in der vorliegenden Version speziell auf die aktuelle Teststruktur zugeschnitten und müsste bei Veränderungen dieser ebenfalls akkomodiert werden.

Optimierungschancen des Testanalyse*tools* bestehen im Vorgang des Importierens von einem Datensatz in die GUI. Dies geschieht derzeit noch „manuell“ im Quellcode. Denkbar wäre eine in das GUI eingefügte Menüleiste, anhand derer über den Aufruf „Importieren“ das Einlesen der Datensätze geschehen könnte. An dieser Stelle könnte ebenfalls die Option einer direkten Verbindung von einer Datenbank zur GUI eingebracht werden. Das würde den „Umweg“ über eine (Excel-) Tabelle im „txt“ Format ersparen. Bei den angeführten Aspekten handelt es sich lediglich um eine Optimierung der Benutzerfreundlichkeit, die sich in der regelmäßigen Verwendung des *tools* im Astronautentraining ergeben hat.

An dieser Stelle sei darauf hingewiesen, dass das *tool* ausschließlich auf **`Mathematica`** basiert. Diese Gebundenheit limitiert unter Umständen die Übertragbarkeit und den Einsatzbereich. Prinzipiell könnte das Testanalyse*tool* auch in andere Programmiersprachen übersetzt werden.

Index

Literatur

[LiRa] Gustav A. Lienert / U. Raatz, *Testaufbau und Testanalyse*, BELTZ PsychologieVerlagsUnion, 6. Auflage, 1998

[Bortz] Jürgen Bortz, *Statistik für Human- und Sozialwissenschaftler*, Springer, 6. Auflage, 2005

[Henze] Norbert Henze, *Stochastik für Einsteiger*, Eine Einführung in die faszinierende Welt des Zufalls, Vieweg Verlag, 5. Auflage, Dezember 2004

[Ziez] Herbert Ziezold, *BIOMETRIE*, Vorlesungsskript Universität Kassel, 4. Auflage, April 2005

[Litz] Hans Peter Litz, *Statistische Methoden in den Wirtschafts- und Sozialwissenschaften*, Oldenbourg Verlag, 1997

[Wiki] www.wikipedia.de, *Die freie Enzyklopädie*, die jeweiligen Suchworte sind einzugeben

[ButtStroh] Günter Buttler / Reinhold Stroh, *Einführung in die Statistik*, rororo Sachbuch, September 1990

[DuFremd] DUDEN *Das Fremdwörterbuch*, DUDENVERLAG, 6. Auflage, 1997

[MaOn] www.mathe-online.at/materialien/klaus.berger/files/regression/korrelation.pdf

A Anhang

A.1 tool.m

```
001 (*name of the window*)
002 Quellcode=Widget["Frame", {
003   "title" -> "test analysis tool",
004
005 (*select between options*)
006   {WidgetGroup[{
007
008 (*the trainees*)
009     {WidgetGroup[{
010     {Widget["CheckBox", {"text" -> "test score per person",
011         BindEvent["action",
012             Script[RW[]]]}],
013         WidgetAlign[]
014     },
015     {Widget["CheckBox", {"text" -> "average score",
016         BindEvent["action",
017             Script[RWQ[]]]}],
018         WidgetAlign[]
019     },
020     {Widget["CheckBox", {"text" -> "standard deviation of scores",
021         BindEvent["action",
022             Script[SigmaRW[]]]}],
023         WidgetAlign[]
024     },
025     {Widget["CheckBox", {"text" -> "distribution of scores",
026         Script[Needs["Statistics`DataManipulation`"];
027             Needs["Graphics`Graphics`"];
028             Needs["Statistics`StatisticsPlots`"];],
029         BindEvent["action",
030             Script[Distr[]]]}],
031         WidgetAlign[]
032     },
033     {Widget["CheckBox", {"text" ->
034         "test for normal distribution of scores (Chi-square test)",
035         Script[Needs["Statistics`NormalDistribution`"];],
036         BindEvent["action",
037             Script[ChiQ[]]]}],
038         WidgetAlign[]
039     },
040     Widget["Panel", WidgetLayout -> {"Stretching" -> {Maximize, Maximize}}]
041     },
042     WidgetLayout -> {"Border" -> {"... the students"}}
043     ]},
044
```

```
045     {WidgetGroup[{
046
047 (*the items part I*)
048     Widget["Panel",
049         WidgetGroup[{
050
051             {Widget["CheckBox",{"text" -> "number of correct answers per
item",
052         BindEvent["action",
053             Script[NR[]]]}],
054         WidgetAlign[]
055         },
056
057             {Widget["CheckBox",{"text" -> "average of correct answers",
058         BindEvent["action",
059             Script[NRQ[]]]}],
060         WidgetAlign[]
061         },
062
063             {Widget["CheckBox",{"text" -> "number of wrong answers per
item",
064         BindEvent["action",
065             Script[NF[]]]}],
066         WidgetAlign[]
067         },
068
069             {Widget["CheckBox",{"text" -> "average of wrong answers",
070         BindEvent["action",
071             Script[NFQ[]]]}],
072         WidgetAlign[]
073         }
074         }]],
075
076 (*the items part II*)
077         Widget["Panel",
078             WidgetGroup[{
079
080             {Widget["CheckBox", {"text" -> "index of difficulty",
081                 BindEvent["action",
082                 Script[P[]]]}],
083             WidgetAlign[]
084             }
085             }]
086         ],
087         Widget["Panel", WidgetLayout -> {"Stretching" -> {Maximize,
Maximize}}]
088
089     },WidgetLayout -> {"Border" -> {"... the items"}}
090     ]},
091
092   {WidgetGroup[{
093 (*reliability*)
094     {Widget["CheckBox", {"text" -> "reliability",
095         BindEvent["action",
096             Script[RelAll[]]]}],
097         WidgetAlign[]
098     },
```

```
099 (*validity*)
100     {Widget["CheckBox", {"text" -> "validity",
101         BindEvent["action",
102             Script[ValAll[]]]}],
103         WidgetAlign[]
104     },
105     Widget["Panel", WidgetLayout -> {"Stretching" -> {Maximize, Maximize}}]
106     },
107   WidgetLayout -> {"Border" -> "... the quality of the whole test"}
108   ]}
109
110     },
111     WidgetLayout -> {"Border" -> {"analysis of ..."} }]
112   },
113
114   Script[
115 (*-------------------------------------------------------------------------
-*)
116   path = "D:\\Taddi\\Uni\\Bachelorarbeit DLR ESA\\Bachelorarbeit\\CD BA";
117 (*-------------------------------------------------------------------------
-*)
118   SetDirectory[path];
119   Tabelle := Import["BeispielTabelle.txt", "TSV"];
120   (*Tabelle := Import["Beispieldatensatz.txt", "TSV"];*)
121 (*-------------------------------------------------------------------------
-*)
122
123
124 (*delete all items in Tabelle, which include "a"*)
125
126     Tabelle = Transpose[Tabelle];
127     as = Position[Tabelle, "a"];
128     Del = Partition[Union[Table[as[[i]][[2]], {i, Length[as]}]], 1];
129     Tabelle = Transpose[Delete[Transpose[Tabelle], Del]];
130     x = First[Dimensions[Tabelle]];
131     y = Last[Dimensions[Tabelle]];
132     PersID = Tabelle[[Range[2, x], 1]];
133     item[i_] := Tabelle[[1, i]];
134     values[i_] := Tabelle[[Range[2, x], i]];
135     valueList = Table[values[i], {i, 2, y}];
136
```

```
137 (*define legends for the coloured output values*)
138
139   legend = List[StyleForm["good  ", FontColor -> Green, FontWeight ->
"Bold",
140         FontSize -> 15],
141       StyleForm["  no action  ", FontColor -> Blue, FontWeight -> "Bold",
142         FontSize -> 15],
143       StyleForm["  check  ", FontColor -> Orange, FontWeight -> "Bold",
144         FontSize -> 15],
145       StyleForm["  replace", FontColor -> Red, FontWeight -> "Bold",
146         FontSize -> 15]];
147
148   legendo0 = List[StyleForm["good  ", FontColor -> Green, FontWeight ->
"Bold",
149         FontSize -> 15],
150       StyleForm["  no action  ", FontColor -> Blue, FontWeight -> "Bold",
151         FontSize -> 15],
152       StyleForm["  replace", FontColor -> Red, FontWeight -> "Bold",
153         FontSize -> 15]];
154
155   legendoB = List[StyleForm["good  ", FontColor -> Green, FontWeight ->
"Bold",
156         FontSize -> 15],
157       StyleForm["  check  ", FontColor -> Orange, FontWeight -> "Bold",
158         FontSize -> 15],
159       StyleForm["  replace", FontColor -> Red, FontWeight -> "Bold",
160         FontSize -> 15]];
161
162
```

```
(* several calculations for the different functions *)

  RWPers[p_] := Sum[valueList[[i, p]], {i, 1, y - 1}];
  RWe = Table[RWPers[p], {p, 1, x - 1}];

  RWQuer = N[Sum[RWe[[p]], {p, 1, x - 1}]/(x - 1)];

  RWSA = Sqrt[(Sum[(RWe[[p]] - RWQuer)^2, {p, 1, x - 1}])/(x - 1)];
  RWVar = RWSA^2;

  NR[i_] := Sum[valueList[[i, p]], {p, 1, x - 1}];
  NRe = Table[NR[i], {i, 1, y - 1}];
  AlleNR = Transpose[
    Join[{Join[{StyleForm[item,  FontWeight -> "Bold", FontSize -> 20]},
          Table[item[i], {i, 2, y}]]}, {Join[{StyleForm[NC,
              FontWeight -> "Bold", FontSize -> 20]}, NRe]}]];
  NRQuer = N[Sum[NRe[[i]], {i, 1, y - 1}]/(y - 1)];

  NF[i_] := (x - 1) - NRe[[i]];
  NFe = Table[NF[i], {i, 1, y - 1}];
  AlleNF = Transpose[
    Join[{Join[{StyleForm[item,  FontWeight -> "Bold", FontSize -> 20]},
          Table[item[i], {i, 2, y}]]}, {Join[{StyleForm[NW,
              FontWeight -> "Bold", FontSize -> 20]}, NFe]}]];
  NFQuer = N[Sum[NFe[[i]], {i, 1, y - 1}]/(y - 1)];

  P[i_] := 100*(NRe[[i]]/(x - 1));
  Pkurz[i_] := N[IntegerPart[(100*(NRe[[i]]/(x - 1)))*100]/100];
  AlleP = Transpose[
    Join[{Join[{StyleForm[item,  FontWeight -> "Bold", FontSize -> 20]},
          Table[item[i], {i, 2, y}]]}, {Join[{StyleForm[P,
              FontWeight -> "Bold", FontSize -> 20]},
          Table[If[Pkurz[i] == 100,
              StyleForm[Pkurz[i], FontColor -> Blue, FontWeight -> "Bold",
                FontSize -> 15],
              If[Pkurz[i] <= 39,
                StyleForm[Pkurz[i], FontColor -> Red, FontWeight -> "Bold",
                  FontSize -> 15],
                StyleForm[Pkurz[i], FontColor -> Green, FontWeight ->
"Bold",
                  FontSize -> 15]]], {i, 1, y - 1}]]}]];

  rtc1[i_] := N[IntegerPart[
          N[((x - 1)* Sum[RWe[[p]]* valueList[[i, p]], {p, 1, x - 1}] -
(Sum[
            RWe[[p]], {p, 1, x - 1}]* Sum[valueList[[i, p]], {p, 1, x -
1}]))]
            *100]/100];
  rtc2[i_] := N[IntegerPart[
          N[Sqrt[((x - 1)*Sum[RWe[[p]]^2, {p, 1, x - 1}] -
            Sum[RWe[[p]], {p, 1, x - 1}]^2)*((x - 1)* Sum[valueList[[i,
p]]^2,
            {p, 1, x - 1}] - Sum[valueList[[i, p]], {p, 1, x - 1}]^2)]]
            *100]/100];

```

```
214 (*Functions for the action-events*)
215
216 (*personal results*)
217     RW[]:=Module[{},
218             AlleRW = Transpose[Join[{Join[{StyleForm[Tabelle[[1, 1]],
219                 FontWeight -> "Bold", FontSize -> 20]}, PersID]},
220                 {Join[{StyleForm[X,  FontWeight -> "Bold",
221                 FontSize -> 20]}, RWe]}]] // MatrixForm;
222             NotebookWrite[EvaluationNotebook[],
223                 Cell["Test score per person (X)", FontSize -> 25]];
224             NotebookWrite[EvaluationNotebook[], ToBoxes[AlleRW]];
225             SelectionMove[EvaluationNotebook[], Next, Cell];
226     ];
227 (*average of personal results*)
228     RWQ[]:=Module[{},
229             RWQkurz = N[IntegerPart[RWQuer*100]/100];
230             NotebookWrite[EvaluationNotebook[], Cell["Average score (X
cross)",
231                 FontSize -> 25]];
232             NotebookWrite[EvaluationNotebook[], ToBoxes[RWQkurz]];
233             SelectionMove[EvaluationNotebook[], Next, Cell];
234             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
235                     DisplayString[Import["average score.bmp"]]],
"Graphics"]];
236             SelectionMove[EvaluationNotebook[], Next, Cell];
237     ];
238 (*standard deviation of personal results*)
239     SigmaRW[]:=Module[{},
240             RWSAkurz = N[IntegerPart[RWSA*100]/100];
241             NotebookWrite[EvaluationNotebook[],
242                 Cell["Standard deviation of scores (sigma)", FontSize ->
25]];
243             NotebookWrite[EvaluationNotebook[], ToBoxes[RWSAkurz]];
244             SelectionMove[EvaluationNotebook[], Next, Cell];
245             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
246                     DisplayString[Import["sigma.bmp"]]], "Graphics"]];
247             SelectionMove[EvaluationNotebook[], Next, Cell];
248     ];
```

```
249 (*distribution of personal results*)
250     Distr[]:=Module[{},
251
252             HFKT = Frequencies[RWe];
253             l = Length[HFKT];
254             maxe = Max[Table[HFKT[[r, 1]], {r, 1, l}]];
255             TKFH = Map[Reverse, HFKT];
256             HFKTGraph = ListPlot[Map[Reverse, Frequencies[RWe]],
257                 DisplayFunction -> Identity,
258                 AxesLabel -> {score, freq}, (*PlotJoined -> True,*)
259                 PlotRange -> {{Min[RWe] - 5, Max[RWe] + 5}, {0, maxe +
0.5}},
260                 PlotStyle -> {PointSize[0.01], Hue[.6]},
261                 TextStyle -> {FontFamily -> "Times", FontSize -> 14},
262                 PlotLabel -> StyleForm["distribution of scores",FontSize ->
18,
263                 FontWeight -> "Bold"], ImageSize -> 400];
264
265             RWQSA = Show[Graphics[{RGBColor[0, 1, 1],
266                 Rectangle[{RWQuer - RWSA, 0.01}, {RWQuer + RWSA, maxe}],
267                 RGBColor[0, 0, 0], Rectangle[{RWQuer - 0.15, 0},
268                     {RWQuer + 0.15, maxe}]}, DisplayFunction -> Identity]];
269
270             H = Histogram[RWe, DisplayFunction -> Identity,
271                 HistogramCategories -> (y - 1),
272                 HistogramRange -> {Min[RWe] - 5, Max[RWe] + 5}];
273
274             leg = List[StyleForm["average score  ",
275                     FontColor -> RGBColor[0, 0, 0],
276                     FontWeight -> "Bold", FontSize -> 15],
277                     StyleForm["  standard deviation of scores  ",
278                     FontColor -> RGBColor[0, 1, 1], FontWeight -> "Bold",
279                     FontSize -> 15], StyleForm["  frequency of scores",
280                     FontColor -> Red, FontWeight -> "Bold", FontSize ->
15]];
281
282             DistrGraph = Show[Graphics[HFKTGraph], RWQSA, Graphics[H]];
283
284             BWP = BoxWhiskerPlot[RWe, BoxOrientation -> Horizontal,
285                 BoxOutliers -> All, BoxOutlierShapes ->
{PlotSymbol[Diamond]},
286                 ImageSize -> 300, TextStyle -> {FontFamily -> "Times",
287                 FontSize -> 14}, PlotLabel -> StyleForm["Box Whisker Plot",
288                 FontSize -> 18, FontWeight -> "Bold"],
289                 DisplayFunction -> Identity];
290
291             NotebookWrite[EvaluationNotebook[], Cell["Distribution of
scores",
292                 FontSize -> 25]];
293             NotebookWrite[EvaluationNotebook[], Cell[TextData[leg]]];
294             SelectionMove[EvaluationNotebook[], Next, Cell];
295             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
296                 DisplayString[DistrGraph]], "Graphics"]];
297             SelectionMove[EvaluationNotebook[], Next, Cell];
298             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
299                 DisplayString[BWP]], "Graphics"]];
300     ];
```

```
(*chi-square-test*)
    ChiQ[]:=Module[{},

            K = 0;
            h = 0;
            R = 0;
            If[Mod[y - 1, 10] == 0, {K = 10, h = (y - 1)/10},
                If[Mod[y - 1, 11] == 0, {K = 11, h = (y - 1)/11},
                If[Mod[y - 1, 12] == 0, {K = 12, h = (y - 1)/12},
                    If[Mod[y - 1, 13] == 0, {K = 13, h = (y - 1)/13},
                    If[Mod[y - 1, 14] == 0, {K = 14, h = (y - 1)/14},
                        If[Mod[y - 1, 15] == 0, {K = 15, h = (y - 1)/15},
                        {K = 10, R = Mod[y - 1, 10],
                        h = IntegerPart[(y - 1)/10]}]]]]];

            RWKlasse[1] := Join[{0}, {h + R}];
            RWKlasse[i_] := Join[{RWKlasse[i - 1][[2]] + 1},
                {RWKlasse[i - 1][[2]] + h}];
            RWKlassen = MatrixForm[Table[RWKlasse[k], {k, 1, K}]];
            KlM = Table[N[(RWKlassen[[1, k]][[1]]
                + RWKlassen[[1, k]][[2]])/2], {k, 1, K}];

            Table[AnzKl[k_] := 0;
            HF = Table[
                If[RWe[[p]]>=RWKlasse[k][[1]] && RWe[[p]]<=RWKlasse[k][[2]],
                AnzKl[k] += 1, AnzKl[k]], {k, 1, K}], {p, 1, x - 1}];
            A = Table[RWKlassen[[1, k]][[1]], {k, 1, K}];
            B = Table[RWKlassen[[1, k]][[2]], {k, 1, K}];
            F = Table[KlM[[k]] - RWQuer, {k, 1, K}];
            z = Table[F[[k]]/RWSA, {k, 1, K}];
            f[x_] := (1/Sqrt[2*\[Pi]])*\[ExponentialE]^((x^2)/-2);
            Y = Table[N[f[z[[k]]]], {k, 1, K}];
            fe = Table[((h*(x - 1))/RWSA)*Y[[k]], {k, 1, K}];

            hf = HF;
            AB = Transpose[Join[{A}, {B}]];
            k = 1;
            Do[
                (*Print[k];
                Print[MatrixForm[Transpose[Join[{AB}, {hf}, {fe}]]]];*)
            If[hf[[k]] < 5,
                AB[[k, 2]] = AB[[k + 1, 2]];
                AB = Delete[AB, k + 1];
                hf[[k]] += hf[[k + 1]];
                hf = Delete[hf, k + 1];
                fe[[k]] += fe[[k + 1]];
                fe = Delete[fe, k + 1];
                ,
                k = k + 1;
                ];
            , {Length[hf] - 1}];
            If[Last[hf] < 5,
                k = Length[hf];
                AB[[k - 1, 2]] = AB[[k, 2]];
                AB = Delete[AB, k];
                hf[[k - 1]] += hf[[k]];
                hf = Delete[hf, k];
                fe[[k - 1]] += fe[[k]];
                fe = Delete[fe, k];
                ];

```

```
362             ChiQua = Sum[((hf[[j]] - fe[[j]])^2)/fe[[j]], {j, 1,
Length[fe]}];
363             ChiQuaKurz = N[IntegerPart[(100*ChiQua)]/100];
364
365             ChiQuantile = If[(df - 3) > 0,
366                 Quantile[ChiSquareDistribution[df - 3], 0.95]
367             ];
368             Resu = If[NumberQ[ChiQuantile],
369                 If[ChiQua <= ChiQuantile,
370  "OK, the score is nearly normal distributed (level of significance = 5%).",
371  "The scores are not covered by the Normal Distribution
372  (level of significance = 5%)."],
373  "It's not possible to calculate the Quantile of Chi-Square-Distribution,
374  because the degree of freedom is too low (zero)."];
375
376             NotebookWrite[EvaluationNotebook[], Cell[
377                 "Test for normal distribution of scores (Chi-square test)",
378                 FontSize -> 25]];
379             NotebookWrite[EvaluationNotebook[], {ToBoxes[Resu],
380                 ToBoxes[ChiQuaKurz]}];
381             SelectionMove[EvaluationNotebook[], Next, Cell];
382             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
383                     DisplayString[Import["chi2.bmp"]]], "Graphics"]];
384             SelectionMove[EvaluationNotebook[], Next, Cell];
385     ];
386 (*correct answers per item*)
387     NR[]:=Module[{},
388             NRTab=MatrixForm[AlleNR];
389             NotebookWrite[EvaluationNotebook[], Cell[
390                 "Number of correct answers per item (NC)", FontSize -> 25]];
391             NotebookWrite[EvaluationNotebook[], ToBoxes[NRTab]];
392             SelectionMove[EvaluationNotebook[], Next, Cell];
393     ];
```

```
394 (*average of correct answers per item*)
395     NRQ[]:=Module[{},
396             NRQkurz = N[IntegerPart[NRQuer*100]/100];
397             NotebookWrite[EvaluationNotebook[], Cell["Average of NC",
398                 FontSize -> 25]];
399             NotebookWrite[EvaluationNotebook[], ToBoxes[NRQkurz]];
400             SelectionMove[EvaluationNotebook[], Next, Cell];
401             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
402                     DisplayString[Import["NC average.bmp"]]], "Graphics"]];
403             SelectionMove[EvaluationNotebook[], Next, Cell];
404     ];
405 (*wrong answers per item*)
406     NF[]:=Module[{},
407             NFTab=MatrixForm[AlleNF];
408             NotebookWrite[EvaluationNotebook[], Cell[
409                 "Number of wrong answers per item (NW)", FontSize -> 25]];
410             NotebookWrite[EvaluationNotebook[], ToBoxes[NFTab]];
411             SelectionMove[EvaluationNotebook[], Next, Cell];
412     ];
413 (*average of wrong answers per item*)
414     NFQ[]:=Module[{},
415             NFQkurz = N[IntegerPart[NFQuer*100]/100];
416             NotebookWrite[EvaluationNotebook[], Cell["Average of NW",
417                 FontSize -> 25]];
418             NotebookWrite[EvaluationNotebook[], ToBoxes[NFQkurz]];
419             SelectionMove[EvaluationNotebook[], Next, Cell];
420             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
421                     DisplayString[Import["NW average.bmp"]]], "Graphics"]];
422             SelectionMove[EvaluationNotebook[], Next, Cell];
423     ];
424 (*index of difficulty*)
425     P[]:=Module[{},
426             PTab=MatrixForm[AlleP];
427             NotebookWrite[EvaluationNotebook[], Cell[
428                 "Index of difficulty per item (P)", FontSize -> 25]];
429             NotebookWrite[EvaluationNotebook[], Cell[TextData[legendo0]]];
430             SelectionMove[EvaluationNotebook[], Next, Cell];
431             NotebookWrite[EvaluationNotebook[], ToBoxes[PTab]];
432             SelectionMove[EvaluationNotebook[], Next, Cell];
433             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
434                     DisplayString[Import["P.bmp"]]], "Graphics"]];
435             SelectionMove[EvaluationNotebook[], Next, Cell];
436             PGraph = ListPlot[Table[P[i], {i, 1, y - 1}],
437                 DisplayFunction -> Identity,
438                 AxesLabel -> TraditionalForm /@ {item, P}, PlotJoined ->
True,
439                 PlotRange -> {{0, y}, {0, 100}}, PlotStyle -> Hue[.6],
440                 TextStyle -> {FontFamily -> "Times", FontSize -> 14},
441                 PlotLabel -> StyleForm["difficulty of items (100 = easy)",
442                     FontSize -> 18, FontWeight -> "Bold"], ImageSize ->
400];
443             NotebookWrite[EvaluationNotebook[], Cell["Curve of difficulty",
444                 FontSize -> 25]];
445             NotebookWrite[EvaluationNotebook[],
Cell[GraphicsData["PostScript",
446                 DisplayString[PGraph]], "Graphics"]];
447             SelectionMove[EvaluationNotebook[], Next, Cell];
448     ];
```

```
449 (*reliability*)
450     RelAll[]:=Module[{},
451         p[i_] := NRe[[i]]/(x - 1);
452         q[i_] := 1 - p[i];
453         Rel = N[IntegerPart[((y - 1)/((y - 1) -
454                     1))*(1 - (Sum[p[i]*q[i], {i, 1, y - 1}]/RWVar))*100]/
455         100];
456         RelAllKoe = If[Rel >= 0.7,
457             StyleForm[Rel, FontColor -> Green, FontWeight -> "Bold",
458                 FontSize -> 15],
459             If[Rel < 0.6,
460             StyleForm[Rel, FontColor -> Red, FontWeight -> "Bold",
461                 FontSize -> 15],
462             StyleForm[Rel, FontColor -> Orange, FontWeight -> "Bold",
463             FontSize -> 15]]];
464         NotebookWrite[EvaluationNotebook[], Cell["Reliability of test (r)",
465             FontSize -> 25]];
466         NotebookWrite[EvaluationNotebook[], Cell[TextData[legendoB]]];
467         NotebookWrite[EvaluationNotebook[], ToBoxes[RelAllKoe]];
468         SelectionMove[EvaluationNotebook[], Next, Cell];
469         NotebookWrite[EvaluationNotebook[], Cell[GraphicsData["PostScript",
470                     DisplayString[Import["reliability.bmp"]]], "Graphics"]];
471         SelectionMove[EvaluationNotebook[], Next, Cell];
472         ];
473 (*validity*)
474     ValAll[]:=Module[{},
475         Validit = Table[If[rtc1[w] == 0 && rtc2[w] == 0, 0,
476             N[IntegerPart[N[rtc1[w]/rtc2[w]]*100]/100]], {w, 1, y - 1}];
477         ValTab = Transpose[Join[{Join[{StyleForm[item,  FontWeight ->
"Bold",
478             FontSize -> 20]},
479             Table[item[j], {j, 2, y}]]}, {Join[{StyleForm[v,
480                 FontWeight -> "Bold", FontSize -> 20]},
481             Table[If[Validit[[i]] >= 0.5,
482               StyleForm[Validit[[i]], FontColor -> Green,
483                 FontWeight -> "Bold", FontSize -> 15],
484               If[Validit[[i]] < -0.1,
485                 StyleForm[Validit[[i]], FontColor -> Red,
486                   FontWeight -> "Bold", FontSize -> 15],
487                 StyleForm[Validit[[i]], FontColor -> Orange,
488                     FontWeight -> "Bold", FontSize -> 15]]],
489                         {i, 1, y - 1}]]}]]//MatrixForm;
490         NotebookWrite[EvaluationNotebook[], Cell["Validity per item (v)",
491             FontSize -> 25]];
492         NotebookWrite[EvaluationNotebook[], Cell[TextData[legendoB]]];
493         NotebookWrite[EvaluationNotebook[], ToBoxes[ValTab]];
494         SelectionMove[EvaluationNotebook[], Next, Cell];
495         NotebookWrite[EvaluationNotebook[], Cell[GraphicsData["PostScript",
496                     DisplayString[Import["validity.bmp"]]], "Graphics"]];
497         SelectionMove[EvaluationNotebook[], Next, Cell];
498         ];
499
500     ]
501   }
502 ]
```

Ich versichere hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen, Programme und Hilfsmittel benutzt habe.

KATRIN HEUMANN