Grundkurs Statistik für Sozialwissenschaftler

Martin Förster

Letzte Bearbeitung: 15.01.2022

Inhaltsverzeichnis

Teil I

Prolog: Messen und Skalieren

Kapitel 1

Messen

In der empirischen Sozialforschung geht es darum, Sachverhalte zu messen. Etwas technischer ausgedrückt, sprechen wir dann nicht mehr von Sachverhalten, sondern von *Variablen*. Variablen zu messen, heißt also, empirisch *Daten* über die Realität zu erheben. Diese Daten weisen eine bestimmte Struktur auf, die sich als Datenmatrix darstellen lässt.Wenn wir z.B. an Personen die Sachverhalte Geschlecht und Niedergeschlagenheit anhand eines Fragebogens (Abbildung 1.1) messen, lassen sich die erhobenen Daten als Datenmatrix folgendermaßen darstellen (Abbildung 1.2):

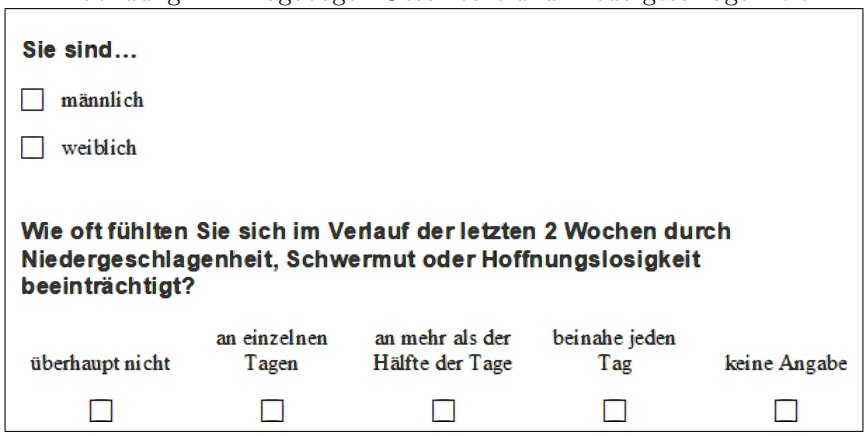
Fälle — Jede *Zeile* stellt eine befragte Person (Fall) dar. Werden also drei Fragebögen ausgefüllt, dann beinhaltet die Datenmatrix drei Zeilen mit jeweils einem Fall. Fälle sind Merkmalsträger: Probanden, Objekte, Länder etc.

Variablen — Jede *Spalte* bedeutet eine Frage (Variable). Die Variablen stehen also für Merkmale, Eigenschaften usw., die für die Fälle erhoben wurden. Insofern hier exemplarisch den Personen lediglich die zwei Fragen (nach dem Geschlecht und der Niedergeschlagenheit) gestellt werden und sonst keine weiteren Merkmale erhoben werden, weist die Datenmatrix mithin 2 Spalten mit jeweils einer Variable auf.

Werte — In den *Zellen* sind die vercodeten Antworten (Merkmalsausprägungen) einer Person auf die entsprechende Frage. Welcher Code für welche Ausprägung steht, muss vorher definiert worden sein. So könnte bei der Variable Geschlecht der Wert 1 für männlich stehen und der Wert 2 für weiblich. Für die Variable Niedergeschlagenheit könnten den Antworten folgende Codes zugeordnet sein: 0=überhaupt nicht, 1=an einzelnen Tagen, 2=an mehr als der Hälfte der Tage, 3=beinahe jeden Tag.

Die Datenmatrix in Abbildung 1.2 zeigt also für 9 Befragte (Fälle) die Antworten (Werte) auf zwei Fragen (Variablen). So ist bspw. die erste Befragte

Abbildung 1.1: Fragebogen Geschlecht und Niedergeschlagenheit



Sie sind...

☐ männlich

☐ weiblich

Wie oft fühlten Sie sich im Verlauf der letzten 2 Wochen durch Niedergeschlagenheit, Schwermut oder Hoffnungslosigkeit beeinträchtigt?

überhaupt nicht	an einzelnen Tagen	an mehr als der Hälfte der Tage	beinahe jeden Tag	keine Angabe
☐	☐	☐	☐	☐

Abbildung 1.2: Beispiel Datenmatrix zum Fragebogen (Abbildung 1.1)

	Geschlecht	Depr2
1	2	1
2	1	2
3	1	1
4	2	1
5	1	0
6	2	3
7	1	0
8	1	3
9	2	1

Person eine *Frau* (Geschlecht=2), die *an einzelnen Tagen* durch Niedergeschlagenheit, Schwermut oder Hoffnungslosigkeit beeinträchtigt ist (Depr2=1). Die zweite befragte Person ist ein *Mann*, welcher *an mehr als der Hälfte der Tage* von dem abgefragten Symptom betroffen ist.

Die hier exemplarisch dargestellten 9 Messungen der Variablen Geschlecht und Depr2 ordnen den jeweiligen Sachverhalten Symbole zu[1]. Messungen resultieren in Variablenwerten, welche als Symbole dargestellte Merkmalsausprägungen sind. Ein essentielles Kriterium für Messungen in diesem Sinne ist, dass die Zuordnung von Symbolen zu Merkmalsausprägungen strukturkonform zu erfolgen hat. D.h. die Relationen zwischen den Merkmalsausprägungen müssen durch entsprechende Relationen zwischen den Variablenwerten abgebildet werden (Stevens, 1951). Schauen wir uns als Beispiel die Variable Depr2 an, mit welcher der Sachverhalt gemessen werden soll, wie oft eine Person an den Symptomen Niedergeschlagenheit, Schwermut oder Hoffnungslosigkeit leidet. Abbildung 1.1 zeigt die vier substantiellen Antwortmöglichkeiten (von „überhaupt nicht“ bis „beinahe jeden Tag“) für die Befragten, außerdem eine residuale Antwortkategorie („keine Angabe“). In der Datenmatrix sind als Variablenwerte nun Codes abgebildet, die jeweils einer bestimmten Antwort auf die entsprechende Frage zugeordnet sind. Neben den oben bereits aufgeführten Codes für die substantiellen Ausprägungen, könnte der residualen Antwortkategorie „keine Angabe“ der Code „m“ zugeordnet werden. Die Codierung der Ausprägungen ist hier strukturkonform, weil die Ordnung der substantielle Antworten (von „überhaupt nicht“ bis „beinahe jeden Tag“) durch die zahlenmäßige Ordnung der Codes (von 0 bis 3) abgebildet wird. An wie vielen Tagen genau eine Person an den abgefragten Symptomen leidet, geht aus den Variablenwerten nicht hervor - diese Information wird aber auch nicht abgefragt. Für dieses Beispiel sind mit der Ordnung der Variablenwerte die Merkmalsausprägungen strukturkonform abgebildet. Welche Informationen für die strukturkonforme Messung berücksichtigt werden müssen, wird durch das Skalenniveau der jeweiligen Variable bestimmt (s. nächstes Kapitel).

[1] Symbole sind alphanumerische Repräsentationen der jeweiligen Sachverhalte: Zahlen, Buchstaben, Zeichenketten.

Kapitel 2

Skalieren

Wenn messen also heißt, für ein bestimmtes Merkmal die möglichen Ausprägungen mit Symbolen darzustellen, dann hat jede Variable eine Menge von Symbolen, welche die möglichen Ausprägungen repräsentiert. Diese Menge möglicher Messwerte wird als *Skala* bezeichnet. Weiterhin wurde im vorangegangenen Kapitel darauf hingewiesen, dass die Messung strukturkonform erfolgen soll. Das impliziert, dass die Ausprägungen des zu messenden Merkmals in einer bestimmten Relation (="Struktur") zueinander stehen - und diese Relationen müssen durch die Skala abgebildet werden. D.h. die Messwerte (Symbole) müssen die selben Relationen untereinander aufweisen wie die entsprechenden Merkmalsausprägungen (empirischen Relative). Unter welchen Aspekten die Relationen zwischen den Messwerten bestimmt werden können, wird durch das *Skalenniveau* beschrieben.

Dass diese strukturellen Unterschiede als Niveaus bezeichnet werden, hängt damit zusammen, dass mit den skalen-strukturellen Unterschieden auch Unterschiede im Informationsgehalt (hinsichtlich der Relationen der Merkmalsausprägungen) verbunden sind. Aufsteigend nach ihrem Informationsgehalt werden vier Skalenniveaus unterschieden: Nominalskala, Ordinalskala, Intervallskala und Ratioskala (Stevens, 1946). Skalen mit nominalem Niveau haben also den geringsten Informationsgehalt, Ratioskalen haben den höchsten Informationsgehalt.

Auf *Nominalskalen* können die Merkmalsausprägungen lediglich klassifiziert werden. Über die Relation der Merkmalsausprägungen kann also nur gesagt werden, dass sie nicht gleich sind. Bezüglich der Ausprägungsrelationen gibt es hier lediglich die Information über die Verschiedenheit bzw. Gleichheit der Merkmalsausprägungen. Ein Beispiel ist das Merkmal der Konfession von Personen. Neben dem Variablenwert, der anzeigt, welche konkrete Konfession eine Person hat, kann hinsichtlich der Relation der Merkmalsausprägungen lediglich festgestellt werden, ob sich die Konfession einer Person von der Konfession einer anderen Person unterscheidet oder nicht.

Beim nächsthöheren Skalenniveau, dem *Ordinalskalenniveau*, kann man schon etwas mehr über die Relationen der Merkmalsausprägungen sagen. Hier weisen

die Ausprägungen nämlich eine natürliche Rangordnung auf. Dabei kommt also - zusätzlich zur Information der Verschiedenheit von Merkmalsausprägungen - die Information zur Reihenfolge hinzu. Als Beispiel soll die Antwortskala der Variable Depr2 aus dem ersten Kapitel dienen: für die Frage „Wie oft fühlten Sie sich im Verlauf der letzten 2 Wochen durch Niedergeschlagenheit, Schwermut oder Hoffnungslosigkeit beeinträchtigt?“ wurde eine Antwortskala mit den substantiellen Ausprägungen *überhaupt nicht*, *an einzelnen Tagen*, *an mehr als der Hälfte der Tage*, *beinahe jeden Tag* vorgegeben (Abbildung 1.1), wobei die zugeordneten Variablenwerte von 0 (=überhaupt nicht) bis 3 (=beinahe jeden Tag) reichen. Mit Blick auf die exemplarische Datenmatrix (ebenfalls Kapitel 1, Abbildung 1.2) kann festgestellt werden, dass die erste befragte Person mit Depr2=1 an einzelnen Tagen an den abgefragten Symptomen leidet. Die zweite befragte Person leidet an mehr als der Hälfte der Tage (Depr2=2) an diesen Symptomen, während die dritte befragte Person wieder an einzelnen Tagen (Depr2=1) davon betroffen ist. Neben der Information, dass sich die zweite befragte Person von der ersten und der dritten hinsichtlich dieser Variable unterscheidet, kann weiterhin festgestellt werden, dass die zweite Person häufiger von den Symptomen betroffen ist als die erste und die zweite Person. D.h. die Ausprägungen dieser Variable weisen eine natürliche Reihenfolge bzw. Rangordnung auf, welche auch - strukturkonform (s. Kapitel 1) - durch die jeweiligen Variablenwerte abgebildet wird.

Auf das Ordinalskalenniveau folgt das *Intervallskalenniveau*. Für die Merkmalsausprägungen konkreter Messungen mit derart skalierten Variablen kann festgestellt werden, ob sie unterschiedlich sind (wie bei einer Nominalskala), in welcher Reihenfolge sie stehen (entsprechend einer Ordinalskala) - und hinzu kommt die Information der Abstände zwischen den Werten. Die Abstände zwischen den Merkmalsausprägungen intervallskalierter Variablen haben eine Bedeutung, welche durch die zahlenmäßige Differenz der Variablenwerte genau wiedergegeben wird. Beispiel: die Grad-Celsius-Temperaturskala. Wird in einem Raum eine Temperatur von 18°C gemessen und in einem anderen Raum eine Temperatur von 24°C, dann können folgende Informationen daraus ermittelt werden: a) die Temperatur in dem einen Raum ist anders als die Temperatur in dem anderen Raum (Unterschiedlichkeit); b) in dem Raum mit 24°C ist es wärmer als in dem Raum mit 18°C (Rangordnung); c) darüber hinaus kann genau angegeben werden, wie groß der Temperaturunterschied ist, dass es also in dem Raum mit 24°C um 6°C wärmer ist als in dem Raum, in welchem die Temperatur 18°C beträgt.

Den größten Informationsgehalt hat das *Ratioskalenniveau*. Derartige Skalen haben - neben allen Informationen, die eine Intervallskala aufweist (Unterschiedlichkeit, Reihenfolge, und Abstände der Ausprägungen) - einen absoluten und natürlichen Nullpunkt mit entsprechender Bedeutung. Die Skalenwerte entstammen der Menge der nicht-negativen rationalen Zahlen, wobei per definitionem die Null der Anfang einer solchen Skala ist. Hier kann ebenfalls eine Temperaturskala als Beispiel dienen - nämlich die Kelvin-Skala. Der relevante Unterschied zur Grad-Celsius-Temperaturskala ist der, dass 0 Kelvin die Bedeutung hat, dass es keine kältere Temperatur geben kann - das ist physikalisch un-

Tabelle 2.1: Skalenniveaus und erlaubte Operationen mit den jeweiligen Werten

	Operationen mit Merkmalsausprägungen			
Skalenniveau	**Rangfolge („Ordnung“) bestimmbar**	**Differenzen („Intervalle“) bestimmbar**	**Verhältnisse bestimmbar**	**Daten**
Nominal	-	-	-	kategorial
Ordinal	✔	-	-	kategorial
Intervall	✔	✔	-	metrisch
Ratio	✔	✔	✔	metrisch

möglich. Dagegen ist 0°C mehr oder weniger beliebig - jedenfalls nicht absolut. „Absoluter Nullpunkt“ bedeutet eben, dass es keinen Wert kleiner als Null gibt.

Tabelle 2.1 zeigt die Operationen, welche mit den Werten der jeweiligen Skalenniveaus unter inhaltlichem Aspekt erlaubt sind. Hier kann ein erneuter Blick auf den Unterschied zwischen Intervall- und Ratioskala - wieder repräsentiert durch die Grad-Celsius- und die Kelvin-Skala respektive - erhellend sein. Allgemein formuliert trifft auf beide Skalen zu, dass mit den jeweiligen Werten gerechnet werden kann - und zwar inhaltlich sinnvoll interpretierbar. Allerdings gibt es Differenzen zwischen beiden Skalen, welche Rechenoperationen erlaubt sind. Intervall-bestimmende Rechenoperationen (Subtraktion, Addition) sind mit den Werten beider Skalenniveaus erlaubt (s. Beispiel oben: Temperaturunterschiede). Relation-bestimmende Rechenoperationen (z.B. das Ermitteln von Anteilen durch Division) dürfen jedoch lediglich mit den Werten einer Ratioskala durchgeführt werden. So kann z.B. die Frage, welche Temperatur, gemessen in Grad-Celsius, doppelt so warm ist wie 2°C nicht mit arithmetischen Mitteln beantwortet werden. Übersetzt in eine mathematische Operation entspricht „doppelt-so-viel“ der Multiplikation mit 2. Das Ergebnis wird durch eine Relation bestimmt: „doppelt-so-viel“ ist „zwei-mal-so-viel“. Jedoch ist 4°C (=2 mal 2°C) eben nicht doppelt so warm wie 2°C, weil die Grad-Celsius-Skala keine Ratioskala ist und somit keinen natürlichen Nullpunkt besitzt, welcher als die für eine Relation notwendige Referenz dienen kann. Hingegen ist bei einer Kelvin-Skala diese Bestimmung einer Relation erlaubt: 4K ist doppelt so warm wie 2K. Noch eindrücklicher ist hier womöglich der (unzulässige) Versuch, mit arithmetischen Mitteln zu bestimmen, welche Temperatur auf der Grad-Celsius-Skala doppelt so warm ist wie -4°C. -8°C (= 2 mal -4°C) ist sogar kälter als -4°C! Auf der Kelvin-Skala gibt es hingegen keine negativen Temperaturwerte (ansonsten wäre die Temperatur von 0K keine natürlicher Nullpunkt).

Die Daten nominal- und ordinal-skalierter Variablen werden als *kategorial* bezeichnet; Daten von intervall- und ratio-skalierten Variablen sind *metrisch*. Die Skala kategorialer Variablen ist immer diskret; d.h. zwischen zwei beliebigen (aber verschiedenen) Skalenwerten kann die Variable eine abzählbare (nicht unendliche) Anzahl unterschiedlicher Werte annehmen. Metrische Variablen können hingegen diskret oder stetig sein, wobei stetig bedeutet, dass zwischen zwei

Tabelle 2.2: Fragen aus dem PHQ (Gesundheitsfragebogen für Patienten)

Wie oft fühlten Sie sich im Verlauf der letzten 2 Wochen durch die folgenden Beschwerden beeinträchtigt?	(0) Überhaupt nicht	(1) An einzelnen Tagen	(2) An mehr als der Hälfte der Tage	(3) Beinahe jeden Tag
Wenig Interesse oder Freude an Ihren Tätigkeiten	☐	☐	☐	☐
Niedergeschlagenheit, Schwermut oder Hoffnungslosigkeit	☐	☐	☐	☐
Schwierigkeiten ein- oder durchzuschlafen oder vermehrter Schlaf	☐	☐	☐	☐
Müdigkeit oder Gefühl, keine Energie zu haben	☐	☐	☐	☐

beliebigen Skalenwerten theoretisch unendlich viele verschiedene Werte liegen.

In den Sozialwissenschaften werden häufig Daten transformiert, d.h. die Werte einer Variable werden in Werte einer anderen Variable verändert. Hat dabei die Zielvariable ein anderes Skalenniveau als die ursprüngliche Variable, muss beachtet werden, dass eine derartige Datentransformation nur durch Reduktion des Informationsgehalts erlaubt ist. D.h. das Skalenniveau der Zielvariable muss gleich oder niedriger dem Skalenniveau der ursprünglichen Variable sein.

2.1 Operationalisieren und Skalieren

In den Sozialwissenschaften hat man es oft mit Merkmalen zu tun, die nicht direkt gemessen werden können. Entsprechende Variablen werden als *latent* bezeichnet - in Abgrenzung zu *manifesten* Variablen, welche direkt gemessen werden können (Opp, 2013). Um latente Merkmale dennoch empirisch zugänglich zu machen, müssen diese *operationalisiert*, d.h. indirekt gemessen werden. Nehmen wir z.B. das Merkmal „Depression“ mit den Ausprägungen „ja“ (liegt vor) und „nein“ (liegt nicht vor). Untersuchungsobjekten (Personen) eine bestimmte Ausprägung der Eigenschaft Depression zuzuschreiben bedeutet: für konkrete Personen zu messen, ob sie eine Depression haben oder nicht. Nun kann man das Vorliegen einer Depression aber nicht direkt messen; es gibt kein Teststreifen oder Ähnliches. Das Merkmal Depression ist also latent. Die Operationalisierung als (indirekte) Messbarmachung bedeutet nun, das latente Merkmal auf manifeste, d.h. unmittelbar messbare, Sachverhalte zu reduzieren.

Zur Demonstration soll der Gesundheitsfragebogen PHQ (Kroenke et al., 2001) herangezogen werden. Das zweite Item dieses Fragebogens („Niedergeschlagenheit, Schwermut oder Hoffnungslosigkeit“) wurde hier bereits exemplarisch verwendet. Der komplette PHQ-Fragebogen umfasst 9 Items, welche in

der Summe Depression diagnostizieren sollen. Zur Veranschaulichung genügt der Rückgriff auf lediglich 4 Items, welche als Indikatoren für Depression fungieren; d.h. Depression wird auf die Items als manifeste Variablen reduziert. Diese Reduktion erschöpft sich allerdings nicht mit der Messung jener manifester Sachverhalte, die mit den Items jeweils abgefragt werden. Immerhin ist es hier das beispielhafte Ziel, *eine* latente Variable zu messen - nicht vier manifeste Variablen. Die Messung der vier Indikatoren ist lediglich ein erforderlicher Schritt dabei. Letztlich bedarf es aber noch einer Vorschrift, wie die vier einzelnen Messwerte der Indikatoren zu einem Messwert für die latente Variable Depression zusammengefasst werden. Eine solche Vorschrift wird als operationale Definition bezeichnet und könnte hier lauten: „Wenn mindestens zwei der hier aufgeführten Sachverhalte an mindestens der Hälfte der Tage auftreten, dann hat die betreffende Person eine Depression".

Mit ihrer konkreten operationalen Definition wird eine latente Variable gleichsam skaliert. Für das eben besprochene Beispiel formuliert die operationale Definition eine Bedingung, die - insofern sie erfüllt ist - darauf verweist, dass eine Depression vorliegt; ist die Bedingung nicht erfüllt, liegt auch keine Depression vor. Die operationale Definition führt hier also dazu, dass die latente Variable durch eine nominal skalierte Variable mit den Ausprägungen „Depression liegt nicht vor" und „Depression liegt vor" repräsentiert wird.

Teil II

Deskriptive Statistik

Kapitel 3

Univariate Verteilungen

3.1 Univariate Verteilungen

Die Aufgabe der univariaten deskriptiven Statistik ist das Beschreiben der Daten mit Zahlen. Es soll eine übersichtliche Darstellung der wesentlichen in den erhobenen Daten enthaltenen Informationen erreicht werden. Dabei bedeutet „univariat“, dass die untersuchten Variablen jeweils einzeln für sich betrachtet werden. Es geht also hier noch nicht um Zusammenhänge zwischen Variablen.

Wenn in der univariaten deskriptiven Statistik von Verteilungen die Rede ist, dann sind damit Häufigkeitsverteilungen gemeint. Die (Häufigkeits-) Verteilung einer Variable ist die Anzahl des Vorkommens der Merkmalsausprägungen (Variablenwerte) der jeweiligen Variable. Die Darstellung univariater Häufigkeitsverteilungen kann in Häufigkeitstabellen oder grafisch erfolgen.

Prinzipiell können Häufigkeitstabellen unabhängig vom Skalenniveau einer Variable erstellt werden - jedoch wird diese Statistik nur für Merkmale mit einer nicht zu großen, *überschaubaren Anzahl von Merkmalsausprägungen* empfohlen, i.d.R. mithin für *diskrete Skalen*.

Die *absolute Häufigkeit* der Merkmalsausprägung j des Merkmals X sei durch

$$f\left(X=j\right)$$

angegeben. Die *relative Häufigkeit* gibt den Anteil der Fälle an, welche diese Merkmalsausprägung aufweisen:

$$p\left(X=j\right)=\frac{f\left(X=j\right)}{n}$$

wobei n die Anzahl aller Fälle mit Messwerten für die jeweilige Variable X ist.

Ist die Variable *mindestens ordinal skaliert* (Kapitel 2), ist weiterhin die Bestimmung der *kumulierten Häufigkeiten* möglich, d.h. das Aufsummieren der Häufigkeiten von Merkmalsausprägungen, die kleiner oder gleich j sind.

Tabelle 3.1: Häufigkeitsverteilung *Depression*

Depression	**f**	**%**
Nein	889	77.24
Ja	262	22.76
n	1151	

Datenquelle: GRD (Schmidl, 2014).

Die *kumulierten absoluten Häufigkeiten* ergeben sich also mit

$$f\left(X \leq j\right) = \sum_{k=min(x)}^{j} f\left(X = k\right)$$

und die *kumulierten relativen Häufigkeiten* mit

$$p\left(X \leq j\right) = \frac{f\left(X \leq j\right)}{n}$$

Als Beispiel für eine in einer Häufigkeitstabelle dargestellten Verteilung greifen wir als erstes auf die in Abschnitt 2.1 exemplarisch operationalisierte Variable Depression zurück. Die dichotome Skala dieser Variable umfasst die Ausprägungen „Nein“ (Depression liegt nicht vor) und „Ja“ (Depression liegt vor), ihr Skalenniveau ist somit nominal[1].

Tabelle 3.1 stellt die Häufigkeitsverteilung der Variable Depression tabellarisch dar. Weil die Variable nominal skaliert ist, ist die Reihenfolge der Merkmalsausprägungen in der Tabelle beliebig. Daher können hier auch keine kumulierten Häufigkeiten sinnvoll angegeben werden. Die relativen Häufigkeiten sind als Prozente (=relative Häufigkeit multipliziert mit 100) wiedergegeben.

Ordinales Skalenniveau hat hingegen eine Variable, die das Item aus dem PISA 2015 Lehrerfragebogen (OECD, 2015) mit der Frage „What is your current employment status as a teacher?“ abbildet (Employment status). Die entsprechende Antwortskala ist: 1=*Part-time (less than 50% of full-time hours)*, 2=*Part-time (50-70%)*, 3=*Part-time (71-90%)*, 4=*Full-time (more than 90%)*. Daher kann die entsprechende Häufigkeitstabelle (Tabelle 3.2) auch die kumulierten relativen Häufigkeiten enthalten.

Als letztes Beispiel für eine tabellarische Darstellung der Häufigkeiten sei die Variable Äquivalenzhaushaltseinkommen (gemessen in der Einheit Euro) verwendet. Die so gemessenen Daten sind metrisch und ratio-skaliert. Insofern diese Variable in der Praxis als stetig charakterisiert wird, ist das Anlegen einer Häufigkeitstabelle ohne weitere Maßnahmen nicht adäquat. Jedoch ist eine Klassierung - d.h. eine Gruppierung der Daten in Messwertklassen - möglich. Das Klassieren einer metrischen Variable bedeutet die Transformation in ordinales

[1] Eine Skala wird als *dichotom* bezeichnet, wenn sie genau zwei Ausprägungen umfasst. Das Skalenniveau dichotomen Variable ist immer nominal.

Tabelle 3.2: Häufigkeitsverteilung *Employment status*

Employment status	**f**	**%**	**kum. %**
Part-time <50%	3618	4.16	4.16
Part-time 50-70%	5573	6.41	10.57
Part-time 71-90%	5194	5.97	16.54
Full-time >90%	72611	83.46	100
n	86996		

Datenquelle: PISA 2015 (OECD, 2015).

Tabelle 3.3: Häufigkeitsverteilung *Äquivalenzhaushaltseinkommen*

Äquivalenzhaushaltseinkommen	**f**	**%**	**kum. %**
bis unter 1000 €	345	28.09	28.09
1000 € bis unter 2000 €	632	51.47	79.56
2000 € bis unter 3000 €	209	17.02	96.58
3000 € bis unter 4000 €	35	2.85	99.43
4000 € oder mehr	7	0.57	100
n	1228		

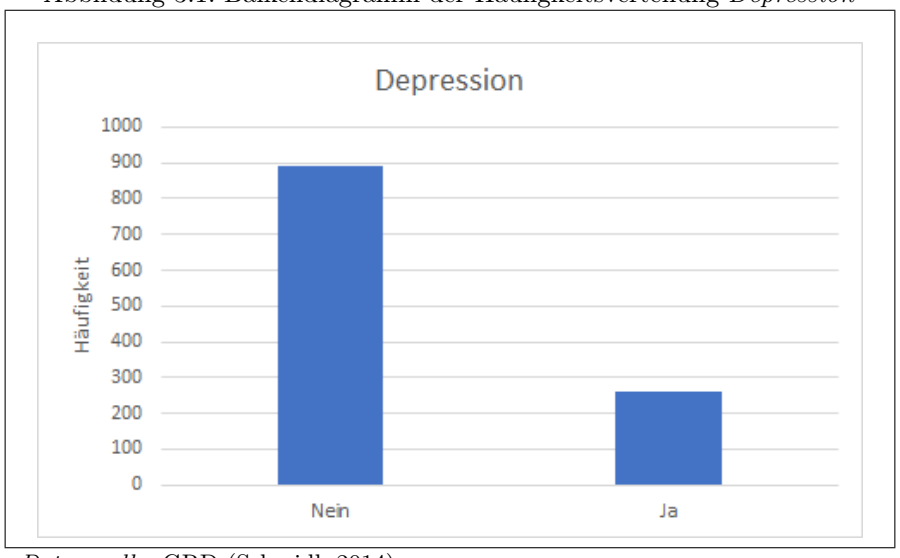
Datenquelle: GRD (Schmidl, 2014).

Skalenniveau; für entsprechend transformierte Variablen ist dann die Darstellung der Häufigkeitsverteilung in einer Tabelle angemessen. Tabelle 3.3 zeigt tabellarisch die (fiktive) Häufigkeitsverteilung des Äquivalenzhaushaltseinkommens.

Die grafische Darstellung einer Häufigkeitsverteilung ist ebenfalls abhängig vom Daten- bzw. Skalenniveau (Tabelle 2.1). Die Verteilung kategorialer Daten (nominal oder ordinal skalierter Variablen) sollte in Balkendiagrammen vermittelt werden, für metrische Daten (intervall- oder ratio-skalierte Variablen) ist hingegen das Histogramm die geeignete Visualisierung. In einem Balkendiagramm wird jeder Variablenwert mit einem Balken dargestellt, welcher von den anderen Balken im Diagramm abgetrennt ist. Dagegen stoßen die Balken in einem Histogramm aneinander. Die Werte stetig-skalierter Variablen werden für die Verwendung in einem Histogramm zu Messwertklassen zusammengefasst - allerdings nur für die grafische Darstellung; die metrischen Daten behalten ihr ursprüngliches Skalenniveau und werden nicht ordinal re-skaliert. Abbildung 3.1 und Abbildung 3.2 zeigen beispielhaft die Verteilung der Variablen Depression (als Balkendiagramm) und Äquivalenzhaushaltseinkommen (als Histogramm) respektive.

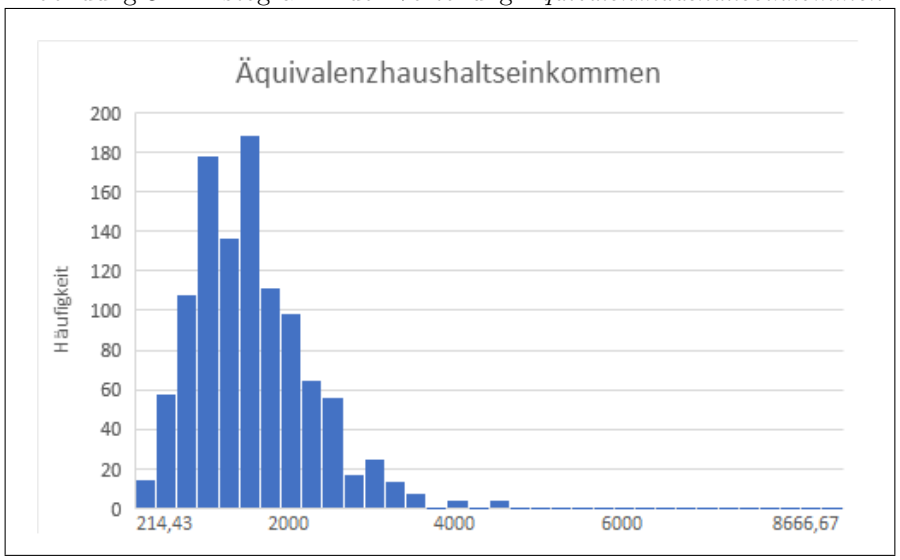
Für mindestens ordinal skalierte Variablen ist die Form ihrer Verteilung hinsichtlich der Anzahl der Gipfel, Schiefe, und Wölbung charakterisierbar. Bezüglich der Anzahl der Gipfel werden üblicherweise Verteilungen mit einem, zwei,

Abbildung 3.1: Balkendiagramm der Häufigkeitsverteilung *Depression*



Datenquelle: GRD (Schmidl, 2014).

Abbildung 3.2: Histogramm der Verteilung *Äquivalenzhaushaltseinkommen*



Datenquelle: GRD (Schmidl, 2014).

oder mehreren Gipfeln unterschieden und entsprechend als unimodal, bimodal, oder multimodal respektive bezeichnet. Schiefe und Wölbung lassen sich sinnvoll nur für unimodale Verteilungen interpretieren. Dabei gibt die Schiefe die Abweichung einer zum Gipfel symmetrischen Verteilungsform an. Eine Häufigkeitsverteilung ist demnach dann symmetrisch, wenn der Gipfel in der Mitte liegt; asymmetrische Verteilungsformen werden als „schief“ bezeichnet. Die relative Höhe des Gipfels wird durch die Wölbung beschrieben: demnach ist die Verteilungsform „schmal“, wenn extremale Werte sehr selten vorkommen und somit der „Höhenunterschied“ zwischen Gipfel und „Tal“ sehr groß ist; ist dieser Unterschied hingegen klein, so spricht man von einer „flachen“ Verteilungsform. Die Form der Häufigkeitsverteilung des Äquivalenzhaushaltseinkommens (Tabelle 3.3 und Abbildung 3.2) kann hier beispielhaft als unimodal, (rechts-) schief, und schmalgipflig umschrieben werden.

3.2 Mittelwerte

Mittelwerte (auch als „Lagemaße“ bezeichnet) kennzeichnen die zentrale Lage bzw. zentrale Tendenz einer Verteilung. Vier Mittelwerte sollen im Folgenden vorgestellt werden: Modalwert (D), Median (Z), arithmetisches Mittel ($\bar{x}$), harmonisches Mittel ($\bar{x}_H$).

Der *Modalwert D* ist geeignet für Merkmale mit einer nicht zu großen, *überschaubaren Anzahl von Merkmalsausprägungen*, i.d.R. für *kategoriale Daten*; er bezeichnet die Kategorie mit der größten Häufigkeit bzw. den am häufigsten auftretenden Messwert einer Verteilung. Mehrere Modalwerte bei einer Verteilung sind möglich - z.B. zwei Modalwerte in einer bimodalen Verteilung. Der Modalwert kann interpretiert werden als der Wert einer Variable, der „am ehesten“ zu beobachten ist. Er ist unempfindlich gegenüber Ausreißern (extremen Werten mit geringer Häufigkeit), jedoch relativ unzuverlässig in Wertebereichen großer Dichte bzw. bei flachen Verteilungsformen. Die Dichte eines Werte-Intervalls in der Verteilung einer metrischen Variable ist durch die kumulierte relative Häufigkeit innerhalb dieses Intervalls gegeben. Bereiche großer Dichte sind also Intervalle, in denen die auf der Skala benachbarten Werte eine ähnlich große Häufigkeit aufweisen. Die Gleichverteilung ist eine extreme Variante einer derart flachen Verteilung, an welcher die Unzuverlässigkeit des Modalwertes bei solchen Verteilungen demonstriert werden kann. Eine Gleichverteilung liegt vor, wenn alle Messwerte mit der gleichen Häufigkeit vorkommen. Angenommen, es wurden bei 100 Fällen 10 verschiedene Werte jeweils 10 mal gemessen: dann hat diese Verteilung 10 Modalwerte. Wenn jedoch jeder Wert ein Modalwert ist, dann lässt sich diese Information kaum mit Mehrwert interpretieren. Allerdings kommen vollständig gleich-verteilte Variablen selten vor - jedenfalls seltener als *annähernd* gleich-verteilte Variablen. Unzuverlässigkeit von Modalwerten meint nun, dass bei solchen flach- bzw. annähernd gleich-verteilten Variablen nur geringfügige Veränderungen der Häufigkeiten zu beachtlichen Unterschieden bezüglich der Lage der Modalwerte führen können. Abschließend sei darauf hingewiesen, dass der Modalwert empfindlich gegenüber den Klassengrenzen

gruppierter Daten ist. D.h., dass sich die Entscheidung, welche Werteintervalle ggf. zu Klassen zusammengefasst werden, auf die Lage des Modalwerts oder der Modalwerte auswirkt. Ungeachtet dieser Vorbehalte ist der Modalwert für eingipflige und nicht zu flache Verteilungen ein brauchbarer und anschaulicher Mittelwert, indem er auf den Verteilungsgipfel zeigt. Darüber hinaus ist der Modalwert das einzige adäquate Lagemaß für nominal skalierte Variablen.

Für die Variable *Employment status* (mit der Verteilung aus Tabelle 3.2) ist der Modalwert $D = 4$ (*Full time >90%*) - die Antwortkategorie mit der im Vergleich zu den anderen Antwortmöglichkeiten größten Häufigkeit ($f = 72611 \hat{=} 83.46\%$).

Ist eine Variable mindestens *ordinal skaliert*, dann kann der *Median* Z als adäquater Mittelwert bestimmt werden. Voraussetzung dafür ist nämlich, dass sich die Beobachtungswerte x_i in einer *Rangreihe* (nach der Größe der Variablenwerte) ordnen lassen[2]. Der Median ist nun der „mittlere Beobachtungswert", der die Rangreihe „halbiert"; also der 50%-Punkt. Bei einer einer *ungeraden* Anzahl von n Fällen ist der Median der Beobachtungswert mit dem Rang $Rg\,(med) = \frac{n+1}{2}$, für n *gerade* ist der Median das arithmetische Mittel der „mittleren" Beobachtungen („Interpolation") mit den Rängen $Rg\,(med_f) = \frac{n}{2}$ und $Rg\,(med_c) = \frac{n}{2} + 1$. Oberhalb und unterhalb des Medians befinden sich gleich viele Beobachtungswerte.

$$Z = \begin{cases} x_{Rg(med)} & \text{für n ungerade} \\ \frac{x_{Rg\left(med_f\right)} + x_{Rg(med_c)}}{2} & \text{für n gerade} \end{cases}$$

Interpretiert werden kann der Median als zentraler Wert bei mindestens ordinal skalierten Variablen. Er gibt - unempfindlich gegenüber Ausreißern - das Zentrum der Verteilung an. Wird der Median für metrische Daten ermittelt, zeigt sich zudem seine Minimaleigenschaft bezüglich der absoluten Abweichungen zu den Beobachtungswerten: $\sum_{i=1}^{n} |x_i - \zeta| \rightarrow min$ hat die Lösung $\zeta = Z$.

Exemplarisch für die Bestimmung des Medians verwenden wir hier die Daten der Variable *Äquivalenzhaushaltseinkommen*. Die Häufigkeitsverteilung in tabellarischer Form (Abbildung 3.1) - ergänzt um die Rangplätze der Beobachtungswerte - liefert Tabelle 3.4. Weil hier eine gerade Anzahl von Fällen beobachtet wurde, müssen die Rangplätze der beiden mittleren Fälle ermittelt werden: $Rg\,(med_f) = \frac{n}{2} = \frac{1228}{2} = 614$ und $Rg\,(med_c) = \frac{n}{2} + 1 = 615$. Beide Rangplätze fallen in die Kategorie $[1000, 2000)$('1000 € bis unter 2000 €'), mithin ist diese Kategorie die Medianklasse: $Z = [1000, 2000)$. Mit den Informationen der Häufigkeitstabelle ist auch eine Schnellbestimmung möglich, da in der Medianklasse die kumulierten 50% liegen.

[2] Während die Skalenwerte die möglichen Messwerte einer Variable sind, handelt es sich bei den Beobachtungswerten um die für die Fälle tatsächlich gemessenen Werte der entsprechenden Variable. Der Beobachtungswert x_i ist also der Wert der Variable X, welcher für den Fall i gemessen wurde.

Tabelle 3.4: Häufigkeitsverteilung und Ränge der Beobachtungswerte von *Äquivalenzhaushaltseinkommen*

Äquivalenz-haushalts-einkommen (in €)	**f**	**%**	**kum. %**	**kum. f**	**Rangplätze**
1: $[0, 1000)$	345	28.09	28.09	345	$1 \ldots 345$
2: $[1000, 2000)$	632	51.47	79.56	977	$346 \ldots 977$
3: $[2000, 3000)$	209	17.02	96.58	1186	$978 \ldots 1186$
4: $[3000, 4000)$	35	2.85	99.43	1221	$1187 \ldots 1221$
5: $[4000, \infty)$	7	0.57	100	1228	$1222 \ldots 1228$
n	1228				

Datenquelle: GRD (Schmidl, 2014).

Das *arithmetische Mittel* $\bar{x}$ kann für *metrisch skalierte* Variablen berechnet werden. Es ergibt sich als Summe der singulären Messwerte x_i geteilt durch die Anzahl der Fälle.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Mit der Eigenschaft $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ lässt sich das arithmetische Mittel als Schwerpunkt der jeweiligen Verteilung interpretieren. Aussagekräftig als „Mitte“ der Verteilung ist dieser Kennwert jedoch lediglich für eingipflige und symmetrische Verteilungen. Zudem ist das arithmetische Mittel empfindlich gegenüber Ausreißern. Eine wichtige statistische Eigenschaft ist die Minimaleigenschaft bezüglich der quadratischen Abweichungen: $\sum_{i=1}^{n} (x_i - \zeta)^2 \rightarrow min$ hat die Lösung $\zeta = \bar{x}$.

Die folgende Liste mit Klausurergebnissen einer Gruppe von Studierenden ($n = 11$) in einer Statistik-Klausur ist die Datenbasis für eine beispielhafte Berechnung des arithmetischen Mittels.

Student i	1	2	3	4	5	6	7	8	9	10	11
Punkte Statistik x	39	34	31	48	46	23	17	12	16	28	10

$$\bar{x} = \frac{39 + 34 + 31 + 48 + 46 + 23 + 17 + 12 + 16 + 28 + 10}{11} = 27.6364$$

Die Gruppe von Studierenden hat in der Statistik-Klausur eine durchschnittliche Punktzahl von $\bar{x} = 27.6$.

Ein adäquater Mittelwert für *ratio-skalierte* Variablen kann (neben dem arithmetischen Mittel und dem Median) das *harmonische Mittel* $\bar{x}_H$ sein - nämlich dann, wenn die Messwerte *Quotienten* sind.

Tabelle 3.5: Haushaltseinkommen pro Kopf, fiktive Datenreihe 1

Haushalts-einkommen (Y)	**Haushaltsgröße (K)**	**Haushaltseinkommen pro Kopf (X)**
3448	3	1149.33
3045	3	1015.00
3554	2	1777.00
2124	2	1062.00
3427	4	856.75
2709	1	2709.00
2216	3	738.67

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Es sei die Variable X ein Quotient aus der ratio-skalierten Variable Y im Zähler und der metrischen Variable K im Nenner. Das Haushaltseinkommen pro Kopf (X) ist eine derartige Variable mit dem Haushaltseinkommen (Y) als Zähler und der Haushaltsgröße, d.h. der Anzahl der im Haushalt lebenden Personen (K), im Nenner: $x_i = \frac{y_i}{k_i}$. Tabelle 3.5 enthält die Daten für das Beispiel.

Das arithmetische Mittel hier ist
$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1149.33+1015+1777+1062+856.75+2709+738.67}{7} = 1329.68$. Problematisch ist das arithmetische Mittel hier dann, wenn es um das durchschnittliche verfügbare Einkommen von *Personen* geht. In diesem Beispiel wird mit dem arithmetischen Mittel allerdings das pro-Kopf-Einkommen von Haushalten gemittelt. Dabei wird nicht berücksichtigt, wie viele Personen denn von dem jeweiligen pro-Kopf-Einkommen leben. Eine statistische Korrektur ist mit einem gewichteten arithmetischen Mittel erzielbar:
$\bar{x}_w = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} k_i} = \frac{3448+3045+3554+2124+3427+2709+2216}{3+3+2+2+4+1+3} = \frac{20523}{18} = 1140.17$. Der Vorteil gegenüber dem (ungewichteten) arithmetischen Mittel ist, dass die Haushaltsgrößen mit in die Berechnung einbezogen werden. Die Nachteile sind: für die Variablen Y und K müssen die jeweiligen Messwerte vorliegen. Zudem reagiert das gewichtete arithmetische Mittel auf die konkreten Werte von Y und K. D.h., dass z.B. für eine identische X-Datenreihe von Quotienten - welcher jedoch andere Y- und K-Datenreihen von Zählern und Nennern respektive zugrunde liegen - ein anderes gewichtetes arithmetisches Mittel berechnet werden könnte. Diese Reaktion des gewichteten arithmetischen Mittels auf die Y- und K-Werte sei beispielhaft mit Datenreihe 2 (Tabelle 3.6) illustriert, bei welcher die X-Werte identisch mit jenen von Datenreihe 1 sind: $\bar{x}_{(2)_w} = \frac{35952.09}{30} = 1198.40$.

Die Reaktivität des gewichteten arithmetischen Mittels ist insofern problematisch, weil ein Mittelwert für die Variable X gefragt ist, und das gewichtete

Tabelle 3.6: Haushaltseinkommen pro Kopf, fiktive Datenreihe 2

Haushalts-einkommen (Y)	**Haushaltsgröße** (K)	**Haushaltseinkommen pro Kopf** (X)
5746.67	5	1149.33
3045.00	3	1015.00
5331.00	3	1777.00
4248.00	4	1062.00
4283.75	5	856.75
8127.00	3	2709.00
5170.67	7	738.67

arithmetische bei gleichen X-Werten - aber anderen Y- und K-Werten - einen anderen Mittelwert liefert. Darüber hinaus sind die für das gewichtete arithmetische Mittel notwendigen Informationen - nämlich die Messwerte für die Variablen Y und K - nicht unbedingt in den Daten vorhanden. Dieser Problemkomplex wird durch das harmonische Mittel behoben. Mit dem harmonischen Mittel wird berücksichtigt, dass es sich bei den Messwerten der Variable X um Quotienten handelt. Zudem sind dafür lediglich die Messwerte von X erforderlich; insofern die Messwerte für Y und K unberücksichtigt bleiben, kann das harmonische Mittel nicht darauf reagieren. Das harmonische Mittel der Variable X ist sowohl für Datenreihe 1 als auch für Datenreihe 2 (mit identischen X-Werten in Tabelle 3.5 und Tabelle 3.6 respektive):

$$\bar{x}_H = \frac{7}{\frac{1}{1149.33} + \frac{1}{1015} + \frac{1}{1777} + \frac{1}{1062} + \frac{1}{856.75} + \frac{1}{2709} + \frac{1}{738.67}} = 1120.04$$

Ein Nachteil des harmonischen Mittels ist, dass es inhaltlich etwas schwerer zu interpretieren ist als das gewichtete arithmetische Mittel. Für das hier herangezogene Beispiel des Haushaltseinkommens pro Kopf beschreibt das harmonische Mittel das durchschnittliche verfügbare Einkommen von Personen, wenn ein bestimmtes (konstantes) Einkommen auf diejenige Anzahl von Personen verteilt werden müsste, die sich aus dem empirischen pro-Kopf-Einkommen der Haushalte ergibt[3]. Somit gilt, dass harmonisches Mittel und gewichtetes arithmetisches Mittel gleich sind, wenn der Zähler Y konstant ist. Die Mittelwerte für Datenreihe 3 (Tabelle 3.7) verdeutlichen diese Verbindung: $\bar{x}_{(3)_w} = 1120.04$, $\bar{x}_H = 1120.04$.

Abschließend sollen die Mittelwerte Modus, Median, und arithmetisches Mittel anschaulich verglichen werden. Dazu begreifen wir Häufigkeiten als Gewicht der jeweiligen Merkmalsausprägungen. Im Lichte dieser Analogie bestimmen wir den Modalwert, indem wir für jede Merkmalsausprägung einzeln eine Ge-

[3]Zum Vergleich die inhaltliche Interpretation des gewichteten arithmetischen Mittels des Haushaltseinkommens pro Kopf für Datenreihe 1 (Tabelle 3.5): 20523€ werden auf 18 Personen verteilt; dieses durchschnittliche pro-Kopf-Einkommen der Personen beträgt 1140.17€.

Tabelle 3.7: Haushaltseinkommen pro Kopf, fiktive Datenreihe 3

Haushalts-einkommen (Y)	**Haushaltsgröße** (K)	**Haushaltseinkommen pro Kopf** (X)
1330	1.16	1149.33
1330	1.31	1015.00
1330	0.75	1777.00
1330	1.25	1062.00
1330	1.55	856.75
1330	0.49	2709.00
1330	1.80	738.67

wichtskraftmessung mit z.B. einer Federwaage durchführen und diejenige Ausprägung als Modalwert bezeichnen, welche das größte Gewicht hat. Für den Median hingegen werden alle Merkmalsausprägungen auf einmal mit einer Balkenwaage gemessen. Dazu werden alle Messwerte nach ihrer Größe (nicht ihrem Gewicht bzw. ihrer Häufigkeit!) in einer Rangliste angeordnet. Es wird diejenige Merkmalsausprägung als Median bezeichnet, an welcher diese Rangliste geteilt werden muss, damit die Massen der beiden Teile gleich groß sind. In beiden Waagschalen einer Balkenwaage befindet sich dann das gleiche Gewicht. Um das arithmetische Mittel zu bestimmen, ist eine Laufgewichtswaage ein geeignetes Instrument zur Veranschaulichung: hierbei wird die Verteilung über der Achse mit den Merkmalsausprägungen ausbalanciert; das arithmetische Mittel ist dann genau der Punkt (die Merkmalsausprägung) auf der Achse, an welcher die Verteilung ausbalanciert ist.

3.3 Streuwerte

Streuwerte beschreiben die Variabilität oder die „Breite“ einer Verteilung. Als eine Auswahl der für die univariate Statistik bedeutendsten Streuwerte seien hier relativer Informationsgehalt (h), mittlerer Quartilsabstand (MQD), durchschnittliche absolute Abweichung (e), Standardabweichung (s), und Variationskoeffizient (V) thematisiert.

Geeignet zur Beschreibung der Streuung *nominal-skalierter* Daten ist der relative *Informationsgehalt* h.

$$h = \frac{k}{n}\left(\frac{ln\,(n)\,n - \left(\sum_{j=1}^{k} n_j ln\,(n_j)\right)}{ln\,(k)\,k}\right)$$

Dabei ist k die Anzahl der Kategorien der jeweiligen Skala. Der relative Informationsgehalt hat das theoretisches Wertespektrum $0 \leq h \leq 1$, wobei $h = 0$

bedeutet, dass keine Streuung in Daten vorliegt und das entsprechende Merkmal mithin konstant ist; dagegen verweist das gegenteilige Extrem $h = 1$ auf die größtmögliche Streuung: das Merkmal ist gleich-verteilt, d.h. alle Merkmalsausprägungen bzw. Kategorien der Skala kommen mit der gleichen Häufigkeit vor. Zur Illustration des relativen Informationsgehalts werden die Beispieldaten der Variable *Depression* herangezogen (Tabelle 3.1, Abbildung 3.1). Demnach ist $h = \frac{2}{1151}\left(\frac{ln(1151)1151-(ln(889)889+ln(262)262)}{ln(2)2}\right) = 0.7739$.

Die Streuung mindestens *ordinal-skalierter* Daten kann mit dem *mittleren Quartilsabstand* MQD erfasst werden. Dabei handelt es sich quasi um eine Streuung um den Median. Dementsprechend basiert dieser Streuwert auf der Rangreihe der Beobachtungswerte (s. Median). Soll der mittlere Quartilsabstand für ordinale Daten ermittelt werden, ist es für eine sinnvolle Interpretation erforderlich, dass die Kategorien ganzzahlig und fortlaufend mit einem Abstand von 1 codiert sind. Für die Berechnung müssen zunächst die Quartile ermittelt werden, deren mittlerer Abstand zum Median der MQD angibt. Quartile sind Beobachtungswerte, welche die Rangreihe „vierteln“. Somit gibt es drei Quartile: Q_1 (unteres Quartil) ist der Beobachtungswert mit dem Rang $Rg\,(0.25)$, Q_3 (oberes Quartil) ist der Beobachtungswert mit dem Rang $Rg\,(0.75)$, und Q_2 (mittleres Quartil) entspricht dem Median $Z = Rg\,(0.5)$. Allgemein lässt sich der Beobachtungswert $x_{Rg(q)}$ bestimmen mit

$$x_{Rg(q)} = \begin{cases} x_{\lceil qn \rceil} & \text{wenn } \lfloor qn \rfloor \neq qn \\ \frac{x_{qn}+x_{qn+1}}{2} & \text{wenn } \lfloor qn \rfloor = qn \end{cases}$$

Der mittlere Quartilsabstand ist formal folgendermaßen definiert:

$$MQD = \frac{1}{2}\left|Q_3 - Q_1\right|$$

Für die Beispieldaten des Äquivalenzhaushaltseinkommens (Tabelle 3.4) ergeben sich zunächst die Quartils-Kategorien $Q_1 = \frac{x_{307}+x_{308}}{2} = 1$ und $Q_3 = \frac{x_{921}+x_{922}}{2} = 2$. In Worten: jenes Viertel der Fälle mit dem geringsten Äquivalenzhaushaltseinkommen, erzielt bis unter 1000 € (Kategorie 1); das Viertel der Fälle mit dem höchsten Äquivalenzeinkommen erzielt mindestens 2000 € (Kategorie 2). Hier ist der mittlere Quartilsabstand $MQD = 0.5\,(2-1) = 0.5$. Damit liegt die mittlere Hälfte der Fälle durchschnittlich eine halbe Kategorie vom Median entfernt. Im obigen Beispiel zum Median konnte festgestellt werden, dass Kategorie 2 die Median-Kategorie ist: das untere Quartil liegt um eine Kategorie davon entfernt, das obere Quartil liegt null Kategorien davon entfernt - gemittelt also eine halbe Kategorie.

Zur Berechnung der *durchschnittlichen absoluten Abweichung* e sind *metrische* Daten erforderlich. Dieser Streuwert gibt die mittlere Differenz der Beobachtungswerte zu einem Mittelwert (i.d.R. arithmetisches Mittel oder Median) an. Sei ζ ein Mittelwert, dann ist die durchschnittliche absolute Abweichung der Beobachtungswerte zu diesem Bezugswert

$$e_\zeta = \frac{1}{n} \sum_{i=1}^{n} |x_i - \zeta|$$

Folgende Tabelle enthält die Beispieldaten, welche zur Erläuterung des arithmetischen Mittels benutzt wurden - hier allerdings in einer Rangreihe nach der Größe der Beobachtungswerte geordnet:

Rang Rg	1	2	3	4	5	6	7	8	9	10	11
Punkte Statistik x	10	12	16	17	23	28	31	34	39	46	48

Es soll nun beispielhaft die durchschnittliche absolute Abweichung dieser Beobachtungswerte zum Median als Bezugswert ermittelt werden. Der Median ist in diesem Fall $Z = x_6 = 28$. Mithin ergibt sich für die durchschnittliche absolute Abweichung zum Median

$$e_Z = \frac{1}{11} \left(\begin{array}{c} |10-28| + |12-28| + |16-28| + |17-28| \\ + |23-28| + |28-28| + |31-28| + |34-28| \\ + |39-28| + |46-28| + |48-28| \end{array} \right) = 10.9091$$

Zum arithmetischen Mittel beträgt die durchschnittliche absolute Abweichung $e_{\bar{x}} = 10.9421$. Der Median ist derjenige Bezugswert, für welchen die durchschnittliche absolute Abweichung minimal ist (d.h. es gibt keinen anderen Bezugswert ζ, zu welchem e_ζ kleiner ist als e_Z, siehe Minimaleigenschaft Median; Hartung, 2005).

Ein weiterer Streuwert für *metrische* Daten ist die *Standardabweichung* s.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Durch das Quadrieren haben größere Abweichungen der jeweiligen Beobachtungswerte vom arithmetischen Mittel einen stärkeren Einfluss. Die quadrierte Standardabweichung ist bekannt als *Varianz* s^2. Werden hier erneut die Beispieldaten verwendet, welche bereits beim arithmetischen Mittel und der durchschnittlichen absoluten Abweichung zum Einsatz gekommen sind, so erhalten wir für die Standardabweichung

$$s = \sqrt{\frac{1}{11-1} \left(\begin{array}{c} (39-27.636)^2 + (34-27.636)^2 + (31-27.636)^2 \\ + (48-27.636)^2 + (46-27.636)^2 + (23-27.636)^2 \\ + (17-27.636)^2 + (12-27.636)^2 + (16-27.636)^2 \\ + (28-27.636)^2 + (10-27.636)^2 \end{array} \right)} = 13.261$$

Der letzte hier zu besprechende Streuwert, der *Variationskoeffizient* V, erfordert *ratio-skalierte* Daten, weil hier die Standardabweichung zum arithmetischen Mittel relativiert und als Quotient - Anteil der Standardabweichung am arithmetischen Mittelwert - ausgedrückt wird:

Tabelle 3.8: Beispiel Variationskoeffizient

	Entfernung vom Baumstamm (gemessen in Metern)		
Frucht	**arithmetisches Mittel**	**Standard-abweichung**	**Variations-koeffizient**
Apfel	2.5	1.0	40
Birne	2.0	0.5	25
Walnuss	7.5	3.0	40

$$V = \frac{s}{\bar{x}} 100$$

Diese Relativierung ermöglicht die Vergleichbarkeit der Streuung, insofern der Variationskoeffizient unabhängig von der Skala der jeweiligen Variablen ist.

Für das Beispiel wird eine Plantage angenommen, auf welcher 3 Sorten von Bäumen mit den jeweiligen Früchten stehen: Äpfel, Birnen, Walnüsse. Nun soll die Streuung der abgefallenen Früchte um den Baumstamm vergleichend für die jeweiligen Fruchtsorten untersucht werden. Die erforderlichen Werte und die daraus berechneten Variationskoeffizienten können Tabelle 3.8 entnommen werden.

Kapitel 4

Bivariate Verteilungen

4.1 Bedingte Häufigkeitsverteilungen

Bedingte Häufigkeitsverteilungen können mit *Kreuztabellen* dargestellt werden. Kreuztabellen sind Häufigkeitstabellen für zwei Variablen zugleich. Dabei werden also die Kombinationen von Ausprägungen zweier Merkmale gezählt bzw. die Häufigkeitsverteilung einer Variable (Y) unter den Bedingungen - d.h. Ausprägungen - einer anderen Variable (X). Diese Art der Darstellung einer bivariaten Verteilung ist dann geeignet, wenn beide Variablen eine diskrete Skala mit wenigen Merkmalsausprägungen aufweisen. Üblicherweise werden die Ausprägungen der Variable X in den Zeilen und die Ausprägungen der Variable Y in den Spalten gezählt. Mithin entsprechen die Zeilensummen der univariaten Häufigkeitsverteilung von X und die Spaltensummen der univariaten Häufigkeitsverteilung von Y. Der Mehrwert an Information, welcher sich mit Kreuztabellen erschließen lässt, ist die Beschreibung des statistischen Zusammenhangs der beiden Variablen. Wird dabei von einem Kausalzusammenhang (Ursache-Wirkungs-Zusammenhang) ausgegangen, so soll X die unabhängige Variable (Ursache) und Y die abhängige Variable (Wirkung) sein. Tabelle 4.1 zeigt den Aufbau einer Kreuztabelle für ein Merkmal X mit k Ausprägungen und einem Merkmal Y mit m Ausprägungen.

Neben den in Tabelle 4.1 formal angezeigten absoluten Häufigkeiten lassen sich in Kreuztabellen zwei Varianten von relativen Häufigkeiten darstellen. Bei den *globalen* relativen Häufigkeiten $p\,(X = j, Y = l) = \frac{f(X=j,Y=l)}{n}$ ist die Gesamtzahl der Fälle n der Bezug, zu welchem der Anteil der Wertekombinationen ermittelt wird. Die *bedingten* relativen Häufigkeiten können zu den jeweiligen Ausprägungen sowohl des Merkmals X („zeilenweise“) als auch Y („spaltenweise“) berechnet werden, sodass dann entsprechend die Zeilen- oder die Spaltensumme als Bezug für die Anteile gilt: $p\,(Y = l\,|X = j\,) = \frac{f(X=j,Y=l)}{f(X=j)}$ oder $p\,(X = j\,|Y = l\,) = \frac{f(X=j,Y=l)}{f(Y=l)}$ respektive.

Ein Indiz für den Zusammenhang der beiden Merkmale lässt sich ggf. durch die Begutachtung der bedingten relativen Häufigkeiten erkennen. Wenn sich

Tabelle 4.1: Schematischer Aufbau einer Kreuztabelle für ein Merkmal X mit k Ausprägungen und ein zweites Merkmal Y mit m Ausprägungen

	Y_1	$\cdots$	Y_l	$\cdots$	Y_m	$\sum_X$
X_1	$f(X=1,Y=1)$	$\cdots$	$f(X=1,Y=l)$	$\cdots$	$f(X=1,Y=m)$	$f(X=1)$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_j	$f(X=j,Y=1)$	$\cdots$	$f(X=j,Y=l)$	$\cdots$	$f(X=j,Y=m)$	$f(X=j)$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_k	$f(X=k,Y=1)$	$\cdots$	$f(X=k,Y=l)$	$\cdots$	$f(X=k,Y=m)$	$f(X=k)$
$\sum_Y$	$f(Y=1)$	$\cdots$	$f(Y=l)$	$\cdots$	$f(Y=m)$	n

Tabelle 4.2: Kreuztabelle Glauben an Himmel und Glauben an Hölle

			Glauben an Hölle		
			nein	**ja**	$\sum$
Glauben an Himmel	**nein**	*f*	40693	979	41672
		% (Glauben an Himmel)	97.7	2.3	
	ja	*f*	17108	84417	101525
		% (Glauben an Himmel)	16.9	83.1	
	$\sum$	*f*	57801	85396	143197
		% (Glauben an Himmel)	40.4	59.6	

Datenquelle: World Values Survey 1981-2008 (Inglehart et al., 2014).

nämlich die Verteilung der bedingten relativen Häufigkeiten zwischen den Ausprägungen des Bezugsmerkmals hinreichend deutlich unterscheiden, dann ist dies ein Hinweis auf einen stochastischen Zusammenhang beider Variablen.

Die mit Tabelle 4.2 dargestellte Kreuztabelle, in welcher die Variablen *Glauben an Himmel* („Believe in: heaven“) und *Glauben an Hölle* („Believe in: hell“) - jeweils mit den Ausprägungen *nein* und *ja* - kreuz-tabelliert wurden, zeigt einen deutlichen Zusammenhang dieser Merkmale. Während ca. 98% von denen, die nicht an den Himmel glauben, ebenfalls nicht an die Hölle glauben, sind nur rund 17% von den Himmels-gläubigen nicht von der Existenz der Hölle überzeugt. Unter der Bedingung *Glauben an Himmel=nein* ist die Verteilung von *Glauben an Hölle* deutlich anders als unter der Bedingung *Glauben an Himmel=ja*.

4.2 Bivariate Streuungen

Die bivariate Verteilung metrisch skalierter Variablen kann mit der *Korrelation* beschrieben werden. Ebenso wie mit der Kreuztabelle kann mit der Korrelation der Zusammenhang der jeweiligen Variablen untersucht werden. Anders als bei der Kreuztabelle können hier aber nur lineare („Je-Desto-“) Zusammenhän-

ge ermittelt werden. Ein Kennwert der Korrelation ist der Produkt-Moment-Korrelationskoeffizient

$$r = \frac{cov\,(X,Y)}{s_X s_Y}$$

Der Produkt-Moment-Korrelationskoeffizient r ist die durch das Produkt der Standardabweichungen der Variablen X und Y normierte Kovarianz

$$cov\,(X,Y) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})\right)$$

Damit ist r ein normiertes Zusammenhangsmaß mit einem theoretischen Wertespektrum $-1 \leq r \leq 1$, wobei negative Werte auf einen entsprechend negativen („inversen“) Zusammenhang verweisen und positive Werte auf einen entsprechend positiven („gleichgerichteten“) Zusammenhang. Für die beispielhafte Berechnung und Interpretation des Produkt-Moment-Korrelationskoeffizienten werden erneut die Daten vom Beispiel des arithmetischen Mittels verwendet - nun allerdings werden die Daten zu den Statistik-Punkten (Variable X) durch Messwerten zu Mathe-Punkten (Variable Y) ergänzt:

Student i	1	2	3	4	5	6	7	8	9	10	11
Punkte Statistik (X)	39	34	31	48	46	23	17	12	16	28	10
Punkte Mathe (Y)	38	47	44	51	35	29	22	14	12	19	9

Für die hier relevanten univariaten Kennwerte - arithmetisches Mittel und Standardabweichung - lassen sich für die Variable X (Punkte Statistik) ermitteln $\bar{x} = 27.6364$ und $s_X = 13.261$; für die Variable Y (Punkte Mathe) $\bar{y} = 29.0909$ und $s_Y = 14.8758$. Die Kovarianz als (unstandardisiertes) bivariates Streuungsmaß beträgt

$$cov\,(X,Y) = \frac{1}{11-1}\left(\begin{array}{c}(39-27.6)\,(38-29.1)\\+\,(34-27.6)\,(47-29.1)\\+\,(31-27.6)\,(44-29.1)\\+\,(48-27.6)\,(51-29.1)\\+\,(46-27.6)\,(35-29.1)\\+\,(23-27.6)\,(29-29.1)\\+\,(17-27.6)\,(22-29.1)\\+\,(12-27.6)\,(14-29.1)\\+\,(16-27.6)\,(12-29.1)\\+\,(28-27.6)\,(19-29.1)\\+\,(10-27.6)\,(9-29.1)\end{array}\right) = 168.1364$$

Daraus wiederum kann als standardisiertes bivariates Streuungsmaß der Produkt-Moment-Korrelationskoeffizient mit $r = \frac{168.1364}{13.261 \cdot 14.8758} = 0.8523$ berechnet werden, welcher für dieses konkrete Beispiel auf einen stark positiven Zusammenhang der Variable X (Punkte Statistik) mit der Variable Y (Punkte Mathe)

Abbildung 4.1: Streudiagramm der Variablen Punkte Statistik (X) und Punkte Mathe (Y)

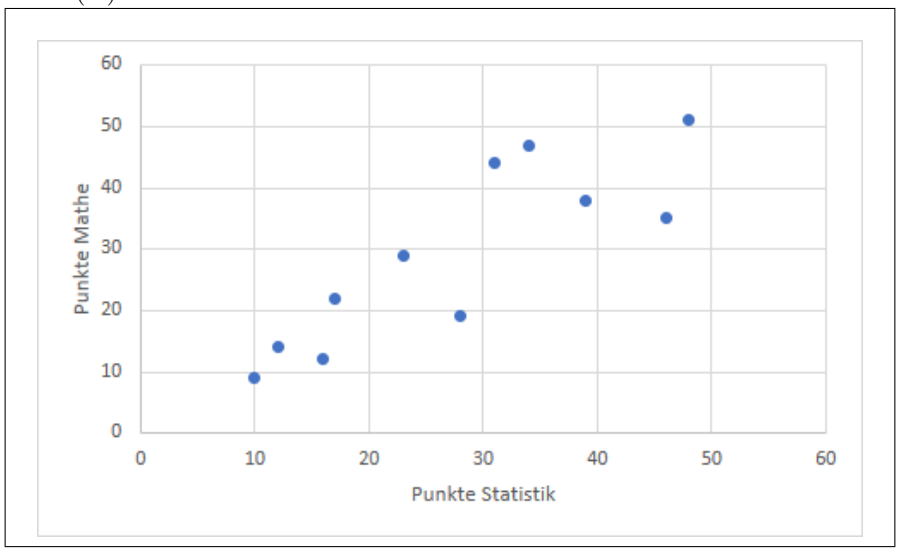
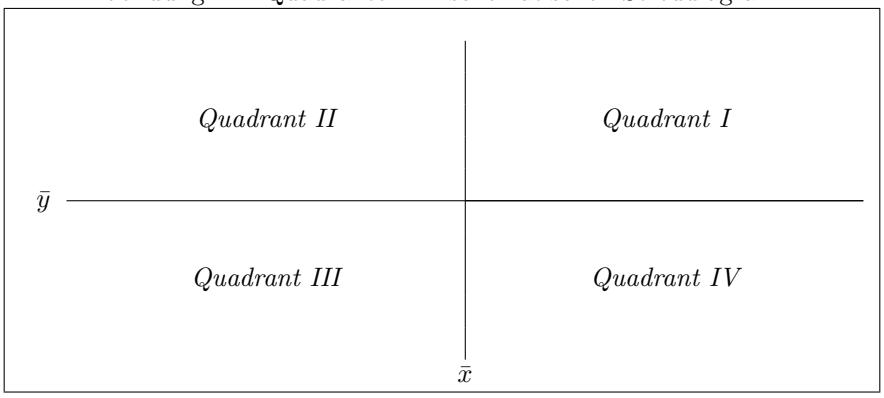


Abbildung 4.2: Quadranten im schematischen Streudiagramm



verweist. D.h. je mehr Punkte ein Student in Statistik erzielt, desto mehr Punkte erreicht er auch in Mathe - oder vice versa (Abbildung 4.1).

Mit der schematischen Darstellung eines Streudiagramms (Abbildung 4.2) kann gut nachvollziehbar gezeigt werden, wie die Richtung eines linearen Zusammenhanges i.S. des Vorzeichens der Kovarianz und des Korrelationskoeffizienten zustande kommt und zu interpretieren ist. Wie an der Formel der Kovarianz leicht zu sehen ist, besteht der Kern der Berechnung darin, für jeden Fall (repräsentiert durch einen Punkt im Streudiagramm) den Abstand zum arithmetischen Mittel für jeweils beide Variablen X und Y zu berechnen und diese beiden Abstände fallweise zu multiplizieren. Anschließend werden die so für jeden Fall erhaltenen Produkte summiert. Mit dieser Summe steht das Vorzeichen der Kovarianz bereits fest, weil sich die Division durch den immer positiven Term $n-1$ nicht auf das Vorzeichen auswirkt. Wird nun die Lage eines Punktes im Streudiagramm reduziert auf die Zuordnung zu einem der vier Quadranten, so lässt sich auf die Variable X bezogen sagen, dass jeder Punkt in den Quadranten I und IV einen positiven Abstand zum arithmetische Mittel aufweist, aber jeder Punkt in den Quadranten II und III einen negativen Abstand. Hinsichtlich der Variable Y bedeutet die Lage eines Punktes in den Quadranten I und II einen positiven Abstand zum arithmetischen Mittel, aber die Lage in den Quadranten III und IV einen negativen Abstand. Das fallweise Produkt dieser Abstände zum jeweiligen arithmetischen Mittel ist nun für alle Punkte in den Quadranten I und III positiv, und für alle Punkte in den Quadranten II und IV negativ. Insofern die Kovarianz die Summe dieser fallweisen Produkte bildet, ergibt sich das Vorzeichen daraus, in welchen Quadranten die Punkte im Streudiagramm überwiegend liegen (wobei das Gewicht eines Punktes umso größer ist, je weiter der Abstand zu den jeweiligen arithmetischen Mitteln ist). Somit wird ein eindeutiger linearer Zusammenhang durch eine Streuung beschrieben, die überwiegend in den Quadranten I und III liegt (positives Vorzeichen) oder in den Quadranten II und IV (negatives Vorzeichen). Dabei bedeutet eine positive Korrelation also: Fälle, deren X-Werte größer sind als das arithmetische Mittel $\bar{x}$, weisen üblicherweise auch Y-Werte oberhalb des arithmetischen Mittels $\bar{y}$ auf; Fälle mit X-Werten unterhalb des arithmetischen Mittels $\bar{x}$ sind dann entsprechend mit Y-Werten unterhalb des arithmetischen Mittels $\bar{y}$ assoziiert. Mit anderen Worten: je größer die X-Werte, desto größer die Y-Werte. Hingegen bedeutet eine negative Korrelation, dass Fälle mit größeren X-Werten als dem arithmetischen Mittel $\bar{x}$ eher Y-Werte aufweisen, die kleiner sind als das arithmetische Mittel $\bar{y}$; und für Fälle mit X-Werten unterhalb des arithmetischen Mittels $\bar{x}$ sind Y-Werte oberhalb des arithmetischen Mittels $\bar{y}$ charakteristisch: je größer die X-Werte, desto kleiner die Y-Werte.

Teil III

Inferentielle Statistik

Kapitel 5

Theoretische Verteilungen

5.1 Zufall und Zufallsvariablen

Im Rahmen der Statistik bedeutet das Konzept *Zufall*, dass Sachverhalte (Ereignisse) kaum vollständig determiniert sind. Vor diesem Hintergrund ist es für die Statistik bedeutsam, dass die (sozial-) wissenschaftliche Praxis i.d.R. mit Stichproben arbeitet: d.h. die auszuwertenden Daten stammen nicht von allen Fällen der jeweiligen Population, sondern nur von einer als (Zufalls-) Stichprobe bezeichneten (annähernd) repräsentativen Teilmenge der Population. Derartige Daten sind mithin als Realisierung von Zufallsexperimenten (zufällige Versuche) zu begreifen. Dabei ist ein zufälliger Versuch ein Vorgang, der - zumindest gedanklich - beliebig oft wiederholbar und dessen Ergebnis innerhalb einer Menge möglicher Ergebnisse Ω zufällig ist. Daten mit dem Umfang n können demnach als n-malige Durchführung eines zufälligen Versuches interpretiert werden, wobei das jeweilige Versuchsergebnis dem entsprechenden Messergebnis entspricht. So kann z.B. die Frage danach, wie häufig die Variable X den Wert j angenommen hat, im Lichte von Zufallsexperimenten reformuliert werden als Frage danach, wie häufig als Versuchsergebnis das Ereignis $x = j$ eingetreten ist. Eine *Zufallsvariable* X ist nun eine Funktion, die jedem elementaren Versuchsergebnis $\omega \in \Omega$ eine reelle Zahl als Funktionswert $x = X(\omega)$ zuordnet (vgl. Kolmogoroff, 1933).

Beispielhaft kann das Konzept des Zufallsexperiments mit Würfeln veranschaulicht werden. Im ersten Zufallsexperiment wird mit einem Würfel gewürfelt. D.h. ein Wurf ist ein Versuch; das Zufallsexperiment besteht insgesamt aus n Würfen. Die Funktion X sei hier die gewürfelte Augenzahl. Mithin ist die Menge möglicher Versuchsergebnisse $\Omega = \{\omega_1 = 1, \omega_2 = 2, \omega_3 = 3, \omega_4 = 4, \omega_5 = 5, \omega_6 = 6\}$. Wird bei einem Versuch z.B. eine 3 gewürfelt, dann ist $X(3) = 3$. In einem zweiten Zufallsexperiment werde mit zwei Würfeln gewürfelt und die Zufallsvariable Y sei die Summe der Augenzahlen des jeweiligen Wurfes. Wird eine 3 und eine 4 gewürfelt, dann ist $Y((3,4)) = 7$.

Um mit einem weiteren Beispiel näher an die sozialwissenschaftliche Praxis

heranzurücken, wenden wir uns ab vom Würfeln und hin zu Teilnehmern einer Befragung. Wir stellen uns das Szenario vor, dass in dieser Befragung von n Teilnehmern u.a. die Frage nach der Häufigkeit von Schlafstörungen (während der letzten zwei Wochen) gestellt wird, mit den Antwortmöglichkeiten 1=*nie*, 2=*manchmal*, 3=*immer*. Hier entspricht dem Zufallsexperiment die gesamte Befragung, einem konkreten Versuch die Befragung eines konkreten Teilnehmers, und die Anzahl der Versuche der Anzahl n der Befragungsteilnehmer. Die Menge möglicher Versuchsergebnisse entspricht den Antwortmöglichkeiten $\Omega = \{\omega_1 = 1, \omega_2 = 2, \omega_3 = 3\}$. Wenn ein Befragter auf die Frage z.B. mit 'immer' antwortet, so ist $x = X(3) = 3$; die Zufallsvariable X nimmt in diesem Fall die Realisierung $x = 3$ an.

Dass die hier für eine stichprobenmäßige Datenerhebung exemplarisch dargestellte Befragung als Zufallsexperiment aufgefasst wird, hat nun nicht die absurde Bedeutung, dass die Befragten zufällig eine Antwort auf die entsprechende Frage geben. In diesem Kontext bezieht sich der Zufall auf die Auswahl der Befragungsteilnehmer: diese werden zufällig aus der Population ausgewählt, sodass jede Person die gleiche Chance hat, Befragungsteilnehmer zu sein. Derartige Zufallsstichproben sind hinsichtlich aller Merkmale (egal ob gemessen oder nicht) theoretisch annähernd repräsentativ für die jeweilige Population.

5.2 Wahrscheinlichkeit

In der inferentiellen Statistik interessiert insbesondere, welche Werte x eine (Zufalls-) Variable X mit welcher *Wahrscheinlichkeit* annehmen kann. Die Wahrscheinlichkeit P - definiert für Teilmengen aus Ω - bestimmt die Wahrscheinlichkeit für X über Ω. Sei A eine Teilmenge von Ω, dann ist die (statistische) Wahrscheinlichkeit (v. Mises, 1919) für das Eintreten von A

$$P(A) = \lim_{n\to\infty} p(A) = \lim_{n\to\infty} \frac{f(A)}{n}$$

Bezogen auf das zuvor verwendete Beispiel der Umfrage mit der Frage nach der Häufigkeit von Schlafstörungen, könnte das Ereignis A z.B. die Antwort *immer* ($X = 3$) sein. Die Wahrscheinlichkeit, dass eine Person aus der Population auf die Frage nach der Häufigkeit von Schlafstörungen mit *immer* antwortet, entspricht der relativen Häufigkeit dieser Ausprägung - allerdings nur unter der Bedingung, dass die Stichprobe annähernd repräsentativ für die Population ist, was durch eine Zufallsstichprobe weitestgehend gewährleistet ist. Die Wahrscheinlichkeit der verschiedenen Variablenwerte lässt sich also durch die empirische Verteilung beschreiben.

Mit *theoretischen Verteilungen* können ebenso Wahrscheinlichkeiten der Ausprägungen von Zufallsvariablen bestimmt werden. Im Unterschied zu empirischen Häufigkeitsverteilungen werden die Wahrscheinlichkeiten jedoch nicht empirisch (als relative Häufigkeiten) ermittelt, sondern durch eine *Verteilungsfunktion*.

5.3 Parameterschätzung

In Abschnitt 5.1 wurde dargelegt, dass in der empirischen Sozialforschung die Daten i.d.R. auf Stichproben basieren. Für solche Stichproben lassen sich nun - wie in Kapitel 3 behandelt - Kennwerte (Mittelwerte, Streuwerte) ermitteln. Sollen diese Kennwerte lediglich die jeweilige Stichprobe beschreiben, genügt die deskriptive Statistik. In den empirischen Sozialwissenschaften begegnet uns allerdings fast immer der Anspruch, anhand von Stichprobendaten die Population zu beschreiben. Für diesen Anspruch ist es notwendig, die deskriptive Statistik mit der inferentiellen (=schließenden, i.S.v. „schlussfolgernden“) Statistik zu ergänzen. Denn bei der Übertagung statistischer Ergebnisse von einer Stichprobe auf die Population gilt es zu beachten, dass die Repräsentativität einer Stichprobe lediglich annähernd, aber kaum vollständig erreicht wird. Daher entsprechen die Kennwerte, wie sie aus den Stichprobendaten ermittelt wurden, auch nur annähernd den jeweiligen Parametern der Population. D.h. die aus den Stichprobendaten ermittelten Kennwerte können lediglich fehlerbehaftete *Schätzungen* der Parameter liefern. Diese Schätzfehler bedeuten nun eine gewisse Unsicherheit bezüglich der Übertragung von der Stichprobe auf die Population - und bei der inferentiellen Statistik geht es darum, diese Unsicherheit zu quantifizieren, indem die Wahrscheinlichkeit angegeben wird, mit welcher eine konkrete Parameterschätzung zutrifft.

Diese Quantifizierung inferentieller Unsicherheiten ist möglich, weil und insofern eine annähernd repräsentative Stichprobe als zufälliger Versuch interpretierbar ist und *Kennwerte als Zufallsvariablen*. D.h.: würden nun aus der Population mehrere Stichproben gezogen und jeweils der interessierende Kennwert ermittelt, dann würde dieser Kennwert um den Parameter streuen. Die Verteilung des Kennwertes könnte in diesem Szenario empirisch festgestellt werden und damit auch die Wahrscheinlichkeit, dass eine bestimmte Ausprägung des Kennwertes dem Parameterwert entspricht. Die Realität empirischer sozialwissenschaftlicher Studien ist jedoch die, dass lediglich *eine* Stichprobe aus der Population gezogen wird[1]. Somit lässt sich die Fehlerwahrscheinlichkeit der Parameterschätzung nicht unmittelbar empirisch bestimmen[2]. Allerdings lässt sich für viele Parameter zeigen, dass die Streuung der *Kennwerte (um ihren jeweiligen Parameter) einer bestimmten theoretischen Verteilung folgt*. Die Wahrscheinlichkeit der Richtigkeit einer solchen Parameterschätzung kann also mit der jeweiligen theoretischen Verteilungsfunktion ermittelt werden.

[1] Zwar werden zu einer sozialwissenschaftlichen Fragestellung mehrere Studien durchgeführt, in deren Rahmen oft auch jeweils eine Datenerhebung aus der gleichen Population durchgeführt wird. Allerdings können bereits kleinste Abweichungen bei der Stichprobenziehung, Datenerhebung und Populationsdefinition dazu führen, dass diese Stichproben nicht vergleichbar sind und mithin die interessierenden Kennwerte nicht als Realisierung *derselben* Zufallsvariable (Parameter) aufgefasst werden können.

[2] Es gibt jedoch durchaus indirekte datenbasierte Verfahren zur Bestimmung der Fehlerwahrscheinlichkeit, die aber i.d.R. erst mit großen Stichproben reliable Schätzungen liefern. Als ein allgemeines Verfahren mit weitreichenden Anwendungsmöglichkeiten sei hier die Bootstrap-Methode (Efron, 1994) erwähnt. Obgleich Bootstrapping sehr rechenintensiv sein kann, ist die Möglichkeit zur Anwendung in den meisten gängigen Statistik-Programmen implementiert.

Tabelle 5.1: Punktschätzungen üblicher Parameter

Parameter	**Bezeichnung**	**Stichprobenfunktion**	**Punktschätzer**
arithmetisches Mittel	μ	$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	$\hat{\mu} = \bar{x}$
Standardabweichung	σ	$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$	$\hat{\sigma} = s$
Anteil	π	$p(X = j) = \frac{f(X=j)}{n}$	$\hat{\pi} = p$

5.3.1 Punktschätzung

Die Punktschätzung von Parametern ist eine naive Schätzung, welche noch nicht die zufällige Abweichung des Kennwertes vom Parameter berücksichtigt.

Sei γ der interessierende Parameter und T eine Stichprobenfunktion, welche aus den Daten $(x_1, \ldots, x_n)$ der Stichprobe einen Wert für γ berechnet. So ist $\hat{\gamma} = T(x_1, \ldots, x_n)$ die konkrete Punktschätzung für γ.

Tabelle 5.1 zeigt die Stichprobenfunktionen der Punktschätzer für drei übliche Parameter (s. Kapitel 3).

Das Problem, welches sich aus der „Naivität“ einer Punktschätzung ergibt, ist, dass diese Schätzung nur mit geringer Wahrscheinlichkeit zutrifft, weil der jeweilige Kennwert als Schätzer für den entsprechenden Parameter die Realisierung einer Zufallsvariable ist. Üblicherweise ist also $P(\hat{\gamma} = \gamma) < 1$, für stetige Zufallsvariablen (wie z.B. dem arithmetischen Mittel) ist diese Wahrscheinlichkeit, dass der Kennwert als Punktschätzer gleich dem Parameter ist, sogar $P(\hat{\gamma} = \gamma) = 0$.

5.3.2 Intervallschätzung

Der Ausweg aus dem Dilemma der Punktschätzung - nämlich das eine zwar sehr (punkt-) genaue Schätzung mit nur geringer Wahrscheinlichkeit zutrifft - bietet die Intervallschätzung. Hierbei erfolgt die Berechnung eines Intervalls aus den konkreten Stichprobendaten mit dem Ziel, dass einerseits das Intervall möglichst klein (=möglichst genau) ist, andererseits das Intervall mit möglichst großer Wahrscheinlichkeit γ enthält. Solche Intervalle heißen *Konfidenzintervalle* (CI) oder „Vertrauensintervalle“.

Die zwei mit einem Konfidenzintervall angestrebten Ziele (Genauigkeit und große Wahrscheinlichkeit) sind nun aber widersprüchlich: ein Konfidenzintervall ist umso genauer, je kleiner es ist - aber die Wahrscheinlichkeit, dass es γ enthält, steigt, je größer es ist. Die Abwägung dieser beiden Ziele erfolgt über die Vorgabe der *Irrtumswahrscheinlichkeit* α. D.h. es wird ein Konfidenzintervall berechnet, in welchem sich der Parameter γ mit einer Wahrscheinlichkeit $1 - \alpha$ befindet. Diese Berechnung ist möglich, insofern die theoretische Verteilung des Parameters bekannt ist und die Stichprobe als Zufallsexperiment interpretiert werden

Tabelle 5.2: Theoretische Verteilung üblicher Parameter

Parameter	**theoretische Verteilung**	**Approximation**
μ	t-Verteilung	ab $n > 30$: Normalverteilung
σ^2	χ^2-Verteilung	*Standardabweichung* ab $n > 100$: Normalverteilung
π	Binomialverteilung	ab $n > 30$: Normalverteilung

kann. Für drei übliche Parameter nennt Tabelle 5.2 die jeweilige theoretische Verteilung. Darüber hinaus lassen sich viele theoretische Verteilungen ab einem hinreichend großen Stichprobenumfang mit der Normalverteilung approximieren. D.h.: je größer die Stichprobe, desto stärker nähern sich die entsprechenden theoretischen Verteilungen der Normalverteilung an.

Die Absicht bei der Intervallschätzung besteht also darin, ein Intervall $CI_\alpha = \hat{\gamma} \pm c$ zu schätzen. Dabei ist α die Irrtumswahrscheinlichkeit, $\hat{\gamma}$ der Punktschätzer für den Parameter γ, und c ist ein Term, welcher - basierend auf der theoretischen Verteilung von $\hat{\gamma}$ - unter gegebenem α die maximale Abweichung von der Punktschätzung liefert. Somit ist $1 - \alpha = P\left(\hat{\gamma} - c \leq \gamma \leq \hat{\gamma} + c\right)$. Die konkreten Schritte zur Schätzung eines Konfidenzintervalls sind nun folgende:

1. $\hat{\gamma}$ aus der Stichprobe ermitteln
2. α festlegen (übliche Niveaus der Irrtumswahrscheinlichkeit sind: 0.1, 0.05, 0.01)
3. theoretische Verteilung des Schätzers $\hat{\gamma}$ bestimmen
4. c und CI berechnen

Die Interpretation einer solchen Intervallschätzung lautet: Das CI_α überdeckt mit $(1 - \alpha)\,100\%$-iger Sicherheit den unbekannten Wert des Parameters γ in der Population.

Konfidenzintervalle für große Zufallsstichproben - in denen die theoretische Verteilung der Punktschätzer einiger Parameter durch die Normalverteilung approximiert werden kann - werden wie in Tabelle 5.3 notiert berechnet.

Im folgenden soll Berechnung und Interpretation von Konfidenzintervallen mit drei Beispielen dargestellt werden. Für das erste Beispiel werden die Daten einer Bevölkerungsumfrage in Deutschland (ALLBUS 2016, GESIS-Leibniz-Institut für Sozialwissenschaften, 2017) verwendet. In diesem Daten-Set wurden $n = 796$ Personen mit einem Alter bis zu 40 Jahren erfasst, die bereit waren, Angaben zum monatlichen Nettoeinkommen zu machen. 200 dieser Personen gaben an, ein Einkommen von mindestens €2000 zu haben. Es soll nun ein Intervall für den Anteil von Personen (in der Population) mit einem Einkommen von

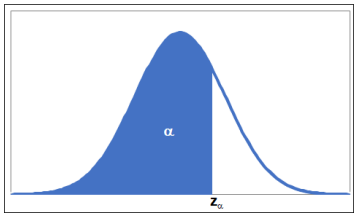
Tabelle 5.3: Berechnung approximierter Konfidenzintervalle $CI_\alpha = \hat{\gamma} \pm c$

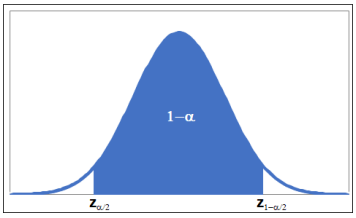
arithmetisches Mittel	**Standardabweichung**	**Anteilswert**
$\gamma := \mu$	$\gamma := \sigma$	$\gamma := \pi$
$c = z_{\alpha/2}\frac{s}{\sqrt{n}}$	$c = z_{\alpha/2}\frac{s}{\sqrt{2n}}$	$c = z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
wenn $n > 30$	wenn $n > 100$	wenn $n > 30$

Dabei ist z_α das α-Quantil der Standardnormalverteilung (z-Verteilung):
$P(u \leq z_\alpha) = \alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_\alpha} e^{-\frac{\nu^2}{2}} d\nu$.

Tafel Verteilungsfunktion der Standardnormalverteilung (z-Verteilung):

α	z_α	α	z_α	α	z_α	α	z_α
0.000	$-\infty$	0.250	-0.674	0.500	0.000	0.750	0.674
0.025	-1.960	0.275	-0.598	0.525	0.063	0.775	0.755
0.050	-1.645	0.300	-0.524	0.550	0.126	0.800	0.842
0.075	-1.440	0.325	-0.454	0.575	0.189	0.825	0.935
0.100	-1.282	0.350	-0.385	0.600	0.253	0.850	1.036
0.125	-1.150	0.375	-0.319	0.625	0.319	0.875	1.150
0.150	-1.036	0.400	-0.253	0.650	0.385	0.900	1.282
0.175	-0.935	0.425	-0.189	0.675	0.454	0.925	1.440
0.200	-0.842	0.450	-0.126	0.700	0.524	0.950	1.645
0.225	-0.755	0.475	-0.063	0.725	0.598	0.975	1.960
						1.000	∞





mindestens €2000 ermittelt werden. Gemäß dem ersten Schritt zur Berechnung eines Konfidenzintervalls wird zunächst der Kennwert als Punktschätzer für den Parameter aus der Stichprobe ermittelt: $\hat{\pi} = \frac{200}{796} = 0.2513$. Im zweiten Schritt wird die Irrtumswahrscheinlichkeit auf das übliche Niveau von fünf Prozent festgelegt: $\alpha = 0.05$. Drittens wird bestimmt, welcher theoretischen Verteilung der Punktschätzer folgt. Weil die Fallzahl mit $n = 796$ deutlich größer ist als 30, kann die theoretische Verteilung des Anteils - welcher einer Binomialverteilung folgt - durch die Normalverteilung approximiert werden: $\hat{\pi} \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$. Der letzte Schritt ist die Berechnung des Konfidenzintervalls. Zunächst werden aus der Tafel die z-Quantile der Standardnormalverteilung abgelesen, zwischen denen die kumulierte relative Häufigkeit bzw. die Wahrscheinlichkeit 95% beträgt: $1-\alpha = P\left(z_{\alpha/2} \leq u \leq z_{1-\alpha/2}\right)$. Weil die Standardnormalverteilung symmetrisch um Null verteilt ist, müssen nicht beide Quantile aus der Tafel abgelesen werden; es genügt die Bestimmung des unteren Quantils, welches ein negatives Vorzeichen aufweist - das obere Quantil hat den gleichen Betrag, jedoch ein positives Vorzeichen: $z_{\alpha/2} = -z_{1-\alpha/2}$ bzw. $\left|z_{\alpha/2}\right| = z_{1-\alpha/2}$. Mithin ist $z_{\alpha/2=0.025} = -1.96$. Entsprechend Tabelle 5.3 kann schließlich das Konfidenzintervall berechnet werden: $CI_{0.05} = \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.2513 \pm 1.96\sqrt{\frac{0.2513(1-0.2513)}{796}} = [0.2212\,,\, 0.2814]$. Mit 95%-iger Wahrscheinlichkeit liegt in der Population der Anteil von Personen mit einem Einkommen von mindestens €2000 zwischen 22.12% und 28.14%: $P\left(0.2212 \leq \pi \leq 0.2814\right) = 0.95$.

Die Datenbasis für die nächsten beiden Beispiele ist die PISA-Schülerbefragung (2015). Für diese Datenerhebung wurde unter anderem abgefragt, wie groß der Lernaufwand (in Minuten pro Woche) der Schüler für das Fach Mathematik ist. Im Durchschnitt der Stichprobe (Schüler in Deutschland, $n = 5224$) beträgt der Lernaufwand 192.68 Minuten pro Woche. Es soll ein Intervall für den durchschnittlichen Lernaufwand der Population (Schüler in Deutschland) ermittelt werden. Wie Tabelle 5.3 zu entnehmen ist, erfordert die Berechnung des Konfidenzintervalls die Information über die Standardabweichung, welche in den Stichprobendaten $s = 64.47$ sei. Ohne die einzelnen Schritte erneut ausführlich zu beschreiben, kann die Aufgabe wie folgt gelöst werden:

$$\hat{\mu} = 192.68 \qquad s = 64.47 \qquad \alpha = 0.05$$

$$\text{da } n > 30\text{: } \hat{\mu} \sim N\left(\mu, \tfrac{\sigma}{\sqrt{n}}\right) \qquad z_{\alpha/2=0.025} = -1.96$$

$$CI_{0.05} = \hat{\mu} \pm z_{\alpha/2}\tfrac{s}{\sqrt{n}} = 192.68 \pm 1.96\tfrac{64.47}{\sqrt{5224}} = [190.93\,,\, 194.43]$$

$$P\left(190.93 \leq \mu \leq 194.43\right) = 0.95$$

Im letzten Beispiel schließlich wird ein Intervall für die Standardabweichung des Lernaufwands in der Population ermittelt:

$\hat{\sigma} = s = 64.47$ $\alpha = 0.05$

da $n > 100$: $\hat{\sigma} \sim N\left(\sigma, \frac{\sigma}{\sqrt{2n}}\right)$ $z_{\alpha/2=0.025} = -1.96$

$CI_{0.05} = \hat{\sigma} \pm z_{\alpha/2}\frac{s}{\sqrt{2n}} = 64.47 \pm 1.96\frac{64.47}{\sqrt{2\times 5224}} = [63.23\,,\, 65.71]$

$P\left(63.23 \leq \sigma \leq 65.71\right) = 0.95$

Kapitel 6

Statistische Tests

6.1 Statistische Hypothesen

Im Fokus statistischer Tests steht - wie bei der Parameter-Intervallschätzung - das Verhältnis von Stichprobe und Population bzw. die Stichprobenverteilung von Kennwerten als Schätzer für die jeweiligen Parameter. Jedoch ändert sich mit statistischen Tests die Fragestellung etwas: während wir bei der Parameter-Intervallschätzung nach einem Wertebereich fragen, in welchem der Wert des Parameters mit einer bestimmten Wahrscheinlichkeit liegt, haben wir bei statistischen Tests eine Hypothese hinsichtlich des Parameters. Dabei fragen wir nun nicht mehr unmittelbar nach dem Parameter bzw. dessen Wertebereich, sondern wir formulieren eine hypothetische Aussage über den Wert des Parameters und überprüfen, inwiefern diese Hypothese zutrifft.

Die zu testende Hypothese, welche die Annahme über eine Verteilung bzw. einen Parameter in der Population formuliert, ist die *Nullhypothese* (H_0). Sie behauptet in ihrer mathematischen Formulierung immer Homogenität. Das bedeutet, dass die Nullhypothese immer eine positive bzw. konkrete Annahme über eine Verteilung oder einen Parameter macht. Ihr steht als negatives Pendant die *Alternativhypothese* (H_A) gegenüber. Wenn also mit der Nullhypothese die Annahme aufgestellt wird, dass „es“ in der Population „so“ ist, dann behauptet die Alternativhypothese, dass „es“ in der Population „nicht so“ ist. Mit einem statistischen Test kann nun ermittelt werden, ob die empirische Realität (=Stichprobendaten) die theoretische Annahme entweder der Nullhypothese oder der Alternativhypothese stützt. Statistisch getestet wird dabei immer die Nullhypothese. Ziel eines statistischen Tests ist die Entscheidung über Annahme oder Ablehnung der Nullhypothese.

Als Thema der inferentiellen Statistik birgt die Entscheidung über Annahme oder Ablehnung der Nullhypothese das Risiko, nicht die Realität der Population zu treffen. Mithin sind bei statistischen Tests zwei Arten von Fehlern denkbar. So kann es passieren, dass ein Test auf die Ablehnung der jeweiligen Nullhypothese verweist - obwohl diese in Wirklichkeit zutrifft. Die Wahrscheinlichkeit

für diesen sogenannten *Alphafehler* ist die bereits im Zusammenhang mit Konfidenzintervallen (Abschnitt 5.3.2) besprochene Irrtumswahrscheinlichkeit, die, insofern sie vorgegeben wird, mit α bezeichnet ist. Andererseits ist es möglich, dass der Test die Nullhypothese bestätigt - aber tatsächlich die Alternativhypothese zutrifft. Dieser zweite Fehler heißt *Betafehler*.

6.2 Prüfgröße

Im Kern bedeutet ein statistischer Test das Berechnen einer sogenannten *Prüfgröße* und eine Beurteilung dieser Prüfgröße auf der Grundlage ihrer *theoretischen Verteilung*. Die Prüfgröße ist ein Maß dafür, wie stark sich ein Parameter oder eine Verteilung von der in H_0 formulierten Annahme unterscheidet.

Sei $\hat{\theta}$ eine Prüfgröße, dann ist $\hat{\theta} = T(X_1, \ldots, X_n)$ eine Stichprobenfunktion und $E\left(\hat{\theta}\right)$ der Erwartungswert der Prüfgröße. Der Erwartungswert einer Prüfgröße ist der Wert, den eine Prüfgröße annimmt, wenn H_0 zutrifft. Für einen statistischen Test muss die theoretische Verteilung der Prüfgröße - auch „Prüfverteilung“ genannt - unter H_0 bekannt sein. Damit gibt die Prüfverteilung eine bedingte kumulierte Wahrscheinlichkeit an: nämlich die kumulierte Wahrscheinlichkeit bestimmter Werte der Prüfgröße unter der Bedingung, dass H_0 zutrifft. Die Prüfverteilung unter H_0 sei mit P_0 bezeichnet.

Mit den kumulierten Wahrscheinlichkeiten aus einer Prüfverteilung kann also angegeben werden, mit welcher Wahrscheinlichkeit (immer unter Bedingung H_0) die jeweilige Prüfgröße in einem bestimmten Wertebereich liegt. Sei die Wahrscheinlichkeit dieses Wertebereichs $1-\alpha$, dann ist α also die Irrtumswahrscheinlichkeit des Alphafehlers. Im Rahmen statistischer Tests („Signifikanztests“) wird α auch als „Signifikanzniveau“ bezeichnet. Insofern für einen statistischen Test das Signifikanzniveau als höchste akzeptable Irrtumswahrscheinlichkeit vorgegeben wird, gelten hier die üblichen Niveau-Schwellen von $\alpha = 0.05$ oder $\alpha = 0.01$. Aus der Prüfverteilung einer Prüfgröße lässt sich ein „kritischer Bereich“ K für die Werte der Prüfgröße bestimmen, sodass $P_0\left(\hat{\theta} \in K\right) \leq \alpha$. Dabei wird K durch „kritische Werte“ begrenzt, welche die entsprechenden Quantile der Prüfverteilung sind. Die Entscheidung, welche ein statistischer Test ermöglicht, kommt letztlich so zustande: Wenn $\hat{\theta} \in K$, dann wird H_0 verworfen und H_A akzeptiert, ansonsten umgekehrt (H_0 akzeptieren und H_A verwerfen, wenn $\hat{\theta} \notin K$). H_0 zu verwerfen heißt, die (empirische) Abweichung von der in H_0 (hypothetisch) formulierten Situation - d.h. die Differenz der Prüfgröße zu ihrem Erwartungswert - ist so groß, dass sie kaum zufällig sein kann; die Abweichung ist signifikant.

Folgt die Verteilung der Prüfgröße einer symmetrischen theoretischen Verteilung, wie z.B. der Normalverteilung, gibt es je nach genauer Formulierung von H_A drei Varianten der Lokalisation von K (s. Abbildung 6.1).

Für die Durchführung eines statistischen Tests „per Hand“ - also ohne Computerunterstützung - wird i.d.R. die berechnete Prüfgröße mit den kritischen Werten (zweiseitig) oder dem kritischen Wert (einseitig) verglichen. Die Frage

Abbildung 6.1: Varianten der K-Lokalisation in symmetrischer Prüfverteilung

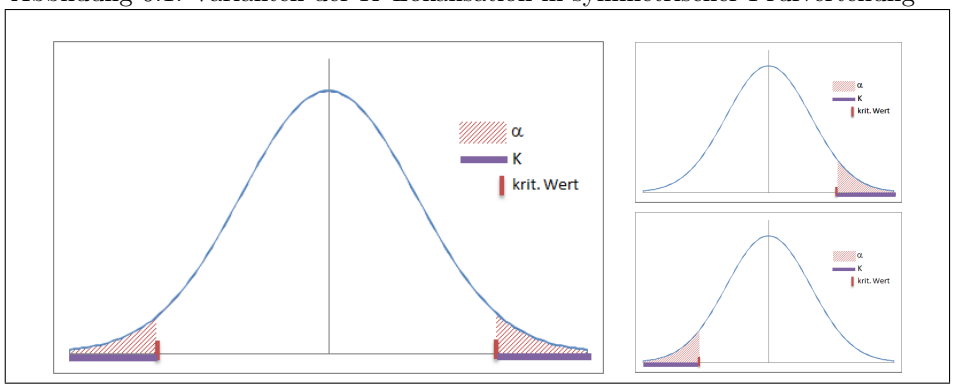
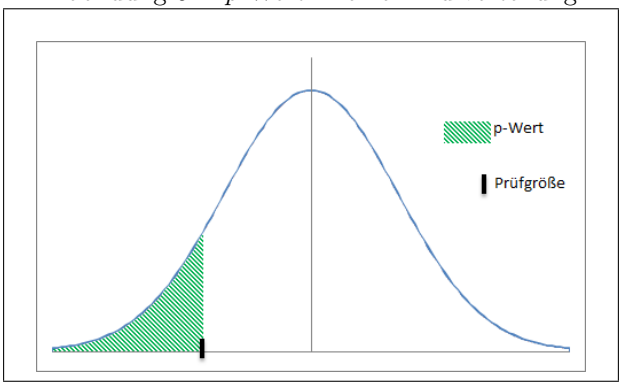


Abbildung 6.2: p-Wert in einer Prüfverteilung



ist dann also, ob die Prüfgröße im kritischen Bereich liegt. Falls die Prüfgröße im kritischen Bereich liegt und damit mindestens so weit von ihrem Erwartungswert entfernt ist wie ein kritischer Wert, dann wird die Nullhypothese verworfen. Dabei lassen sich die kritischen Werte mit gegebenem α unmittelbar aus der Tafel der jeweiligen Prüfverteilung ablesen. Mit Computerunterstützung erfolgt die Durchführung statistischer Tests etwas anders: für die berechnete Prüfgröße wird die Wahrscheinlichkeit ermittelt, dass die Prüfgröße unter H_0 so oder noch weiter vom Erwartungswert entfernt liegt. Diese Wahrscheinlichkeit wird von der Statistik-Software mit dem p-Wert ausgegeben (s. Abbildung 6.2). Die Interpretation der Statistik bzw. die Testentscheidung erfolgt dann durch den Vergleich des p-Werts mit (vorgegebenem) α: ist $p \leq \alpha$, dann wird H_0 abgelehnt; ist hingegen $p > \alpha$, dann wird H_0 beibehalten.

6.3 Anpassungstests

Die Fragestellung für Anpassungstests lässt sich als die Frage formulieren, ob eine Stichprobe hinsichtlich eines bestimmten Kennwertes oder der Verteilung repräsentativ für eine Population mit bekanntem Parameter bzw. bekannter Verteilung ist. Insofern in der Praxis die Parameter einer Population kaum bekannt sind, wird es sich dabei um eine theoretisch angenommene Population handeln. Weil die Stichprobe allerdings aus einer tatsächlichen Population stammt, kann die Fragestellung wie folgt übersetzt werden: entspricht die tatsächliche Population, aus welcher die Stichprobe stammt, bezüglich des zu untersuchenden Parameters oder der Verteilung der theoretisch angenommenen Population? Im folgenden wird eine kleine Auswahl von Anpassungstests exemplarisch dargestellt: χ^2-Test für die (univariate) Häufigkeitsverteilung einer kategorialen Variable, z-Test für den Anteil, t-Test für das arithmetische Mittel (einer metrischen Variable) und t-Test für den Produkt-Moment-Korrelationskoeffizienten (zweier metrischer Variablen).

6.3.1 Häufigkeitsverteilung: χ^2-Anpassungstest

Der χ^2-*Anpassungstest* wird angewendet, wenn überprüft werden soll, ob sich die empirische Verteilung der Variable X mit k Kategorien von einer theoretisch vorgegebenen Verteilung unterscheidet. Die theoretische Verteilung wird dabei als relative Erwartungshäufigkeiten $\tilde{p}$ der jeweiligen Kategorien angegeben. Dieser statistische Test erfordert, dass folgende Bedingungen erfüllt sind: $min\left(n\tilde{p}\left(X=1\right),\ldots,n\tilde{p}\left(X=k\right)\right) \geq 5$ und $n \geq 40$; d.h. die absoluten Erwartungshäufigkeiten müssen jeweils mindestens 5 betragen und die Stichprobengröße darf nicht weniger als 40 Messwerte umfassen. Die Prüfgröße heißt „Chi-quadrat“ und wird berechnet mit $\chi^2 = \sum_{j=1}^{k} \frac{(f(X=j)-n\tilde{p}(X=j))^2}{n\tilde{p}(X=j)}$. Der gleichnamigen Prüfverteilung müssen als Parameter die Freiheitsgrade $df = k-1$ übergeben werden: $\chi^2 \sim \chi^2_{df=k-1}$. Hierbei behauptet die Nullhypothese, dass es keine Abweichung der empirischen von den theoretisch erwarteten Häufigkeit gibt, also $H_0 : \tilde{p}\left(X=1\right) = p\left(X=1\right),\ldots,\tilde{p}\left(X=k\right) = p\left(X=k\right)$. Die Alternativhypothese $H_A : \tilde{p}\left(X=1\right) \neq p\left(X=1\right)$ *oder* $\tilde{p}\left(X=2\right) \neq p\left(X=2\right)$ *oder* ... hingegen behauptet entsprechende Abweichungen. Weil die Prüfverteilung nicht symmetrisch ist, wird der Test einseitig durchgeführt, das Kriterium für die Ablehnung von H_0 ist damit $\chi^2 \geq \chi^2_{df,1-\alpha}$.

Als Basis für die beispielhafte Berechnung werden die Daten des World Values Survey, Welle 6 (2013), für Deutschland, verwendet. Mit dieser Befragung wurde unter anderem erhoben, was die Befragten aus einer Liste mit vier politischen Zielen des Landes für das wichtigste halten. Die Häufigkeit der Antworten zeigt Tabelle 6.1.

Für dieses Beispiel sollen die Erwartungshäufigkeiten der Gleichverteilung folgen. D.h. wir überprüfen, ob die zur Auswahl stehenden Ziele in der Population mit gleicher Häufigkeit genannt werden würden. Weil die Skala vier Kategorien abbildet, ist die relative Erwartungshäufigkeit für jede Kategorie

Tabelle 6.1: Häufigkeitsverteilung *wichtigstes Ziel des Landes*

wichtigstes Ziel	**f**	**%**
hohes Wirtschaftswachstum	993	49.18
starke Landesverteidigung	75	3.71
mehr individuelle Mitbestimmung	807	39.98
schönere Städte und Landschaften	144	7.13
n	2019	

Datenquelle: World Values Survey, Deutschland Welle 6 (Inglehart et al., 2014).

$\tilde{p} = \frac{1}{k} = 0.25$, als absolute Erwartungshäufigkeit ergibt sich mithin $n\tilde{p} = 2019 \times 0.25 = 504.75$ für jede Kategorie. Für die Prüfgröße wird

$$\chi^2 = \frac{(993 - 504.75)^2}{504.75} + \frac{(75 - 504.75)^2}{504.75} + \frac{(807 - 504.75)^2}{504.75} + \frac{(144 - 504.75)^2}{504.75}$$

$$= 1277.01$$

berechnet; sie folgt der χ^2-Prüfverteilung mit hier konkret $df = k - 1 = 3$ Freiheitsgraden. Der entsprechende kritische Wert bei $\alpha = 0.05$ ist $\chi^2_{df=3} = 7.81$, der kritische Bereich mithin $K = [7.81\,, \infty)$. Da $\chi^2 > \chi^2_{3,0.95}$ - oder alternativ ausgedrückt: $\chi^2 \in K$ - wird H_0 verworfen: die Verteilung der Ziele weicht also signifikant von einer Gleichverteilung ab.

6.3.2 Anteilswert: z-Anpassungstest

Wenn getestet werden soll, ob sich ein Anteilswert p von einem theoretisch vorgegebenen Anteilswert p_0 unterscheidet, dann kann der *z-Anpassungstest* zum Einsatz kommen. Die Nullhypothese lautet, dass der Anteilswert nicht vom theoretisch vorgegebenen Anteilswert abweicht: $H_0 : p = p_0$. Für die Formulierung der Alternativhypothese sind drei Varianten denkbar. Bleibt die Richtung der hypothetischen Abweichung unspezifisch - d.h.: wenn es irrelevant ist, ob der Anteilswert größer oder kleiner als der theoretisch vorgegebene Anteilswert ist, dann lautet die Alternativhypothese $H_A : p \neq p_0$. Allerdings ist es bei diesem statistischen Test auch möglich, die Richtung der Abweichung zu spezifizieren mit der Annahme, dass der Anteilswert größer oder kleiner als der theoretisch vorgegebene Anteilswert ist. Die entsprechenden Alternativhypothesen lauten dann $H_A : p > p_0$ oder $H_A : p < p_0$ respektive. Welche Variante der Alternativhypothese Verwendung findet, muss letztlich auf inhaltliche Erwägungen und die jeweils konkrete Fragestellung zurückgeführt werden.

Die Prüfgröße $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ ist normal-verteilt mit Erwartungswert $E(z) = 0$ und Varianz $var(z) = 1$. Somit ist die Prüfverteilung der Prüfgröße also die Standardnormalverteilung: $z \sim N(0, 1)$. Die Bedingungen für diesen Test sind,

Tabelle 6.2: Alternativhypothesen und Lokalisation der kritischen Werte für z-Anpassungstest

H_A	Kriterium für Ablehnung von H_0	kritischer Bereich
$p \neq p_0$	$\lvert z\rvert \geq z_{1-\alpha/2}$	$K = \left(-\infty\,,\, z_{\alpha/2}\right] \cup \left[z_{1-\alpha/2}\,,\, \infty\right)$
$p > p_0$	$z \geq z_{1-\alpha}$	$K = \left[z_{1-\alpha}\,,\, \infty\right)$
$p < p_0$	$z \leq -z_{1-\alpha}$	$K = \left(-\infty\,,\, z_{\alpha}\right]$

dass $n \geq 60$ und $np_0 \geq 5$. Die Lage des kritischen Bereichs bzw. das Kriterium für die Ablehnung der Nullhypothese muss in Abhängigkeit von der konkreten Alternativhypothese bestimmt werden.

Erneut verwenden wir für ein Beispiel die Daten des World Values Survey, und zwar eine Variable, welche die Selbsteinschätzung der Handlungsautonomie abfragt. Es soll auch hier wieder untersucht werden, ob die Variable über die Ausprägungen 1=*ja, sehe mich als autonom handelndes Individuum* und 2=*nein, sehe mich nicht als autonom handelndes Individuum* gleich-verteilt ist. Bei einer dichotom skalierten Variable ergibt sich unter der Gleichverteilungsannahme $p_0 = 0.5$ für beide Ausprägungen. Insofern die Behauptung der Alternativhypothese - nämlich dass keine Gleichverteilung vorliegt - unspezifisch ist, wird zweiseitig getestet: $H_A : p_1 \neq p_0$. In diesem Szenario, mit einer dichotom skalierten Variable und einer unspezifischen Alternativhypothese, ist es bedeutungslos, für welche der beiden Ausprägungen der entsprechende Anteil auf eine Anpassung an p_0 getestet wird. Für dieses Beispiel sei p_1 der Anteil der Ausprägung 1 (Personen, die sich als autonom handelndes Individuum sehen). Aus den Daten des World Values Survey lässt sich ermitteln, dass von $n = 1982$ Befragten (mit gültigem Antwortwert für diese Variable) sich $n_1 = 1639$ Personen als autonom handelnde Individuen sehen. Somit beträgt der empirische Anteil für diese Antwortkategorie $\hat{p}_1 = \frac{1639}{1982} = 0.8269$. Als Prüfgröße wird $z = \frac{0.8269-0.5}{\sqrt{\frac{0.5\times(1-0.5)}{1982}}} = 29.107$ ermittelt. Für eine Irrtumswahrscheinlichkeit $\alpha = 0.05$ erhalten wir aus der Standardnormalverteilung als Prüfverteilung bei einem zweiseitigen Test zwei kritische Werte: $z_{\alpha/2} = -1.96$ und $z_{1-\alpha/2} = 1.96$. Weil $\lvert z\rvert > z_{1-\alpha/2}$ bzw. $z \in K$, wird H_0 verworfen. Die empirische Verteilung weicht also signifikant von einer Gleichverteilung ab.

6.3.3 Arithmetisches Mittel: t-Anpassungstest

Den Unterschied zwischen einem arithmetischen Mittel μ einer Population und einem theoretisch vorgegebenen arithmetischen Mittel μ_0 untersuchen wir mit dem *t-Anpassungstest*. Die Nullhypothese behauptet mit $H_0 : \mu = \mu_0$, dass die beiden arithmetischen Mittel gleich sind. Wie beim z-Anpassungstest, kann hier die Alternativhypothese unspezifisch oder spezifisch formuliert und entsprechend zwei- oder einseitig getestet werden. Die Prüfgröße ist $t = \frac{\hat{\mu}-\mu_0}{\hat{\sigma}}\sqrt{n}$, wobei $\hat{\sigma}$ die geschätzte Varianz der Population ist. Die Prüfverteilung ist die t-Verteilung mit $df = n - 1$ Freiheitsgraden, die ab $n > 30$ durch die Standard-

Tabelle 6.3: Alternativhypothesen und Lokalisation der kritischen Werte für t-Anpassungstest

H_A	Kriterium für Ablehnung von H_0	kritischer Bereich
$\mu \neq \mu_0$	$\lvert t\rvert \geq t_{df,1-\alpha/2}$	$K = \left(-\infty\,,\, t_{df,\alpha/2}\right] \cup \left[t_{df,1-\alpha/2}\,,\, \infty\right)$
$\mu > \mu_0$	$t \geq t_{df,1-\alpha}$	$K = \left[t_{df,1-\alpha}\,,\, \infty\right)$
$\mu < \mu_0$	$t \leq -t_{df,1-\alpha}$	$K = \left(-\infty\,,\, t_{df,\alpha}\right]$

normalverteilung approximiert werden kann.

Ein für diesen Test geeignetes Item im World Values Survey ist die Frage danach, wie fair die Befragten die meisten Menschen einschätzen: „Glauben Sie, die meisten Menschen würden Sie ausnutzen, sobald sich ihnen eine Möglichkeit bietet? Oder glauben Sie, die meisten Menschen würden sich fair und korrekt verhalten? [...]“. Die metrische Antwortskala reicht von 1=*Die Menschen nutzen einen aus* bis 10=*Die Menschen verhalten sich fair*. Eine beispielhafte und für den t-Anpassungstest adäquate Frage könnte lauten, ob die durchschnittliche Einschätzung der Fairness signifikant von $\mu_0 = 5.5$ abweicht. Dieser nicht weiter spezifizierten Frage entspricht die Alternativhypothese $H_A : \mu \neq \mu_0$. Der empirische Durchschnitt in der Stichprobe ($n = 2037$) liegt bei $\hat{\mu} = 5.66$, die Standardabweichung beträgt $\hat{\sigma} = 2.064$. Für die Prüfgröße ergibt sich $t = \frac{5.66-5.5}{2.064}\sqrt{2037} = 3.4987$. Bei dieser großen Fallzahl kann die Prüfverteilung mit der Standardnormalverteilung approximiert werden, die entsprechenden kritischen Werte (bei $\alpha = 0.05$) sind somit $t_{df,\alpha/2} \approx z_{\alpha/2} = -1.96$ und $t_{df,1-\alpha/2} \approx z_{1-\alpha/2} = 1.96$. Da $\lvert t\rvert > z_{1-\alpha/2}$ und mithin $t \in K$, kann H_0 verworfen werden: der empirische Durchschnitt der Fairness-Einschätzung unterscheidet sich signifikant von $\mu_0 = 5.5$.

6.3.4 t-Test für Produkt-Moment-Korrelationskoeffizienten

Um zu überprüfen, ob ein (Produkt-Moment-) Korrelationskoeffizient von einem theoretisch vorgegebenen Korrelationskoeffizienten abweicht, kann ebenfalls ein t-Test zur Anwendung kommen. Dabei ist es üblich, den theoretisch vorgegebenen Korrelationskoeffizienten mit $\varrho_0 = 0$ anzunehmen. Denn die Nullhypothese $H_0 : \varrho = \varrho_0 = 0$ kann dann auch inhaltlich substantiell interpretiert werden als „kein Zusammenhang zwischen den beiden Variablen (X und Y)“. Die vorzeichenmäßige Gerichtetheit des Korrelationskoeffizienten ermöglicht dementsprechend (neben einer ungerichteten, zweiseitig zu testenden) auch eine gerichtete (einseitig zu testende) Formulierung der Alternativhypothese. Als Prüfgröße wird $t = \left(\left(0.5ln\left(\frac{1+r}{1-r}\right)\right) - \left(0.5ln\left(\frac{1+\varrho_0}{1-\varrho_0}\right)\right)\right)\sqrt{n-3}$ berechnet, wobei $r = \hat{\varrho}$ und die Prüfverteilung die t-Verteilung mit $df = n - 2$ Freiheitsgraden ist.

In Abschnitt 4.2 wurde für $n = 11$ Fälle die Korrelation der Variablen *Punkte Statistik* und *Punkte Mathe* mit $r = 0.8523$ ermittelt. Die Frage danach, ob es einen positiven Zusammenhang zwischen den beiden Variablen gibt, lässt sich als Alternativhypothese $H_A : \varrho > 0$ formulieren. Die Prüf-

Tabelle 6.4: Alternativhypothesen und Lokalisation der kritischen Werte für t-Test eines Produkt-Moment-Korrelationskoeffizienten

H_A	Kriterium für Ablehnung von H_0	kritischer Bereich
$\varrho \neq \varrho_0$	$\lvert t\rvert \geq t_{df,1-\alpha/2}$	$K = (-\infty\,,\, t_{df,\alpha/2}] \cup [t_{df,1-\alpha/2}\,,\, \infty)$
$\varrho > \varrho_0$	$t \geq t_{df,1-\alpha}$	$K = [t_{df,1-\alpha}\,,\, \infty)$
$\varrho < \varrho_0$	$t \leq -t_{df,1-\alpha}$	$K = (-\infty\,,\, t_{df,\alpha}]$

größe $t = \left(0.5ln\left(\frac{1+0.8523}{1-0.8523}\right)\right)\sqrt{11-3} = 3.5765$ wird hier einseitig und mit $df = n-2 = 9$ mit dem kritischen Wert (bei $\alpha = 0.05$) $t_{df=9,1-\alpha} = 2.2622$ verglichen. Das Resultat dieses Vergleichs $t > t_{df,1-\alpha}$ bedeutet $t \in K$ und lässt sich in die Aussage übersetzten, dass es einen statistisch signifikanten positiven Zusammenhang zwischen den Variablen *Punkte Statistik* und *Punkte Mathe* gibt.

6.4 Unterschiedstests

Bei Unterschiedstests fragen wir, ob mehrere Sub-Stichproben bezüglich eines Kennwerts oder der Verteilung aus sich gleichenden (Sub-) Populationen stammen. Die Annahme ist, dass sich die Sub-Stichproben unterscheiden, wenn sich die jeweiligen Populationen unterscheiden. Die Sub-Stichproben gehören dabei zu *einer* umfassenden Stichprobe und sind durch die Ausprägungen einer kategorialen Variable definiert. Statistische Unterschiedstests können immer auch als Tests auf statistischen Zusammenhang zweier Variablen verstanden werden. Sollte nämlich ein Test das Ergebnis liefern, dass die Sub-Stichproben aus sich unterscheidenden Populationen stammen, dann ist die Ausprägung des Kennwertes oder die Verteilung eben *abhängig* von der Population. Es wird also der Zusammenhang zwischen der Variable, welche die Sub-Stichproben definiert, und der Variable, deren Kennwert oder Verteilung Gegenstand des Tests ist, untersucht. Kein Unterschied zwischen den Sub-Stichproben deutet darauf hin, dass es keinen Zusammenhang der Variablen gibt. Bei jedem Unterschiedstest behauptet die Nullhypothese, dass es keine Unterschiede bzw. keinen Zusammenhang gibt.

Exemplarisch werden hier zwei übliche Unterschiedstests vorgestellt: χ^2-Test für Verteilung (kategoriale Variable) und für das arithmetische Mittel einen F-Test.

6.4.1 Häufigkeitsverteilung: χ^2-Unabhängigkeitstest

Soll untersucht werden, ob es einen stochastischen Zusammenhang zwischen zwei kategorialen Variablen gibt, dann kann der χ^2-Unabhängigkeitstest zur Anwendung kommen. Dabei sei k die Kategorienzahl der Variable X, und m die Anzahl der Kategorien von Variable Y. Die Aufgabenstellung besteht hier darin, zu testen, ob sich die durch X bedingten Verteilungen der Variable Y signifikant unterscheiden - d.h. die selbe Fragestellung, die in Abschnitt 4.1 nach

Tabelle 6.5: Kreuztabelle Glauben an Himmel und Glauben an Hölle (mit Erwartungshäufigkeiten)

			Glauben an Hölle		
			nein	**ja**	$\sum$
Glauben an Himmel	**nein**	f	40693	979	41672
		% (Glauben an Himmel)	97.7	2.3	
		$n\tilde{p}$	16820.8	24851.2	
	ja	f	17108	84417	101525
		% (Glauben an Himmel)	16.9	83.1	
		$n\tilde{p}$	40980.2	60544.8	
	$\sum$	f	57801	85396	143197
		% (Glauben an Himmel)	40.4	59.6	

Datenquelle: World Values Survey 1981-2008 (Inglehart et al., 2014).

„Augenmaß“ beurteilt wurde, nun mit einem statistischen Test zu beantworten. Die Nullhypothese lautet, dass es keine Unterschiede zwischen den X-bedingten Verteilungen von Y gibt, $H_0 : p(Y|X=1) = \ldots = p(Y|X=k)$. Wenn die Nullhypothese zutrifft und die Verteilungen von Y also unabhängig von X sind, dann ist unter jeder Bedingung von X die Verteilung von Y gleich ihrer (unbedingten) Randverteilung, $p(Y|X) = p(Y)$. Somit sind die bedingten relativen Erwartungshäufigkeiten $\tilde{p}(Y=l|X=j) = p(Y=l)$ und die globalen relativen Erwartungshäufigkeiten demnach $\tilde{p}(X=j, Y=l) = p(X=j)\,p(Y=l)$. Mit den Voraussetzungen, dass $n \geq 60$ und $min(n\tilde{p}(X=j, Y=l)) \geq 5$, kann der Test mit der $\chi^2_{df,1-\alpha}$-verteilten Prüfgröße $\chi^2 = \sum_{j=1}^{k}\sum_{l=1}^{m} \frac{(f(X=j,Y=l)-n\tilde{p}(X=j,Y=l))^2}{n\tilde{p}(X=j,Y=l)}$ und $df = (k-1)(m-1)$ durchgeführt werden.

Ein passendes exemplarisches Szenario findet sich in Abschnitt 4.1 mit den Daten in Tabelle 4.2, ergänzt durch die Erwartungshäufigkeiten in Tabelle 6.5. Damit kann die Prüfgröße

$$\chi^2 = \frac{(40693 - 16820.8)^2}{16820.8} + \frac{(979 - 24851.2)^2}{24851.2} + \frac{(17108 - 40980.2)^2}{40980.2}$$

$$+ \frac{(84417 - 60544.8)^2}{60544.8} = 80130.4626$$

berechnet und mit der χ^2-Prüfverteilung verglichen werden. Der kritische Wert bei $\alpha = 0.05$ und mit $df = (2-1)(2-1) = 1$ ist $\chi^2_{df,1-\alpha} = 3.84$. Da $\chi^2 > \chi^2_{df,1-\alpha}$, wird die Unabhängigkeit behauptende Nullhypothese verworfen und das bereits in Abschnitt 4.1 erzielte Ergebnis bestätigt: Es gibt einen Zusammenhang zwischen *Glaube an Himmel* und *Glaube an Hölle*.

6.4.2 Arithmetisches Mittel: F-Test auf Unterschied

Eine Möglichkeit, den Zusammenhang zwischen einer kategorialen Variable (X mit k Kategorien) und einer metrischen Variable (Y) zu untersuchen, ist der Vergleich der X-bedingten arithmetischen Mittel von Y. Dafür kann getestet werden, ob sich die durch X bedingten arithmetischen Mittel der Variable Y signifikant unterscheiden. Die Nullhypothese $H_0 : \mu(Y|X=1) = \ldots = \mu(Y|X=k)$ behauptet, dass es keine Unterschiede zwischen den X-bedingten arithmetischen Mittelwerten von Y gibt. Falls die Nullhypothese zutrifft, so ist Y unabhängig von X und die arithmetischen Mittelwerte von Y somit unter jeder Bedingung von X gleich dem unbedingten arithmetischen Mittel. Die Prüfgröße $F = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left(f(X=j)(\hat{\mu}(Y|X=j)-\hat{\mu}(Y))^2\right)}{\frac{1}{n-k}\left(\sum_{j=1}^{k}\sum_{i=1}^{f(X=j)}\left((y_{ij}-\hat{\mu}(Y|X=j))^2\right)\right)}$ ist $F_{df_1,df_2,1-\alpha}$-verteilt mit $df_1 = k-1$ und $df_2 = n-k$ Freiheitsgraden (Winer, 1962).

Tabelle 6.6 listet für die Variable der Fairness-Einschätzung (s. Abschnitt 6.3.3) sowohl für die Bundesländer als auch gesamt die arithmetischen Mittel, Standardabweichungen, Häufigkeiten, und die mit den jeweiligen Häufigkeiten gewichteten quadrierten Differenzen der arithmetischen Mittel in den Bundesländern zum arithmetischen Mittel der Gesamtstichprobe. Eine für den F-Test geeignete Fragestellung ist nun, ob sich die Personen in den Bundesländern bezüglich der durchschnittlichen Fairness-Einschätzung (signifikant) unterscheiden. Mit anderen Worten: hat das Bundesland, in dem eine Person lebt, Einfluss auf die Fairness-Einschätzung? Mit den gegebenen Daten kann die Prüfgröße $F = \frac{\frac{1}{16-1}142.9293}{\frac{1}{2037-16}((2.064^2\times(2037-1))-142.9293)} = 2.2574$ berechnet und mit dem kritischen Wert aus der F-Verteilung mit $df_1 = 16-1 = 15$, $df_2 = 2037-16 = 2021$ und $\alpha = 0.05$, $F_{df_1,df_2,1-\alpha} = 1.6713$, verglichen werden. Da in diesem Fall $F > F_{df_1,df_2,1-\alpha}$,kann H_0 verworfen werden; d.h. die durchschnittliche Fairness-Einschätzung unterscheidet sich signifikant zwischen den Bundesländern.

6.5 Weitere Kriterien zur Beurteilung statistischer Testresultate

Die Entscheidung über Annahme oder Ablehnung der (Null-) Hypothese erfolgt in der klassischen inferentiellen Statistik allein anhand der ermittelten (Irrtums-) Wahrscheinlichkeit des Alphafehlers bzw. der berechneten Prüfgröße. Und tatsächlich ist die Wahrscheinlichkeit des Alphafehlers ein notwendiges - und unter „rein inferenzstatistischem“ Aspekt auch hinreichendes - Kriterium für entsprechende Fragestellungen. Allerdings liefert die inferenzstatistische Interpretation solcher Testresultate nicht per se die praktische Bedeutung. Dafür können allerdings Effektstärken herangezogen werden (Cohen, 1994). Besonders dienlich sind in dieser Hinsicht vergleichbare Parameter mit einem definierten und normalisierten Wertespektrum. Bezüglich solcher statistischer Tests, mit denen der Zusammenhang von zwei Variablen untersucht werden kann, liegt

Tabelle 6.6: Einschätzung der Fairness in Bundesländern

Bundesland j	$\hat{\mu}(Y \mid X = j)$	$s(Y \mid X = j)$	$f(X = j)$	$f(X = j)\left(\hat{\mu}(Y \mid X = j) - \hat{\mu}(Y)\right)^2$
Schleswig-Holstein	5.73	1.909	71	0.3546
Hamburg	6.16	2.002	45	11.2801
Niedersachsen	6.01	1.841	197	24.2247
Bremen	5.00	1.566	17	7.3902
Nordrhein-Westfalen	5.74	1.674	445	2.8957
Hessen	5.13	2.033	148	41.4685
Rheinland-Pfalz	5.93	1.649	100	7.3261
Baden-Württemberg	5.47	2.274	264	9.4635
Bayern	5.62	2.468	311	0.4811
Saarland	5.71	2.035	25	0.0642
Berlin	5.81	2.511	87	1.9750
Brandenburg	5.70	1.983	64	0.1058
Mecklenburg-Vorpommern	5.47	2.191	42	1.5056
Sachsen	5.89	2.198	107	5.6932
Sachsen-Anhalt	4.95	1.955	57	28.6797
Thüringen	5.64	1.965	57	0.0213
$k = 16$	$\hat{\mu}(Y) = 5.66$	$s(Y) = 2.064$	$n = 2037$	$\sum = 142.9293$

ein solches Maß z.B. mit dem bereits besprochenen Koeffizienten der Produkt-Moment-Korrelation vor. Zwar kann der Korrelationskoeffizient nicht für jeden Test unmittelbar berechnet werden. Aber viele Prüfgrößen lassen sich in einen vorzeichenlosen Korrelationskoeffizienten umrechnen; der Betrag ist dann ein Maß für die Effektstärke. Ist $|r| = 0$, dann gibt es keinen Effekt, $|r| = 1$ repräsentiert den stärksten möglichen Effekt. Beispielhaft sei hier die Umrechnung von χ^2 des entsprechenden Unabhängigkeitstests mit den Daten aus Abschnitt 6.4.1 demonstriert. Eine Variante der Gleichung für die Umrechnung (Stuart, 2010) ist $|r| = \sqrt{\frac{\chi^2}{\chi^2+n}}$, für das konkrete Beispiel also $|r| = \sqrt{\frac{80130.4626}{80130.4626+143197}} = 0.599$. Demnach ist der Zusammenhang zwischen *Glaube an Himmel* und *Glaube an Hölle* nicht nur statistisch signifikant, sondern darüber hinaus auch mit $|r| > 0.5$ als stark einzuschätzen.

Weiterhin kann die Güte eines Tests mit der auf dem Betafehler (s. Abschnitt 6.1) basierenden statistischen Power beschrieben werden (Cohen, 1988). Dabei ist es allerdings problematisch, dass die Wahrscheinlichkeit β, den Betafehler zu begehen, nicht unmittelbar aus den Daten bestimmt werden kann, da die Alternativhypothese nicht hinreichend spezifisch dafür ist. D.h. es gibt keinen konkreten Erwartungswert für $\hat{\theta}$ unter der Bedingung, dass H_A zutrifft. Für entsprechend spezifizierte Szenarien kann die Wahrscheinlichkeit β jedoch ermittelt werden. Für diese Bestimmung der statistischen Power muss also eine maximal hinnehmbare Wahrscheinlichkeit α des Alphafehlers gesetzt werden. Dafür kann auf das übliche Signifikanzniveau $\alpha = 0.05$ rekurriert werden. Schließlich kann die Power eines statistischen Tests berechnet werden als $1 - \beta = P_0\left(\hat{\theta} - \theta_{1-\alpha}\right)$. Um das angefangene Beispiel nun auch unter dem Aspekt der statistischen Power darzustellen, ergibt sich $1 - \beta = P_{\chi^2_{df=1}}(80130.4626 - 3.84) = 0.9999$. Insofern für einen akzeptablen Test eine statistische Power $1 - \beta \geq 0.8$ erwartet wird, kann auch in dieser Hinsicht das signifikante Ergebnis des besprochenen Beispieltests als robust angesehen werden.

Teil IV

Statistische Modelle

Kapitel 7

Schätzmodelle

7.1 Erklärung, Prädiktion und Residuen

Die grundsätzliche Funktion von statistischen Modellen ist die Prädiktion: die Werte einer (abhängigen) Variable Y sollen für konkrete Fälle *vorhergesagt* werden. Eine derartige Vorhersage durch ein Modell wird als *Prädiktion* bezeichnet. Dafür muss das theoretische Modell als Funktions- bzw. Modellgleichung formuliert werden. Konkrete vorhergesagte Werte $\hat{y}_i$ sind somit jeweils das Ergebnis der Modellfunktion. Insofern solche Funktionen Parameter der Population verwenden, die i.d.R. unbekannt sind, müssen die auf den empirischen Daten beruhenden Lösungen dieser Gleichungen als *Schätzungen* begriffen werden - daher die Bezeichnung entsprechender Modelle als *Schätzmodelle*.

Mit Schätzmodellen können zwei Ziele verfolgt werden. Das häufigste Anliegen in Verbindung mit Schätzmodellen ist die quantifizierte Erklärung von Zusammenhängen. Quantifizierung in diesem Sinne meint die Schätzung der Gleichungsparameter der Modellfunktion $\hat{Y}$ und damit die Information, welchen konkreten Wert der Variable Y das Modell für einen bestimmten Fall i mit konkreten Ausprägungen des (unabhängigen) Variablenvektors X vorhersagt: $\hat{y}_i = \hat{Y}(X_i)$. Dieses Ziel scheint womöglich auf den ersten Blick keinen brauchbaren Mehrwert zu bringen, insofern die konkreten Ausprägungen y_i in den Daten bekannt und gemessen sind und daher gar nicht vorhergesagt werden müssen. Bei der Lösung der Modellgleichung werden die konkreten Y-Werte auch tatsächlich als gegeben behandelt - die Lösung betrifft die unbekannten Parameter des Modells. Die „Vorhersage des bereits Geschehenen“ ist hierbei also quasi das Mittel für das Ziel der Modellschätzung. Und mit einem Modell bzw. den jeweiligen Parametern können dann eben Zusammenhänge präziser abgebildet werden.

Das zweite Ziel in Verbindung mit statistischen Schätzmodellen betrifft ihre Anwendung im Sinne der Vorhersage tatsächlich nicht gemessener Werte. Für die wissenschaftlich-empirische Forschung ist diese Anwendung statistischer Schätzmodelle jedoch eher trivial, ein Mehrwert solcher Prädiktionen ergibt sich hier

für die (nicht-wissenschaftliche) Praxis. Allerdings steht dieses Anwendungsziel mit dem statistischen Ziel in Verbindung: ein statistisches Schätzmodell, mit welchem Prädiktionen vorgenommen werden können, kommt eben nicht wie Manna vom Himmel gefallen - es muss auf einer empirischen Grundlage geschätzt werden.

Sowohl für das statistische Primärziel als auch für das sekundäre Anwendungsziel wird ein Schätzmodell angestrebt, welches die Realität so gut wie möglich trifft. D.h. die Prädiktionen (der bekannten Y-Werte) durch das Modell sollen in der Summe mit möglichst wenigen bzw. kleinen Fehlern erfolgen. Die als *Residuen* (e) bezeichneten Prädiktionsfehler beschreiben die Differenz einer Prädiktion zum jeweils empirischen (tatsächlich gemessenen) Y-Wert: $e_i = y_i - \hat{y}_i$. Die Güte eines Modells („Goodness-of-Fi“) ist also umso größer, je kleiner die Residuen sind. Die gängigen Algorithmen zur Schätzung der Modellparameter zielen im Kern darauf ab, die Modellgleichung durch Schätzung der Parameter so zu lösen, dass die Residuen minimal sind.

7.2 Varianzkomponenten

Ein Maß dafür, wie groß in der Summe die Residuen sind, ist die Varianz der Residuen um $E(e) = 0$: $s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. Jedoch ist diese Residuen-Varianz allein und mit Blick auf die Güte des Schätzmodells schwer bzw. gar nicht substantiell interpretierbar. Denn es gibt zwar eine theoretische Untergrenze der Varianz - aber keine Obergrenze. Mithin lässt sich die Modellgüte anhand der Varianz lediglich im Vergleich mit der Varianz eines alternativen Modells (mit derselben abhängig Variable und der Schätzung anhand der gleichen Datenbasis) bestimmen: welches Schätzmodell führt zu einer geringeren Varianz und hat damit eine größere Güte? Der Vergleich der Varianz mit einem beliebigen alternativen Modell liefert aber immer noch kein standardisiertes Maß für die Modellgüte.

Allerdings lässt sich ein Modellgütemaß auf der Grundlage der Residuen-Varianz bestimmen, wenn als Referenz statt irgendeines alternativen Schätzmodells ein sogenanntes *Nullmodell* verwendet wird. Das Besondere an einem solchen Nullmodell ist, dass es keine Funktion unabhängiger Variablen X ist, sondern eine Funktion allein der abhängigen Variable Y. Zudem ist die Prädiktion des Nullmodells eine Konstante, d.h. für jeden konkreten Fall i eines Samples wird der gleiche Y-Wert vorhergesagt. Mit dem Nullmodell wird also genau *ein* Parameter geschätzt. Nichtsdestotrotz gilt für das Nullmodell der Anspruch minimaler Residuen. Mithin muss der geschätzte Parameter, der als Prädiktion für Y fungiert, die Eigenschaft aufweisen, dass es - für den konkreten Datenbestand - keinen anderen Wert als Prädiktion gibt, für welchen die Varianz der Residuen noch kleiner wäre. Nur Schätzmodelle, die zusätzliche Parameter für den X-Vektor schätzen, können eine geringere Varianz aufweisen. Diese Minimaleigenschaft nun erfüllt das arithmetische Mittel (s. Abschnitt 3.2):

weil $\sum_{i=1}^{n} (y_i - \zeta)^2 \rightarrow min$ die Lösung $\zeta = \bar{y}$ hat, ist das Nullmodell mit der geringsten Varianz $\hat{y}_i = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. Mit Blick auf die Residuen ist dabei also $y_i = \bar{y} + e_i$, d.h. $e_i = y_i - \bar{y}$. Somit entspricht die Varianz des Nullmodells $s_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ der Varianz der Variable Y (s. Abschnitt 3.3). Die Varianz des Nullmodells wird dann als Referenz verwendet, denn bei gleichem Sample und gleicher Variable gibt es kein Modell welches eine größere Varianz und damit eine schlechtere Modellgüte aufweist. Sei die Varianz der Residuen eines Modells mit X-Vektor s_e^2, dann ist $s_M^2 = s_0^2 - s_e^2$ der Teil der Varianz des Nullmodells, welcher durch das finale Modell mit X-Vektor erklärt werden kann. Das finale Modell ist quasi um s_M^2 besser als das „schlechteste“ aller Modelle, das Nullmodell. Nun ist es möglich, anhand dieser Zerlegung der Varianz des Nullmodells in *Varianzkomponenten* ($s_0^2 = s_e^2 + s_M^2$) - der durch das finale Modell gegenüber dem Nullmodell *erklärten Streuung* der Residuen s_M^2 und der *nicht erklärten Streuung* s_e^2 - die Modellgüte des finalen Modells mit einem normierten Maß zu beschreiben. Als Anteil der erklärten Varianz an der Varianz des Nullmodells liegt $R^2 = \frac{s_M^2}{s_0^2}$ im Intervall $[0, 1]$, also $0 \leq R^2 \leq 1$.

7.3 Regression: Modelle, Algorithmen, Beispiele

Eine weit verbreitete Klasse von Verfahren zur Schätzung statistischer Modell firmiert unter dem Begriff *Regressionsanalyse*. Wörtlich bedeutet *Regression* die *Zurückführung* - im Kontext der Statistik die Zurückführung der Ausprägung einer abhängigen Variable auf die Ausprägungen eines Vektors unabhängiger Variablen, was eben genau der Beschreibung eines statistischen Schätzmodells entspricht. Allerdings gibt es diverse Arten von Funktionsgleichungen, deren Lösung ggf. spezielle Schätzalgorithmen erfordern. In diesem Kapitel werden lediglich die prominentesten Regressionsverfahren besprochen.

7.3.1 Lineare Regression: OLS

Wie die Bezeichnung der *linearen Regression* womöglich vermuten lässt, basiert dieses Verfahren auf einem Modell, welches die statistischen Beziehungen zwischen abhängiger Variable und unabhängiger Variable(n) als linear formuliert. Um dies zu veranschaulichen, greifen wir zunächst auf das Beispiel zur Korrelation aus Abschnitt 4.2 zurück. Auf einen linearen Zusammenhang hin untersucht wurden die Variablen Punkte Statistik (X) und Punkte Mathe (Y). Anhand der berechneten Korrelation ($r = 0.8523$) konnte ermittelt werden, dass es einen starken und positiven linearen Zusammenhang zwischen den beiden Variablen gibt. D.h. je größer die Punktzahl eines Studierenden in Mathe, desto größer seine Punktzahl in Statistik. Zwar suggeriert dabei die Zuordnung der Symbole X und Y, dass Punkte Statistik die unabhängige Variable ist und Punkte Mathe die abhängige Variable. Für die Berechnung des Korrelationskoeffizienten ist diese Zuordnung jedoch ohne Belang und theoretisch nicht erforderlich. Aber für

ein Schätzmodell, welches die Prädiktion allein der abhängigen Variable liefert, ist die Unterscheidung von abhängiger Variable und unabhängiger Variable(n) essentiell. Um für dieses Beispiel ein statistisches Modell zu formulieren und seine Parameter zu schätzen, legen wir explizit Punkte Statistik als unabhängige Variable (X) fest und Punkte Mathe als abhängige Variable (Y)[1].

Das allgemeine Modell einer linearen Regression mit k unabhängigen Variablen ist formuliert mit

$$\hat{y}_i = \hat{b}_0 + \sum_{j=1}^{k} \left(\hat{b}_j x_{ji} \right)$$

Dabei sind $\hat{b}_0$ und $\hat{b}_j$ die statistisch geschätzten Parameter des Modells. Zwar sind beide Parameter im Modell konstant. Allerdings wird nur $\hat{b}_0$ auch als Konstante bezeichnet, weil die Produkte der sogenannten Regressionskoeffizienten $\hat{b}_j$ mit dem jeweiligen Wert der unabhängigen Variable x_{ji} eben nicht konstant sind. Es wird deutlich, dass die Prädiktion $\hat{y}_i$ abhängig von den jeweiligen Werten der unabhängigen Variable(n), x_{ji}, ist.

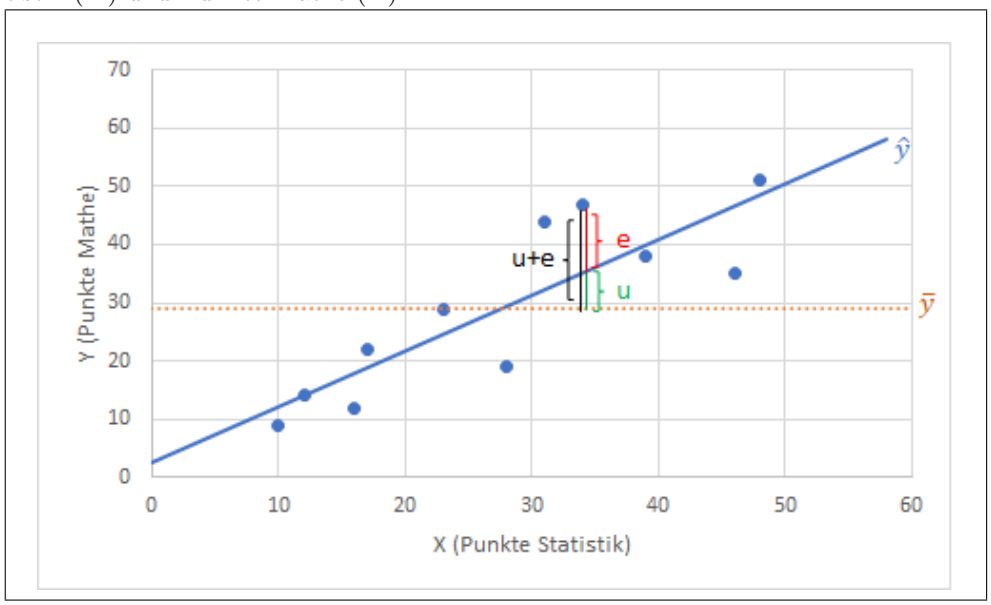
Der OLS-Algorithmus zur Lösung einer solchen Modellgleichung lässt sich elegant in Matrixschreibweise darstellen. Sei $\mathbf{X}$ eine $(n, k+1)$-Matrix, die einer Datenmatrix ähnlich in den Zeilen die Fälle enthält und in den Spalten 2 bis $k+1$ die unabhängigen Variablen des Modells. Die erste Spalte ist mit 1 aufgefüllt und repräsentiert den konstanten Vektor. Der n-dimensionale Spaltenvektor $\mathbf{y}$ beinhaltet die Werte der abhängigen Variable für die jeweiligen Fälle. Die statistischen Schätzungen der Regressionskoeffizienten $\hat{b}_j$ (inklusive der Konstante $\hat{b}_0$) werden im $(k+1)$-dimensionalen Vektor $\mathbf{b}$ abgebildet. Die Gleichung zur Lösung der Koeffizienten lautet somit

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Für das begonnene Beispiel wird nun das Modell einer linearen Regression formuliert, welches sich hier für die Prädiktion von Punkten in Mathe (Y) den Zusammenhang dieser Variable mit Punkten in Statistik (X) zunutze macht: $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$. Als Ergebnis erhalten wir hier $\hat{b}_0 = 2.6675$ und $\hat{b}_1 = 0.9561$; die konkrete Regressionsgleichung für dieses Beispiel ist also $\hat{y}_i = 2.6675 + 0.9561 x_i$. Somit ist nun auch leicht ersichtlich, inwieweit das Modell den Zusammenhang von Y mit X als linear formuliert: im Streudiagramm kann das Modell als eine Gerade durch die bivariate Streuung dargestellt werden (Abbildung 7.1). In diesem Lichte lässt sich der Regressionskoeffizient $\hat{b}_1$ als Steigung dieser Gerade beschreiben: wenn X um einen Punkt steigt, dann verändert sich der Schätzwert für Y um $\hat{b}_1$. Die Konstante $\hat{b}_0$ fungiert als Achsenabschnitt (d.h. Schnittpunkt mit der Y-Achse): die Prädiktion für Y ist also $\hat{b}_0$, wenn $x_i = 0$.

[1] Von Ausnahmen abgesehen, folgt in der Praxis die Zuordnung der Variablen als *abhängig* und *unabhängig* (kausal-) theoretischen Erwägungen. Für dieses Beispiel gibt es jedoch keine theoretische Begründung; die Zuordnung von *Punkte Statistik* zu X und *Punkte Mathe* zu Y ist hier arbiträr.

Abbildung 7.1: Streudiagramm, Modell und Residuen der Variablen Punkte Statistik (X) und Punkte Mathe (Y)



Residuen: $e_i = y_i - \hat{y}_i$, $u_i = \hat{y}_i - \bar{y}$, $u_i + e_i = y_i - \bar{y}$

In Abbildung 7.1 ist zudem die Zerlegung der Varianz in ihre Komponenten anhand eines konkreten Falles dargestellt. Demnach kann die Varianz des Nullmodells $s_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (u_i + e_i)^2$ in die durch das Modell erklärte Varianz $s_M^2 = \frac{1}{n-1} \sum_{i=1}^{n} (u_i)^2$ und die vom Modell nicht erklärte Varianz $s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} (e_i)^2$ zerlegt werden. Die Varianz des Nullmodells ist die quadrierte (bereits in Abschnitt 4.2 ermittelte) Standardabweichung, $s_0^2 = s_Y^2 = 14.8758^2 = 221.2894$. Die Ermittlung der nicht erklärten Varianz ist detailliert mit der um die Prädiktionen und die Residuen erweiterten Datenreihe verdeutlicht:

Student i	1	2	3	4	5	6	7	8	9	10	11
Pkte. Statistik (X)	39	34	31	48	46	23	17	12	16	28	10
Pkte. Mathe (Y)	38	47	44	51	35	29	22	14	12	19	9
Modell ($\hat{Y}$)	39.55	34.77	31.91	48.16	46.25	24.26	18.52	13.74	17.56	29.04	11.83
Res. $e_i = y_i - \hat{y}_i$	-1.55	12.23	12.09	2.84	-11.23	4.74	3.48	0.26	-5.5644	-10.04	-2.83

$$s_e^2 = \frac{1}{11-1}\begin{pmatrix} 1.5547^2 \\ +12.2259^2 \\ +12.0942^2 \\ +2.8405^2 \\ +11.2474^2 \\ +4.7430^2 \\ +3.4796^2 \\ +0.2601^2 \\ +5.5644^2 \\ +10.0376^2 \\ +2.8278^2 \end{pmatrix} = 60.7109$$

Die erklärte Varianz ergibt sich aus der Differenz zwischen der Varianz des Nullmodells und der residualen Varianz: $s_M^2 = s_0^2 - s_e^2 = 221.2894 - 60.7109 = 160.5785$. Somit beträgt der Anteil der erklärten Varianz (an der Varianz des Nullmodells) $R^2 = \frac{s_M^2}{s_0^2} = \frac{160.5785}{221.2894} = 0.7256$.

7.3.2 Logistische Regression

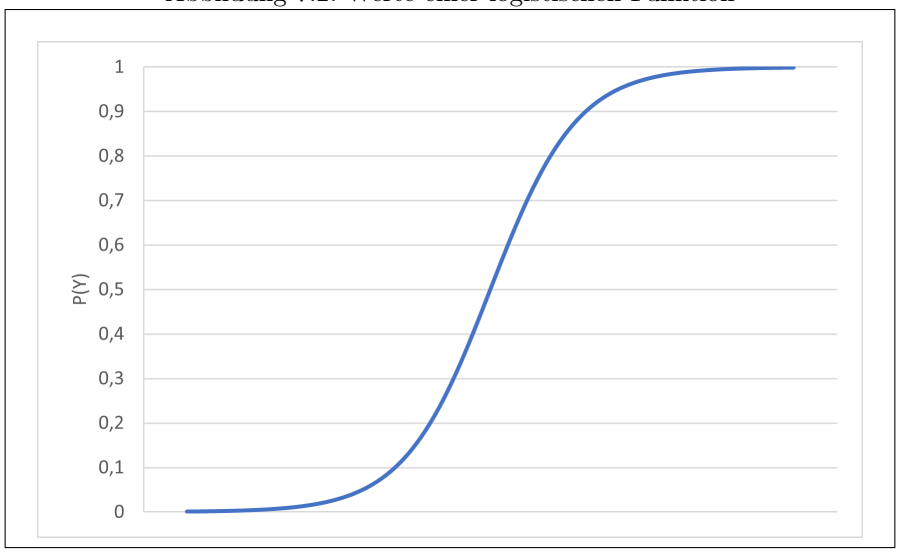
Die Modellgleichung einer logistischen Regression ist nicht linear, sondern logistisch. Logistische Modellgleichungen eignen sich besonders gut zur Modellierung abhängiger Variablen, die in ihrem Wertespektrum begrenzt sind. So werden z.B. Variablen, die ein dichotomes Ereignis repräsentieren, häufig mit einer logistischen Regression statistisch modelliert. Um dies zu illustrieren, greifen wir erneut auf die Beispieldaten zurück, welche bereits für die Themen *Korrelation* und *lineare Regression* verwendet wurden - allerdings mit einer entscheidenden Modifikation der abhängigen Variable. Y sei nicht länger die Punktzahl in einer Mathematik-Klausur, sondern die Zuordnung, ob ein Studierender die Mathematik-Klausur bestanden hat ($Y = 1$) oder nicht ($Y = 0$). Als Schwelle wollen wir festlegen, dass Studierende mit mindestens 20 Punkten in Mathe die Klausur bestanden haben.

$$\text{Punkte Mathe} \begin{cases} < 20 & y = 0 \quad \text{(Mathe-Klausur } \textit{nicht} \text{ bestanden)} \\ \geq 20 & y = 1 \quad \text{(Mathe-Klausur } \textit{bestanden}\text{)} \end{cases}$$

Student i	1	2	3	4	5	6	7	8	9	10	11
Punkte Statistik (X)	39	34	31	48	46	23	17	12	16	28	10
Mathe-Klausur bestanden (Y)	1	1	1	1	1	1	1	0	0	0	0

Die OLS-Modellierung von Y würde hier zu unmöglichen Werten führen, insofern das Wertespektrum der linearen Funktionsform unbegrenzt und stetig ist - während die Skala der hier verwendeten Variante von Y begrenzt und diskret ist: $\{0, 1\}$. Eine Möglichkeit der Modellierung einer solchen dichotomen Variable besteht nun darin, nicht mehr die Variable selbst - sondern die Wahrscheinlichkeit ihrer Ausprägungen als abhängig zu betrachten: $\hat{P}(Y = 1)$. Eine sich daraus

Abbildung 7.2: Werte einer logistischen Funktion



unmittelbar ergebende Konsequenz ist, dass eine Eintrittswahrscheinlichkeit als abhängige Variable stetig ist. Eine lineare Modellierung von Wahrscheinlichkeiten ist allerdings trotz der stetigen Skalierung inadäquat, weil das Wertespektrum mit $[0, 1]$ weiterhin begrenzt ist und OLS weiterhin unmögliche Werte vorhersagen könnte. Hingegen wird mit einer logistischen Modellierung genau diese Anpassung an das Wertespektrum von Wahrscheinlichkeiten erreicht (s. Darstellung in Abbildung 7.2).

Das allgemeine Modell einer logistischen Regression mit k unabhängigen Variablen lautet

$$\hat{P}\left(y_i = 1\right) = \frac{1}{1 + \exp^{-\left(\hat{b}_0 + \sum_{j=1}^{k}\left(\hat{b}_j x_{ji}\right)\right)}}$$

Wie leicht zu sehen ist, handelt es sich dabei eigentlich um die logistische Transformation des linearen Schätzmodells $\hat{Y}_i$: $\hat{P}\left(y_i = 1\right) = \frac{1}{1+\exp^{-\hat{Y}}}$. Der entscheidende Unterschied des logistischen Schätzmodells zum OLS-Modell ist die Bestimmung der Residuen, insofern die lineare Modellkomponente eben nicht adäquat die abhängige Variable vorhersagt. Die Vorhersage ist letztlich die Wahrscheinlichkeit. Daher werden die Residuen im logistischen Modell als $e_i = y_i - \hat{P}\left(y_i = 1\right)$ ermittelt. Als Schätzmodell wird auch von der logistischen Regression gefordert, dass die Varianz der Residuen minimiert wird - eine Anforderung, die nun jedoch die OLS-Methode für die abhängige Variable hier nicht mehr erreichen kann. Stattdessen erfolgt die Residuen minimierende Schätzung der Modellparameter iterativ mit dem Maximum-Likelihood-Verfahren (ML). Dabei wird die residuale Varianz über einen Umweg minimiert: nämlich durch

die Maximierung der Likelihood. Die Likelihood ist ein Maß, welches die Ähnlichkeit zwischen Modell (Prädiktion) und (empirischen) Daten beschreibt. Das Schema des iterativen ML-Verfahrens ist, dass mit den in der vorangegangenen Iteration ermittelten Modellparametern die Prädiktion der abhängigen Variable ($\hat{P}$) durchgeführt und die Ähnlichkeit zur abhängigen Variable (Y) fallweise mit $C_i = \hat{P}\left(y_i = 1\right)^{y_i} \left(1 - \hat{P}\left(y_i = 1\right)\right)^{1-y_i}$ berechnet wird. Die Likelihood ergibt sich dann aus der Gesamtheit dieser fallweisen Beiträge: $L = \prod_{i=1}^{n} C_i$.

Als Startwert für die erste Iteration kann den Modellparametern Null zugewiesen werden. Der erste Schritt einer ML-Iteration besteht in einer Prädiktion der $\hat{P}\left(y_i = 1\right)$ und der sich daraus ergebenden Streuung

$$W = \sum_{i=1}^{n} \left(\hat{P}\left(y_i = 1\right)\left(1 - \hat{P}\left(y_i = 1\right)\right)\right)$$

Mit dem Skalar W wird die Matrix X quasi gewichtet:

$$\tilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$$

Mit dem nächsten Schritt wird die Matrix

$$\mathbf{V} = \left(\mathbf{X}'\tilde{\mathbf{X}}\right)^{-\mathbf{1}} \mathbf{X}'$$

ermittelt, deren Elemente die Grundlage für

$$d_{ij} = v_{ij}\left(y_i - \hat{P}\left(y_i = 1\right)\right)$$

sind, womit die Modellparameter angepasst werden können:

$$\hat{b}_j = \hat{b}_j^{\bullet} + n\sum_{i=1}^{n} d_{ij}$$

wobei $\hat{b}_j^{\bullet}$ der Schätzwert des jeweiligen Modellparameters aus der vorangegangenen Iteration ist. Das Iterieren kann abgebrochen werden, wenn sich die Schätzwerte der Modellparameter durch die iterative Anpassung kaum noch ändern, wenn also die Elemente der Matrix $\mathbf{D}$ sehr kleine Beträge aufweisen (z.B. $\max\left(|d_{ij}|\right) < 0.001$). Weiterhin kann eine maximale Anzahl von Iterationen (z.B. 20) vorgegeben werden.

Mit der logistischen Regression erhalten wir hier folgende Schätzungen: $\hat{b}_0 = -3.9294$ und $\hat{b}_1 = 0.1875$.

Student i	1	2	3	4	5	6	7	8	9	10	11
Punkte Statistik (X)	39	34	31	48	46	23	17	12	16	28	10
Mathe bestanden (Y)	1	1	1	1	1	1	1	0	0	0	0
lineares Modell ($\hat{Y}$)	3.38	2.45	1.88	5.07	4.70	0.38	-0.74	-1.68	-0.93	1.32	-2.05
Wahrsch.-smodell ($\hat{P}$)	0.97	0.92	0.87	0.99	0.99	0.59	0.32	0.15	0.28	0.79	0.11
fallw. Likelihood (C_M)	0.97	0.92	0.87	0.99	0.99	0.59	0.32	0.84	0.72	0.21	0.89
Nullmodell ($\bar{y}$)	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
Likelihood Nullmod. (C_0)	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.36	0.36	0.36	0.36

Üblicherweise (u.a. um nicht mit sehr kleinen Beträgen rechnen zu müssen) wird statt der Likelihood die logarithmierte Likelihood („LogLikelihood“) LL berichtet. Für das finale Modell erhalten wir hier $LL_M = -4.1203$, und für das Nullmodell $LL_0 = -7.2106$. Mit dem ML-Verfahren wurde also die (logarithmierte) Likelihood maximiert - und insofern jedes statistische Modell mit Prädiktoren mindestens so viel Varianz aufklärt wie das Null-Modell, ist dementsprechend $LL_M \geq LL_0$. Nun ist allerdings die Differenz $LL_M - LL_0$ ein ziemlich schwer zu interpretierendes Maß für die Modellgüte. Jedoch ist die mit -2 multiplizierte Differenz der logarithmierten Likelihoods χ^2-verteilt (mit $df = k$ Freiheitsgraden): $\chi^2_{LL} = -2\,(LL_0 - LL_M)$. Insofern nun χ^2 in die Effektgröße r umgerechnet werden kann (vgl. Abschnitt 6.5), lässt sich weiterhin das Gütemaß pseudo-$R^2 = \frac{\chi^2_{LL}}{\chi^2_{LL}+n}$ berechnen.

Teil V

Anwendungen: Probleme und Notizen

Für die statistischen Auswertungen in diesem Teil wurde die Software DASMod (Förster, 2022) verwendet.

V.1 Statistische Drittvariablenkontrolle

In Teil IV wurde thematisiert, dass mit statistischen Modellen der Effekt einer unabhängigen Variable X auf eine abhängige Variable Y geschätzt werden kann. Um einen solchen Effekt unverzerrt zu schätzen, muss ggf. eine hohe Komplexität berücksichtigt werden: Y wird womöglich nicht lediglich durch X beeinflusst, sondern auch durch andere Variablen Z („Drittvariablen“). Der Effekt einer Drittvariable Z könnte den primär interessierenden Effekt der Variable X überlagern und damit unterdrücken oder verstärken. Statistische Schätzmodelle berücksichtigen diese Komplexität insofern mit ihnen statistisch auf Drittvariablen kontrolliert wird, was die Schätzung bereinigter Nettoeffekte erlaubt.

Hier wird im Folgenden ein beispielhaftes Szenario dargestellt, in welchem eine unterlassene statistische Drittvariablenkontrolle zu einer Unterschätzung des interessierenden Effekts führt.

Abbildung V.1.1: Skizze Versuchsaufbau

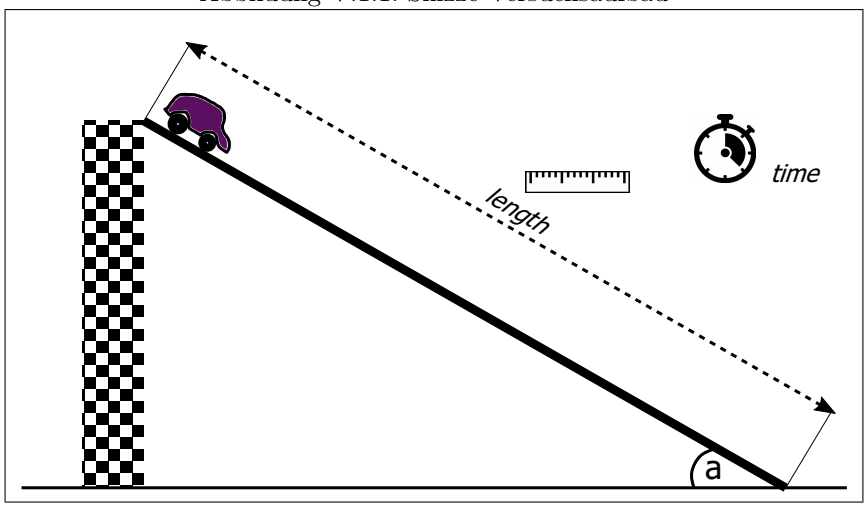
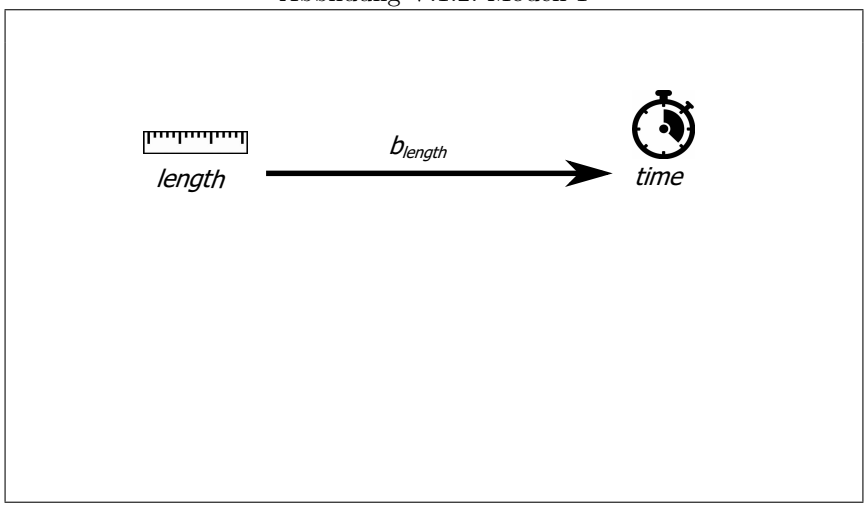


Abbildung V.1.1 zeigt den Versuchsaufbau: Für ein Fahrzeug ohne eigenen Antrieb wird jeweils eine Strecke aufgebaut, welche durch zwei variable Parameter charakterisiert ist - die Länge (*length*) der Strecke und ihre Neigung (α). Nachdem die Strecke entsprechend der vorher festgelegten Werte der beiden Streckenparameter aufgebaut wurde, erfolgte die Positionierung des Fahrzeugs am oberen Anfang der Strecke. Die Zeit (*time*), die das Fahrzeug benötigte, um bis zum unteren Ziel zu rollen, wurde gemessen.

Der primär interessierende Effekt ist der Einfluss der Distanz auf die benötigte Zeit. Mit anderen Worten: *time* ist die abhängige Variable (Y), *length* ist

die unabhängige Variable (X).

Abbildung V.1.2: Modell 1



Mit Modell 1 werden lediglich die unmittelbar interessierenden Variablen berücksichtigt (Abbildung V.1.2).

$$time_i = \hat{b}_0 + \hat{b}_{length} \times length_i + e_i \qquad \text{(Modell 1)}$$

Das korrespondierende statistische Modell schätzt mit $\hat{b}_{length}$ den Effekt der Distanz auf die benötigte Zeit: um wie viele Einheiten sich die benötigte Zeit ändert, wenn die Distanz um eine Einheit erhöht wird.

Die Distanz wurde in Millimetern eingestellt; die benötigte Zeit wird in Sekunden gemessen. Insgesamt wurden 26 Versuche durchgeführt. D.h. das Fahrzeug wurde 26 mal am Start positioniert, rollen gelassen und die benötigte Zeit gemessen. Vor einigen der Versuche wurden die Einstellung rekonfiguriert - *length* und α wurden variiert.

Jeder Punkt im Streudiagramm (Abbildung V.1.3) zeigt die für die Schätzung von Modell 1 relevante Konfiguration und die entsprechenden Resultate hinsichtlich *time*.

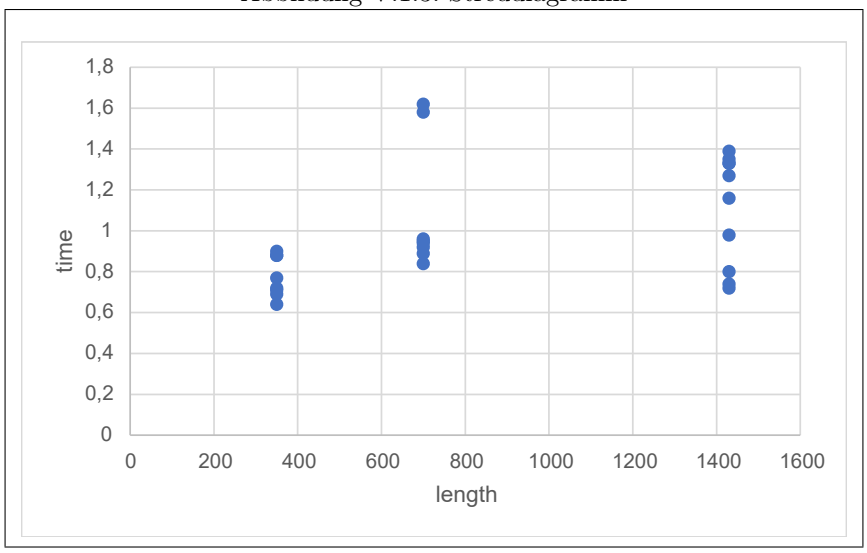
$$\hat{b}_0 = 0.774$$

(Modell 1)

$$\hat{b}_{length} = \frac{0.257}{1000}$$

Bezüglich des interessierenden Effekts der Distanz auf die benötigte Zeit, liefert das Schätzmodell eine Antwort: ein zusätzlicher Meter Streckenlänge erhöht die benötigte Zeit um 0.257 Sekunden.

Abbildung V.1.3: Streudiagramm



Somit kann die benötigte Zeit durch das Modell vorhergesagt werden. Allerdings ist diese Prädiktion nicht fehlerfrei. In jedem Versuch i resultiert ein Fehler e_i aus der Differenz zwischen tatsächlich gemessener Zeit und der durch das Modell vorhergesagten Zeit (Tabelle V.1.1).

$$\hat{time}_i = time_i - e_i$$

In der Praxis der empirischen Forschung sind Schätzmodelle kaum ohne Fehler - mindestens zufällige Messfehler sind oft zu erwarten. Jedoch kann auch eine inadäquate Modellspezifikation die Fehler erhöhen.

Um unterschiedlich spezifizierte Modelle hinsichtlich ihrer Qualität miteinander zu vergleichen, wird die Summe der quadrierten Fehler berechnet. Je kleiner diese Summe, desto besser passt das Modell zu der mit den Daten abgebildeten Realität.

$$\sum e^2 = 1.664 \qquad \text{(Modell 1)}$$

$$time_i = \hat{b}_0 + \hat{b}_{length} \times length_i + \hat{b}_\alpha \times \alpha_i + e_i \qquad \text{(Modell 2)}$$

Das zweite Modell (Abbildung V.1.4) ist etwas komplexer als das erste. Damit wird nun berücksichtigt, dass zusätzlich zur Streckenlänge auch die Neigung die benötigte Zeit beeinflusst. Der Alpha-Gradient ist eine Drittvariable, auf deren Effekt nun kontrolliert wird.

Tabelle V.1.1: Werte, Prädiktion und Fehler

		Modell 1			**Modell 2**	
length (mm)	*time* (sec)	$\hat{time}$	e	α	$\hat{time}$	e
350	0.88	0.864	-0.016	5	0.969	0.083
350	0.71	0.864	0.154	5	0.969	0.253
350	0.88	0.864	-0.016	5	0.969	0.083
350	0.90	0.864	-0.036	5	0.969	0.063
350	0.77	0.864	0.094	7	0.811	0.041
350	0.69	0.864	0.174	7	0.811	0.121
350	0.64	0.864	0.224	7	0.811	0.171
350	0.72	0.864	0.144	7	0.811	0.091
700	1.58	0.954	-0.626	5	1.240	-0.340
700	1.62	0.954	-0.666	5	1.240	-0.380
700	0.95	0.954	0.004	10	0.860	-0.090
700	0.94	0.954	0.014	10	0.860	-0.080
700	0.89	0.954	0.064	10	0.860	-0.030
700	0.96	0.954	-0.006	10	0.860	-0.100
700	0.84	0.954	0.114	12	0.708	-0.132
700	0.92	0.954	0.034	12	0.708	-0.212
1430	1.33	1.142	-0.188	10	1.437	0.107
1430	1.33	1.142	-0.118	10	1.437	0.107
1430	1.35	1.142	-0.208	10	1.437	0.087
1430	1.27	1.142	-0.128	12	1.285	0.015
1430	1.16	1.142	-0.018	12	1.285	0.125
1430	1.39	1.142	-0.248	12	1.285	-0.105
1430	0.98	1.142	0.162	18	0.829	-0.151
1430	0.72	1.142	0.422	18	0.829	0.109
1430	0.74	1.142	0.402	18	0.829	0.089
1430	0.80	1.142	0.342	18	0.829	0.029

$$\hat{b}_0 = 1.066$$

$$\hat{r} = 0.770$$

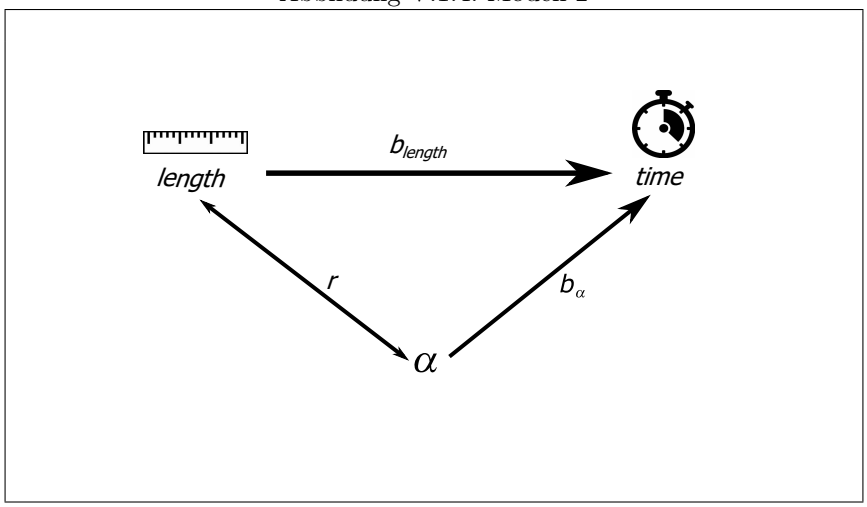
(Modell 2)

$$\hat{b}_{length} = \frac{0.791}{1000}$$

$$\hat{b}_\alpha = -0.076$$

Tatsächlich liefert Modell 2 andere Schätzungen. Die Tatsache, dass sich die Resultate so deutlich von den mit Modell 1 ermittelten Schätzungen un-

Abbildung V.1.4: Modell 2



terscheiden, ist teilweise auf die Korrelation der Streckenlänge mit der Neigung (je länger die Distanz, desto größer die Neigung) zurückzuführen. Natürlich ist diese Korrelation nicht theoretisch zwingend - sie ist hier im Zuge der jeweiligen Rekonfiguration des Versuchsaufbaus zustande gekommen. Im Rahmen dieses eher simplen Versuchsaufbaus ist das nicht zufällig geschehen. Aber es wurde damit ein realistisches Szenario simuliert, insofern in der „echten“ empirischen Forschung die Komplexität i.d.R. deutlich größer ist: Alles hängt mit Allem zusammen.

Der Effekt der Streckenlänge auf die benötigte Zeit ist deutlich größer in Modell 2: jeder zusätzliche Meter Streckenlänge erhöht die benötigte Zeit um 0.791 Sekunden.

Mithin sagt Modell 2 die benötigte Zeit nicht lediglich basierend auf der Streckenlänge voraus, sondern es wird zudem die Neigung berücksichtigt.

$$\sum e^2 = 0.585 \qquad \text{(Modell 2)}$$

Die Summe der quadrierten Fehler ist für Modell 2 deutlich kleiner verglichen mit Modell 1. Modell2 erreicht eine bessere Anpassung und hat daher eine höhere Güte als Modell 1.

Zusammenfassend lässt sich feststellen, dass Modell 1 nicht auf Drittvariablen kontrolliert und in diesem Sinne unter-spezifiziert ist. Demgegenüber kontrolliert Modell 2 auf die offenbar relevante Neigung als eine Drittvariable. In der Konsequenz wurde der Effekt der Streckenlänge auf die benötigte Zeit mit Modell 1 unterschätzt. Modell 2 schätzt diesen Effekt weniger verzerrt. Dies zeigt sich auch in der größeren Güte von Modell 2 im Vergleich zu Modell 1.

Literatur

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J: L. Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. doi:10.1037/0003-066x.49.12.997

Efron, B. (1994). *An introduction to the bootstrap*. New York: Chapman & Hall.

Förster, M. (2022). DASMod. Verfügbar unter https://seastar158.wordpress.com/dasmod/

GESIS-Leibniz-Institut für Sozialwissenschaften. (2017). Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2016. doi:10.4232/1.12796

Hartung, J. (2005). *Statistik*. Oldenbourg.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... et al. (2014). World Values Survey 6.

Kolmogoroff, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. doi:10.1007/978-3-642-49888-6

Kroenke, K., Spitzer, R. L. & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. doi:10.1046/j.1525-1497.2001.016009606.x

OECD. (2015). Programme for International Student Assessment (PISA).

Opp, K.-D. (2013). *Methodologie der Sozialwissenschaften*. VS Verlag für Sozialw. Verfügbar unter https://www.ebook.de/de/product/34418358/karl_dieter_opp_methodologie_der_sozialwissenschaften.html

Schmidl, B. (2014). GRD: Geschlechtsspezifische Risikofaktoren für Depressionen. doi:10.7802/71

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680. doi:10.1126/science.103.2684.677

Stevens, S. S. (1951). Mathematics, Measurement, and Psychophysics. In *Handbook of Experimental Psychology*. Wiley.

Stuart, A. (2010). *Kendall's Advanced Theory of Statistics, Set*. PAPERBACKSHOP UK IMPORT. Verfügbar unter https://www.ebook.de/de/product/16567721/alan_stuart_kendall_s_advanced_theory_of_statistics_set.html

v. Mises, R. (1919). Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, *5*(1-2), 52–99. doi:10.1007/bf01203155

Winer, B. J. (1962). *Statistical principles in experimental design*. Includes bibliography. New York: McGraw-Hill.