Inhaltsverzeichnis

1.0 Die Determinante einer Matrix

Zunächst eine kurze Wiederholung aus den Pflichtmodulen:

Eine Determinante ist eine Funktion, die einer quadratischen Matrix eine Zahl zuordnet.

Wichtig sind zunächst die Determinanten für 2 x 2 und 3 x 3 Matrizen.

Die Determinante einer 2 x 2 Matrix $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ berechnet sich nach

$$Det\ \boldsymbol{A} = a_{11}a_{22} - a_{12}a_{21}$$

Die Komponenten der Hauptdiagonalen werden miteinander multipliziert und das Produkt der Komponenten der Nebendiagonalen wird davon subtrahiert. Da das schwer verständlich ist, hier konkret:

Die Determinante einer 3 x 3 Matrix $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$ berechnet sich nach

$$Det\ \boldsymbol{A} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31+}a_{13}a_{21}a_{32} - (a_{13}a_{22}a_{31} + a_{12}a_{21}a_{33} + a_{11}a_{23}a_{32})$$

Um sich diese Rechenregel leichter zu merken, solltest du dir die einzelnen Produkte grafisch veranschaulichen. Dazu setzen wir die ersten zwei Spalten hinten an die Matrix ran:

Die ersten drei Produkte sind die Produkte aus den folgenden drei eingezeichneten Diagonalen.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} & a_{31} & a_{32} \end{pmatrix}$$

Die zweiten drei Produkte sind die Produkte aus den folgenden drei eingezeichneten Diagonalen.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} & a_{31} & a_{32} \end{pmatrix}$$

Die Subtraktion der 3 Nebendiagonalen der erweiterten Matrix von den Hauptdiagonalen nennt man die **Regel von Sarrus**.

Achtung: Die Regel von Sarrus gilt nur für 3 x 3 Matrizen!

Für den Kurs Vertiefung Wirtschaftsmathematik Du Statistik solltest du zusätzlich noch folgendes beherrschen:

Determinante einer 4x4 Matrix (Laplace Entwicklungssatz)

Zur Berechnung der Determinante wählt man eine beliebige Zeile (am besten die mit möglichst vielen Nullen) und geht nach folgendem Schema vor:

1. Schritt: Wähle das erste Element der gewählten Zeile und streiche die Zeile und Spalte dieses Elementes. Multipliziere nun das Element mit der Determinante der neuen 3x3 Matrix (die durch Streichung der Zeile und Spalte übrig geblieben ist).

2. Schritt: Es geht weiter mit der Ausgangsmatrix. Wähle das zweite Element der gewählten Zeile und streiche die Zeile und Spalte dieses Elementes. Multipliziere nun das Element mit der Determinante der neuen 3x3 Matrix.

3. Schritt: Wie 2.Schritt nur mit dem 3ten Element der gewählten Zeile.

4. Schritt: Wie 2.Schritt nur mit dem 4ten Element der gewählten Zeile.

5. Schritt: Verknüpfe die Ergebnisse aus Schritt 1-4 und nutze dabei folgendes Vorzeichenschema:

$$\begin{pmatrix} + & - & + & - & .. \\ - & + & - & + & .. \\ + & - & + & - & .. \\ - & + & - & + & .. \\ + & - & + & - & .. \end{pmatrix}$$

Je nachdem, welche Zeile man gewählt hat müssen also entsprechende Vorzeichen gewählt werden. Im Fall der ersten Zeile müsste man also rechnen: Ergebnis des ersten Schritts minus das des 2.Schritts plus das des dritten Schrittes, usw.

Beispiel:

Berechne die Determinante folgender Matrix:

$$\begin{pmatrix} 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \\ 1 & 2 & 2 & 1 \\ 0 & 1 & 2 & 2 \end{pmatrix}$$

Antwort:

Ich wähle die erste Zeile und beginne mit dem ersten Element:

$$\begin{pmatrix} [1] & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \\ 1 & 2 & 2 & 1 \\ 0 & 1 & 2 & 2 \end{pmatrix}$$

Es muss nun die Determinante der folgenden Matrix berechnet werden:

$$\begin{pmatrix} 1 & 1 & 0 \\ 2 & 2 & 1 \\ 1 & 2 & 2 \end{pmatrix}$$

Im zweiten Schritt werden die Zeilen und Spalten des zweiten Elementes der ersten Zeile gestrichen und es muss die Determinante der folgenden Matrix berechnet werden:

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 2 & 2 \end{pmatrix}$$

Nach 2 weiteren Schritten ergibt sich die Determinante der 4x4 Matrix wie folgt:

$$1 * \det\begin{pmatrix} 1 & 1 & 0 \\ 2 & 2 & 1 \\ 1 & 2 & 2 \end{pmatrix} - 2 * det\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 2 & 2 \end{pmatrix} + 0 * det\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} - 1 * det\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 2 \\ 0 & 1 & 2 \end{pmatrix}$$

Interpretation der Determinante

Der Betrag einer Determinante einer 2x2 Matrix gibt den Flächeninhalt des Parallelogramms an, das durch die Zeilenvektoren aufgespannt wird.

Der Betrag einer Determinante einer 3x3 Matrix gibt das Volumen des erweiterten Parallelogramms (Spat) an, das durch die Zeilenvektoren aufgespannt wird.

Rechenregeln für Determinanten

1) Nur quadratische Matrizen haben Determinanten

2) $det(\boldsymbol{A}^{-1}) = \frac{1}{\det(\boldsymbol{A})}$

3) $\det(\boldsymbol{A} * \boldsymbol{B}) = \det(\boldsymbol{A}) * \det(\boldsymbol{B})$

4) $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^T)$

5) Vertauscht man zwei Zeilen einer Matrix, so ändert sich das Vorzeichen der Determinante.

6) Addiert man zu einer Zeile der Matrix das vielfache einer anderen Zeile, so ändert sich die Determinante nicht.

7) Wird eine Zeile mit einer beliebigen Zahl λ multipliziert so muss auch die Determinante mit dieser Zahl multipliziert werden.

Lösung eines LGS mit Hilfe der Determinantenrechnung (Cramersche Regel)

Zu lösen sei folgendes Gleichungssystem:

$$\boldsymbol{Ax} = \boldsymbol{b}$$

Die Cramersche Regel möchte ich direkt an einem Beispiel erläutern:

Gegeben sei ein LGS der Form

$$\boldsymbol{Ax} = \boldsymbol{b}$$

mit

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

$$\boldsymbol{b} = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$$

Ausgeschrieben ergibt sich folgendes LGS:

$$1x_1 + 2x_2 + 1x_3 = 2$$

$$0 + 1x_2 + 3x_3 = 2$$

$$2x_1 + 3x_2 + 1x_3 = 2$$

Um x_1 zu berechnen, vertauscht man den x_1-Vektor mit dem b Vektor und berechnet die Determinante der neuen Matrix. Diese teilt man durch die Determinante der Matrix A. Genauso verfährt man mit x_2 und x_3.

Konkret erhält man für x_1 die Matrix:

$$\begin{pmatrix} 2 & 2 & 1 \\ 2 & 1 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

Die Determinante ist $2 + 12 + 6 - 2 - 4 - 18 = -4$

Die Determinante der Matrix A ist: $1 + 12 - 2 - 9 = 2$

Damit ist $x_1 = -2$

Aufgaben zu 1.0

Aufgabe 1.1

Berechne die Determinante zu folgender Matrix:

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 1 \\ 2 & 2 & 1 \end{pmatrix}$$

Aufgabe 1.2

Berechne die Determinante zu folgender Matrix:

$$A = \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 2 & 1 & 2 \\ 2 & 2 & 1 & 3 \\ 1 & 2 & 2 & 3 \end{pmatrix}$$

Lösungen zu 1.0

Lösung zu 1.1

Nutzung der Regel von Sarrus:

$$\det A = (1 * 2 * 1 + 2 * 1 * 2) - (1 * 1 * 2) = 4$$

Lösung zu 1.2

Hier müssen drei „Unterdeterminanten“ berechnet werden:

$$det\begin{pmatrix} 2 & 1 & 2 \\ 2 & 1 & 3 \\ 2 & 2 & 3 \end{pmatrix} = 20 - 22 = -2$$

$$det\begin{pmatrix} 0 & 1 & 2 \\ 2 & 1 & 3 \\ 1 & 2 & 3 \end{pmatrix} = 11 - 8 = 3$$

$$det\begin{pmatrix} 0 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 2 & 2 \end{pmatrix} = 6 - 10 = -4$$

$$\det A = 1 * (-2) - 2 * 3 - 1 * (-4) = -4$$

2.0 Eigenwerte und quadratische Formen

Bei sogenannten Eigenwerten geht es um folgende Fragestellung:

Für welches λ gilt:

$$\boldsymbol{Ax} = \lambda \boldsymbol{x}$$

für $\boldsymbol{x} \neq 0$

Das λ, für das die oben angegebene Gleichung gilt, heißt dann Eigenwert von A. Die Bedingung für den Eigenwert lässt sich mathematisch auch anders formulieren:

$$(\boldsymbol{A} - \lambda \mathbf{1})x = 0$$

Diese Bedingung ist ein lineares homogenes Gleichungssystem. Gibt es für dieses Gleichungssystem eine Lösung, die nicht die triviale Lösung $x = 0$ ist, so ist die Determinante gleich Null. Durch Berechnung der Determinante von $(\boldsymbol{A} - \lambda \mathbf{1})$ erhält man das „charakteristische Polynom“. Dies verdeutlicht man am besten anhand eines

Beispiels:

Zu berechnen sind die Eigenwerte der Matrix

$$A = \begin{pmatrix} 8 & 3 \\ 3 & 0 \end{pmatrix}$$

Antwort:

Man stellt zunächst die Matrix $(\boldsymbol{A} - \lambda \mathbf{1})$ auf:

$$\begin{pmatrix} 8 & 3 \\ 3 & 0 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 8-\lambda & 3 \\ 3 & 0-\lambda \end{pmatrix}$$

Dann berechnet man die Determinante:

$$(8 - \lambda) * (-\lambda) - 9 = 0$$

$$\lambda^2 - 8\lambda - 9 = 0$$

Durch Lösung dieser Gleichung erhält man die Eigenwerte:

In diesem Fall lassen sich die Nullstellen der Funktion leicht ablesen.

$$\lambda_1 = -1$$

$$\lambda_2 = 9$$

Ermittlung der Eigenvektoren

Die zu den Eigenwerten gehörenden Eigenvektoren werden ermittelt, indem man die Eigenwerte in die Gleichung

$$(\boldsymbol{A} - \lambda \mathbf{1})\boldsymbol{x} = 0$$

einsetzt. Für den ersten Eigenvektor erhält man:

$$\begin{pmatrix} 8-\lambda_1 & 3 \\ 3 & -\lambda_1 \end{pmatrix}\boldsymbol{x} = \begin{pmatrix} 9 & 3 \\ 3 & 1 \end{pmatrix}\boldsymbol{x} = 0$$

Dies ist ein Gleichungssystem mit 2 Unbekannten und 2 Gleichungen.

$$9x_1 + 3x_2 = 0$$

$$3x_1 + x_2 = 0$$

Aus Gleichung I folgt:

$$3x_1 = -x_2$$

Ein möglicher Eigenvektor wäre also $\boldsymbol{x}^1 = \begin{pmatrix} 1 \\ -3 \end{pmatrix}$.

Entsprechend berechnet man den Eigenvektor zum zweiten Eigenwert.

Achtung: Wie Du siehst, kann es zu jedem Eigenwert mehrere Eigenvektoren geben.

Quadratische Formen

Eine Matrix **Q** hat quadratische Form, wenn sie wie folgt darstellbar ist:

$$\boldsymbol{Q} = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}$$

,wobei **A** eine symmetrische Matrix ist.

Beispiel:

Bestimme die quadratische Form der Matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$$

Antwort:

$$Q = (x_1 \quad x_2) * \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix} * \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$= \quad (x_1 + 2x_2 \quad 2x_1 + 2x_2) * \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 + 2x_1x_2 + 2x_1x_2 + 2x_2^2$$

Definitheit

Bei der Definitheit geht es um die Frage, ob die quadratische Form für alle x nur positive oder nur negative Werte annimmt.

Eine quadratische Form heißt:

-positiv definit, wenn für alle x gilt: $Q > 0$,

-positiv semidefinit, wenn für alle x gilt: $Q \geq 0$,

-negativ definit, wenn für alle x gilt: $Q < 0$,

-negativ semidefinit, wenn für alle x gilt: $Q \leq 0$,

-indefinit, in allen anderen Fällen.

Für Matrizen gilt: Eine Matrix ist

-positiv definit, wenn alle Eigenwerte positiv sind,

-positiv semidefinit, wenn keiner der Eigenwerte negativ und mindestens einer positiv ist,

-negativ definit, wenn alle Eigenwerte negativ sind,

-negativ semidefinit, wenn keiner der Eigenwerte positiv und mindestens einer negativ ist,

-indefinit, in allen anderen Fällen.

Aufgaben 2.0

Aufgabe 2.1

Berechne die Eigenwerte und Eigenvektoren zu folgender Matrix:

$$A = \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix}$$

Aufgabe 2.2

Bestimme die quadratische Form der folgenden Matrix:

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 1 \\ 2 & 2 & 1 \end{pmatrix}$$

Lösungen zu 2.0

Lösung zu 2.1

$$A - \lambda\mathbf{1} = \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} = \begin{pmatrix} 1-\lambda & -1 \\ 2 & 4-\lambda \end{pmatrix}$$

Berechnung der Determinante:

$$\det\begin{pmatrix} 1-\lambda & -1 \\ 2 & 4-\lambda \end{pmatrix} = (1-\lambda) * (4-\lambda) + 2$$

Nullsetzen:

$$(1-\lambda) * (4-\lambda) + 2 = 0$$

$$\lambda^2 - 5\lambda + 6 = 0$$

$$\lambda_1 = 3$$

$$\lambda_2 = 2$$

Eigenvektor zu $\lambda_1 = 3$:

$$\begin{pmatrix} 1-3 & -1 \\ 2 & 4-3 \end{pmatrix} * \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

Als Gleichungssystem geschrieben:

$$-2x_1 - 1x_2 = 0$$

$$2x_1 + 1x_2 = 0$$

Daraus folgt:

$$x_2 = -2x_1$$

Ein Eigenvektor von $\lambda_1 = 3$ wäre also

$$x_{\lambda_1} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

Eigenvektor zu $\lambda_2 = 2$:

$$\begin{pmatrix} 1-2 & -1 \\ 2 & 4-2 \end{pmatrix} * \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

Als Gleichungssystem geschrieben:

$$-1x_1 - 1x_2 = 0$$

$$2x_1 + 2x_2 = 0$$

Daraus folgt:

$$x_2 = -x_1$$

Ein Eigenvektor von $\lambda_1 = 3$ wäre also

$$x_{\lambda_1} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Lösung zu 2.2

$$Q = (x_1 \quad x_2 \quad x_3) * \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 1 \\ 2 & 2 & 1 \end{pmatrix} * \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$(x_1 + 2x_2 \quad 2x_2 + x_3 \quad 2x_1 + 2x_2 + x_3) * \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$= x_1^2 + 2x_2x_1 + 2x_2^2 + x_2x_3 + 2x_1x_3 + 2x_2x_3 + x_3^2$$

3.0 Lineare Planungsrechnung

Dieses Kapitel beschäftigt sich mit der linearen Optimierung. Bei der linearen Optimierung wird eine Zielfunktion unter Einhaltung einer oder mehrerer Nebenbedingungen optimiert. Mathematisch wird also eine Funktion $z = f(x_1, x_2, \dots, x_n)$ gegeben und zu den x_i werden Nebenbedingungen aufgestellt. Da es sich um lineare Optimierung handelt, ist $f(x_1, x_2, \dots, x_n)$ eine Funktion erster Ordnung.

Beispiel für ein lineares Optimierungsproblem:

Ein Unternehmen bietet zwei verschiedene Produkte mit folgenden Kennzahlen an.

	Produkt 1	Produkt 2
Erlös pro Stück	8	22
Gesamtkosten pro Stück	3	8
Arbeitseinsatz in h	2	4
Kapitaleinsatz in €	4	1

Die Gesamtkapazitäten für Arbeit und Kapital sind folgendermaßen beschränkt:

- Arbeit max. 7.000h

- Kapital max. 6.000€

Ziel der linearen Optimierung ist es nun, den Gewinn zu maximieren und dabei die Faktorbeschränkungen einzuhalten.

Die Zielfunktion ist $max\ z\ =\ 5x_1 + 14x_2$

(5 und 14 sind der Gewinn, der pro verkauftem Produkt 1 bzw. Produkt 2 anfällt)

Die Nebenbedingungen sind:

$$2x_1 + 4x_2 \leq 7.000$$

(Beschränkung für den Arbeitseinsatz)

und

$$4x_1 + x_2 \leq 6.000$$

(Beschränkung für den Kapitaleinsatz)

Außerdem wird gefordert, dass der Absatz nicht negativ werden kann:

$$x_1 \geq 0$$

$$x_2 \geq 0$$

Um die $\leq$ Bedingungen in = Bedingungen zu überführen, werden nun sogenannte Schlupfvariablen eingeführt.

Aus

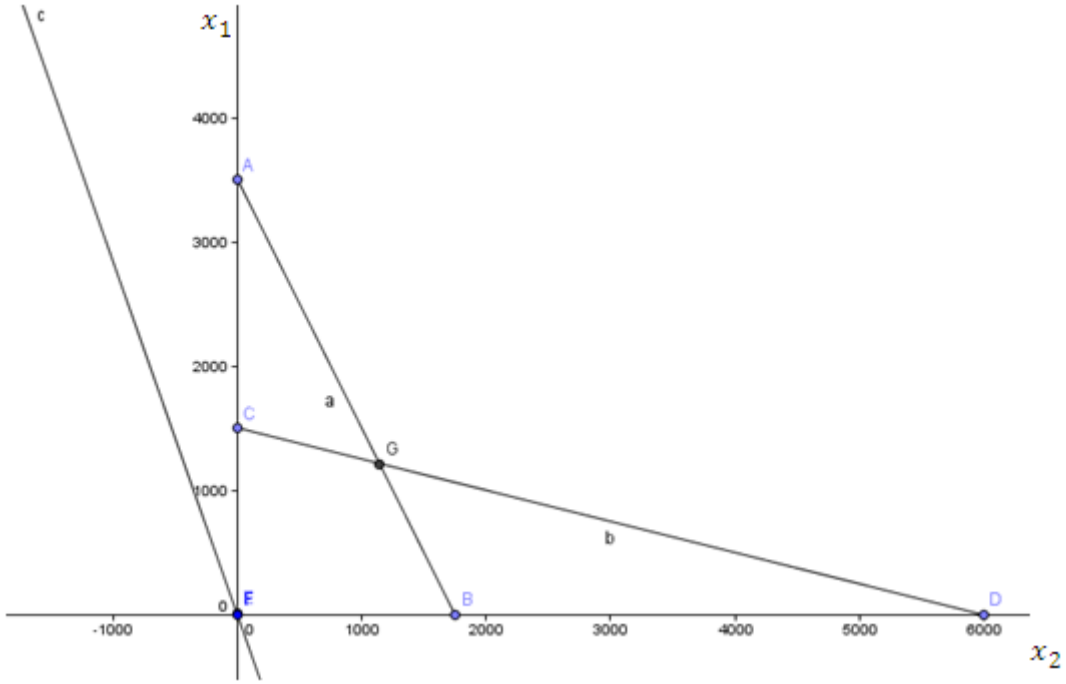
$$ax_1 + bx_2 \leq c$$

wird so

$ax_1 + bx_2 + s_1 = c$

Grafische Lösung eines linearen Optimierungsproblems (LOP)

Für 2 unabhängige Variablen x_1 und x_2 lassen sich Zielfunktion und Nebenbedingungen einfach zeichnen (Gleichungen nach x_1 umstellen):



Die erste Restriktion ist die Strecke a, die zweite Restriktion ist die Strecke b und die Gerade durch den Nullpunkt ist die Zielfunktion. Diese gilt es soweit wie möglich nach oben/rechts zu verschieben. Der Definitionsbereich wird durch die Punkte 0,B,G,C begrenzt. Es ist zu erkennen, dass die Zielfunktion im Punkt B maximal wird.

Achtung: Bei Minimierungsproblemen muss die Gerade der Zielfunktion Richtung Nullpunkt verschoben werden.

Lösen eines LOP mit dem Simplex-Algorithmus

Der sogenannte Simplex - Algorithmus löst das LOP in Tabellenform. Wie ein LOP in Tabellenform überführt und gelöst wird, versteht man am besten durch ein **Beispiel**:

Folgendes LOP soll mit dem Simplex Algorithmus gelöst werden:

$$\max x_0 = x_1 + 2x_2$$

Unter den Nebenbedingungen (u.d.N.)

$$-x_1 + x_2 \leq 1$$

$$x_1 \leq 6$$

$$x_2 \leq 5$$

$$x_1, x_2 \geq 0$$

Zunächst muss für jede ≤ Gleichung eine Schlupfvariable eingeführt werden. Wir nennen diese 3 Variablen x_3, x_4, x_5.

$$-x_1 + x_2 + x_3 = 1$$

$$x_1 + x_4 = 6$$

$$x_2 + x_5 = 5$$

$$x_1, x_2 \geq 0$$

Das Simplex Tableau hat nun folgendes Schema:

x_0	x_1	x_2	x_3	x_4	x_5	b
1	−1	−2	0	0	0	0
0	−1	1	1	0	0	1
0	1	0	0	1	0	6
0	0	1	0	0	1	5

Zur Notation: Wie du siehst, gibt es in dem Tableau 4 Einheitsvektoren. Die zugehörigen Variablen nennt man Basisvariablen und sagt auch, sie sind „in der Basis“. In der Ausgangslösung ist der Zielfunktionswert Null und die beiden Schlupfvariablen sind in der Basis. Die Nichtbasisvariablen x_1, x_2 sind damit minimal, woraus auch der minimale Zielfunktionswert resultiert. Aus dem Tableau kann direkt die Basislösung abgelesen werden. Alle Nichtbasisvariablen sind 0 und die Werte der Basisvariablen können direkt auf der rechten Seite abgelesen werden (x_5 ist 5, x_4 ist 6, der Zielfunktionswert x_0 ist 0 und x_3 ist 1.)

In die erste Zeile (Zielfunktionszeile) wird immer die 1 für x_0 eingetragen und die Koeffizienten aus der Zielfunktion werden mit negativen Werten in die erste Zeile eingetragen. Alle anderen Koeffizienten werden direkt in das Tableau übernommen.

Ist das Ausgangstableau aufgestellt, so kann der Simplex-Algorithmus starten:

1. Schritt: Wähle als Pivotspalte die Spalte mit dem größten negativen Wert in der Zielfunktionszeile. Sind alle Werte positiv, so ist die optimale Lösung gefunden.

2. Schritt: Ermittle die Quotienten aus den Werten der rechten Seite(b) und den entsprechenden Elementen der Pivotspalte. Wähle das Element als Pivotelement, für das der Quotient am kleinsten, aber positiv ist.

3. Schritt: Erzeuge an der Stelle des Pivotelementes eine 1 und darüber und darunter Nullen. Hierbei gelten alle Rechenregeln, die du aus dem Gauß'schen Eliminationsverfahren kennst. Weiter bei 1)

Beispiel:

1. Schritt: Als größtes negatives Element der Zielfunktionszeile identifizieren wir -2. Die Spalte x_2 wird Pivotspalte.

2. Schritt: Die Quotienten aus der rechten Seite(b) und den entsprechenden Elementen der Pivotspalte betragen: 5/1, 6/0, 1/1. Das Minimum ist 1.Damit wird die 1 in der Spalte x_2 und der dritten Zeile zum Pivotelement.

3. Schritt: Wie aus dem Gauß'schen Eliminationsverfahren bekannt, erzeugen wir nun die 1 an der Position des Pivotelementes (ist bereits 1) und ansonsten Nullen in der Pivotspalte. Wir erhalten nach einem Simplex Schritt das Tableau:

x_0	x_1	x_2	x_3	x_4	x_5	b
1	−3	0	2	0	0	2
0	−1	1	1	0	0	1
0	1	0	0	1	0	6
0	1	0	−1	0	1	4

Durch den ersten Simplex Schritt ist der Zielfunktionswert von 0 auf 2 angestiegen. Da aber in der Zielfunktionszeile noch ein negativer Wert steht, sind wir nicht fertig und führen einen weiteren Simplex Schritt durch.

1. Schritt: Als größtes negatives Element wird -3 identifiziert und x_1 wird zur Pivotspalte.

2. Schritt: Die Quotienten aus der rechten Seite (b) und den entsprechenden Elementen der Pivotspalte betragen: 4/1, 6/1, 1/-1. Der kleinste positive Quotient ist 4 und daher wird die 1, also das letzte Elemente der Pivotspalte, zum Pivotelement.

3. Schritt: Wir erzeugen die 1 an der Stelle des Pivotelementes (ist bereits 1) und ansonsten Nullen in der Pivotspalte. Wir erhalten nach einem Simplex Schritt das Tableau:

x_0	x_1	x_2	x_3	x_4	x_5	b
1	0	0	−1	0	3	14
0	0	1	0	0	1	5
0	0	0	1	1	−1	2
0	1	0	−1	0	1	4

Da in der Zielfunktionszeile noch immer ein negatives Element steht, führen wir noch einen Simplex Schritt durch und erhalten

x_0	x_1	x_2	x_3	x_4	x_5	b
1	0	0	0	1	2	16
0	0	1	0	0	1	5
0	0	0	1	1	−1	2
0	1	0	0	1	0	6

Aus diesem Tableau lassen sich die Lösungen ablesen. Die zum Einheitsvektor gehörenden Variablen sind in der Basis. Die Lösung lautet: $x_1 = 6$, $x_2 = 5$, $x_3 = 2$, $x_0 = 16$

Übungsaufgaben zu 3.0

Aufgabe 3.1

Zur Produktion der Güter x_1 und x_2 werden drei Rohstoffe benötigt. Der Bedarf wird durch folgende Rohstoffbedarfsmatrix angegeben:

$$\boldsymbol{R} = \begin{pmatrix} 1 & 4 \\ 2 & 1 \\ 5 & 2 \end{pmatrix}$$

Die maximal zur Verfügung stehenden Rohstoffmengen sind:

$$\boldsymbol{v} = \begin{pmatrix} 100 \\ 80 \\ 120 \end{pmatrix}$$

Das Gut x_1 und x_2 werden jeweils zu 1 Euro pro Stück verkauft.

Stelle das Problem grafisch dar.

Lösungen zu 3.0

Lösung zu 3.1

Die Zielfunktion ist natürlich, den Gewinn zu maximieren.

$$\max z = 1x_1 + 1x_2$$

Aus der ***R*** und ***v*** ergeben sich die Nebenbedingungen:

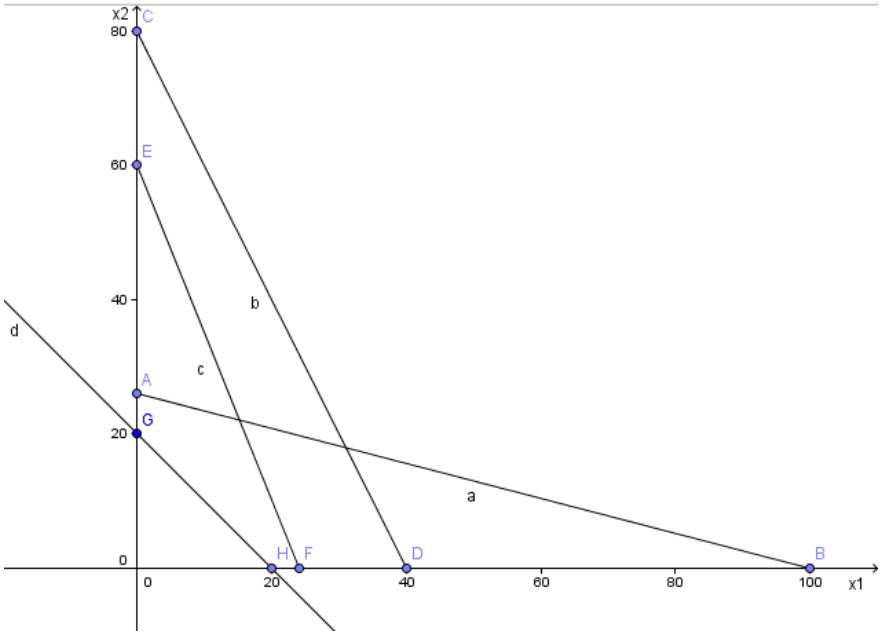
$$x_1 + 4x_2 \leq 100$$

$$2x_1 + 1x_2 \leq 80$$

$$5x_1 + 2x_2 \leq 120$$

Außerdem müssen x_1 und x_2 ≥ sein.

Grafisch erhält man:



Das Optimum ist im Schnittpunkt der Strecken a und c erfüllt.

Die Strecke b ergibt sich beispielsweise aus

$$2x_1 + 1x_2 \leq 80$$

Für $x_1 = 0$ erhält man $x_2 = 80$ und für $x_2 = 0$ erhält man $x_1 = 40$. Da es sich um lineare Gleichungen handelt, werden die Punkte mit einer Geraden verbunden. Die Zielfunktion ist die Gerade d und muss nach rechts oben verschoben werden, bis eine der Nebenbedingungen erreicht wird. Alle Nebenbedingungen sind $\leq$ Bedingungen, daher liegt der zulässige Bereich links unter den Nebenbedingungen.

4.0 Funktionen mehrerer Variabler

Aus der Schule kennst du wahrscheinlich nur Funktionen, die von einer Variablen abhängen. Funktionen können aber auch von mehreren Variablen abhängen.

Beispiel: Der Absatz dieses Skriptes ist davon abhängig, wie viele Studenten das Modul belegen und vom Preis, den ich dafür verlange.

Mathematisch schreibt man:

$$f(x_1, x_2)$$

für den Fall von zwei unabhängigen Variablen. Dabei ist x_1 beispielsweise die Anzahl der Studenten, die das Modul belegen und x_2 der Preis, den ich für das Skript verlange. Die abhängige Variable ist also gleichzeitig von zwei unabhängigen variablen abhängig.

Ein Beispiel für so eine Funktion wäre:

$$f(x_1, x_2) = 2x_1 + 3x_2 + 2$$

Im folgenden werden wir einige Eigenschaften dieser Funktionen mehrerer Variabler besprechen.

Homogenität

Ich beginne mit der mathematischen Definition der Homogenität:

Eine Funktion ist homogen vom Grad r, wenn für jede reelle Zahl $\lambda > 0$ gilt:

$$f(\lambda x_1, \lambda x_2, \lambda x_3, \dots, \lambda x_n) = \lambda^r * f(x_1, x_2, x_3, \dots, x_n)$$

Multipliziert man also alle Variablen mit λ, dann nimmt der Funktionswert den $\lambda^r -$ fachen Wert an.

Du kannst die Homogenität nach den Formeln des Skriptes berechnen. Bei den eher einfachen Funktionen, die in Klausuren geprüft werden, kannst du die Homogenität aber auch direkt ablesen. Dabei gilt:

- Sind x_1 und x_2 durch Multiplikation verbunden, so sind sie homogen. Der Grad der Homogenität berechnet sich aus der Summe der beiden Exponenten.

- Sind x_1 und x_2 durch Addition oder Subtraktion verbunden, so sind sie homogen, wenn sie denselben Exponenten haben, ansonsten nicht. Die Homogenität entspricht den 2 Exponenten (nicht der Summe dieser).

-Sollte die Homogenität nicht so leicht abgelesen werden können, so muss diese durch Umformung ermittelt werden.

Beispiel:

Gegeben sei die Funktion:

$$f(x_1, x_2) = 2x_1 + 4x_1x_2$$

Bestimme die Homogenität.

Antwort: Die Funktion ist nicht homogen, da die beiden Summanden verschiedene (einzelne) Homogenitäten haben ($2x_1$ ist homogen vom Grad 1 und $4x_1x_2$ ist homogen vom Grad 2).

Partielle Ableitung

Bei der partiellen Ableitung wird ermittelt, wie sich der Zielfunktionswert ändert, wenn nur eine Variable minimal erhöht wird und alle anderen Variablen konstant gehalten werden. Daher müssen die anderen Variablen auch als Konstanten behandelt werden!

Beispiel:

Die erste partielle Ableitung der Funktion

$$f(x_1, x_2) = 2x_1 + 4x_1^2x_2$$

nach $x_1$ist:

$$f_{x_1}(x_1, x_2) = 2 + 8x_1x_2$$

x_2 wird hier also als Konstante behandelt.

Leitet man die erste Ableitung nach x_1 bzw. x_2 nach x_2 bzw. $x_1$ab, so erhält man die Kreuzableitung. Dabei ist es egal, nach welcher Variable man zuerst ableitet.

Beispiel:

$$f_{x_1,x_2}(x_1, x_2) = 8x_1$$

Gradient

Der Gradient ist der Vektor der beiden ersten Ableitungen einer Funktion mit 2 unabhängigen Variablen.

Mathematisch schreibt man:

$$Grad\ f(x_1, x_2) = \begin{matrix} f_{x_1}(x_1, x_2) \\ f_{x_2}(x_1, x_2) \end{matrix}$$

Interpretation: Die Richtung des Gradientenvektors ist die Richtung des steilsten Anstiegs der Funktion und die Länge des Gradientenvektors ist ein Maß für die Stärke des Anstiegs der Funktion in Richtung des Gradienten.

Hesse-Matrix

Die Hesse-Matrix besteht aus den zweiten partiellen Ableitungen.

Mathematisch schreibt man:

$$H(x_1, x_2) = \begin{pmatrix} f_{x_1,x_1}(x_1, x_2) & f_{x_1,x_2}(x_1, x_2) \\ f_{x_2,x_1}(x_1, x_2) & f_{x_2,x_2}(x_1, x_2) \end{pmatrix}$$

Mit Hilfe der Hesse-Matrix kann man Aussagen über die Konvexität/Konkavität der Funktion machen:

-Ist die Hesse-Matrix **positiv semidefinit**, so ist die Funktion konvex.

- Ist die Hesse-Matrix **streng positiv semidefinit**, so ist die Funktion streng konvex.

- Ist die Hesse-Matrix **negativ semidefinit**, so ist die Funktion konkav.

- Ist die Hesse-Matrix **streng negativ semidefinit**, so ist die Funktion streng konkav.

-In allen anderen Fällen kann keine Aussage über das Krümmungsverhalten gemacht werden (die Matrix ist **indefinit**).

Die Hesse-Matrix wird auch genutzt, um Aussagen über die kritischen Punkte der Funktion zu machen. Dabei gilt:

-Ist die Determinante der Hesse-Matrix an einem kritischen Punkt positiv, so liegt ein Extremwert vor. Ist das Element $a_{11} > 0$, so handelt es sich um ein Minimum. Ist das Element $a_{11} < 0$, so handelt es sich um ein Maximum.

-Ist Hesse-Matrix an einem kritischen Punkt negativ, so liegt ein Sattelpunkt vor.

Hauptunterdeterminantenkriterium

Die Definitheit einer Matrix läßt sich auch anhand der Hauptunterdeterminanten (Minoren) bestimmen:

Sind alle Hauptunterdeterminanten positiv, so ist die Matrix positiv definit.

Alternieren die Vorzeichen der Hauptunterdeterminanten, so ist die Matrix negativ definit.

Die Semidefinitheit kann über die Hauptunterdeterminanten nicht bestimmt werden.

Das totale Differential

Beim totalen Differential wir die Änderung des Funktionswertes bei einer infinitisimal kleinen Erhöhung beider bzw. aller unabhängigen Variablen ermittelt. Diese ist die Summer der beiden/aller partiellen Ableitungen.

Mathematisch schreibt man:

$$df(x_1, x_2) = f'_{x1} * dx_1 + f'_{x2} * dx_2$$

Beispiel:

$$y = 3x + 4z^2$$

Zur Bestimmung des totalen Differentials bildet man zunächst beide Ableitungen:

$$\frac{dy}{dx} = 3$$

$$\frac{dy}{dz} = 8z$$

Das totale Differential ist dann:

$$dy = 3 * dx + 8z * dz$$

$$df(x_1, x_2) = f'_{x1} * dx_1 + f'_{x2} * dx_2$$

Isoquanten

An dieser Stelle empfiehlt es sich das Thema Isoquanten und Grenzrate der Substitution aus Einführung in die Wirtschaftswissenschaft zu widerholen. Ich habe daher den entsprechenden Abschnitt aus meinem Skript einkopiert:

Bei substitutionalen Produktionsfunktionen mit 2 Produktionsfaktoren kann man den Einsatz des einen Faktors verringern und den Einsatz des anderen Faktors erhöhen, ohne dass die Ausbringungsmenge sich ändert. Eine gegebene Ausbringungsmenge kann also mit verschiedenen Faktorkombinationen produziert werden. Grafisch kann dieses Verhältnis in einem Diagramm mit den beiden Produktionsfaktoren r_1 und r_2 dargestellt werden. Die Linie, die alle möglichen Faktorkombinationen für eine gegebene Ausbringungsmenge darstellt, nennt man Isoquante.

Mathematisch erhält man die Isoquantengleichung aus der Produktionsfunktion, indem man die Produktionsfunktion nach $r_2$auflöst.

Beispiel:

Für die Produktionsfunktion

$$M = 2r_1 * \sqrt{r_2}$$

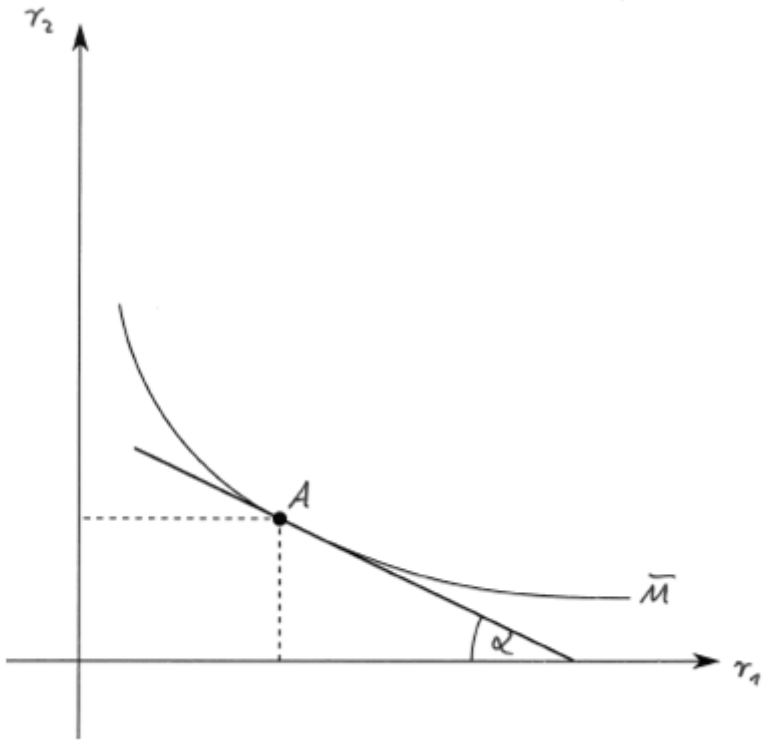
ergibt sich die Isoquantengleichung:

$$r_2 = \left(\frac{M}{2r_1}\right)^2$$

Die Durchschnittsrate der technischen Substitution bezüglich zweier Punkte auf der Isoquante erhält man, indem man die Änderung des einen Faktors durch die Änderung des anderen Faktors teilt.

Grenzrate der Substitution

Wie man an der Isoquante sieht, ist das Tauschverhältnis zwischen den beiden Einsatzfaktoren nicht konstant, sondern abhängig von dem jeweiligen Faktoreinsatzverhältnis. Die Grenzrate der Substitution gibt für jeden Punkt der Isoquante an, welche Menge des einen Faktors notwendig ist, um eine sehr kleine Einheit des anderen Faktors zu ersetzen. Grafisch wird dieses Tauschverhältnis für einen bestimmten Punkt auf der Isoquante durch die Tangente durch diesen Punkt dargestellt.



Mathematisch beträgt die Grenzrate der Substitution:

$$GRS_{2,1} = \frac{dr_2}{dr_1} = -\frac{\frac{\partial M}{\partial r_1}}{\frac{\partial M}{\partial r_2}}$$

,wobei $\frac{\partial M}{\partial r_1}$ die erste Ableitung der Gesamtertragskurve nach dem Faktor r_1 darstellt.

Man kann die Grenzrate der Substitution ermitteln, indem man die Produktionsfunktion nach r_2 umstellt und dann nach r_1 ableitet oder indem man das totale Differential der Produktionsfunktion bildet und nach $\frac{dr_2}{dr_1}$ umstellt. Beide Herleitungen können sehr gut in der Klausur vorkommen.

Grenzrate der Substitution - Herleitung Variante 1 (totales Differential):

Gegeben sei eine Produktionsfunktion der Form $M = f(r_1, r_2)$

Das totale Differential ist dann:

$$dM = \frac{\partial M}{\partial r_1} dr_1 + \frac{\partial M}{\partial r_2} dr_2$$

Da die Grenzrate der Substitution für eine konstante Ausbringungsmenge berechnet wird, gilt $dM = 0$.

$$dM = \frac{\partial M}{\partial r_1} dr_1 + \frac{\partial M}{\partial r_2} dr_2 = 0$$

Aufgelöst nach $\frac{dr_2}{dr_1}$ erhält man:

$$\frac{dr_2}{dr_1} = -\frac{\frac{\partial M}{\partial r_1}}{\frac{\partial M}{\partial r_2}}$$

Grenzrate der Substitution - Herleitung Variante 2 (Isoquantengleichung):

Die zweite Möglichkeit, die Grenzrate der Substitution herzuleiten besteht darin die Produktionsfunktion nach r_2 umzustellen (Isoquantengleichung) und nach r_1 abzuleiten:

Beispiel:

Gegeben sei die substitutionale Produktionsfunktion:

$$M = 2r_1 r_2$$

Zur Ermittlung der Grenzrate der Substitution löst man diese Gleichung nach r_2 auf:

$$r_2 = \frac{M}{2r_1}$$

und leitet nach r_1 ab:

$$\frac{dr_2}{dr_1} = -\frac{M}{2r_1^2} = -\frac{2r_1 r_2}{2r_1^2} = -\frac{r_2}{r_1}$$

Dies entspricht

$$\frac{dr_2}{dr_1} = -\frac{\frac{\partial M}{\partial r_1}}{\frac{\partial M}{\partial r_2}}$$

denn

$$-\frac{\frac{\partial M}{\partial r_1}}{\frac{\partial M}{\partial r_2}} = -\frac{2r_2}{2r_1} = -\frac{r_2}{r_1}$$

Änderungsraten und Elastizitäten

Partielle Änderungsrate

Die partielle Änderungsrate gibt die relative Änderung des Funktionswertes auf eine Änderung der einzelnen unabhängigen Variablen an.

Mathematisch schreibt man:

$$A_{x_1} f(x_1, x_2) = \frac{f'_{x_1}(x_1, x_2)}{f(x_1, x_2)}$$

Also einfach nur die partielle Ableitung geteilt durch die Funktion.

Elastizität

Für die Elastizität gilt:

$$E_{x_1} f(x_1, x_2) = x_{1*} A_{x_1} f(x_1, x_2) = x_1 \frac{f'_{x_1}(x_1, x_2)}{f(x_1, x_2)}$$

Extrema bei Funktionen mehrerer Variabler

Die Extrema von Funktionen mehrerer Variabler werden wie folgt ermittelt:

1) Es werden die partiellen Ableitungen erster und zweiter Ordnung berechnet.

2) Es werden die Nullstellen der partiellen Ableitungen erster Ordnung ermittelt. Hierzu muss ein Gleichungssystem aufgestellt und aufgelöst werden.

3) Es wird die Hesse-Matrix aufgestellt, die Nullstellen werden eingesetzt und die Determinante berechnet.

Es gelten nun folgende Bedingungen:

- Wenn $detH > 0$, dann liegt ein Extremwert vor. Ist das Element $a_{11} > 0$, dann Minimum, ist $a_{11} < 0$, dann Maximum.

- Wenn $detH < 0$, dann liegt ein Sattelpunkt vor.

- Wenn $detH = 0$, dann ist keine Aussage möglich.

Beispiel:

Zu ermitteln sind die Extremstellen der folgenden Funktion:

$$f(x_1, x_2) = \frac{1}{4}x_1^3 - x_1 + \frac{1}{2}x_1x_2$$

Antwort:

Zunächst bildest du die partiellen Ableitungen:

$$f'_{x_1} = \frac{3}{4}x_1^2 - 1 + \frac{1}{2}x_2$$

$$f''_{x_1,x_1} = \frac{3}{2}x_1$$

$$f'_{x_2} = \frac{1}{2}x_1$$

$$f''_{x_2,x_2} = 0$$

$$f''_{x_1,x_2} = \frac{1}{2}$$

Die Nullstellen der ersten partiellen Ableitungen sind:

$$\frac{3}{4}x_1^2 - 1 + \frac{1}{2}x_2 = 0$$

$$\frac{1}{2}x_1 = 0$$

Es folgt:

$$x_1 = 0$$

$$x_2 = 2$$

Aufstellen der Hesse-Matrix:

$$H(x_1, x_2) = \begin{pmatrix} \frac{3}{2}x_1 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$$

Für $x_1 = 0$ und $x_2 = 2$ gilt

$$detH = -\frac{1}{4}$$

Daher liegt an der Stelle

$$x_1 = 0$$

$$x_2 = 2$$

ein Sattelpunkt vor.

Extremwerte unter Nebenbedingungen

Bei der Berechnung von Extremwerten unter Nebenbedingungen ist eine Zielfunktion unter Einhaltung von verschieden vielen Nebenbedingungen zu maximieren/minimieren. Dieses Modell ist aus der linearen Optimierung bekannt.

Zur Berechnung von Extremwerten unter Nebenbedingungen gibt es zwei Methoden:

- die Substitutionsmethode

- die Lagrangemethode

Zur Substitutionsmethode:

Es wird zunächst eine Nebenbedingung nach einer Variablen aufgelöst und diese dann in die Zielfunktion eingesetzt. In Klausuraufgaben erhält man meist eine Funktion in Abhängigkeit einer Variablen und kann diese auf Extremwerte untersuchen.

Beispiel:

Es sollen die Extremwerte der Funktion

$$f(x_1, x_2) = -x_1^2 - 2x_2^2 + 20$$

unter der Nebenbedingung

$$x_1 + x_2 = 4$$

ermittelt werden.

Lösung:

Auflösung der Nebenbedingung nach einer Variablen:

$$x_1 = -x_2 + 4$$

Einsetzen in die Zielfunktion

$$f(x_1, x_2) = -(-x_2 + 4)^2 - 2x_2^2 + 20$$

Zu dieser Funktion können nun leicht die Extremstellen berechnet werden

$(x_2 = 1{,}33,\ woraus\ x_1 = 2{,}67\ folgt)$.

Zur Lagrangemethode:

In der Nebenbedingung werden zunächst alle Terme auf die linke Seite gebracht, sodass auf der anderen Seite der Gleichung die Null steht. Nun wird die Lagrangefunktion aufgestellt (Zielfunktion$+\lambda *$ „linke Seite der umgeformten Nebenbedingung“).

Nun müssen noch die ersten partiellen Ableitungen (auch nach λ!) gebildet werden und gleich Null gesetzt werden.

Beispiel:

Es sollen die Extremwerte der Funktion

$$f(x_1, x_2) = -x_1^2 - 2x_2^2 + 20$$

unter der Nebenbedingung

$$x_1 + x_2 = 4$$

ermittelt werden.

Lösung

Umformung der Nebenbedingung:

$$x_1 + x_2 - 4 = 0$$

Aufstellen der Lagrangefunktion:

$$L = f(x_1, x_2, \lambda) = -x_1^2 - 2x_2^2 + 20 - \lambda(x_1 + x_2 - 4)$$

Bilden der partiellen Ableitungen:

$$f'_{x_1} = -2x_1 - \lambda$$

$$f'_{x_2} = -4x_2 - \lambda$$

$$f'_{\lambda} = x_1 + x_2 - 4$$

Ermitteln der Nullstellen:

$$-2x_1 - \lambda = 0$$

$$-4x_2 - \lambda = 0$$

$$x_1 + x_2 - 4 = 0$$

Aus der ersten Gleichung folgt

$$\lambda = -2x_1$$

eingesetzt in Gleichung 2:

$$-4x_2 + 2x_1 = 0$$

woraus folgt:

$$x_1 = 2x_2$$

Eingesetzt in Gleichung 3 folgt:

$$x_2 = 1{,}33$$

und somit

$$x_1 = 2{,}67$$

Aufgaben zu 4.0

Aufgabe 4.1

a) Berechne die partiellen Ableitungen der 1. Und 2. Ordnung für folgende Funktion:

$$f(x_1, x_2) = x_1 x_2 + x_1^2 + \frac{1}{x_2^1}$$

b) Berechne den Wert des Gradienten an der Stelle (1,2).

Aufgabe 4.2

Bilde die Kreuzableitung zu der Funktion

$$f(x_1, x_2) = x_1 + \frac{x_1^2}{x_2}$$

Aufgabe 4.3

Berechne das totale Differential zu der folgenden Funktion:

$$f(x_1, x_2) = x_1 + \frac{x_1^2}{x_2}$$

Aufgabe 4.4

Gegeben sei die folgende Funktion:

$$f(x_1, x_2) = 2x_1^2 - x_1 x_2 + 2x_2^2$$

Ist die Funktion konvex?

Aufgabe 4.5

Berechne die Extremstellen der folgenden Funktion:

$$f(x_1, x_2) = 2x_1^2 + 4x_2^2 - 8x_1 x_2$$

Aufgabe 4.6

Berechne den Betrag des Gradientenvektors für folgende Funktion an der Stelle (1,2):

$$f(x_1, x_2) = x_1^2 + x_2^4 - 4x_1x_2 + 5$$

Aufgabe 4.7

Berechne die Extremstellen der folgenden Produktionsfunktion mit den Faktoreinsatzmengen x_1 und x_2:

$$f(x_1, x_2) = x_1^3 - 3x_1^2 + 3x_2^3 - 36x_2 + 10$$

Aufgabe 4.8

Gegeben sei die Produktionsfunktion

$$f(x_1, x_2) = e^{x_1x_2}$$

,wobei x_1x_2 Faktoreinsatzmengen bezeichnen.

Die Nebenbedingung ist

$$x_1 + x_2 = 4$$

Berechne mit Hilfe der Substitutionsmethode das Maximum der Funktion.

Aufgabe 4.9

Berechne zu folgenden Produktionsfunktionen die Isoquantengleichung:

a) $M = 2r_1 + 4r_2$

b) $M = 2 * \sqrt{r_1 r_2}$

c) $M = \sqrt{r_1} * r_2$

Aufgabe 4.10

Skizziere eine überlinear homogene, eine unterlinear homogene, eine linear homogene und eine inhomogene Produktionsfunktion in einem Koordinatensystem. Zeichne auch das Niveau λ ein.

Aufgabe 4.11

Gib zu den folgenden Funktionen an, ob sie homogen sind und gib an, ob sie überlinear, unterlinear oder linear homogen sind.

a) $M = r_1^{\frac{1}{2}} * r_2^2$

b) $M = 2r_1^{\frac{1}{3}} * r_2^{\frac{1}{2}}$

c) $M = r_1^2 + r_2$

d) $M = 2r_1^2 + r_2^2$

Aufgabe 4.12

Leite die Grenzrate der Substitution für die substitutionale Profuktionsfunktion

$$M = 2r_1 * r_2^2$$

a) aus dem totalen Differential her.

(Totales Differential siehe Anhang!)

b) her, indem du die entsprechende Isoquante nach r_1 ableitest.

Aufgabe 4.13

Ein Unternehmen verfügt über folgende Produktionsfunktion:

$$f(x_1, x_2) = \frac{x_1^2 x_2^2}{x_1 + x_2}$$

a) Berechne die Produktionselastizität des ersten Faktors.

b) Bestimme den Homogenitätsgrad.

Lösungen zu 4.0

Lösung zu 4.1

a)

$$f'_{x_1}(x_1, x_2) = x_2 + 2x_1$$

$$f''_{x_1x_1}(x_1, x_2) = 2$$

$$f'_{x_2}(x_1, x_2) = x_1 - \frac{1}{x_2^2}$$

$$f''_{x_2x_2}(x_1, x_2) = 2\frac{1}{x_2^3}$$

b)

Der Gradiente ist der Vektor aus den beiden ersten partiellen Ableitungen.

$$grad\, f(x_1, x_2) = \left(x_2 + 2x_1 \quad x_1 - \frac{1}{x_2^2}\right)$$

$$grad\, f(1{,}2) = \left(2 + 2 \quad \frac{3}{4}\right) = (4 \quad 0{,}75)$$

Lösung zu 4.2

$$f'_{x_1}(x_1, x_2) = 1 + \frac{2x_1}{x_2}$$

$$f''_{x_1x_2}(x_1, x_2) = -\frac{2x_1}{x_2^2}$$

Lösung zu 4.3

$$df(x_1, x_2) = \left(1 + \frac{2x_1}{x_2}\right)dx_1 + \left(-\frac{x_1^2}{x_2^2}\right)dx_2$$

Lösung zu 4.4

Ermittlung der Hesse-Matrix:

$$f'_{x_1}(x_1, x_2) = 4x_1 - x_2$$

$$f''_{x_1x_1}(x_1, x_2) = 4$$

$$f'_{x_2}(x_1, x_2) = -x_1 + 4x_2$$

$$f''_{x_2x_2}(x_1, x_2) = 4$$

$$f''_{x_1x_2}(x_1, x_2) = -1$$

$$Hf(x_1, x_2) = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$

Da die Hessematrix positiv definit ist, ist $f(x_1, x_2)$ streng konvex.

Lösung zu 4.5

Zunächst ermittelt man die ersten partiellen Ableitungen:

$$f'_{x_1}(x_1, x_2) = 4x_1 - 8x_2$$

$$f'_{x_2}(x_1, x_2) = 8x_2 - 8x_1$$

Diese müssen zu Null gesetzt werden:

$$4x_1 - 8x_2 = 0$$

$$8x_2 - 8x_1 = 0$$

Das Gleichungssystem ist nicht lösbar. Es existieren keine Extremstellen.

Lösung zu 4.6

Bildung der partiellen Ableitungen erster Ordnung:

$$f'_{x_1}(x_1, x_2) = 2x_1 - 4x_2$$

$$f'_{x_2}(x_1, x_2) = 4x_2^3 - 4x_1$$

Der Gradientenvektor ist dann:

$$grad\, f(x_1, x_2) = (2x_1 - 4x_2 \quad 4x_2^3 - 4x_1)$$

Der Betrag dieses Vektors (bekannt aus Grundlagen der Wirtschaftsmathematik und Statistik) an der Stelle (1,2) ist dann:

$$|grad\, f(x_1, x_2)| = \sqrt{820}$$

Lösung zu 4.7

Zunächst bildet man die partiellen Ableitungen erster Ordnung:

$$f'_{x_1}(x_1, x_2) = 3x_1^2 - 6x_1$$

$$f'_{x_2}(x_1, x_2) = 9x_2^2 - 36$$

Diese werden zu Null gesetzt und aufgelöst.

$$3x_1^2 - 6x_1 = 0$$

$$9x_2^2 - 36 = 0$$

Aus der ersten Gleichung folgt:

$$x_1^2 = 2x_1$$

$$x_2^2 = 4$$

Die Nullstellen sind entsprechend:

$$x_{1_1} = 0$$

$$x_{1_2} = 2$$

$$x_{2_1} = -2$$

$$x_{2_2} = 2$$

Da Faktoreinsatzmengen nicht negativ sein können, erhält man die Extremstellen (0,2) und (2,2).

Lösung 4.8

Die Nebenbedingung wird nach einer Variablen umgeformt und in die Funktion eingesetzt:

$$x_1 = 4 - x_2$$

$$f(x_2) = e^{(4-x_2)x_2}$$

Nun werden die ersten beiden Ableitungen der Funktion gebildet (Kettenregel)

$$f'(x_2) = (-2x_2 + 4)e^{4x_2 - x_2^2}$$

$$f''(x_2) = -2e^{4x_2 - x_2^2} + (-2x_2 + 4)^2 e^{4x_2 - x_2^2}$$

Die Extremstellen sind die Nullstellen der ersten Ableitung.

$$(-2x_2 + 4)e^{4x_2 - x_2^2} = 0$$

Da die e-Funktion keine Nullstellen hat, gilt für die Nullstelle:

$$-2x_2 + 4 = 0$$

$$x_2 = 2$$

Anhand der zweiten Ableitung bestimmt man, ob es sich um ein Maximum oder Minimum handelt:

$$f''(2) = -2e^{4*2-2^2} + (-2 * 2 + 4)^2 e^{4*2-2^2}$$

Der Wert ist negativ und damit handelt es sich um ein Maximum.

Lösung zu 4.9

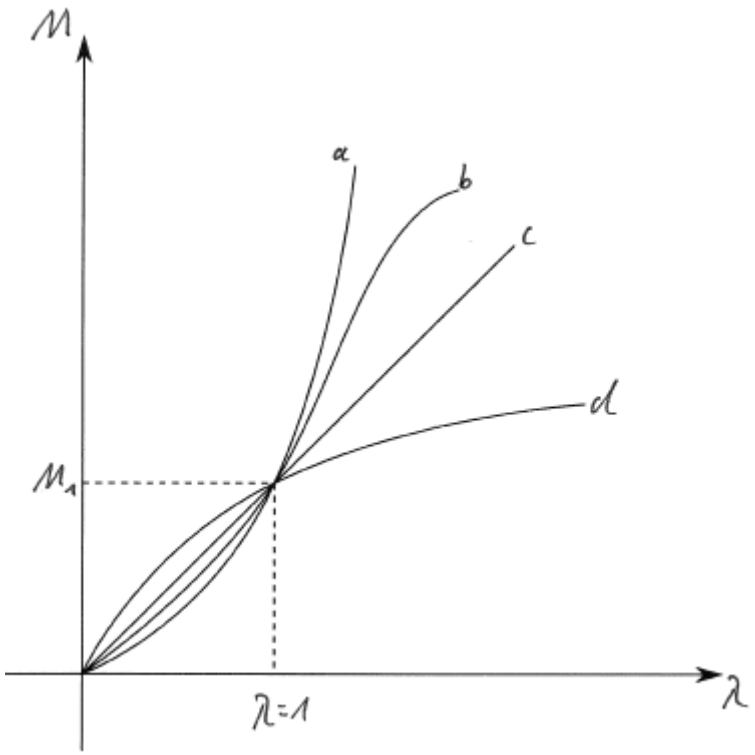
Die Produktionsfunktionen müssen nach r_2 umgestellt werden, um die Isoquantengleichung zu erhalten.

a) $r_2 = 0{,}25M - 0{,}5r_1$

b) $r_2 = \frac{M^2}{4r_1}$

c) $r_2 = \frac{M}{\sqrt{r_1}}$

Lösung zu 4.10



a) überlinear homogen

b) nicht homogen

c) linear homogen

d) unterlinear homogen

Lösung zu 4.11

Du kannst die Homogenität nach den Formeln des Skriptes berechnen. Bei den eher einfachen Formen der Produktionsfunktionen, die in Klausuren geprüft werden, kannst du die Homogenität aber auch direkt ablesen. Dabei gilt:

- Sind r_1 und r_2 durch Multiplikation verbunden, so sind sie homogen. Der Grad der Homogenität berechnet sich aus der Summe der beiden Exponenten.

- Sind r_1 und r_2 durch Addition oder Subtraktion verbunden, so sind sie homogen, wenn sie denselben Exponenten haben, ansonsten nicht. Die Homogenität entspricht den 2 Exponenten (nicht der Summe dieser).

a) homogen vom Grad 2,5, also überlinear (da größer als 1).

b) homogen vom Grad $\frac{5}{6}$, also unterlinear (da kleiner als 1).

c) inhomogen, da verschiedene Exponenten und durch Addition verbunden.

d) homogen vom Grad 2, da 2 der Exponent bei beiden Faktoren ist, also überlinear.

Lösung zu 4.12

a) Das totale Differential lautet:

$$dM = 2r_2^2 dr_1 + 4r_1 r_2 dr_2$$

Da die Ausbringungsmenge als konstant angenommen wird, gilt:

$$2r_2^2 dr_1 + 4r_1 r_2 dr_2 = 0$$

$$2r_2^2 dr_1 = -4r_1 r_2 dr_2$$

$$\frac{2r_2^2}{-4r_1 r_2} = \frac{dr_2}{dr_1}$$

$$-\frac{r_2}{2r_1} = \frac{dr_2}{dr_1}$$

b) Die Gleichung für die Isoquante lautet:

$$r_2 = \sqrt{\frac{M}{2r_1}}$$

Ableiten nach r_1:

Kettenregel: äußere Ableitung mal innere Ableitung

$$\frac{dr_2}{dr_1} = \frac{1}{2}\left(\frac{M}{2r_1}\right)^{-\frac{1}{2}} * \left(-\frac{M}{2r_1^2}\right)$$

$M = 2r_1 * r_2^2$einsetzten und Kürzen.

$$= \frac{1}{2}(r_2^2)^{-\frac{1}{2}} * \left(-\frac{2r_1 * r_2^2}{2r_1^2}\right)$$

$$= \frac{1}{2r_2} * \left(-\frac{2 * r_2^2}{2r_1}\right)$$

$$= -\frac{r_2}{2r_1}$$

Hinweis: Die Kettenregel ist für die Klausur höchstwahrscheinlich nicht relevant.

Weitere mögliche Klausuraufgaben:

Aufgabe 4.13

Ein Unternehmen verfügt über folgende Produktionsfunktion:

$$f(x_1, x_2) = \frac{x_1^2 x_2^2}{x_1 + x_2}$$

a) Berechne die Produktionselastizität des ersten Faktors.

b) Bestimme den Homogenitätsgrad.

Lösung zu 4.13

a)Für die Elastizität gilt:

$$E_{x_1} f(x_1, x_2) = x_{1*} A_{x_1} f(x_1, x_2) = x_1 \frac{f'_{x_1}(x_1, x_2)}{f(x_1, x_2)}$$

Bei der Ermittlung der ersten Ableitung muss jeweils die Quotientenregel genutzt werden:

Zur Erinnerung, kurz die **Quotientenregel:**

$$f(x) = \frac{g(x)}{h(x)}$$

$$f'(x) = \frac{g'(x) * h(x) - g(x) * h'(x)}{(h(x))^2}$$

$$f'_{x_1}(x_1, x_2) = \frac{2x_1 x_2^2 * (x_1 + x_2) - x_1^2 x_2^2}{(x_1 + x_2)^2}$$

$$E_{x_1} f(x_1, x_2) = \frac{(2x_1 x_2^2 * (x_1 + x_2) - x_1^2 x_2^2) * (x_1 + x_2)}{(x_1 + x_2)^2 * x_1^2 x_2^2}$$

Nun kann gekürzt werden:

$$E_{x_1} f(x_1, x_2) = \frac{(2(x_1 + x_2) - 1) * (x_1 + x_2)}{(x_1 + x_2)^2}$$

b)

$$f(\lambda x_1, \lambda x_2) = \frac{\lambda^2 x_1^2 \lambda^2 x_2^2}{\lambda x_1 + \lambda x_2} = \lambda^3 f(x_1, x_2)$$

5.0 Differentialgleichungen

Eine Differentialgleichung (kurz DGL) ist eine Gleichung, in der eine Ableitung der Funktion y vorkommt.

Beispiel:

$$y = 2x + y'$$

Für eine Differentialgleichung gibt es die **allgemeine Lösung.** Dies sind Funktionen, die mit ihren Ableitungen die Differentialgleichung erfüllen.

Beispiel:

Die Differentialgleichung

$$y = y'$$

hat die allgemeine Lösung

$$y = e^x + c$$

Sind zur DGL weitere Informationen gegeben (z.B. einzelne Punkte der Kurve), so spricht man von einem Anfangswertproblem und erhält eine **spezielle Lösung** (c kann eindeutig ermittelt werden).

Es gibt die verschiedenen Formen von DGLs und keinen universellen Lösungsweg für alle Typen von DGLs. Wir werden nun einige spezielle Typen von DGLs kennenlernen, für die es Lösungsverfahren gibt.

Lösungsverfahren für Differentialgleichungen für die Klausur:

Ich möchte nun kurz ein Lösungsverfahren vorstellen, welches in der Klausur angewendet werden kann, aber ansonsten kaum praktische Relevanz hat.

Bei Differentialgleichungen wird nach einer Funktion gesucht, die die DGL erfüllt. In Klausuraufgaben ist es oft so, dass die DGL gegeben ist und dann als Multiple Choice Aufgabe mehrere Funktionen gegeben sind, von denen die genannt werden soll, die die DGL erfüllt. Anstatt nun aus der DGL die Funktion zu ermitteln, die die DGL erfüllt, kann man einfach jede mögliche Lösung und deren Ableitung in die DGL einsetzen und prüfen ob die DGL erfüllt ist. Da es wesentlich einfacher ist, die Ableitung zu bilden, als eine DGL aufzulösen, kann man so viel Zeit sparen.

Das ist in der Praxis natürlich nicht möglich, da dann keine Auswahl an Lösungen gegeben ist!

Beispiel:

Es ist folgende DGL gegeben:

$$y'' + y' - 2xe^x - 3e^x = 0$$

Welche der folgenden Funktionen erfüllen die DGL?

a)

$$y = x - e^x$$

b)

$$y = x^2e^x$$

c)

$$y = xe^x$$

d)

$$y = e^{2x}$$

Zu jeder dieser Funktionen kann man relativ leicht die ersten beiden Ableitungen ermitteln und diese in die DGL einsetzen. Ich werde das für die korrekte Lösung (Lösung c)) einmal ausführlich machen:

$$y' = e^x + xe^x$$

$$y'' = e^x + e^x + xe^x$$

Eingesetzt in die DGL:

$$e^x + e^x + xe^x + e^x + xe^x - 2xe^x - 3e^x = 0$$

Korrekt!

Typen von Differentialgleichungen

- Unterscheidung nach der Ordnung

DGLs werden nach ihrer Ordnung klassifiziert. Die Ordnung einer DGL ist die höchste Ableitung von y, die in der DGL vorkommt.

Beispiel:

Die DGL

$$y = 2xy' + x^2$$

hat die Ordnung 1.

- Unterscheidung nach der Linearität

DGLs heißen linear, wenn y und deren Ableitungen nur in einfacher Potenz vorkommen.

Beispiel:

$$y = 2xy' + x^2$$

Ist linear,

$$y = \sqrt{2xy'} + x^2$$

Ist nicht linear

- Unterscheidung nach der Homogenität

Eine DGL heißt homogen, wenn sie nur Terme mit y oder deren Ableitungen enthält- es dürfen also keine Terme mit x vorkommen.

Beispiel:

$$y = 2xy' + x^2$$

ist nicht homogen, aber

$$y = 2y'$$

ist homogen.

- Partiell/gewöhnlich

Kommen in einer DGL nur x und y, sowie die Ableitungen von y vor und keine weiteren Variablen, so heißt sie **gewöhnlich**, ansonsten **partiell**.

Beispiel:

$$y = 2xy' + x^2 * z$$

ist eine partielle DGL.

- Explizit/implizit

Eine DGL, die nach y‘ aufgelöst ist, heißt **explizit**, ansonsten heißt sie **implizit**.

Lösungsverfahren 1: Trennung der Variablen

Ist es möglich die DGL so umzuformen, dass auf der einen Seite eine Funktion von x (ausschließlich) und auf der anderen Seite eine Funktion von y (ausschließlich) steht, so kann die DGL relativ einfach durch Integration gelöst werden. Hierbei ist es wichtig, zu beachten, dass gilt:

$$y' = \frac{dy}{dx}$$

Beispiel:

$$y' = 2y + 4$$

Lösung:

Zunächst isoliert man alle Terme in Abhängigkeit von y auf einer Seite:

$$\frac{dy}{dx} = 2y + 4$$

$$\frac{dy}{(2y+4)} = dx$$

Jetzt kann integriert werden:

$$\int \frac{dy}{(2y+4)} = \int dx$$

$$\frac{1}{2}\ln|2y+4| = x + c$$

Diese Gleichung muss nun noch nach y aufgelöst werden:

$$\ln|2y+4| = 2x + 2c$$

$$|2y+4| = e^{2x+2c}$$

Die e-Funktion nimmt nur positive Werte an. Wenn man den Betrag wegläßt, muss vor der e-Funktion ein $\pm$ eingefügt werden.

$$2y + 4 = \pm e^{2x+2c}$$

$$y = \pm\frac{1}{2} * e^{2x+2c} + 2$$

Diese Funktion ist eine **allgemeine Lösung** zu dem gegebenen DGL. Der Wert c entscheidet über die spezielle Lösung. Wird in der Aufgabenstellung eine Anfangsbedingung gegeben, so kann der Wert C bestimmt werden.

Beispiel:

Zusätzlich zu der letzten Aufgabenstellung soll gelten:

$$y(0) = 1$$

Nun kann durch Einsetzen eine spezielle Lösung ermittelt werden:

$$1 = \pm\frac{1}{2} * e^{2*0+2c} + 2$$

Da $2y + 4$ für $y = 1$ positiv ist, kann das $\pm$ weggelassen werden.

$$e^{2*0+2c} = 2$$

$$2c = \ln 2$$

$$c = \frac{1}{2} * \ln 2$$

Die spezielle Lösung lautet also:

$$y = \frac{1}{2} * e^{2x+\ln 2} + 2$$

Lösungsverfahren 2: Lösen von exakten (bzw. totalen) Differentialgleichungen

Eine exakte bzw. totale Differentialgleichung hat die Form

$$y' = -\frac{g(x,y)}{h(x,y)}$$

mit

$$f'_x(x,y) = g(x,y)$$

und

$$f'_y(x,y) = h(x,y)$$

Um zu testen, ob eine exakte DGL vorliegt, nutzt man folgende Eigenschaft der Kreuzableitung:

$$f''_{xy}(x,y) = f''_{yx}(x,y)$$

Es wird also der Zähler nach y und der Nenner nach x abgeleitet. Sind die Ergebnisse gleich, so liegt eine exakte DGL vor.

Beispiel:

Gegeben sei folgende DGL:

$$y' = \frac{4x + y^2}{2xy - 3}$$

Dies ist eine exakte DGL, da gilt:

$$\frac{d(4x + y^2)}{dy} = 2y = \frac{d(2xy - 3)}{dx}$$

Lösung einer exakten DGL

Liegt eine exakte DGL vor, so gilt folgende Lösung:

$$f(x, y) = G(x, y) + \int \big(h(x, y) - G'(x, y)\big) dy = c$$

,wobei G die Stammfunktion von g bezüglich x bezeichnet.

Beispiel:

Zu ermitteln ist die allgemeine Lösung der folgenden DGL:

$$y' = \frac{5x^3 + 6xy}{3x^2 + 3y^2}$$

Zunächst prüft man, ob die DGL exakt ist:

$$\frac{d(5x^3 + 6xy)}{dy} = 6x = \frac{d(3x^2 + 3y^2)}{dx}$$

Die DGL ist exakt!

Für die Lösung

$$f(x,y) = G(x,y) + \int \Big(h(x,y) - G'_y(x,y)\Big)\, dy = c$$

benötigen wir folgende Terme:

$$G(x,y) = \frac{5}{4}x^4 + 3x^2y$$

$$G'_y(x,y) = 3x^2$$

Nun wird eingesetzt:

$$f(x,y) = \frac{5}{4}x^4 + 3x^2y + \int (3x^2 + 3y^2 - 3x^2)dy = c$$

$$f(x,y) = \frac{5}{4}x^4 + 3x^2y + 3x^2y + y^3 - 3x^2y = c$$

Auflösen nach y:

$$y = \frac{-\frac{5}{4}x^4 - y^3 + c}{3x^2}$$

Lösungsverfahren 3: Lösen von Ähnlichkeits-DGLs

Ähnlichkeits-DGLs haben die Form:

$$y' = g\left(\frac{y}{x}\right)$$

Beispiel für eine Ähnlichkeits-DGL:

$$y' = 2\frac{y}{x} + \frac{1}{\cos\frac{y}{x}}$$

Die Ähnlichkeits-DGL wird gelöst, indem man $\frac{y}{x}$ durch z substituiert.

Aus

$$z = \frac{y}{x}$$

folgt dann

$$y = z * x$$

und

$$y' = z + xz'$$

(hier wird die Produktregel angewendet, da z von x abhängig ist).

Eingesetzt in die DGL folgt:

$$y' = g\left(\frac{y}{x}\right)$$

$$z + xz' = g(z)$$

Als nächstes werden die Variablen getrennt (Lösungsverfahren 1)

$$\frac{1}{x} = \frac{z'}{g(z) - z}$$

Jetzt kann integriert werden

Die Lösung ist:

$$\ln|x| + c = \int \frac{1}{g(z) - z} dz$$

Integrieren, Zurücksubstituieren und nach y auflösen führt zur allgemeinen Lösung.

Beispiel:

Zu Lösen ist folgende DGL:

$$y' = \frac{y}{x} - \left(\frac{y}{x}\right)^2$$

Lösung:

$$z \coloneqq \frac{y}{x}$$

$$y' = x * z' + z$$

Gleichsetzen der beiden Ausdrücke für y‘:

$$z - z^2 = x * z' + z$$

Dies wird umgeformt zu:

$$xz' = -z^2$$

$$\frac{dx}{x} = -\frac{dz}{z^2}$$

Integrieren:

$$\frac{1}{z} = \ln|x| + c$$

Zurücksubstituieren:

$$\frac{x}{y} = \ln|x| + c$$

Auflösen nach y:

$$y = \frac{x}{\ln|x| + c}$$

Lineare DGLs erster Ordnung

Lineare DGLs haben die mathematische Form:

$$y' + p(x)y = q(x)$$

Es wird zwischen den folgenden verschiedenen Formen von linearen DGLs erster Ordnung unterschieden:

1) Homogene lineare DGLs: $q(x) = 0$
2) Homogene lineare DGL mit konstanten Koeffizienten: $q(x) = 0$ und $p(x)$ ist konstant.
3) Inhomogene lineare DGLS: $q(x) \neq 0$
4) Inhomogene lineare DGL mit konstanten Koeffizienten: $q(x) \neq 0$ und $p(x)$ ist konstant.

Zu jeder Form gibt es bestimmte Lösungswege!

Zu 1) und 2) Homogene lineare DGLs

Diese können durch Trennung der Variablen gelöst werden.

Es ergibt sich die folgende Lösung:

$$y = c * e^{-\int p(x)dx}$$

Beispiel:

Zu Lösen ist die folgende DGL:

$$y' + 2x^3y = 0$$

Lösung:

Zunächst identifiziert man die Funktionen:

$$q(x) = 0$$

$$p(x) = 2x^3$$

Als Integral von $p(x)$ erhält man:

$$\int p(x)dx = \frac{1}{2}x^4$$

Eingesetzt in die Lösungsgleichung:

$$y = c * e^{-\frac{1}{2}x^4}$$

Zu 3) und 4) inhomogene lineare DGLs

Bei inhomogenen linearen DGLs ist die Trennung der Variablen nicht möglich. Daher wird das Verfahren namens „Variation der Konstanten“ genutzt.

Der Ansatz dieses Verfahrens ist die Lösung für homogene DGLs:

$$y = c * e^{-\int p(x)dx}$$

Nun wird für c eine Funktion in Abhängigkeit von x gewählt. Diese Funktion wird dann so variiert, dass die inhomogene DGL erfüllt wird.

Die Herleitung ist nicht klausurrelevant, daher lasse ich sie hier weg. Im Ergebnis ist die Lösung der inhomogenen DGL:

$$y = e^{-P(x)} * \int \left(c + q(x) * e^{P(x)}\right)dx$$

,wobei gilt:

$$P(x) = \int p(x)dx$$

Beispiel:

Zu Lösen ist die folgende LDG:

$$y' = 2xy + x$$

Lösung:

Umstellen der Funktion auf Standardform:

$$y' - 2xy = x$$

Nun werden die Funktionen identifiziert:

$$p(x) = -2x$$

$$q(x) = x$$

Durch Integration erhält man $P(x)$:

$$P(x) = -x^2$$

Nun wird eingesetzt:

$$y = e^{x^2} * \int \left(c + x * e^{-x^2}\right)dx$$

$$y = e^{x^2}\left(cx - \frac{e^{-x^2}}{2}\right)$$

Aufgaben zu 5.0

Aufgabe 5.1

Welche der folgenden Funktionen ist eine Lösung der folgenden Differentialgleichung:

$$y'' = -\sin x$$

a)

$$y = \cos x$$

b)

$$y = 1 - \cos x$$

c)

$$y = \sin x + 1$$

d)

$$y = \sin x$$

e) keine der Lösungen

Aufgabe 5.2

Bestimme eine Lösung der folgenden DGL:

$$y' * e^y = x^2$$

Aufgabe 5.3

Berechne eine Lösung zu der folgenden DGL:

$$y' = -\frac{2x + 4y + 2}{4x + 12y + 8}$$

Aufgabe 5.4

Berechne eine Lösung zu der folgenden DGL:

$$y' = \frac{y^2 + xy}{\frac{1}{2} * x^2 + 2yx}$$

Aufgabe 5.5

Berechne eine Lösung zu der folgenden DGL:

$$y' = \frac{9x^2 + 3y^2}{2xy}$$

Aufgabe 5.6

Berechne eine Lösung zu der folgenden DGL mit der Anfangsbedingung $y(0) = 1$:

$$y' + 2x * y = x * e^{-x^2}$$

Lösungen zu 5.0

Lösung zu 5.1

Hier heißt es: ableiten und einsetzen!

Die Lösungen c) und d) sind richtig.

Lösung zu 5.2

Zunächst muss man erkennen, welche Form die DGL hat. In diesem Fall können die Variablen getrennt werden.

$$\frac{dy}{dx} * e^y = x^2$$

$$dy * e^y = x^2 * dx$$

Nun kann man integrieren:

$$\int dy * e^y = \int x^2 * dx$$

$$e^y = \frac{1}{3}x^3 + c$$

$$y = \ln\frac{1}{3}x^3 + c$$

Lösung zu 5.3

Zunächst muss man erkennen, welche Form die DGL hat. Diese ist eine exakte DGL, da gilt:

$$\frac{d(2x+4y+2)}{dy} = 4 = \frac{d(4x+12y+8)}{dx}$$

Für die Lösung

$$f(x,y) = G(x,y) + \int \Big(h(x,y) - G'_y(x,y)\Big)\, dy = c$$

benötigen wir folgende Terme:

$$G(x,y) = x^2 + 4yx + 2x$$

$$G'_y(x,y) = 4x$$

Einsetzen in die Lösungsgleichung:

$$f(x,y) = x^2 + 4yx + 2x + \int (4x + 12y + 8 - 4x)dy = c$$

$$f(x,y) = x^2 + 2x + 4xy + 6y^2 + 8y = c$$

Lösung zu 5.4

Zunächst muss man erkennen, welche Form die DGL hat. Diese ist eine exakte DGL, da gilt:

$$\frac{d(y^2+yx)}{dy} = 2y + x = \frac{d\left(\frac{1}{2} * x^2 + 2yx\right)}{dx}$$

Für die Lösung

$$f(x,y) = G(x,y) + \int \Big(h(x,y) - G'_y(x,y)\Big)\, dy = c$$

benötigen wir folgende Terme:

$$G(x,y) = xy^2 + \frac{1}{2}x^2y$$

$$G'_y(x,y) = 2xy + \frac{1}{2}x^2$$

Einsetzen in die Lösungsgleichung:

$$f(x,y) = xy^2 + \frac{1}{2}x^2y + \int \left(\frac{1}{2} * x^2 - 2xy + \frac{1}{2}x^2\right) dy = c$$

$$f(x,y) = xy^2 + \frac{1}{2}x^2y + \frac{1}{2} * x^2y - y^2x - \frac{1}{2}x^2y = c$$

Lösung zu 5.5

Zunächst muss man erkennen, welche Form die DGL hat. Dies ist eine Ähnlichkeits-DGL, da man sie auch wie folgt schreiben kann:

$$y' = \frac{9x^2 + 3y^2}{2xy} = \frac{9 + 3\left(\frac{y}{x}\right)^2}{2\frac{y}{x}}$$

Nun wird substituiert:

$$z \coloneqq \frac{y}{x}$$

$$\frac{9 + 3z^2}{2z} = y'$$

Die weiteren Schritte des Lösungsweges geschehen nach dem bekannten Lösungsschema.

$$\frac{dz}{\frac{9 + 3z^2}{2z} - z} = \frac{dx}{x}$$

$$\frac{2zdz}{9 + z^2} = \frac{dx}{x}$$

$$\int \frac{2zdz}{9 + z^2} = \int \frac{dx}{x}$$

Daraus folgt dann:

$$\ln|9 + z^2| = \ln|x| + C$$

$$z = \pm\sqrt{Cx - 9}$$

Rücksubstituieren:

$$y = \pm x\sqrt{Cx - 9}$$

Lösung zu 5.6

Auch wenn in der Klausur die Form der DGL oft in der Aufgabenstellung gegeben sein wird, halte ich es für sinnvoll, wenn du die DGL selber auf ihre Form untersuchst. Diese ist eine inhomogene lineare DGL erster Ordnung.

Die Lösung der inhomogenen DGL ist:

$$y = e^{-P(x)} * \int \left(c + q(x) * e^{P(x)}\right) dx$$

,wobei gilt:

$$P(x) = \int p(x) dx$$

Es sind folgende Funktionen zu identifizieren:

$$p(x) = 2x$$

$$q(x) = x * e^{-x^2}$$

Für $P(x)$ folgt:

$$P(x) = x^2$$

Einsetzen:

$$y = e^{-x^2} * \left(C + \int \left(x * e^{-x^2} * e^{x^2}\right) dx\right) = e^{-x^2} * \left(C + \frac{1}{2} x^2\right)$$

Dies ist die allgemeine Lösung.

Die spezielle Lösung erhält man durch Einsetzen der Anfangsbedingung.

$$1 = e^{-0^2} * \left(C + \frac{1}{2} 0^2\right)$$

Daraus folgt:

$$C = 1$$

6.0 Differenzengleichungen

In der Wirtschaft gibt es viele Größen, die nicht stetig sind sondern nur zu bestimmten Zeitpunkten ermittelt werden. Für die Funktionen dieser Größen gibt es keine Ableitungen. Es ist aber möglich die Differenzen in Gleichungsform darzustellen. Dabei entstehen Differenzengleichungen (DiffGL).

Beispiel:

Das Bruttosozialprodukt wird monatlich ermittelt und es ergeben sich folgende Werte für das erste Halbjahr:

Monat	1	2	3	4	5	6
BSP	1001	1003	1005	1006	1007	1010

Die Differenz zwischen zwei Monaten läßt sich nun schreiben als:

$$\Delta y = y_{n+1} - y_n$$

Wobei y für das BSP steht und n für den jeweiligen Vormonat.

Für Δy schreibt man auch allgemein r und erhält so die Grundform einer DiffGL:

$$y_{n+1} - py_n = r$$

Für r=0 ist die DiffGL homogen, ansonsten inhomogen.

Lösen von homogenen linearen DiffGL

Das „Lösen“ von homogenen linearen DiffGL ist denkbar einfach. Man zunächst stellt die Gleichung

$$y_{n+1} - py_n = 0$$

nach py_n um und erhält

$$py_n = y_{n+1}$$

In jedem Schritt nimmt y also um den Faktor p zu. Mathematisch ausgedrückt:

$$y_n = y_0 * p^n$$

Beispiel:

Gib die Lösung der folgenden DiffGL an:

$$y_{n+1} - 2y_n = 0$$

mit $y_0 = 1$.

Antwort:

In jedem Schritt wird y verdoppelt. Die Lösung lautet also:

$$y_n = y_0 * 2^n$$

Und für $y_0 = 1$ ergibt sich:

$$y_n = 2^n$$

Lösen von inhomogenen linearen DiffGL

Bei inhomogenen linearen DiffGL kommt in jedem Schritt noch ein konstanter Faktor r hinzu. Das sieht man besser, wenn man die Gleichung

$$y_{n+1} - py_n = r$$

nach y_{n+1} umstellt.

$$y_{n+1} = py_n + r$$

Achtung, die Lösung der DiffGL ist aber nicht

$$y_n = p^n y_0 + r * n$$

Da die r's der Vorperioden auch mit dem Faktor p verfielfacht werden, wäre diese Lösung nur für p=1 richtig. Die richtige Lösung lautet:

$$y_n = y_0 + r\frac{1 - p^n}{1 - p}$$

7.0 Grundlagen der induktiven Statistik

Ich beginne dieses Kapitel mit einigen einfachen Begriffen, die du kennen solltest. Diese werden gelegentlich in der ersten Aufgabe der Klausur abgefragt und bringen auch ohne statistisches Verständnis einfache Punkte.

Deskriptive Statistik: Auch beschreibende Statistik genannt. Es werden aus gegebenen Daten heraus verschiedene Maßzahlen, wie Mittelwerte oder Standardabweichungen berechnet.

Induktive Statistik: In der induktiven Statistik werden Rückschlüsse aus den Maßzahlen auf die zugrundeliegende Verteilung geschlossen.

Grundgesamtheit: Auch Population genannt. Die Menge aller relevanten Merkmalsträger.

Stichprobe: Eine Untermenge der Grundgesamtheit.

Stichprobenvariable: Zufallsvariablen, die verschiedene Realisationen annehmen kann. Beispiel: „Ergebnis eines Würfelwurfes“ ist eine Zufallsvariable mit den möglichen Realisationen 1 bis 6.

Realisation: Konkrete Werte, die eine Zufallsvariable annehmen kann.

Einfache Stichprobe: Eine Stichprobe mit uneingeschränkter Zufallsauswahl und voneinander unabhängigen Stichprobenvariablen.

Verfahren für konkrete Stichprobe: Der Fernuni Kurs nennt folgende Verfahren:

> **-Originalverfahren:** Stichprobenziehung anhand von Zufallszahlentabellen oder Computergenerierten Zufallszahlen (pseudo zufällig).
>
> **-Ersatzverfahren:** Buchstabenverfahren (In die Stichprobe gelangen alle Menschen, deren Nachname mit einem bestimmten Buchstaben beginnen), Geburtstagsverfahren (pseudo zufällig): In die Stichprobe gelangen die Menschen zuerst, die in einem bestimmten Zeitraum Geburtstag haben.

Pseudo zufällig: Ein Ereignis, dass zwar nicht zufällig ist, aber als zufällig wahrgenommen wird und statistisch nicht von wahrer Zufälligkeit unterschieden werden kann. Beispiel: Ist eine Münze in der Luft, so ist es theoretisch möglich zu bestimmen ob sie auf Kopf oder auf Zahl landen wird. Dennoch wird das Ergebnis als zufällig wahrgenommen.

Arten von Grundgesamtheiten: Das Fernuni-Skript nennt hier drei Arten, beschreibt diese aber nicht weiter:

> -konkret vorliegende Grundgesamtheit: Die gesamte Grundgesamtheit steht konkret zur Verfügung. Es kann also theoretisch eine komplette Erhebung durchgeführt werden.
>
> -teilweise konkret vorliegende Grundgesamtheit: Die gesamte Grundgesamtheit steht nur teilweise konkret zur Verfügung.

-hypothetische Grundgesamtheit: Eine Menge von fiktiven Objekten, die sich erst im Zuge der Stichprobe konkretisieren.

Beispiel: alle möglichen Ziehungen im Lotto, alle Wirtschaftssubjekte.

Verschiebungssatz: Der Verschiebungssatz bezeichnet folgende Umformung für die Berechnung der Varianz:

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\sum_{i=1}^{N} x_i^2\right) - N\bar{x}^2 = \left(\sum_{i=1}^{N} x_i^2\right) - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2$$

, dabei zu beachten ist:

$$\sum_{i=1}^{N} 2\, x_i * \bar{x} = 2\, N\bar{x}^2$$

Zum Schluß noch eine grundliegende Anmerkung zur Berechnung der Stichprobenvarianz: Hier wird zwischen der korrigierten und der unkorrigierten Stichprobenvarianz unterschieden (bei der korrigierten wird durch N-1 geteilt, bei der unkorrigierten durch N).

Hierzu merke Dir: Wird der Mittelwert aus der Stichprobe geschätzt, so nutze die korrigierte Form.

8.0 Wahrscheinlichkeitsverteilungen

In diesem Kapitel werde ich die wichtigsten Wahrscheinlichkeitsverteilungen kurz vorstellen- Diese sind schon aus dem Grundlagen der Wirtschaftsmathematik und Statistik-Modul bekannt. Wichtig ist hier, dass du nicht versuchst die Ausdrücke zu verstehen. Hier geht es darum Formeln als gegeben hinzunehmen. Es macht zum Beispiel keinen Sinn, zu versuchen die Formel für die Varianz einer F-verteilten Zufallsvariable $\left(Var(F) = \frac{2N^2(N+M-2)}{M(N-4)(M-2)^2}\right)$ verstehen zu wollen oder gar herleiten zu wollen-jedenfalls wäre dies nicht wichtig für das Bestehen der Klausur. Daher mein Rat: Bis zur Klausur solltest Du nur das verstehen, was Du verstehen musst-nach der Klausur kannst Du Dich dann dem Verständnis der einzelnen Themen widmen.

8.1 Die Normalverteilung

Die Normalverteilung ist die wichtigste Verteilung - sowohl für die Klausur, als auch in der Praxis.

Ein bekanntes Bespiel für die Normalverteilung sind die Aktienmarktreturns.

Merkmale der Normalverteilung:

- Erwartungswert der Normalverteilung:

$$E(X) = \mu$$

- Varianz der Normalverteilung:

$$Var(X) = \sigma^2$$

Man schreibt auch für „Die Zufallsvariable X ist normalverteilt mit Mittelwert μ und Varianz σ^2“:

$$X \sim N(\mu, \sigma^2)$$

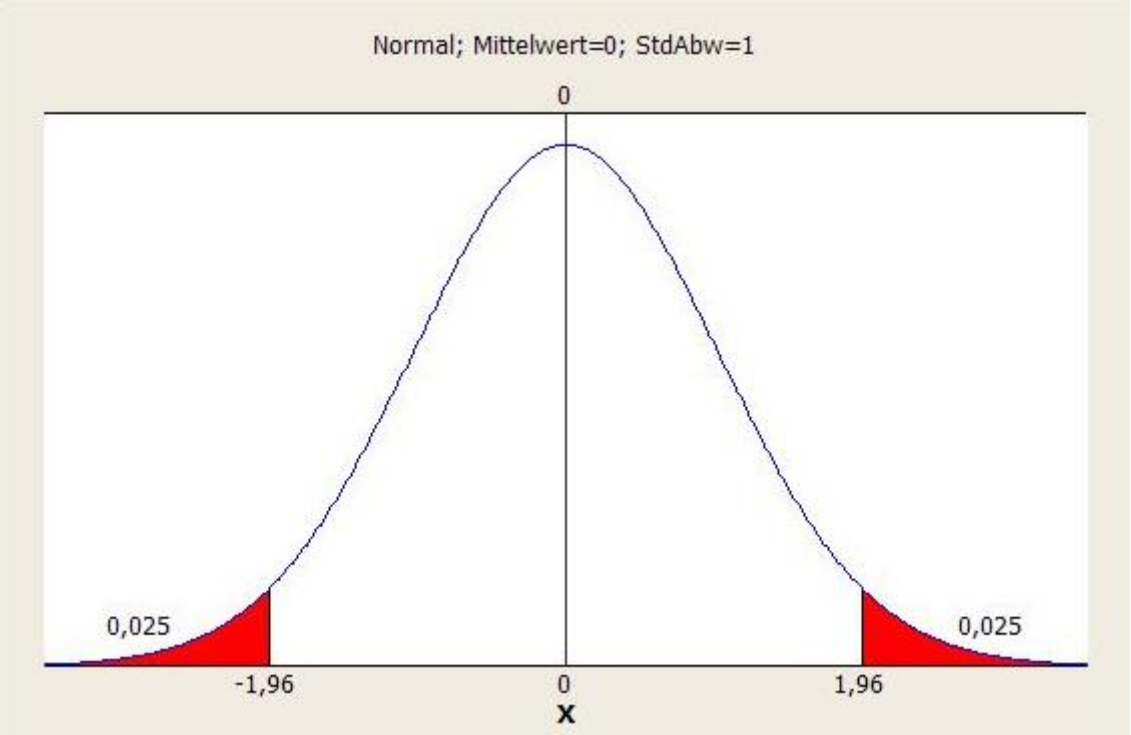
Eigenschaften der Normalverteilung:

- Die Normalverteilung ist symmetrisch.

- Mit steigender Varianz wird die Normalverteilung flacher und breiter. Mit steigendem Erwartungswert wird sie auf der X-Achse nach rechts verschoben.

- Eine Normalverteilung mit Erwartungswert 0 und Varianz 1 heißt **Standardnormalverteilung.** In diesem Zusammenhang ist auch interessant, wie eine Zufallsvariable standardisiert wird, nämlich indem man den Erwartungswert subtrahiert (auf der X-Achse zum Nullpunkt hin verschiebt) und das Ergebnis durch die Standardabweichung teilt (die Höhe und Breite an die Standardnormalverteilung mit Varianz 1 anpasst).

- Grafik der Normalverteilung (Gauß-Kurve):



Dichtefunktion der Normalverteilung

Die Dichtefunktion der Standardnormalverteilung ist ein komplexer mathematischer Ausdruck und sollte nicht klausurrelevant sein. Die Verteilungsfunktion lässt sich mathematisch nicht als Funktion darstellen, daher wird zur Ermittlung der Verteilungsfunktion auf Tabellen zurückgegriffen.

Den **Umgang mit der Tabelle der Standardnormalverteilung** musst du sicher beherrschen!

Die Tabelle der Standardnormalverteilung an der Stelle z gibt die Wahrscheinlichkeit an, mit der x einen kleineren Wert als z annimmt. Formal schreibt man:

$$\Phi_{0;1}(z)$$

Beispiele:

1) Eine standardnormalverteilte Zufallsvariable nimmt mit einer Wahrscheinlichkeit von 0,63307 einen kleineren Wert als 0,34 an.

2) Eine Zufallsvariable ist $N\sim(0,1)$. Mit welcher Wahrscheinlichkeit nimmt sie einen Wert unter 1 an?

Antwort: $0{,}84134$

Achtung! Die Tabelle gibt Wahrscheinlichkeiten für $x \geq 0$.Wird nach negativen X-Werten gefragt, so musst du $1 - \Phi_{0;1}(z_1)$ berechnen (Die Normalverteilung ist symmetrisch).

Beispiel: Eine Zufallsvariable ist $N\sim(0,1)$. Mit welcher Wahrscheinlichkeit nimmt sie einen Wert unter -1 an?

Antwort: $1 - 0{,}84134$

Um die Wahrscheinlichkeit zu finden, mit der X in ein bestimmtes Intervall z_1, z_2 fällt, muss die Differenz

$$\Phi_{0;1}(z_2) - \Phi_{0;1}(z_1)$$

gebildet werden.

Beispiel: Eine Zufallsvariable ist $N\sim(0,1)$. Mit welcher Wahrscheinlichkeit nimmt sie einen Wert zwischen 1 und 2 an?

Antwort:

$$\Phi_{0;1}(2) - \Phi_{0;1}(1) = 0{,}97725 - 0{,}84134 = 0{,}13591$$

Bestimmung von Wahrscheinlichkeiten für normalverteilte Zufallsvariablen

Bei der Bestimmung von Wahrscheinlichkeiten geht es um die Frage mit welcher Wahrscheinlichkeit eine normalverteilte Zufallsvariable in ein bestimmtes Intervall fällt.

Dabei macht man sich folgenden Zusammenhang zu Nutze:

Sei Z eine standardnormalverteilte Zufallsvariable und X eine normalverteilte Zufallsvariable mit Erwartungswert μ und Varianz σ^2, dann gilt:

Die Wahrscheinlichkeit, dass Z zwischen z_1 und z_2 liegt, ist genau so groß wie die Wahrscheinlichkeit, dass X zwischen $\mu + z_1\sigma$ und $\mu + z_2\sigma$ liegt.

Um die Wahrscheinlichkeit zu ermitteln, dass X in das Intervall z_1 bis z_2 fällt, muss also X zunächst standardisiert werden.

$$Z = \frac{x - \mu}{\sigma}$$

Z ist dann standardnormalverteilt.

Die Wahrscheinlichkeit kann dann an der Tabelle der Standardnormalverteilung abgelesen werden.

Achtung: Wenn die Wahrscheinlichkeit gegeben ist und nach dazugehörenden X-Wert gefragt ist, musst du den X-Wert von der standardnormalverteilten Zufallsvariablen in eine normalverteilte Variable transformieren.

Beispiel 1:

Eine Zufallsvariable X ist $N\sim(2,3)$veteilt. Mit welcher Wahrscheinlichkeit nimmt sie einen Wert kleiner als 4 an?

Antwort: Standardisieren der Zufallsvariable:

$$Z = \frac{4 - 2}{\sqrt{3}} = \frac{2}{\sqrt{3}} \approx 1{,}15$$

Z ist nun standardnormalverteilt. Nun muss ermittelt werden, mit welcher Wahrscheinlichkeit Z einen Wert kleiner als $1{,}15$ annimmt.

$$\Phi_{0;1}(1{,}15) = 0{,}87493$$

Beispiel 2:

X ist $N(2,4)$ verteilt. Mit welcher Wahrscheinlichkeit nimmt X einen Wert **größer** als 3 an?

Antwort: 1.Schritt: Zufallsvariable standardisieren!

$$Z = \frac{3-2}{\sqrt{4}} = 0{,}5$$

$$\Phi_{0;1}(0{,}5) = 0{,}69146$$

Da aber nach der Wahrscheinlichkeit für X größer als 3 gefragt war, beträgt die Wahrscheinlichkeit

$$1 - 0{,}69146 = 0{,}30854$$

Die Aufgabenstellung kann natürlich auch lauten: Unter welchem X-Wert liegen 90% der Ausprägungen wenn $X \sim N(2,4)$ verteilt ist? Nun musst du aus der Wahrscheinlichkeit der Tabelle der Standardnormalverteilung den entsprechenden X-Wert ablesen.

Achtung: Wenn die Wahrscheinlichkeit gegeben ist und nach dazugehörenden X-Wert gefragt ist, musst du den X-Wert von der standardnormalverteilten Zufallsvariablen in eine normalverteilten Variable transformieren.

Beispiel: Unter welchem X-Wert liegen 90% der Ausprägungen, wenn $X \sim N(2,4)$ verteilt ist?

$$\Phi^{-1}{}_{0;1}(90\%) \approx 1{,}28$$

Transformation des Wertes 1,28 zu einer $N(2,3)$ –verteilten Zufallsvariablen:

$$X = 1{,}28\sigma + \mu = (1{,}28 * 2) + 2 = 4{,}56$$

Die Chi-Quadrat-Verteilung

Quadriert man mehrere normalverteilte Zufallsvariablen, so ist deren Summe Chi-Quadrat verteilt.

Mathematisch schreibt man:

$$\chi^2 = \sum_{i=1}^{N} X_i^2$$

Die Chi-Quadrat Verteilung hat nur einen Parameter, nämlich N. Diesen nennt man auch „Freiheitsgrad“.

$$\chi^2_{a;N}$$

Ist also die Chi Quadrat Verteilung mit N Freiheitsgraden zum Niveau a.

Approximation:

Für $N > 30$ kann das Quantil $z(a)$ wie folgt approximiert werden:

$$z(a) = \sqrt{2\chi^2_{a;N}} - \sqrt{2N-1}$$

Merke: Die Stichprobenvarianzen von normalverteilten Stichprobenvariablen sind Chi-Quadrat verteilt.

Tabelle der Chi-Quadrat Verteilung:

	Wahrscheinlichkeit p										
Freiheitsgrade	**0,005**	**0,01**	**0,025**	**0,05**	**0,1**	**0,5**	**0,9**	**0,95**	**0,975**	**0,99**	**0,995**
1	0,00	0,00	0,00	0,00	0,02	0,45	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	1,39	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	2,37	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	3,36	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	4,35	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	5,35	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	6,35	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	7,34	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	4,17	8,34	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	9,34	15,99	18,31	20,48	23,21	25,19

Die t-Verteilung

Eine Zufallsvariable der Form

$$T = \frac{x}{\sqrt{\frac{Y}{N}}}$$

,wobei X standardnormalverteilt und Y Chi-Quadrat verteilt ist, ist t-verteilt.

Eigenschaften der t-Verteilung:

$$E(T) = 0; N > 0$$

$$VAR(T) = \frac{N}{N-2}; N > 2$$

Für $N \to \infty$ strebt die t-Verteilung gegen die Standardnormalverteilung. Für $N > 30$ kann statt der t-Verteilung die Standardnormalverteilung genutzt werden.

Nutzung der t-Verteilung

Die t-Verteilung wird meist genutzt, wenn die Standardabweichung der Grundgesamtheit unbekannt ist. Ist dies der Fall, so muss sie aus der Stichprobe geschätzt werden und es muss statt der Standardnormalverteilung die t-Verteilung genutzt werden.

Berechnung der Standardabweichung S der Grundgesamtheit

Ist die Standardabweichung der Stichprobe berechnet, so berechnet man den Schätzwert für die Standardabweichung der Grundgesamtheit nach folgender Formel:

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$$

Die t –Verteilung hat neben der Wahrscheinlichkeit noch den Parameter „Freiheitsgrade“. Dieser wird mit v bezeichnet und errechnet sich durch

$$v = N - 1$$

Beispiel:

Der Wert der t-Verteilung für 90% und 10 Freiheitsgrade ist $1{,}812$.

Tabelle der t-Verteilung:

Anzahl Freiheitsgrade n	P für zweiseitigen Vertrauensbereich							
	0,5	0,75	0,8	0,9	0,95	0,98	0,99	0,998
	P für einseitigen Vertrauensbereich							
	0,75	0,875	0,90	0,95	0,975	0,99	0,995	0,999
1	1,000	2,414	3,078	6,314	12,706	31,821	63,657	318,309
2	0,816	1,604	1,886	2,920	4,303	6,965	9,925	22,327
3	0,765	1,423	1,638	2,353	3,182	4,541	5,841	10,215
4	0,741	1,344	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,301	1,476	2,015	2,571	3,365	4,032	5,893
6	0,718	1,273	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,254	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,240	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,230	1,383	1,833	2,262	2,821	3,250	4,297
10	0,700	1,221	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,214	1,363	1,796	2,201	2,718	3,106	4,025
12	0,695	1,209	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,204	1,350	1,771	2,160	2,650	3,012	3,852
14	0,692	1,200	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,197	1,341	1,753	2,131	2,602	2,947	3,733
16	0,690	1,194	1,337	1,746	2,120	2,583	2,921	3,686
17	0,689	1,191	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,189	1,330	1,734	2,101	2,552	2,878	3,610
19	0,688	1,187	1,328	1,729	2,093	2,539	2,861	3,579
20	0,687	1,185	1,325	1,725	2,086	2,528	2,845	3,552

Die F-Verteilung

Die F-Verteilung kommt bei Test über Varianzen zum Einsatz.

Eine Zufallsvariable der Form

$$F = \frac{\frac{X}{M}}{\frac{Y}{N}}$$

Ist F-verteilt mit den Freiheitsgraden M und N, wobei X Chi-Quadrat verteilt ist mit M Freiheitsgraden und Y Chi-Quadrat verteilt ist mit N Freiheitsgraden.

Eigenschaften der F-Verteilung

$$E(F) = \frac{N}{N-2}; N \geq 3$$

$$Var(F) = \frac{2n^2(N+M-2)}{M(N-4)(M-2)^2}; N \geq 5$$

Mehr musst Du für die Klausur über die F-Verteilung nicht wissen. Sie sollte für Dich nur ein Werkzeug sein, das Du anwendest.

Für die Quantile der F-Verteilung gilt:

$$f(p,m,n) = \frac{1}{f(1-p,n,m)}$$

Die Binomialverteilung

Eine ebenfalls einfache Verteilung ist die sogenannte Binomialverteilung.

Betrachtet wird ein Zufallsexperiment, bei dem es nur 2 mögliche Ergebnisse, A und $\bar{A}$, gibt, die mit der Wahrscheinlichkeit p bzw. $1-p$ auftreten. Dieses Experiment wird nun x-mal hintereinander ausgeführt (ein sogenanntes Bernoulli-Experiment).

Die Binomialverteilung gibt nun bei n-maliger Wiederholung des Experimentes an, mit welcher Wahrscheinlichkeit das Ereignis *A* **genau** x-mal eintritt.

Beispiel: Eine Münze wird 5 mal geworfen. Mit welcher Wahrscheinlichkeit kommt genau 4 mal Kopf? Die Anzahl an „Köpfen“ ist hier binomialverteilt.

Da die Verteilung von der Wahrscheinlichkeit p und der Anzahl an Wiederholungen N abhängt schreibt man auch:

X ist binomialverteilt mit den Parametern n und p.

$$B(x|N,p)$$

Die Binomialverteilung hat folgende Merkmale:

- Wahrscheinlichkeitsfunktion:

$$B(x|N,p) = \binom{N}{x} p^x (1-p)^{n-x}$$

für $x = 1$ bis n.

- Verteilungsfunktion:

$$F_X(x) = \sum_{k=0}^{x} \binom{N}{k} p^k (1-p)^{n-k}$$

für $x = 1$ bis n. Das k gibt in diesem Fall die Anzahl an positiven Ereignissen an.

- Erwartungswert der Binomialverteilung:

$$E(X) = Np$$

- Varianz der Binomialverteilung:

$$Var(X) = Np(1-p)$$

Approximationsmöglichkeiten

Für $np(1-p) > 9$ ist eine binomialverteilte Zufallsvariable annähernd normalverteilt mit Erwartungswert np und Varianz $np(1-p)$.

Die Multinomialverteilung

Der Binomialverteilung sehr ähnlich ist die Multinomialverteilung. Statt nur 2 möglichen Ausprägungen, können hier mehrere Ausprägungen (Kategorien) angenommen werden. Ein häufiges Beispiel ist die Wahlumfrage mit fest vorgegebenen Parteien.

Parameter der Verteilung sind die jeweiligen Prozentwerte der Kategorien.

Für die Prozentwerte der einzelnen Kategorien schreibt man mathematisch:

$$p_j = \hat{\pi}_j = \frac{x_j}{N}$$

,wobei x_j die Anzahl an Personen ist, die Kategorie j wählen und N der Stichprobenumfang.

Die Varianz berechnet sich nach folgender Formel:

$$Var(\hat{\pi}_j) = \frac{1}{N}\pi_j(1-\pi_j)$$

Entsprechend gilt für die Schätzung der Varianz:

$$\widehat{Var}(\hat{\pi}_j) = \frac{1}{N}\hat{\pi}_j(1-\hat{\pi}_j)$$

Beispiel:

Bei einer Wahlumfrage, bei der 4 Parteien (A, B, C, D) zur Auswahl standen, wurden folgende Ergebnisse erzielt:

A	B	C	D
300	350	250	100

Insgesamt wurden 1000 Personen befragt.

Die Schätzungen für die Anteilswerte entsprechen den Prozentsätzen der Stichprobe:

$$\hat{\pi}_A = 30\%$$

$$\hat{\pi}_B = 35\%$$

usw.

Die geschätzte Varianz des Anteils $\hat{\pi}_A$ beträgt:

$$\widehat{Var}(\hat{\pi}_{A)} = \frac{1}{N}\hat{\pi}_A(1-\hat{\pi}_A) = \frac{1}{1000}0{,}3 * 0{,}7 = 0{,}00021$$

In diesem Zusammenhang ist noch folgendes wichtig: Die Fehlertoleranz entspricht-wenn nichts anderes gesagt- 5%. Da 95% der Ausprägungen in das Intervall von ± 2 Standardabweichungen fällt, beträgt die Fehlertoleranz in diesem Fall:

$$2\sqrt{\widehat{Var}(\hat{\pi}_{A)}} = 2{,}9\%$$

9.0 Einfache Schätzverfahren

In diesem Kapitel werden Methoden zur Schätzung von Parametern (Meistens Erwartungswert und Varianz) bei unbekannten Wahrscheinlichkeitsverteilungen behandelt.

Einen Schätzer werde ich mit dem Akzent $\hat{}$ kennzeichnen und den zu schätzenden Parameter werde ich, wie auch im Fernuni-Skript, Q nennen.

Schätzfunktionen und Punktschätzungen

Bei einer Schätzung wird eine Stichprobe aus einer Grundgesamtheit gezogen und die Parameter der Stichprobe werden auf die Grundgesamtheit übertragen. Die Rechenvorschrift für den entsprechenden Parameter wird Schätzfunktion genannt.

Beispiel:

Aus 1000 Menschen werden 100 nach ihrem Körpergewicht gefragt. Der Erwartungswert der 100 Befragten wird als Schätzwert für den Erwartungswert der 1000 Menschen genommen. Die Funktion

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} X_i$$

wird Schätzfunktion genannt.

Wird ein Parameter genau geschätzt, so nennt man das eine Punktschätzung. Man kann auch schätzen, in welchem Intervall ein Parameter mit einer gewissen Wahrscheinlichkeit liegt (Intervallschätzung), dazu später mehr.

Eigenschaften von Schätzfunktionen

Zu diesem Kapitel waren die Fragen in Klausuren meist qualitativ. Zu den Eigenschaften von Schätzfunktionen solltest du folgendes wissen:

- Eine Schätzfunktion heißt erwartungstreu oder **unverzerrt (unbiased)**, wenn der Erwartungswert des Schätzers dem wahren Wert des zu schätzenden Parameters entspricht.

- Eine Schätzfunktion heißt **effizient oder wirksam**, wenn der Schätzwert eine endliche Varianz hat und es keine andere erwartungstreue Schätzfunktion gibt, die einen Schätzwert mit geringerer Varianz liefert.

- Der Mittlere quadratische Fehler **(Mean Squared Error-MSE)** einer Schätzfunktion ist der Erwartungswert der quadrierten Differenz zwischen dem Schätzwert und dem tatsächlichen Wert des Parameters.

Mathematisch schreibt man:

$$E(\hat{Q} - Q)^2$$

Eine Schätzfunktion ist effizient, wenn der mittlere quadratische Fehler minimal ist. Für eine erwartungstreue Schätzfunktion ist der mittlere quadratische Fehler gleich der Varianz.

Ein erwartungstreuer Schätzer für die Varianz ist beispielsweise $S^2 * \frac{n}{n-1}$.

-Konsistenz: Eine Schätzfunktion ist konsistent, wenn der Schätzfehler für $N \to \infty$ gegen Null strebt. Das leuchtet irgendwie auch ein: Wenn der Stichprobenumfang der Grundgesamtheit entspricht, so sollte der zu schätzende Parameter exakt dem tatsächlichen Parameter entsprechen.

Momentenmethode

Im Fernuni Skript wird auf wenigen Zeilen kurz die Formel für die Momentenmethode angegeben. Da dieses Thema extrem kurz gefasst ist, kann man über die Klausurrelevanz streiten. Du solltest Dir zumindest folgendes merken:

- Für das k-te Moment gilt mathematisch:

-im diskreten Fall:

$$M_k = E(X^k) = \frac{1}{N}\sum_{i=1}^{n} x_i^k f_x(x_i)$$

-im stetigen Fall:

$$M_k = E(X^k) = \int_{-\infty}^{\infty} x^n f(x)dx$$

Das erste Moment $(k = 1)$ ist – wie an der Formel zu erkennen ist- ein Schätzer für den Mittelwert der Verteilung.

Für uns besonders interessant ist die besondere Form des **zentralen Momentes**. In diesem Fall wird statt x_i der Ausdruck $(x_i - \bar{x})$ verwendet.

Es gilt nun mathematisch:

-im diskreten Fall:

$$M_k = E(x_i - \bar{x})^k = \frac{1}{N}\sum_{i=1}^{n} (x_i - \bar{x})^k f_x(x_i)$$

-im stetigen Fall:

$$M_k = E(x_i - \bar{x})^k = \int_{-\infty}^{\infty} (x - \bar{x})^k f(x)dx$$

Das zweite Moment $(k = 2)$ ist – wie an der Formel zu erkennen ist- ein Schätzer für die Varianz der Verteilung. Da beim Momentenschätzer aber statt durch N-1 nur durch N getelit wird, ist der Schätzer nicht erwartungstreu, aber asymptotisch erwartungstreu.

Schätzer nach der Momentenmethode sind gemäß Fernuni Skript „wenig effizient“.

Maximum-Likelihood-Schätzung

Für die Klausur solltest du anschaulich verstanden haben wie die Maximum-Likelihood-Schätzung funktioniert und wie man einen konkreten ML Schätzer berechnet. Auf die Herleitung werde ich hier nicht eingehen.

Anschauliche Erklärung der ML-Schätzung

Bei einer ML-Schätzung ist die Art der Verteilung gegeben und die Parameter sind zu schätzen. Die Frage ist nun: Für welche Parameter ist die gezogene Stichprobe am wahrscheinlichsten?

Beispiel: In New York sind alle Taxis von 1 bis n nummeriert. Du siehst zufällig ein Taxi und es hat die Nummer 10123. Was ist ein ML-Schätzer für die Anzahl an Taxis in NY? (Dieses Beispiel soll Dir nur beim Verständnis helfen-Klausuraufgaben werden anders aussehen)

Antwort: Man muss sich hier überlegen, welche Zahl es „am wahrscheinlichsten“ macht, dass wir zufällig das Taxi mit der Nummer 10123 sehen. Je höher die Zahl der Taxis, desto geringer die Wahrscheinlichkeit, die Nummer 10123 zu sehen. Da es aber mindestens 10123 Taxis gibt, ist die Antwort: 10123 Taxis gibt es wahrscheinlich in NY.

Berechnung eines ML-Schätzers

Vorweg ein Hinweis: In den bisherigen Klausuraufgaben war es nicht notwendig eine ML-Schätzfunktionen herzuleiten und es reichte aus, die ML-Schätzfunktionen für den Erwartungswert der bekannten Verteilungen zu kennen. Diese sind:

$\mu = \bar{x}$ für die Normalverteilung

p für die Binomialverteilung

Die ML Schätzer berechnen sich nach der allgemeinen Schätzfunktion:

$$L(x_1, x_2, \dots x_n; q) = \prod_{i=1}^{n} f_{X_i}(x_i, q)$$

Die ML-Schätzfunktion nimmt für den ML- Schätzer ein Maximum an. Dieses kann durch Ableitung der ML Funktion nach dem gesuchten Parameter –und Nullsetzen dieser Ableitung- gefunden werden. (In bisherigen Klausuren mußte keine Ableitung berechnet werden.)

Merke:

ML-Schätzer sind:

-konsistent,

-asymptotisch effizient,

-asymptotisch normalverteilt.

Maximum a posteriori Schätzer

Die Maximum a posteriori Schätzung ähnelt stark der Maximum-Likelihood Schätzung, nur dass der Modalwert zur Schätzung des entsprechenden Parameters verwendet wird.

Kleinste-Quadrate-Schätzung

Nach der Methode der kleinsten Quadrate ergibt sich der Schätzer für μ für den Wert, bei dem die Summe der quadrierten Abweichungen minimal wird.

Dieses Prinzip werden wir später im Kapitel über die Regressionsanalyse behandeln.

Punktschätzer als Zufallsvariablen

Zum Schluss noch ein paar Worte zu den Schätzwerten des Mittelwertes: Der Schätzwert des Mittelwertes ist eine Zufallsvariable, die sich aus der Summe vieler Zufallsvariablen ergibt. Der Schätzwert ist also selber auch eine Zufallsvariable (Zieht man auch einer Grundgesamtheit von 100 nacheinander mehrmals eine Stichprobe vom Umfang 10, so werden sich die Stichproben sehr wahrscheinlich unterscheiden-sie sind zufällig zusammengesetzt). Für diese Zufallsvariable kann man nach mehrmaligem Ziehen wieder die bekannten Parameter Mittelwert und Varianz berechnen. Diese werden zwar in den Kapiteln im Skript nicht benötigt, aber sie werden erwähnt und in den Aufgaben geprüft. Mathematisch gilt:

Erwartungswert des Mittelwertes:

$$E(\bar{X}) = \mu$$

(Der Erwartungswert des Mittelwertes der Stichprobe entspricht dem wahren Mittelwert der Grundgesamtheit).

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N}$$

(Die Varianz des Stichprobenmittelwertes entspricht der Varianz der Grundgesamtheit geteilt durch den Stichprobenumfang.)

Sind mehrere Stichproben durchgeführt worden, so können Erwartungswert und Varianz natürlich aus den Daten der Stichproben errechnet werden.

Beispiel: Es werden 10 Stichproben gezogen und die Mittelwerte betragen:

10,11,10,9,12,10,8,8,11,13

Mittelwert und Varianz dieser Daten werden jetzt nach den bekannten Formeln berechnet.

Aufgaben zu 9.0

Aufgabe 9.1

Welche der folgenden Aussagen sind richtig?

a) Eine Schätzfunktion ist genau dann erwartungstreu, wenn sie immer den wahren Parameterwert liefert.

b) Eine Schätzfunktion ist unverzerrt, wenn der Erwartungswert der Schätzfunktion dem wahren Wert des zu schätzenden Parameters entspricht.

c) Das Stichprobenmittel ist ein erwartungstreuer Schätzer für den Erwartungswert der Grundgesamtheit.

d) Von 2 gegebenen erwartungstreuen Schätzfunktionen ist die mit der größeren Varianz effizienter.

e) Eine effiziente Schätzfunktion minimiert den mittleren quadratischen Fehler aller unverzerrten Schätzfunktionen.

f) Eine erwartungstreue Schätzfunktion ist effizient.

Aufgabe 9.2

Welche der folgenden Aussagen sind richtig?

a) Parameter, die aus Stichproben ermittelt werden, sind Zufallsvariablen.

b) Für erwartungstreue Schätzfunktionen ist der mittlere quadratische Fehler gleich der Varianz.

c) Die Stichprobenvarianz $S^2 * \frac{n}{n-1}$ ist ein erwartungstreuer Schätzer für die Varianz der Grundgesamtheit.

d) Bei endlichen Grundgesamtheiten braucht zwischen „Ziehen mit Zurücklegen“ und „Ziehen ohne Zurücklegen“ nicht unterschieden zu werden.

Aufgabe 9.3

Die Stichprobenvariablen $X_1, X_2, \dots X_n (n > 2$ und gerade) sind unabhängig voneinader und haben dieselbe Verteilung. Gegeben sind die folgenden Schätzfunktionen:

$$T_1 = \frac{1}{2}(X_1 + X_n)$$

$$T_2 = (X_1 + X_2)$$

$$T_3 = \frac{1}{2}(X_1 + X_2 + X_3 - 2X_n)$$

a) Welche Schätzfunktion ist erwartungstreu?

b) Berechne die Varianz des Schätzers aus T_1.

Lösungen zu 9.0

Lösung zu 9.1

a) Falsch. Der Erwartungswert der Schätzfunktion entspricht dem zu schätzenden Parameter.

b) Richtig.

c) Richtig.

d) Falsch. Je kleiner die Varianz, desto effizienter.

e) Richtig.

f) Falsch. Die Bedingungen sind erwartungstreu **und** minimale Varianz.

Lösung zu 9.2

a) Richtig. Die Werte ergeben sich zufällig, da die Stichprobe ja zufällig aus der Grundgesamtheit gezogen wird.

b) Richtig.

c) Richtig.

d) Falsch. Dies ist bei unendlichen Grundgesamtheiten der Fall.

Lösung zu 9.3

a) Hier muss man die Erwartungswerte der Schätzfunktionen berechnen.

$$E(T_1) = \frac{1}{2}(\mu_1 + \mu_n)$$

Da gilt: $\mu_1 = \mu_n$ erhält man:

$$E(T_1) = \frac{1}{2}(\mu_1 + \mu_n) = \mu$$

Damit ist T_1 ein erwartungstreuer Schätzer.

Entsprechend erhält man für

$$E(T_2) = 2\mu$$

Dies ist also kein erwartungstreuer Schätzer.

Für $E(T_3)$ gilt entsprechend:

$$E(T_3) = 1/2\mu$$

b) $Var(T_1) = \frac{1}{4}(2\sigma^2) = \frac{1}{2}\sigma^2$

10.0 Konfidenzintervalle

Im letzten Kapitel haben wir die Punktschätzungen kennengelernt. Da es sehr unwahrscheinlich ist den Parameter einer Verteilung punktgenau zu schätzen, werden wir nun Intervalle schätzen, in denen der zu schätzende Parameter mit gegebener Wahrscheinlichkeit liegen wird. In den meisten Fällen werden wir Mittelwerte schätzen.

Konfidenzintervall für μ bei bekannter Varianz und normalverteilter Grundgesamtheit

In diesem Kapitel geht es um folgende Aufgabenstellung:

Aus einer Grundgesamtheit mit unbekanntem Mittelwert und bekannter Standardabweichung σ wird eine Stichprobe vom Umfang N gezogen. Es ist nun ein Intervall anzugeben, in das der tatsächliche Mittelwert der Grundgesamtheit mit einer gewissen Wahrscheinlichkeit (Konfidenz) fallen wird.

Beispiel: Aus einer Grundgesamtheit vom Umfang 100 mit $N(\mu, 16)$-verteiltem X wird eine Stichprobe von N $= 25$ gezogen. Der Mittelwert der Stichprobe ist 15. Gib ein Intervall an, in dem der Mittelwert der Grundgesamtheit mit einer Wahrscheinlichkeit von 95% liegen wird.

Auf eine anschauliche Erklärung werde ich - genau wie die Fernuni- verzichten und gehe direkt zur **Lösung:**

Das Konfidenzintervall für den Mittelwert der Grundgesamtheit erhälst du nach folgender Formel:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

a ist die Irrtumswahrscheinlichkeit. Da wir ein zweiseitig begrenztes Intervall haben, muß diese halbiert werden. Die Formel sieht etwas unheimlich aus, ist aber ganz einfach. Eigentlich mußt du nur wissen, wie man mit der Tabelle der Standardnormalverteilung umgeht und die Werte in die Formel einsetzen.

$z_{1-\alpha/2}$ erhält man aus der Standardnormalverteilung für gegebenes a.

Als **Beispiel** lösen wir nun das obige Beispiel:

Aus einer Grundgesamtheit mit $N(\mu, 16)$-verteiltem X wird eine Stichprobe von 25 gezogen. Der Mittelwert der Stichprobe ist 15. Gib ein Intervall an, in dem der Mittelwert der Grundgesamtheit mit einer Wahrscheinlichkeit von 95% liegen wird.

Lösung:

Ermittle den z-Wert und setze in die Formel ein:

$$z_{1-\alpha/2} = z_{1-0{,}05/2} = z_{0{,}975} = 1{,}96$$

$$\sigma = 4$$

$$\bar{X} = 15$$

$$N = 25$$

$$15 - 1{,}96\frac{4}{5}\mu \leq 15 + 1{,}96\frac{4}{5}$$

$$13{,}43 \leq \mu \leq 16{,}57$$

Konfidenzintervall für μ bei unbekannter Standardabweichung und normalverteilter Grundgesamtheit

Ist die Standardabweichung der Grundgesamtheit nicht gegeben, so muss sie aus der Stichprobe geschätzt werden und es muss statt der Standardnormalverteilung die t-Verteilung genutzt werden.

Berechnung der Standardabweichung S der Grundgesamtheit

Für die Standardabweichung der Stichprobe gilt:

$$S^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X})^2$$

Achtung: Da es sich um eine Schätzung für die Varianz handelt, wird hier durch $(N-1)$ geteilt, anstatt durch N.

Nutzung der t-Verteilung

Die t –Verteilung hat neben der Wahrscheinlichkeit noch den Parameter „Freiheitsgrade“. Dieser wird mit v bezeichnet und errechnet sich durch

$$v = N - 1$$

Für mehr als 30 Freiheitsgrade, also einen Stichprobenumfang über 31, gleicht die t-Verteilung der Standardnormalverteilung.

Für die Bestimmung eines Konfidenzintervalls bei unbekannter Grundgesamtheit gilt dann

$$\bar{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{N}}$$

Konfidenzintervall für μ bei beliebig verteilter Grundgesamtheit und großem Stichprobenumfang

Ist die Verteilung der Grundgesamtheit unbekannt, so nutzt man folgende Eigenschaft von Zufallsvariablen einer beliebig verteilten Grundgesamtheit:

Summen von beliebig verteilten Zufallsvariablen sind approximativ normalverteilt (nicht verstehen, akzeptieren!).

Für $N > 30$ kann also die Grundgesamtheit als normalverteilt behandelt werden und die entsprechenden Formeln finden Anwendung.

Bestimmung des notwendigen Stichprobenumfangs für eine Intervallschätzung

Aus der Berechnung für ein Konfidenzintervall sieht man, dass das Intervall kleiner und damit die Schätzung genauer wird, je größer der Stichprobenumfang n ist.

Man kann also auch fragen: Wie groß muss der Stichprobenumfang mindestens sein, damit das Intervall eine bestimmte Genauigkeit erreicht? Die Genauigkeit der Schätzung wird mit e bezeichnet und es gilt:

$$e = z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Zur Bestimmung von N in Abhängigkeit von der Genauigkeit e des Konfidenzintervalls, bei gegebener Wahrscheinlichkeit, muss die Gleichung

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} = e$$

nach n aufgelöst werden.

Man erhält:

$$N \geq \frac{\sigma^2}{e^2} z_{1-\alpha/2}^2$$

Wie auch die Fernuni, habe ich die Endlichkeitskorrektur bei der Berechnung des Stichprobenumfangs vernachlässigt.

Beispiel:

Aus einer Grundgesamtheit vom Umfang 10000 mit $N(\mu, 16)$-verteiltem X wird eine Stichprobe gezogen. Der Mittelwert der Stichprobe ist 15. Gib an, wie groß die Stichprobe mindestens sein muß, damit μ mit einer Wahrscheinlichkeit von 95% in dem Konfidenzintervall 14,5 bis 15,5 liegt.

Lösung:

Folgende Werte müssen ermittelt werden:

$$\sigma^2 = 16$$

$$z_{1-\alpha/2}^2 = 1{,}96^2 = 3{,}84$$

$$e^2 = 0{,}25$$

$$N \geq \frac{16}{0{,}25} 3{,}84 = 245{,}76$$

Der Stichprobenumfang müßte also 246 betragen.

Konfidenzintervall für den Anteilswert einer dichotomen Grundgesamtheit bei großem Stichprobenumfang

Für ein dichotomes Merkmal ergibt sich das Konfidenzintervall nach folgender Formel:

$$\hat{\pi} - z_{1-\frac{a}{2}} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \leq \pi \leq \hat{\pi} + z_{1-\frac{a}{2}} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

Wobei π der zu schätzende Anteil in der Grundgesamtheit ist und $\hat{\pi}$ der Anteil in der Stichprobe.

Zur Herleitung: Für große Stichproben gleicht die Binomialverteilung der Normalverteilung. Es gilt:

Wenn

$$N\,\pi \geq 5$$

und

$$N(1-\pi) \geq 5$$

,dann ist π approximativ normalverteilt mit Mittelwert $\hat{\pi}$ und Varianz $\hat{\pi}(1-\hat{\pi})$.

In das Konfidenzintervall für den Mittelwert einer normalverteilten Grundgesamtheit müssen dann nur noch Mittelwert und Varianz eingesetzt werden und man erhält die oben angegebene Formel für das Konfidenzintervall.

Beispiel:

Aus einer binomialverteilten Grundgesamtheit wird eine Stichprobe vom Umfang $N = 1.000$ gezogen. Aus der Stichprobe ergibt sich $\hat{\pi} = 0{,}4$. Das Konfidenzniveau zu $a = 0{,}5$ ist dann:

$$\hat{\pi} - z_{1-\frac{a}{2}} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \leq \pi \leq \hat{\pi} + z_{1-\frac{a}{2}} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

$$0{,}4 - 1{,}96 * \sqrt{\frac{0{,}4(1-0{,}4)}{1.000}} \leq \pi \leq 0{,}4 + 1{,}96 * \sqrt{\frac{0{,}4(1-0{,}4)}{1.000}}$$

$$0{,}37 \leq \pi \leq 0{,}43$$

Konfidenzintervall für die Varianz einer normalverteilten Grundgesamtheit

Als Testgröße für die Varianz wird die Stichprobenvarianz s^2 verwendet. Das Konfidenzintertvall ergibt sich zu:

$$\frac{(N-1)S^2}{\chi^2_{1-\frac{a}{2};N-1}} \leq \sigma^2 \leq \frac{(N-1)S^2}{\chi^2_{\frac{a}{2};N-1}}$$

Wobei σ^2 die angenommene Varianz und $\chi^2_{a;N-1}$ die Chi-Quadrat Verteilung mit N -1 Freiheitsgraden zum Niveau a ist.

Beispiel:

Es soll ein Konfidenzintervall ermittelt werden für die Varianz einer normalverteilten Grundgesamtheit.

Es wird eine Stichprobe vom Umfang $N = 36$ genommen und das Konfidenzniveau soll $a = 10\%$ betragen. Die Stichprobenvarianz beträgt 24.

Lösung:

Für das Konfidenzintervall gilt:

$$\frac{(N-1)S^2}{\chi^2_{1-\frac{a}{2};N-1}} \leq \sigma^2 \leq \frac{(N-1)S^2}{\chi^2_{\frac{a}{2};N-1}}$$

$$\frac{(35)24}{49{,}8} \leq \sigma^2 \leq \frac{(35)24}{22{,}47}$$

$$16{,}87 \leq \sigma^2 \leq 37{,}38$$

Konfidenzintervall für den Korrelationskoeffizienten *p*

Beim Konfidenzintervall für den Korrelationskoeffizienten wird der Student vom Fernuni Skript mit einem Formelsalat erschlagen. Hier heißt es wieder: Nicht verstehen, sondern wichtige Formeln identifizieren!

Voraussetzung ist: X und Y sind normalverteilt und es liegt eine große Stichprobe vor.

Das Konfidenzintervall für p lautet:

$$\frac{e^A - 1}{e^A + 1} \leq p \leq \frac{e^B - 1}{e^B + 1}$$

,wobei gilt:

$$A = \ln\frac{1+R}{1-R} - \frac{2z}{\sqrt{N-3}}$$

$$B = \ln\frac{1+R}{1-R} + \frac{2z}{\sqrt{N-3}}$$

R ist wiederum:

$$R = \frac{S(X,Y)}{S(X)S(Y)} = \frac{\sum_{n=1}^{N}(X_n - \bar{X})(Y_n - \bar{Y})}{\sqrt{\sum_{n=1}^{N}(X_n - \bar{X})^2 \sum_{n=1}^{N}(Y_n - \bar{Y})^2}}$$

Beispiel:

Gegeben sind zwei Variablen mit Korrelation $r = 0{,}5$. Es soll ein 95% Konfidenzintervall für den unbekannten Parameter *p* ermittelt werden. Der Stichprobenumfang beträgt $N = 100$.

Antwort:

Hier heißt es stupide in Formeln einsetzen!

Für die Formel:

$$\frac{e^A - 1}{e^A + 1} \leq p \leq \frac{e^B - 1}{e^B + 1}$$

Braucht man zunächst A und B:

$$A = \ln\frac{1 + R}{1 - R} - \frac{2z}{\sqrt{N - 3}}$$

Der z-Wert von 97,5% ist 1,96 (beidseitiges Konfidenzintervall, daher 1-a/2).

Durch einsetzen erhält man:

$$A = ln\frac{1 + 0{,}5}{1 - 0{,}5} - \frac{2 * 1{,}96}{\sqrt{100 - 3}} = 0{,}7$$

Entsprechend erhält man für B:

$$B = 1{,}5$$

Für das Konfidenzintervall folgt:

$$\frac{e^A - 1}{e^A + 1} \leq p \leq \frac{e^B - 1}{e^B + 1}$$

$$0{,}34 \leq p \leq 0{,}64$$

Es ist auch möglich, dass Kovarianz und Varianzen gegeben sind, dann muss R zunächst aus diesen berechnet werden (nach der bekannten Formel für die Korrelation).

Aufgaben zu 10.0

Aufgabe 10.1

Berechne für eine normalverteilte Zufallsvariable mit unbekanntem Mittelwert und Standardabweichung $\sigma = 600$ ein 95% Konfidenzintervall für den Mittelwert. Der Stichprobenumfang beträgt 25 mit Erwartungswert $\bar{X} = 4000$.

Aufgabe 10.2

Für eine normalverteilte Zufallsvariable mit unbekannten Lageparametern gibt es beim Ziehen mit Zurücklegen folgende Stichprobenergebnisse:

$$8{,}7{,}8{,}5{,}7{,}9{,}7{,}6{,}6$$

Berechne ein Konfidenzintervall zum Niveau $a = 5\%$ für den Mittelwert.

Aufgabe 10.3

Für eine normalverteilte Zufallsvariable mit Erwartungswert 10 und Varianz 16 wird eine Stichprobe vom Umfang 36 gezogen. Auf Endlichkeitskorrektur soll verzichtet werden.

Mit welcher Wahrscheinlichkeit fällt der Mittelwert in das Intervall 8,69 bis 11.31?

Aufgabe 10.4

Welche Aussagen sind richtig?

a) Mit einer Wahrscheinlichkeit von a überdeckt das Konfidenzintervall den wahren Parameterwert.

b) Je größer das Konfidenzintervall, desto kleiner a.

c) Je größer das Konfidenzniveau, desto kleiner das Konfidenzintervall.

d) Je größer der Stichprobenumfang, desto kleiner das Konfidenzintervall.

e) Die Grenzen der Konfidenzintervalle sind Stichprobenfunktionen

f) μ ist eine Zufallsvariable

Lösungen zu 10.0

Lösung zu 10.1

Diese Aufgaben kommen sehr regelmäßig dran. Es muss in die Formel

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

eingesetzt werden.

Für die Aufgabenstellung ergibt sich:

Untere Grenze:

$$4.000 - 1{,}96 * \frac{600}{\sqrt{25}} = 3.764{,}8$$

Obere Grenze:

$$4.000 - 1{,}96 * \frac{600}{\sqrt{25}} = 4.235{,}8$$

Lösung zu 10.2

Da die Varianz unbekannt ist, muss sie aus der Stichprobe geschätzt werden.

Für den Mittelwert der Stichprobe erhält man:

$$\bar{x} = 7$$

$$s^2 = 1{,}5 \ , \qquad s = 1{,}22$$

Damit ergibt sich

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{N}} = \frac{1{,}22}{3} = 0{,}41$$

Für das Konfidenzintervall ergibt sich entsprechend:

Untere Grenze:

$$\bar{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{N}}$$

Obere Grenze:

$$\bar{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{N}}$$

Der Wert der t-Verteilung für $N - 1 = 8$ Freiheitsgrade zum Niveau $97{,}5\%$ ist $2{,}306$.

Eingesetzt in die obige Formel ergibt sich die untere Grenze zu

$$7 - 2{,}31 * 0{,}41 = 6{,}05$$

und die obere Grenze zu

$$7 + 2{,}31 * 0{,}41 = 7{,}95$$

Lösung zu 10.3

Achte immer darauf, ob die Varianz oder die Standardabweichung gegeben sind. In diesem Fall muss die Formel:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} = 8{,}69$$

nach $z_{1-\alpha/2}$ aufgelöst werden (Dies ist die Formel für die Abweichung nach unten. Man hätte auch die Formel für die Abweichung nach oben nutzen können).

$$10 - \frac{4z_{1-\alpha/2}}{6} = 8{,}69$$

$$z_{1-\alpha/2} = (-8{,}69 + 10)1{,}5 = 1{,}96$$

Für 1,96 folgt aus der Tabelle der Standardnormalverteilung eine Wahrscheinlichkeit von 97,5%. Daraus folgt $a = 5\%$. Die gesuchte Wahrscheinlichkeit beträgt also 95%.

Lösung zu 10.4

a) Falsch. Die genannte Wahrscheinlichkeit ist $1 - a$.

b) Richtig. a ist die Wahrscheinlichkeit, daß das Konfidenzintervall den wahren Parameter nicht überdeckt. Je größer das Intervall, desto eher überdeckt es den wahren Parameter.

c) Falsch. Will man ganz sicher gehen, dass der wahre Parameter von dem Konfidenzintervall überdeckt wird, so muss man das Intervall sehr groß wählen.

d) Richtig. Im Extremfall entspricht die Stichprobe der Grundgesamtheit und man kann den wahren Parameter direkt angeben. (Zu all den Fragen zu diesen Zusammenhängen kannst du einfach in die Formel schauen und ablesen ob z.B. ein hohes n das Konfidenzintervall vergrößert oder verkleinert.)

e) Richtig.

f) Falsch. Es ist eine feste aber unbekannte Größe.

11.0 Grundlagen der Testtheorie

In diesem und den nächsten Kapiteln werden wir uns mit Parametertests befassen. Dabei ist es wichtig, dass Du einmal verstehst, wie ein Parametertest durchgeführt wird und mit den entsprechenden Verteilungen umgehen kannst. Hast Du das einmal verstanden und eingeübt, kannst du eigentlich jeden beliebigen Test durchführen. Egal, was getestet wird, Du musst die Hypothesen formulieren, die Teststatistik aus dem Glossar raussuchen, die Werte aus der Aufgabenstellung einsetzen und den Wert der Teststatistik mit den kritischen Werten der Verteilung vergleichen. Eigentlich ist es also überflüssig, sich mit jedem einzelnen Test zu beschäftigen. Ich gehe in diesem Kapitel kurz auf jeden Test ein.

Einführung in die Systematik eines statistischen Testes

Bei einem Parametertest werden Annahmen (Hypothesen) über einen Parameter anhand einer Stichprobe getestet.

Beispiel: Ich habe einen Würfel und möchte testen, ob er einen Mittelwert von 3,5 hat. Ich führe also eine Stichprobe durch indem ich mehrmals würfel. Es gibt nun zwei Möglichkeiten:

1) Der Mittelwert der Stichprobe liegt in der Nähe von 3,5.

2) Der Mittelwert der Stichprobe liegt weit entfernt von 3,5.

Im ersten Fall, ist es gut möglich, dass der Mittelwert meines Würfels 3,5 ist, er könnte aber auch 3,6 oder 3,4 sein.

Im zweiten Fall kann ist es sehr unwahrscheinlich, dass der Mittelwert des Würfels bei 3,5 liegt. Ich lehne also die Hypothese ab.

Hieraus lernst du eine sehr wichtige Eigenschaft von Hypothesentests: **Man kann eine Hypothese nur widerlegen, nicht bestätigen!**

Um trotzdem einen Nutzen aus Hypothesentests zu ziehen, müssen Hypothesen wie folgt formuliert werden:

Man formuliert immer zwei Hypothesen:

Zunächst die zu testende Hypothese (genannt „Gegenhypothese“) und dann eine Hypothese, die alle anderen Fälle einschließt (genannt „Nullhypothese“). Kann man die Nullhypothese widerlegen, so muss zwangsläufig die Gegenhypothese gelten.

Beispiel: Ich möchte testen ob der Mittelwert meines Würfels außerhalb des Intervalls von 3 bis 4 liegt. Dazu stelle ich folgende Hypothesen auf:

Nullhypothese: Der Mittelwert des Würfels liegt zwischen 3 und 4.

Gegenhypothese: Der Mittelwert des Würfels liegt außerhalb des Intervalls 3-4.

Kann ich aufgrund der Stichprobe die Nullhypothese widerlegen, so muss der wahre Mittelwert zwangsläufig außerhalb des Intervalls 3-4 liegen.

Fehler bei statistischen Tests

Es gibt bei jedem statistischen Test die Möglichkeit, Fehlschlüsse aus der Stichprobe zu ziehen. Es können die folgenden 2 Fehler gemacht werden:

- Ablehnung einer richtigen Nullhypothese („**Fehler 1.Art**“).

- Nicht-Ablehnung einer falschen Nullhypothese („**Fehler 2.Art**“).

Beim Fehler 1. Art wird tatsächlich eine falsche Entscheidung getroffen. Man nimmt die Gegenhypothese an, obwohl diese falsch ist.

Beim Fehler 2. Art wird keine Entscheidung getroffen, obwohl man eine hätte treffen können.

Daher ist es wichtiger den Fehler 1. Art zu vermeiden. Die Wahrscheinlichkeit für einen Fehler 1.Art wird bei jedem Test angegeben. Man nennt sie auch „Konfidenzniveau“.

Berechnung des Fehlers 2ter Art

Der Fehler zweiter Art ist der Fehler, eine falsche Nullhypothese nicht abzulehnen.

Um den Fehler zweiter Art zu bestimmen, wird also die wahre Verteilung benötigt. Der Fehler zweiter Art ist dann die Wahrscheinlichkeit, mit der die Nullhypothese nicht abgelehnt wird vorausgesetzt die wahre Verteilung gilt. Zur Erinnerung: Bei einem einfachen Hypothesentest wird vorausgesetzt, dass die Nullhypothese gilt.

Beispiel:

Eine Variable X ist normalverteilt mit Erwartungswert 100 und Varianz 64. Der Stichprobenumfang ist 36. Es wird die Hypothese geprüft, dass der Mittelwert über 102 liege zum Niveau $a = 5\%$.

Der Fehler 2ter Art berechnet man nun wie folgt:

Zunächst muss bestimmt werden, für welche kritischen Bereiche die Nullhypothese angenommen wird:

$$102 - 1{,}65\frac{8}{\sqrt{36}} = 99{,}8 \leq \bar{X}$$

Es muss also die Wahrscheinlichkeit gesucht werden, mit der (unter Annahme der wahren Verteilung) ein Wert größer oder gleich 99,8 angenommen wird. Dies ist aus Grundlagen der Wirtschaftsmathematik und Statistik bekannt: Es ist nach der Wahrscheinlichkeit gesucht, mit der der Stichprobenmittelwert 0,2 nach unten von seinem wahren Mittelwert abweicht.

Standardisierung der Differenz zwischen kritischem Wert und Mittelwert:

$$\frac{99{,}8 - 100}{\frac{8}{\sqrt{36}}} = -0{,}15$$

Es ist also nach der Wahrscheinlichkeit gesucht, mit der die Standardnormalverteilung einen Wert von -0,15 oder weniger annimmt. Diese kann aus der Tabelle abgelesen werden und beträgt ca. 44%.

Der Aufbau von Parametertests

Der Aufbau eines Parametertests lässt sich folgendermaßen untergliedern:

1) Festlegung der Grundgesamtheit und Typ der Verteilung
   - Handelt es sich um ein quantitatives oder qualitatives Merkmal?
   - Ist die Grundgesamtheit endlich?
   - Wie ist die Verteilung der Grundgesamtheit (meistens normalverteilt)?
2) Formulierung der Nullhypothese (H_0) und der Gegenhypothese (H_1).
3) Festlegung der Testgröße
   - Der zu testende Parameter muss aus der Stichprobe geschätzt werden. In Diesem Schritt wird die Schätzfunktion bestimmt.
4) Bestimmung der Verteilung der Testgröße
   - Testgröße ist meistens in der Stichprobe genauso verteilt wie in der Grundgesamtheit. Hier wird angenommen, dass die Nullhypothese gilt.
5) Vorgabe der Irrtumswahrscheinlichkeit
   - Oft wir $a = 5\%$ gewählt.
6) Bestimmung des Ablehnungsbereichs
   - Nun ermittelt man den Bereich, in den die Testgröße unter den getroffenen Annahmen und der Annahme der Nullhypothese zu $(1 - a)\%$ fallen wird. Der restliche Bereich ist der Ablehnungsbereich. Achte darauf ob es ein einseitiger Test ist (> oder < Bedingung) oder ein 2 seitiger (Intervall für den Parameter).
7) Ziehung der Stichprobe und Bestimmung der Testgröße
8) Testentscheidung
9) Interpretation der Entscheidung
   - **Achtung**: Statistische Tests unterliegen immer dem Risiko, dass ein Fehler 1. oder 2. Art vorliegt! Daher sagt man auch: “Das Testergebnis ist signifikant zum Niveau $(1 - a)\%$ abgesichert“.

Beispiel:

Die Brenndauer von Glühbirnen sei normalverteilt mit $\sigma = 5$ Wochen. Nach Angaben des Herstellers sei der Mittelwert der Brenndauer mindestens 100 Wochen. Zur Prüfung der mittleren Brenndauer zum Niveau 5% wird eine Stichprobe von 25 Stück aus einer Lieferung aus 1000 Glühbirnen gezogen. Die Ziehung der Glühbirnen ergibt einen Mittelwert von 97 Wochen

1) - Merkmal ist quantitativ

- Grundgesamtheit ist begrenzt auf 1.000.

- Zufallsvariable ist normalverteilt.

2) Da zu testen ist, ob die mittlere Brenndauer über 100 Wochen beträgt, wählen wir als Gegenhypothese $H_1: \mu_0 \geq \mu$.

Daraus folgt für die Nullhypothese: $H_0: \mu_0 < \mu$.

3) Testgröße ist der Stichprobenmittelwert $\bar{X}$.

4) Da X normalverteilt ist, ist auch $\bar{X}$ normalverteilt. Der Mittelwert der Stichprobe ist $\mu_0 = 100$ und $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (auf Endlichkeitskorrektur wird verzichtet, da $25/1000 < 0{,}05$)

5) Als Signifikanzniveau wird $a = 0{,}05$ vorgegeben.

6) $\bar{X}$ ist normalverteilt mit $E(X) = 100$ und Varianz $\left(\frac{\sigma}{\sqrt{n}}\right)^2 = 1$.
Unter der Annahme, dass die Nullhypothese wahr ist, werden 95% der Ausprägungen in den Bereich unterhalb von

$$\bar{X} - z_{1-a}\sigma_{\bar{X}} = 100 - 1{,}65 * 1 = 98{,}35$$

Fallen (siehe Kapitel 11.5 bis 11.7 zur Berechnung der Konfidenzintervalle). Der Ablehnungsbereich der Nullhypothese ist damit $\bar{X} \geq 98{,}35$.

7) Nun wird die Stichprobe gezogen. Die mittlere Brenndauer beträgt laut Aufgabenstellung 97 Wochen.

8) Die Testgröße fällt nicht in den Ablehnungsbereich. Daher kann H_0 nicht abgelehnt werden.

9) Da die Nullhypothese nicht abgelehnt werden konnte, kann auch keine signifikante Aussage über die mittlere Brenndauer der Glühbirnen gemacht werden.

Für die Klausur:

Du solltest noch wissen, was das Produzentenrisiko und was das Konsumentenrisiko ist:

Wird eine Lieferung über eine Stichprobe auf ihre Güte überprüft und darf eine bestimmte Fehlerquote nicht überschritten werden, so gibt es die folgenden Möglichkeiten Fehler zu machen:

1) Die Lieferung überschreitet die Fehlerquote, dies wird aber nicht entdeckt.
2) Die Lieferung überschreitet die Fehlerquote nicht, wird aber als „Fehlerquote überschritten“ reklamiert.

Der Fall 1) ist das Konsumentenrisiko, da der Konsument eine fehlerhafte Lieferung kauft.

Der Fall 2) ist das Produzentenrisiko, da eine einwandfreie Lieferung als fehlerhaft reklamiert wird.

Aufgaben zu 11.0

Aufgabe 11.1

1) Gib an, bei welchen der folgenden Probleme ein einseitiger und bei welchen ein zweiseitiger Test zu verwenden ist:

a) Mindestkörpergröße von Polizeibeamten

b) Abweichung der Glasdicke von Windschutzscheiben vom Mittelwert

c) Brenndauer von Wachskerzen sei größer als vier Stunden

d) Untersuchung über die Gewichtszunahme der Deutschen

e) Der HSV hat bei einem Spiel höhere Siegchancen als der FC St. Pauli

2) Was wird bei einem einseitigen Test geprüft?

3) Was wird bei einem zweiseitigen Test geprüft?

4) Was bedeutet das Signifikanzniveau α?

5) Was bedeutet es für die Nullhypothese, wenn diese nicht abgelehnt wird?

6) Ändert man bei einem Test nur das Signifikanzniveau von 5% auf 1%, dann

a) wird der Ablehnungsbereich größer.

b) wird der Ablehnungsbereich kleiner.

7) Bei der Berechnung der Verteilung der Prüfgröße wird

a) die Annahme der Nullhypothese als wahr angenommen.

b) die Annahme der Gegenhypothese als wahr angenommen.

8) Welchen Einfluss hat eine Erhöhung des Stichprobenumfanges auf einen Fehler 2.Art?

9) Der Hersteller von Glühbirnen beliefert seinen Abnehmer mit einer Lieferung von 10.000 Glühbirnen. Der Abnehmer überprüft die Lieferung durch eine Stichprobe von 100 Stück. Wenn mehr als 5% defekt sind sendet er die Lieferung als defekt zurück. Was ist das Produzentenrisiko?

Aufgabe 11.2

Welche der folgenden Aussagen sind richtig?

a) Eine Vervierfachung des Stichprobenumfangs führt zu einer Halbierung der Breite des Konfidenzintervalls.

b) Verwirft man die Nullhypothese zum Niveau 5%, so würde man sie auch zum Niveau 10% verwerfen.

c) Ein Fehler 2.Art bedeutet eine falsche Nullhypothese nicht abzulehnen.

d) Durch einen höheren Stichprobenumfang steigt die Wahrscheinlichkeit eine falsche Hypothese auch abzulehnen.

e) Der Ablehnungsbereich ist unabhängig von der Stichprobe.

f) Wird die Nullhypothese zum Niveau a verworfen, so gilt die Alternativhypothese mit einer Wahrscheinlichkeit von a.

g) Die Fehler erster und zweiter Art steigen und fallen gemeinsam mit einer Veränderung von a.

h) Die nachzuweisende Hypothese sollte stets als Nullhypothese gewählt werden.

i) Ein trennscharfer Test verwirft die falsche Nullhypothese mit großer Wahrscheinlichkeit.

Aufgabe 11.3

Formuliere die Nullhypothese und die Alternativhypothese zu folgendem Test:

Es soll geprüft werden, ob sich die Prüfgröße A im letzten Jahr mindestens verdoppelt hat. A bezeichnet die aktuelle Prüfgröße, A_{-1}die vom letzten Jahr.

Lösungen zu 11.0

Lösung zu 11.1

1) Immer wenn nur die Abweichung in eine Richtung geprüft werden soll, ist ein einseitiger Test gefragt. Soll auf Abweichung nach oben und unten geprüft werden, muss ein 2-seitiger Test genutzt werden.

a) einseitig

b) zweiseitig

c) einseitig

d) einseitig

e) einseitig

2) Es wird geprüft ob ein Parameter mindestens einen gegebenen Wert hat, gegen die Alternative, dass der Wert größer **bzw.** kleiner ist.

3) Es wird geprüft, ob ein Parameter einen gegebenen Wert hat, gegen die Alternative, dass der Wert größer **oder** kleiner ist.

4) Es bedeutet, dass H_0 abgelehnt wird, obwohl es zutrifft.

5) Dies bedeutet nur, dass sie nicht abgelehnt wird - es bedeutet nicht, dass sie angenommen wird!

6) Antwort a ist richtig. Je mehr man ablehnt, desto „feiner“ die Auslese.

7) Antwort a ist richtig.

8) Die Wahrscheinlichkeit für einen Fehler 2ter Art verringert sich. Je höher der Stichprobenumfang, desto genauer die Schätzung.

9) Das Produzentenrisiko besteht darin, dass die Lieferung zurückgesendet wird, obwohl sie nicht defekt ist.

Lösung zu 11.2

a) Richtig. In der Formel steht Wurzel aus n im Nenner.

b) Richtig. Je größer a desto eher verwirft man die Nullhypothese.

c) Richtig.

d) Richtig. Durch höhere Stichprobenumfänge steigt die Qualität des Tests.

e) Richtig. Er hängt von den Annahmen der Nullhypothese ab.

f) Falsch. Die Alternativhypothese gilt dann mit Wahrscheinlichkeit $1 - a$.

g) Falsch. Die Fehler erster und zweiter Art sind gegenläufig bei Veränderungen von a.

h) Falsch. Es sollte die Alternativhypothese gewählt werden, da nur die statistisch belegt werden kann.

i) Richtig.

Lösung zu 11.3

$$H_0: A < 2A_{-1}$$

$$H_1: A \geq 2A_{-1}$$

12.0 Parametertests

Hast du erst einmal die Theorie, bzw. das „Formel identifizieren und Einsetzen“ der Konfidenzintervalle verstanden, so ist für den Parametertest kaum noch etwas zusätzlich zu verstehen. Bevor ich auf die einzelnen Tests eingehe, werde ich einmal exemplarisch den Parametertest für den Mittelwert einer normalverteilten Grundgesamtheit mit bekannter Varianz herleiten. Nach demselben Prinzip können dann die anderen Parametertests aus den entsprechenden Konfidenzintervallen hergeleitet werden.

Vorweg etwas zur Schreibweise der Prüfgrößen und Annahmebereiche: Es gibt zwei verschiedene Schreibweisen der Prüfgrößen und Annahmebereiche. Man kann einerseits für den Parameterwert den Annahmebereich berechnen und prüfen, ob der Parameter in diesen Annahmebereich hineinfällt oder man kann den Parameterwert so umformen, dass man ihn direkt mit einem festen Annahmebereich einer bekannten Verteilung (z.B. der Standardnormalverteilung) vergleichen kann. Ich halte es für verständlicher zunächst den Annahmebereich für den Parameter anzugeben und werde die ersten Tests entsprechend darstellen.

Da der Lehrstuhl eine andere Darstellungsform hat, werde ich später nur noch die Prüfgröße und die zugehörige Verteilung angeben.

Beispiel: Test über den Mittelwert einer normalverteilten Grundgesamtheit mit bekannter Varianz

Das Konfidenzintervall für den Mittelwert einer normalverteilten Grundgesamtheit mit bekannter Varianz ist:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Das bedeutet, daß μ mit der Wahrscheinlichkeit $1 - a$ in das oben angegebene Intervall fallen wird.

Man ermittelt also zunächst $\bar{X}$ und leitet daraus ein Konfidenzintervall für μ ab. Bei einem Test über den Mittelwert trifft man zunächst eine Annahme über μ. Aus diesem angenommenen μ_0 bildet man dann ein Intervall (Annahmebereich) in das $\bar{X}$ bei einer Stichprobe zu $1 - a$ fallen müßte, wenn die Annahme richtig ist.

Wenn also die Annahme über μ richtig, dann müßte $\bar{X}$ mit der Wahrscheinlichkeit $1 - a$ in folgendes Intervall fallen:

$$\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \bar{X} \leq \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

wobei μ_0 der hypothetische Mittelwert ist. (Es ist wichtig, dass Du hier den Zusammenhang zum oben genannten Konfidenzintervall siehst.)

Fällt $\bar{X}$ nicht in dieses Intervall, so kann die Annahme (Hypothese) mit einer Fehlerwahrscheinlichkeit von a abgelehnt werden.

Wie du siehst, besteht kein großer Unterschied zwischen dem Konfidenzintervall und der Herleitung des Annahmebereiches eines Parametertests.

Im Folgenden werde ich nun die verschiedenen Parametertests durchgehen und jeweils die Prüfgröße und den Annahmebereich nennen.

Achtung- wie oben erwähnt zur Schreibweise: Der Lehrstuhl stellt die Formel neuerdings anders dar- und zwar wie folgt:

$$z_{1-\alpha/2} \leq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}} \leq z_{1-\alpha/2}$$

Die Formel wird also nur umgestellt, sodass der Annahmebereich direkt aus der entsprechenden Verteilung abgelesen werden kann.

Test über den Mittelwert einer normalverteilten Grundgesamtheit mit bekannter Varianz

Eben schon einmal exemplarisch hergeleitet, jetzt also nur kurz zur Vollständigkeit:

Die Testgröße ist der Mittelwert der Stichprobe.

Der Annahmebereich für die Nullhypothese, dass der Mittelwert einem bestimmten Wert μ_0 entspricht, ist:

- für einen zweiseitigen Test:

$$\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \bar{X} \leq \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

wobei μ_0 der hypothetische Mittelwert ist.

- für einen einseitigen Test entsprechend

$$\bar{X} \leq \mu_0 + z_{1-a}\frac{\sigma}{\sqrt{N}}$$

- oder

$$\mu_0 - z_{1-a}\frac{\sigma}{\sqrt{N}} \leq \bar{X}$$

Merke:

Die Testgröße ist:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{N}}}$$

Sie ist standardnormalverteilt und es gelten die entsprechenden kritischen Werte.

Beispiel:

Folgende Daten sind für eine normalverteilte Zufallsvariable gegeben:

Mittelwert der Stichprobe: 10

Varianz der Grundgesamtheit: 36

$N = 25$

$a = 10\%$

Es soll getestet werden, ob der Mittelwert der Grundgesamtheit mindestens 9 ist. **Antwort:**

Als Nullhypothese wählt man:

$$H_0: \mu < \mu_0$$

gegen

$$H_1: \mu \geq \mu_0$$

μ_0 ist der hypothetische Wert 9.

Der Annahmebereich für die Nullhypothese ist:

$$\bar{X} \leq \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{N}}$$

$$\bar{X} \leq 9 + 1{,}28 \frac{6}{5}$$

$$\bar{X} \leq 10{,}54$$

Da $\bar{X} = 10$, kann die Nullhypothese nicht abgelehnt werden.

Test über den Mittelwert einer normalverteilten Grundgesamtheit mit unbekannter Varianz

Bei **unbekannter Standardabweichung** der Grundgesamtheit muss diese durch die Stichprobenvarianz geschätzt werden und es muss die t-Verteilung genutzt werden.

Die Prüfgröße ist dann:

Der Annahmebereich für die Nullhypothese, dass der Mittelwert einem bestimmten Wert μ_0 entspricht, ist:

- für einen zweiseitigen Test:

$$\mu_0 - t_{1-\alpha/2}\frac{s}{\sqrt{N}} \leq \bar{X} \leq \mu_0 + t_{1-\alpha/2}\frac{s}{\sqrt{N}}$$

Merke:

Die Prüfgröße ist:

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{N}}}$$

T ist t-verteilt mit N-1 Freiheitsgraden und es gelten die entsprechenden kritischen Werte.

Test über den Mittelwert einer unbekannten Grundgesamtheit mit unbekannter Varianz

Ist auch die Verteilung unbekannt, aber $N > 30$, so wird die Varianz aus der Stichprobe geschätzt und die Prüfgröße gilt als approximativ normalverteilt.

Merke:

Die Prüfgröße ist:

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{N}}}$$

Z ist approximativ standardnormalverteilt. Es gelten die kritischen Werte der Standardnormalverteilung.

Test über den Mittelwert einer unbekannten Grundgesamtheit mit bekannter Varianz

Ist die Verteilung unbekannt, aber $N > 30$ und die Varianz bekannt, so gilt die Prüfgröße als approximativ normalverteilt.

Merke:

Die Prüfgröße ist:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{N}}}$$

Z ist approximativ standardnormalverteilt. Es gelten die kritischen Werte der Standardnormalverteilung.

Tests für die Varianz

Als Testgröße für die Varianz wird die Stichprobenvarianz s^2 verwendet. Der Annahmebereich der Nullhypothese ergibt sich für

- zweiseitige Tests zu:

$$\frac{\sigma_0^2 \chi^2_{\frac{a}{2};N-1}}{N-1} \leq s^2 \leq \frac{\sigma_0^2 \chi^2_{1-\frac{a}{2};N-1}}{N-1}$$

- einseitige Tests zu

$$\frac{\sigma_0^2 \chi^2_{a;N-1}}{N-1} \leq s^2$$

Merke:

Die Prüfgröße ist:

$$s^2 \leq \frac{\sigma_0^2 \chi^2_{1-a;N-1}}{N-1}$$

Wobei σ_0^2 die angenommene Varianz ist und $\chi^2_{a;n-1}$ die

Verteilung mit N -1 Freiheitsgraden zum Niveau a.

Beispiel:

Es soll getestet werden, ob die Varianz einer Grundgesamtheit von 25 verschieden ist.

Es wird eine Stichprobe vom Umfang $N = 36$ genommen und der Test soll zum Niveau $a = 10\%$ durchgeführt werden.

Lösung:

Für die Null- und die Alternativhypothese gilt:

$H_0: \sigma^2 = \sigma_0^2 \quad , \quad H_1: \sigma^2 \neq \sigma_0^2$

Dies ist ein zweiseitiger Test. Der Annahmebereich der Nullhypothese beträgt:

$$\frac{\sigma_0^2 \chi^2_{\frac{a}{2};N-1}}{N-1} \leq s^2 \leq \frac{\sigma_0^2 \chi^2_{1-\frac{a}{2};N-1}}{N-1}$$

$$\frac{25 * 22{,}47}{35} \leq s^2 \leq \frac{25 * 49{,}8}{35}$$

$$16{,}5 \leq s^2 \leq 35{,}57$$

Liegt also die Stichprobenvarianz außerhalb des Annahmebereichs, so kann die Nullhypothese verworfen werden.

Test über den Anteilswert einer dichotomen Grundgesamtheit

Oft kann die Binomialverteilung durch die Normalverteilung approximiert werden.

Zur Erinnerung:

Für große Stichproben gleicht die Binomialverteilung der Normalverteilung. Es gilt:

Wenn

$$N\,\pi \geq 5$$

und

$$N(1-\pi) \geq 5$$

,dann kann die Binomialverteilung durch die Normalverteilung approximiert werden (Mittelwert $\hat{\pi}$ und Varianz $\hat{\pi}(1-\hat{\pi})$).

In den Annahmebereich für den Mittelwert einer normalverteilten Grundgesamtheit müssen dann nur noch Mittelwert und Varianz eingesetzt werden und man erhält die oben angegebene Formel für den Annahmebereich des Anteilswertes.

$$\pi_0 - z_{1-\alpha/2}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{N}} \leq \hat{\pi} \leq \pi_0 + z_{1-\alpha/2}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{N}}$$

Merke:

Die Prüfgröße ist:

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1-\pi_0)}}$$

Z ist approximativ standardnormalverteilt.

Der Vorzeichentest bei einem stetigen Merkmal ohne Bindungen (Wilcoxon-Vorzeichen-Rangtest)

Beim Vorzeichentest (der in der Literatur oft einen ganz anderen Test bezeichnet, also Vorsicht) ist ein Test über den Median.

Als Prüfgröße wählt man:

$$T := \text{“}Anzahl\ der\ Werte, oberhalb\ von\ \delta_0\text{“}$$

,wobei δ_0 den hypothetischen Median bezeichnet.

T ist binomialverteilt mit $B(N; 0{,}5)$.

Unter der Annahme, dass $\delta_0$richtig ist, gilt dann:

$$E(T) = N/2$$

(Das bedeutet nichts anderes, als dass die Hälfte der Werte oberhalb des Medians liegen (Dabei gilt $P(X_i = \delta_0) = 0$, da es sich um ein stetiges Merkmal handelt)).

und

$$VAR(T) = N * 0{,}5(1 - 0{,}5) = \frac{N}{4}$$

Merke:

Die Prüfgröße ist nun:

$$Z = \frac{T - N/2}{\sqrt{N/4}}$$

Z ist approximativ standardnormalverteilt.

Beispiel:

Gegeben sind folgende Daten einer Stichprobe:

1,1	1,6	2	1,4	2,1	2,5	1

Es soll die Hypothese zu $a = 5\%$ getestet werden, daß der Median der Grundgesamtheit 1,5 sei. Unabhängig vom Stichprobenumfang sei davon auszugehen, dass T approximativ normalverteilt sei mit Erwartungswert 3,5 und Varianz 1,75.

Antwort:

Die Prüfgröße berechnet sich zu:

$$Z = \frac{4 - 3{,}5}{\sqrt{1{,}75}} = 0{,}38$$

Die kritischen Werte der Standardnormalverteilung für $a = 5\%$ liegen beim zweiseitigen Test bei $-1{,}96$ und $1{,}96$. Die Prüfgröße liegt daher im Annahmebereich und die Hypothese kann nicht abgelehnt werden.

Vorzeichentest bei beliebigen Merkmalen mit Bindungen

Hier ist es möglich, dass gilt: $P(X_i = \delta_0) \neq 0$.

Daher lässt man alle Werte mit $X_i = \delta_0$ weg und führt den wie oben beschrieben (ohne Bindungen) durch.

Symmetrietest

Beim Symmetrietest wird zunächst folgende Testgröße ermittelt:

$$W^+ = \sum R_n Z_n$$

, dabei ist:

$$R_n = rg\,|X_n - \delta_0|$$

$$Z_n = \begin{cases} 1, falls\ X_n > \delta_0 \\ 0\ sonst \end{cases}$$

Beispiel:

Ermittle W^+ zu dem folgenden Datensatz unter der Hypothese $\delta_0 = 1{,}6$:

1,1	1,5	2	1,4	2,1	2,5	1

Zunächst ermittelt man $|X_n - \delta_0|$ und Z_n :

	1,1	1,5	2	1,4	2,1	2,5	1
δ_0	1,60	1,60	1,60	1,60	1,60	1,60	1,60
$X_n - \delta_0$	-0,50	-0,10	0,40	-0,20	0,50	0,90	-0,60
$\|X_n - \delta_0\|$	0,5	0,1	0,4	0,2	0,5	0,9	0,6
Z_n	0	0	1	0	1	1	0

$$W^+ = 0{,}4 + 0{,}5 + 0{,}9$$

Ende des Beispiels.

Für $N > 20$ ist W^+ normalverteilt mit Erwartungswert

$$E(W^+) = \frac{N(N+1)}{4}$$

und Varianz

$$VAR(W^+) = \frac{N(N+1)(2N+1)}{24}$$

Die Prüfgröße ist dann:

$$Z = \frac{W^+ - E(W^+)}{\sqrt{Var(W^+)}}$$

Z ist approximativ standardnormalverteilt und die kritischen Werte werden aus der Standardnormalverteilung ermittelt.

Chi-Quadrat Anpassungstest

Bei dem sogenannten Chi Quadrat Anpassungstest wird getestet, ob eine Grundgesamtheit einer bestimmten Verteilung folgt. Dafür werden die in der Stichprobe beobachteten Häufigkeiten h_{oi}mit den Häufigkeiten der hypothetischen Verteilung h_{ei} verglichen.

Merke:

Als Prüfgröße nutzt man:

$$\chi_*^2 = \sum_{i=1}^{m} \frac{(h_{oi} - h_{ei})^2}{h_{ei}}$$

Dabei ist m die Anzahl an Merkmalsausprägungen (bzw. Merkmalsklassen bei stetigen Merkmalen). Es muss gelten $h_{ei} > 5$, ansonsten kann die Verteilung nicht verwendet werden.

Dies ist ein einseitiger Test und die obere Annahmegrenze ist der Wert der Chi Quadrat Verteilung mit $m - 1$ Freiheitsgraden und Signifikanzniveau $1 - a$.

Beispiel:

Es sollgeprüft werden, ob eine Verteilung nicht der Gleichverteilung folgt. Als Signifikanzniveau wird $a = 10\%$ bestimmt. Eine Stichprobe liefert folgende Ergebnisse:

Merkmalsausprägung	h_{oi}	h_{ei}	$h_{oi} - h_{ei}$	$(h_{oi} - h_{ei})^2$	$(h_{oi} - h_{ei})^2 / h_{ei}$
1	8	10	-2	4	0,4
2	9	10	-1	1	0,1
3	7	10	-3	9	0,9
4	11	10	1	1	0,1
5	12	10	2	4	0,4
6	10	10	0	0	0
7	9	10	-1	1	0,1
8	9	10	-1	1	0,1
				Summe:	2,1

Als Nullhypothese gilt H_0: Die Verteilung ist eine Gleichverteilung gegen die

Alternativhypothese: H_1: Die Verteilung ist keine Gleichverteilung.

Die Prüfgröße nimmt den Wert 2,1 an. Die obere Grenze für die Nullhypothese ist

$$\chi^2_{(1-a;m-1)} = 12{,}02 > 2{,}1$$

Die Annahme der Gleichverteilung kann also nicht abgelehnt werden.

13.3 Test für den Vergleich zweier Mittelwerte

Es soll getestet werden, ob zwei Grundgesamtheiten denselben Mittelwert haben.

Als Testgröße für die Differenz zweier Mittelwerte wird natürlich die Differenz der beiden Stichprobenmittelwerte genommen:

Merke: Die Prüfgröße ist immer:

$$\frac{\bar{X} - \bar{Y} - \delta_0}{S}$$

,wobei δ_0 die hypothetische Differenz zwischen $\bar{X}$ und $\bar{Y}$ist.

Weiter müssen folgende vier Fälle zur Varianz der beiden Grundgesamtheiten unterschieden werden:

1)Bekannte Standardabweichungen, X und Y normalverteilt

Dies ist der einfachste Fall. Der Annahmebereich der Nullhypothese ist dann:

$$-z_{1-\frac{a}{2}} * \sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}} \leq D \leq z_{1-\frac{a}{2}} * \sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}$$

Merke: Die Prüfgröße ist:

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{S}$$

S ist in diesem Fall

$$S = \sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}$$

Z ist standardnormalverteilt.

Beispiel:

Aus zwei Grundgesamtheiten wird eine Stichprobe von jeweils 16 gezogen. Die Varianz der ersten Grundgesamtheit ist 25, die der zweiten ist 36. Teste mit $a = 5\%$ ob die beiden Mittelwerte verschieden sind.

Lösung:

Als Hypothesen stellt man auf:

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

Der Annahmebereich berechnet sich zu:

$$-1{,}96 * \sqrt{\frac{25}{16} + \frac{36}{16}} \leq D \leq 1{,}96 * \sqrt{\frac{25}{16} + \frac{36}{16}}$$

Fällt D außerhalb von diesem Intervall, so wird H_0 verworfen.

2) Unbekannte aber gleiche Standardabweichungen, X und Y normalverteilt

In diesem Fall gilt für die Varianz S:

$$S = \sqrt{\left(\frac{1}{N_x} + \frac{1}{N_y}\right)\frac{(N_x - 1)S_x^2 + (N_y - 1)S_y^2}{N_x + N_y - 2}}$$

Die Prüfgröße ist

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{S}$$

T ist t-verteilt mit $N_x + N_y - 2$ Freiheitsgraden.

3) Unbekannte und ungleiche Standardabweichungen, X und Y normalverteilt

In diesem Fall gilt für die Varianz S:

$$S = \sqrt{\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}}$$

Die Prüfgröße ist

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{S}$$

T ist t-verteilt mit k Freiheitsgraden. Für k gilt:

$$k = \frac{\left(\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}\right)^2}{\frac{\left(\frac{s_x^2}{N_x}\right)^2}{N_x - 1} + \frac{\left(\frac{s_y^2}{N_y}\right)^2}{N_y - 1}}$$

4) X und Y beliebig verteilt

In diesem Fall gilt für die Varianz S:

$$S = \sqrt{\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}}$$

Die Prüfgröße ist

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{S}$$

T ist approximativ standardnormalverteilt für $N_x, N_y > 30$.

Vergleich der Anteilswerte zwei dichotomer Merkmale

Ohne Herleitung oder Begründung reicht es aus, die gegebenen Werte in die Testgröße einzusetzen. Diese ist:

$$Z = \frac{\hat{\pi}_x - \hat{\pi}_y - \delta_0}{\sqrt{S_x^2 + S_y^2}}$$

, wobei gilt:

$$S_x^2 = \frac{\hat{\pi}_x(1 - \hat{\pi}_x)}{N}$$

$$S_y^2 = \frac{\hat{\pi}_y(1 - \hat{\pi}_y)}{M}$$

Dabei sind $\hat{\pi}_x$, $\hat{\pi}_y$ die Anteilswerte, die aus der Stichprobe geschätzt wurden, N und M die entsprechenden Stichprobenumfänge. δ_0 ist wieder die hypothetische Differenz der beiden Anteilswerte.

Sind N und M jeweils über 30, so werden die Quantile der Standardnormalverteilung als kritische Werte genutzt.

Beispiel:

Es werden zwei Stichproben aus dichotomen Grundgesamtheiten gezogen. Der Stichprobenumfang ist jeweils 50.

Aus der Stichprobe ergeben sich folgende Werte:

$$\hat{\pi}_x = 0{,}45$$

$$\hat{\pi}_y = 0{,}4$$

Es soll durch einen Test zu $a = 5\%$ bestätigt werden, daß $\pi_x > \pi_y$ gilt.

Antwort:

Es werden folgende Hypothesen aufgestellt:

$$H_0: \pi_x - \pi_y \leq 0$$

gegen

$$H_1: \pi_x - \pi_y > 0$$

Die Teststatistik berechnet sich nach:

$$Z = \frac{\hat{\pi}_x - \hat{\pi}_y - \delta_0}{\sqrt{S_x^2 + S_y^2}}$$

mit

$$S_x^2 = \frac{\hat{\pi}_x(1 - \hat{\pi}_x)}{N} = \frac{0{,}45(1 - 0{,}45)}{50} = 0{,}00495$$

$$S_y^2 = \frac{\hat{\pi}_y(1 - \hat{\pi}_y)}{M} = \frac{0{,}4(1 - 0{,}6)}{50} = 0{,}0048$$

$$Z = \frac{0{,}45 - 0{,}4}{\sqrt{0{,}00495 + 0{,}0048}} = 0{,}5$$

Der kritische z-Wert für den einseitigen Test ist 1,65. Da Z unter diesem Wert liegt, kann die Nullhypothese nicht abgelehnt werden.

Test für den Vergleich zweier Varianzen

Testet man die Varianzen von zwei Grundgesamtheiten auf Gleichheit, so ist die Prüfgröße

$$F^* = \frac{s_1^2}{s_2^2}$$

Diese folgt einer F-Verteilung mit $N-1$ und $M-1$ Freiheitsgraden. Was eine F-Verteilung ist, oder wie man sie berechnet, brauchst du nicht zu wissen, nur wie man Werte aus der Tabelle der F-Verteilung abliest!

Der Annahmebereich der Nullhypothese für F^* ist

$$\frac{1}{F_{1-\frac{a}{2};N-1;M-1}} \leq F^* \leq F_{1-\frac{a}{2};N-1;M-1}$$

Beispiel: Die Varianzen zweier Verteilungen sollen auf Ungleichheit geprüft werden. Eine Stichprobe liefert als Ergebnis $\frac{s_1^2}{s_2^2} = 0{,}75$. Getestet werden soll zum Niveau $a = 10\%$. Der Stichprobenumfang war jeweils 20.

Als Nullhypothese formuliert man:

$H_0: \frac{s_1^2}{s_2^2} = 0$, $H_1: \frac{s_1^2}{s_2^2} \neq 0$

Der Annahmebereich beträgt

$$\frac{1}{F_{0{,}95;19;19}} \leq F^* \leq F_{0{,}95;19;19}$$

bzw.

$$0{,}46 \leq F^* \leq 2{,}17$$

Wilcoxon Rangsummen Test

Bei diesem Test wird die Gleichheit der Verteilung der Ränge zweier Stichproben untersucht. Dazu zieht man zwei Stichproben und fügt diese zusammen. Dann verteilt man Ränge in dieser zusammengefassten Menge und ermittelt die Rangsumme der Werte der ersten Stichprobe.

Mathematisch schreibt man:

$$T_W = \sum_n rg(X_i)$$

Beispiel:

Es werden zwei Stichproben gezogen:

1. Stichprobe: {1,4,5,8}

2. Stichprobe:{2,3,5,10}

Die Stichproben werden zu einer Menge zusammengefasst und folgende Ränge werden vergeben:

Ausprägung	1	2	3	4	5	8	10
Rang	1	2	3	4	5	6	7

Gibt es“ Bindungen“, treten also Ausprägungen doppelt auf, so wird der Mittelwert der zu vergebenen Ränge als Rang für beide verwendet.

Die Werte der ersten Stichprobe haben also folgende Ränge:

1,4,5,6

Die Rangsumme ist 16.

Die Prüfgröße ist nun:

$$Z = \frac{T_W - E[T_W]}{\sqrt{Var[T_W]}}$$

,wobei gilt:

$$E[T_W] = \frac{N(N+M+1)}{2}$$

$$\sqrt{Var[T_W]} = \sqrt{\frac{NM(N+M+1)}{12}}$$

N und M sind die Stichprobenumfänge der ersten und zweiten Stichprobe. Die Testgröße ist approximativ standardnormalverteilt und es gelten die entsprechenden kritischen Werte.

Abhängige Stichproben

Sind zwei Merkmale voneinander abhängig, so wird durch den Kovarianzterm die Teststatistik verfälscht. Dies erkennt man schnell an der Formel für die Varianz der Summe zweier Zufallsvariablen:

$$Var(A+B) = Var(A) + Var(B) - 2Cov(A,B)$$

(Sind die beiden Zufallsvariablen unabhängig, so beträgt die Kovarianz Null.)

Um diesen Sachverhalt zu umgehen, nutzt man unter anderem den Pre-Test/Post-Test.

Pre-Test/ Post-Test

Bei diesem Test werden die gleichen Stichproben zu zwei verschiedenen Zeitpunkten durchgeführt. Dies kann z.B. sinnvoll sein, um die Wirkung einer bestimmten Maßnahme zu ermitteln. Es wird dann die Veränderung auf Signifikanz geprüft.

Mathematisch schreibt man:

$$D_n = X_n - Y_n$$

Die Testgröße ist der Mittelwert dieser Differenzen:

$$\overline{D} = \frac{1}{n}\sum_{n} D_n$$

Die Prüfgröße ist dann:

$$T = \frac{\overline{D} - \delta_0}{S_D/\sqrt{N}}$$

, wobei gilt:

$$S_D = \sqrt{S_x^2 + S_Y^2 - 2S_x S_y R_{xy}}$$

R_{xy} ist die Korrelation zwischen X und Y.

δ_0 ist die hypothetische Differenz.

Die Prüfgröße T ist t-verteilt mit N-1 Freiheitsgraden und es gelten die entsprechenden kritischen Werte.

Beispiel:

Ein Unternehmen möchte den Erfolg einer Werbemaßnahme testen und befragt dazu 16 Personen vor und nach der Werbemaßnahme nach der Wahrnehmung der beworbenen Marke (Skala 1-100).

$X:=$ Wahrnehmung vor der Werbemaßnahme

$Y:=$ Wahrnehmung nach der Werbemaßnahme

Aus der Stichprobe ergeben sich folgende Werte:

$$\bar{x} = 58$$

$$\bar{y} = 55$$

$$s_x = 25$$

$$s_y = 20$$

$$\bar{d} = \bar{x} - \bar{y} = 0{,}5$$

$$s_d = 8$$

Es soll zum Niveau 5% geprüft werden, ob die Veränderung der Wahrnehmung von Null verschieden ist.

Die Prüfgröße ergibt sich nun zu:

$$T = \frac{\overline{D} - \delta_0}{S_D/\sqrt{N}} = \frac{3 - 0}{8/\sqrt{16}} = 1{,}5$$

Die kritischen Werte der t-Verteilung für $a = 5\%$ und $N = 15$ sind $-2{,}131$ und $2{,}131$.

Da geprüft werden soll, ob die Wahrnehmung sich geändert hat, ist die Nullhypothese:

$$H_0: \overline{D} - \delta_0 = 0$$

Gegen

$$H_1: \overline{D} - \delta_0 \neq 0$$

Da T aber in den Annahmebereich fällt, kann die Nullhypothese nicht abgelehnt werden.

Überschreitungswahrscheinlichkeit

Zum Ende des Kapitels der verschiedenen Tests noch eine Kleinigkeit zum Verständnis: Die Überschreitungswahrscheinlichkeit p bezeichnet die Wahrscheinlichkeit ausserhalb der gemessenen Teststatistik.

Kann eine Nullhypothese nicht abgelehnt werden, so könnte man fragen, bei welchem a die Nullhypothese abgelehnt worden wäre. Genau dieses a ist die Überschreitungswahrscheinlichkeit p.

Zusammenhangsanalyse

In diesem Abschnitt geht es um die Korrelation zwischen zwei Merkmalen.

An dieser Stelle möchte ich kurz einen Abschnitt aus „Grundlagen der Wirtschaftsmathematik und Statistik wiederholen:

Kovarianz

Ein wichtiger Parameter zweidimensionaler Merkmalsausprägungen ist die sogenannte Kovarianz. Die Kovarianz ist ein Parameter für die positive oder negative Abhängigkeit der beiden Merkmale.

Mathematisch wird die Kovarianz wie folgt ermittelt:

$$Cov(X,Y) = \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{r}(x_i - \bar{x}) * (y_k - \bar{y}) * h(x_j, y_k)$$

Man kann zur Berechnung auch die relativen Häufigkeiten verwenden:

$$Cov(X,Y) = \sum_{j=1}^{m}\sum_{k=1}^{r}(x_i - \bar{x}) * (y_k - \bar{y}) * f(x_j, y_k)$$

Es ist wichtig, dass du verstehst, was hier berechnet wird: Es wird das Produkt aus den Abweichungen der jeweiligen Ausprägungen von ihren Mittelwerten gebildet. Dieses Produkt wird mit seiner relativen Häufigkeit multipliziert und diese Produkte werden aufsummiert. Wenn also das zweite Merkmal immer dann nach oben von seinem Mittelwert abweicht, wenn auch das erste nach oben abweicht (und entsprechend immer nach unten, wenn auch das erste nach unten abweicht), dann ist das Produkt immer positiv und die Summe über alle Produkte wird größer. Weicht das zweite Merkmal nach unten von seinem Mittelwert ab, wenn das erste nach oben abweicht (und entsprechend andersrum), so wird das Produkt negativ und die Summe der Produkte entsprechend kleiner.

Beispiel: Die Kovarianz der folgenden Werte berechnet sich wie folgt:

X	1	5	6
Y	2	4	6

Zunächst werden die Mittelwerte ermittelt:

$$\bar{x} = \frac{1+5+6}{3} = 4$$

$$\bar{y} = \frac{2+4+6}{3} = 4$$

Dann ist die Kovarianz:

$$Cov(X,Y) = \frac{1}{3}\big((1-4)*(2-4)*1 + (5-4)*(4-4)*1 + (6-4)*(6-4)*1\big) = \frac{10}{3}$$

Es ist wichtig zu sehen, dass die absolute Höhe der Kovarianz keine Aussage über das Maß des Zusammenhangs der beiden Merkmale macht. Für die einzelne Stichprobe gilt natürlich: Je höher die Kovarianz, desto höher der Zusammenhang, aber alleine aus der Aussage: „Kovarianz 5“ kann man nicht ablesen ob ein starker oder ein schwacher Zusammenhang besteht.

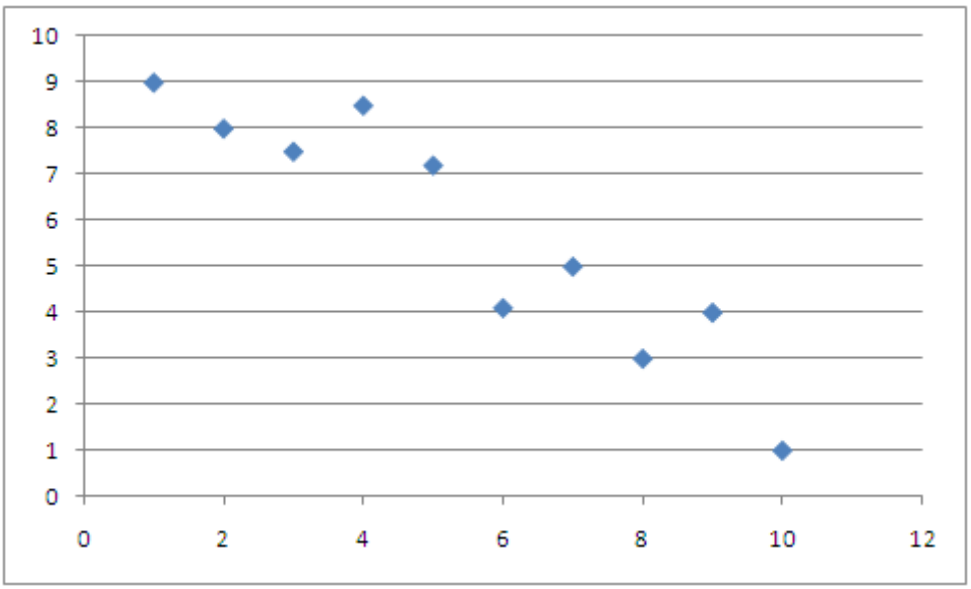
Folgendes **Beispiel** soll dies verdeutlichen:

Wir nehmen die Werte des letzten Beispiels und multiplizieren diese mit 2. Der Zusammenhang bleibt dabei völlig unbeeinflusst und trotzdem steigt der Wert der Kovarianz. (Die Berechnung überlasse ich dir).

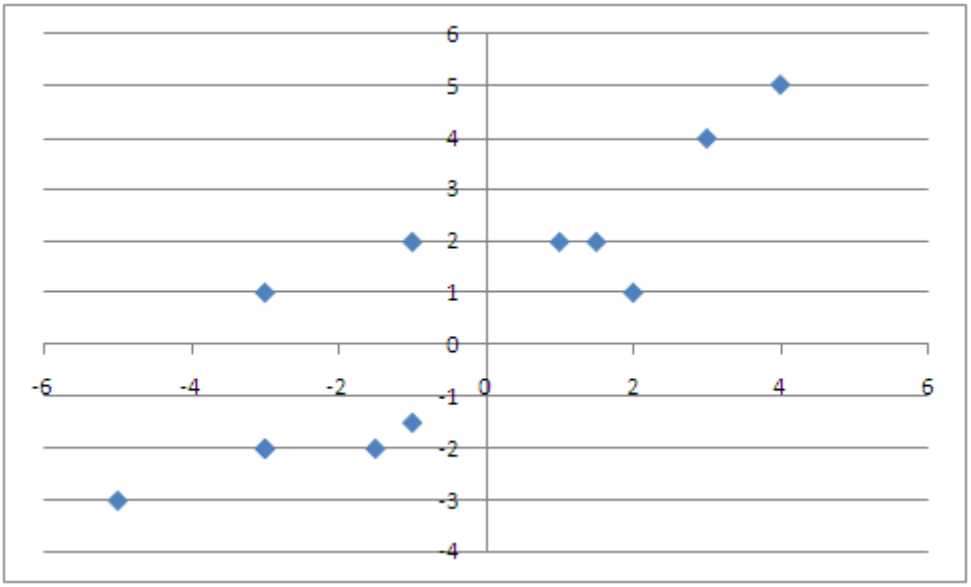
Unabhängige und abhängige Merkmale

Ob zwei Merkmale voneinander abhängig sind, kann man gut grafisch anhand des Streuungsdiagramms erkennen:

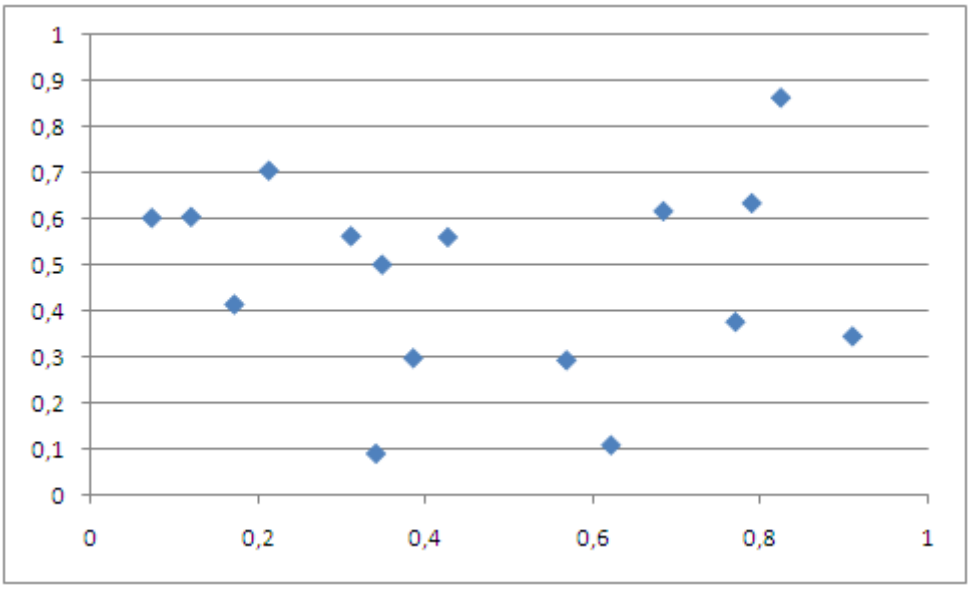
Negative Abhängigkeit:



Positive Abhängigkeit:



Unabhängigkeit:



Die Unabhängigkeit zweier Merkmale kann auch anhand der bedingten Verteilungen überprüft werden. Sollte die Ausprägung des Merkmals A keine Auswirkung auf die Ausprägung des Merkmals B haben, so ist die relative Verteilung des Merkmals B gleich der bedingten relativen Verteilung des Merkmals B für alle Ausprägungen des Merkmals A. Um den Satz zu verstehen, hier ein **Beispiel:**

Angenommen die Abiturnote (Merkmal A) und die Universitätsnote sind voneinander unabhängig. Dann gilt: Betrachte ich alle Universitätsnoten von StudentInnen mit einer Abiturnote von 1 und vergleiche diese mit den Universitätsnoten von StudentInnen mit der Abinote 2, so gibt es keinen Unterschied. Für beliebige Abinoten ändert sich die relative Verteilung der Universitätsnoten nicht!

Wären die beiden Merkmale positiv voneinander abhängig, so würde die relative Verteilung der Universitätsnoten der StudentInnen mit Abinote 1 besser sein als die relative Verteilung der Universitätsnoten der StudentInnen mit Abinote 2.

Formal schreibt man:

$$f(x|y_k)$$

Da die Verteilung von X unabhängig von der speziellen Ausprägung von y_k ist, ist natürlich auch die spezielle Ausprägung von X unabhängig. Formal schreibt man

$$f(x_j|y_k)$$

Korrelationsrechnung

Bestimmung der Stärke des Zusammenhangs.

Bei der Bestimmung der Stärke des Zusammenhangs zweier Merkmale muss zwischen linearem Zusammenhang und nichtlinearem Zusammenhang unterschieden werden.

Korrelation zweier Merkmale mit linearem Zusammenhang

Der lineare Zusammenhang zwischen zwei **quantitativen** Merkmalen wird über den **Pearsonschen Korrelationskoeffizienten** gemessen. Um aus der Kovarianz (siehe Kapitel 7.5) eine aussagefähige Maßzahl zu berechnen, wird diese durch das Produkt der Standardabweichungen der beiden Merkmale geteilt. Wie in Kapitel 7.5 erläutert, steigt die Kovarianz einerseits mit dem Zusammenhang der beiden Merkmale, aber auch proportional mit ihren Standardabweichungen. Da wir uns nur für den Zusammenhang interessieren, werden die Standardabweichungen eliminiert (indem durch die Standardabweichungen geteilt wird). Mathematisch schreibt man:

$$r = \frac{Cov(X,Y)}{\tilde{s}_x * \tilde{s}_y}$$

Wobei $\tilde{s}_y$ für die geschätzte Standardabweichung des Merkmals Y steht.

Nachdem die Standardabweichungen eliminiert wurden, ergibt sich eine Maßzahl für die Stärke des linearen Zusammenhangs, die zwischen -1 und 1 liegt. Ein Pearsonscher Korrelationskoeffizient von 1(-1) bedeutet dabei, dass die beiden Merkmale perfekt positiv (negativ) zusammenhängen (korreliert sind). Bei einem Korrelationskoeffizienten von 0 sind sie unkorreliert.

Beispiel: Folgende Tabelle stellt die wöchentlichen Renditen von 2 Aktien dar. Berechne den Pearsonschen Korrelationskoeffizienten.

Aktie A (A)	5	2	-1	2	12
Aktie B (B)	3	1	-2	-1	4

Hierzu berechnet man zunächst die Standardabweichungen der beiden Merkmale, sowie deren Kovarianz:

$$\tilde{s}_A = 4{,}43$$

$$\tilde{s}_B = 2{,}28$$

$$Cov(A,B) = 9$$

Eingesetzt in $r = \frac{Cov(X,Y)}{\tilde{s}_x * \tilde{s}_y}$ ergibt:

$$\frac{9}{4{,}43 * 2{,}28} = 0{,}89$$

Bestimmtheitsmaß

Das Bestimmtheitsmaß ist das Quadrat des Pearsonschen Korrelationskoeffizienten. Wie man mathematisch zeigen kann (Herleitung siehe Fernuni-Skript), entspricht das Bestimmtheitsmaß dem Quotienten aus der Varianz der durch lineare Regression geschätzten Y-Werte und der Varianz der beobachteten Y-Werte.

$$r^2 = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2}$$

Interpretation des Bestimmtheitsmaßes

Das Bestimmtheitsmaß gibt an, welcher Anteil an Streuung (Varianz) der beobachteten Y-Werte durch die Varianz der geschätzten Y-Werte erklärt wird. Der Ausdruck „erklärt“ ist etwas hier etwas missverständlich. Es soll bedeuten, dass dieser Anteil an Varianz durch den Einfluss des Merkmals X herbeigeführt wird. Angenommen das Merkmal Y hängt zu 50% von dem Merkmal X ab und zu 50% von einem dritten unbekannten Merkmal. In diesem Fall würde das Bestimmtheitsmaß bezüglich X und Y 50% (0,5) betragen.

Hängt Y ausschließlich von X ab, so beträgt das Bestimmtheitsmaß 1. Da die gesamte Änderung des Merkmals Y von der Änderung des Merkmals X abhängt, liegen alle tatsächlich beobachteten Werte auch auf der Regressionsgerade.

Korrelation zweier Merkmale mit nichtlinearem Zusammenhang

Da der Pearsonsche Korrelationskoeffizient für Merkmale mit nichlinearem Zusammenhang **nicht** nach der Formel

$$r = \frac{Cov(X,Y)}{\tilde{s}_x * \tilde{s}_y}$$

berechnet werden kann, muß er aus dem Bestimmtheitsmaß ermittelt werden. Da gilt

$$r^2 = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2},$$

muss dafür $\tilde{s}_{\hat{y}}^2$ durch nichtlineare Kleinste-Quadrate-Regression (siehe Kapitel 7.7.2) ermittelt werden. Der Pearsonsche Korrelationskoeffizient ergibt sich dann aus der Wurzel des Bestimmtheitsmaßes.

Signifikanztest für den Korrelationskoeffizienten

Der folgende Test prüft, ob zwei Merkmale einen Korrelationskoeffizienten ungleich von 0 haben. Hier heißt es wieder: Nicht verstehen sondern stupide anwenden!

Für den Test

$$H_0: p = 0$$

$$H_1: p \neq 0$$

nutzt man die Teststatistik:

$$T = \sqrt{N-2}\frac{R}{\sqrt{1-R^2}}$$

T ist t-verteilt mit N-2 Freiheitsgraden.

Beispiel:

Gegeben sind folgende Daten aus einer Stichprobe vom Umfang 38:

$$r = 0{,}4$$

Ein Test zu $a = 5\%$ ergibt:

$$T = \sqrt{38-2}\frac{0{,}4}{\sqrt{1-0{,}4^2}} = 3{,}56$$

Der Wert liegt deutlich außerhalb des Annahmebereiches der t-Verteilung für $a = 5\%$ und $N = 38$. Die Nullhypothese kann daher abgelehnt werden.

Fisher-Z-Statistik

Möchte man den Korrelationskoeffizienten nicht aus Null, sondern auf einen konkreten Wert p_0 testen, so nutzt man als Prüfgröße:

$$Z = \frac{\sqrt{N-3}}{2}\left(\ln\frac{1+R}{1-R} - \ln\frac{1+p_0}{1-p_0}\right)$$

Z ist für große Stichprobenumfänge approximativ standardnormalverteilt.

Beispiel:

Aus einer Stichprobe vom Umfang 39 ergeben sich r=0,7. Es soll die Hypothese $H_{0:}p = p_0 = 0{,}75$ zu $a = 5\%$ geprüft werden.

Als Prüfgröße erhält man:

$$Z = \frac{\sqrt{39-3}}{2}\left(\ln\frac{1+0{,}7}{1-0{,}7} - \ln\frac{1+0{,}75}{1-0{,}75}\right)$$

Das Ausrechnen überlasse ich Dir. Der Annahmebereich ist -1,96 bis 1,96.

Spearmansche Rangkorrelationskoeffizient

Sollten Merkmale nicht quantitativ, sondern nur ordinal, also dem Rang nach, sortiert werden können, so kann die Korrelation der beiden Merkmale durch den Spearmanschen Rangkorrelationskoeffizienten ermittelt werden.

Der Spearmansche Rangkorrelationskoeffizient ist eigentlich nichts weiter als der Pearsonsche Korrelationskoeffizient, nur dass statt der quantitativen Merkmalsausprägungen die Rangzahlen genommen werden. Man kann beide Koeffizienten auch für beide Arten von Merkmalsausprägungen nutzen - die Aussagekraft ist dann aber geringer.

Mathematisch schreibt man:

$$r_s = \frac{\sum_{i=1}^{n}\left(rg(x_i) - \overline{rg}_X\right)\left(rg(y_i) - \overline{rg}_Y\right)}{\sqrt{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_X)^2 \sum_{i=1}^{n}\left(rg(y_i) - \overline{rg}_Y\right)^2}}$$

Haben mehrere Beobachtungswerte denselben Wert, so nennt man diese **Bindungen**. Für diese Werte wird der durchschnittliche Rang der entsprechenden Ränge vergeben. Das folgende Beispiel verdeutlicht das Vorgehen bei Bindungen:

Beispiel:

5 Studenten treten im 100 Meter Sprint und im Weitsprung gegeneinander an. Es werden folgende Zeiten/Weiten erzielt:

Student Nr.	1	2	3	4	5
Sprint (X)	12,4	12,2	15	12,4	13,2
Weitsprung (Y)	4	4,5	5,2	6	6,2

Der Spearmansche Rangkorrelationskoeffizient berechnet sich nun wie folgt:

Zunächst müssen die Ränge berechnet werden:

Student Nr.	1	2	3	4	5
Sprint (X)	2,5	1	5	2,5	4
Weitsprung (Y)	5	4	3	2	1

Da beim Sprint zwei Studenten dieselbe Zeit haben, liegen Bindungen vor. Beide erhalten als Rangzahl das arithmetische Mittel der beiden zu vergebenen Ränge. Nun müssen die Mittelwerte ermittelt werden:

$$\overline{rg}_X = \overline{rg}_Y = 3$$

Der Unabhängigkeitstest

Mit dem Unabhängigkeitstest wird die Abhängigkeit zweier Merkmale X und Y getestet. Dazu werden wieder die eingetretenen und die zu erwartenden Häufigkeiten miteinander verglichen. Die bei Unabhängigkeit zu erwartende Häufigkeit ergibt sich für die Randhäufigkeiten $h(x_i)$ und $h(y_j)$ zu:

$$h_{eij} = \frac{h(x_i) * h(x_j)}{n}$$

Die Testgröße ist dann:

$$\chi_*^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(h_{oij} - h_{eij})^2}{h_{eij}}$$

wobei r die Anzahl an Ausprägungen von X und s die Anzahl an Ausprägungen von Y bezeichnet.

Die Nullhypothese ist immer die Unabhängigkeit der beiden Merkmale. Der Ablehnungsbereich für die Nullhypothese ist

$$\chi_*^2 > \chi_{(1-a,v)}^2$$

wobei v die Anzahl an Freiheitsgeraden bezeichnet. Diese berechnen sich zu $v = (r-1) * (s-1)$.

H_0 wird also abgelehnt, wenn die Testgröße größer ausfällt als $\chi_{(1-a,v)}^2$.

Beispiel:

Die folgenden Werte sollen auf statistische Abhängigkeit zu $a = 10\%$ geprüft werden:

$Y\backslash X$	x_1	x_2	x_3	Summe
y_1	50	30	20	100
y_2	15	10	35	60
y_3	15	10	15	40
Summe	8	50	70	200

Die bei Unabhängigkeit zu erwartenden Werte wären:

$Y\backslash X$	x_1	x_2	x_3	Summe
y_1	40	25	35	100
y_2	24	15	21	60
y_3	16	10	14	40
Summe	80	50	70	200

Als Differenz erhält man:

$Y\backslash X$	x_1	x_2	x_3
y_1	10	5	-15
y_2	-9	-5	14
y_3	-1	0	1

und für $\frac{(h_{oij}-h_{eij})^2}{h_{eij}}$:

$j\backslash i$	$i=1$	$i=2$	$i=3$	Summe
$j=1$	0,20	0,08	1,13	1,41
$j=2$	0,54	0,25	0,56	1,35
$j=3$	0,01	0,00	0,01	0,01
Summe	0,75	0,33	1,69	2,77

$j\backslash i$	$i=1$	$i=2$	$i=3$	Summe
$j=1$	2,50	1,00	6,43	9,93
$j=2$	3,38	1,67	9,33	14,38
$j=3$	0,06	0,00	0,07	0,13
Summe	5,94	2,67	15,83	24,44

Für χ_*^2 ergibt sich also ein Wert von 24,44. Die Nullhypothese der Unabhängigkeit kann nun abgelehnt werden, wenn dieser Wert größer ist als $\chi^2_{(1-a,v)}$.

Für $v = 2 * 2 = 4$ Freiheitsgrade erhält man:

$$\chi^2_{(0,9;4)} = 7{,}78$$

Daher kann die Nullhypothese abgelehnt werden.

Nun müssen die Werte nur noch in die Formel für den Spearmanschen Rangkorrelationskoeffizienten eingesetzt werden. Im Einzelnen ergibt sich:

$$\sum_{i=1}^{n}\left(rg(x_i)-\overline{rg}_X\right)\left(rg(y_i)-\overline{rg}_Y\right)$$
$$=(2{,}5-3)*(5-3)+(1-3)*(4-3)+(5-3)*(3-3)+(2{,}5-3)$$
$$*(2-3)+(4-3)*(1-3)=-4{,}5$$

$$\sum_{i=1}^{n}\left(rg(x_i)-\overline{rg}_X\right)^2=9{,}5$$

$$\sum_{i=1}^{n}\left(rg(y_i)-\overline{rg}_Y\right)^2=10$$

$$\frac{-4{,}5}{\sqrt{9{,}5*10}}=-0{,}46$$

Also ein negativer Zusammenhang.

Liegen keine Bindungen vor, so vereinfacht sich die Berechnung von r_s zu:

$$r_s=1-\frac{6\sum d_i^2}{n(n^2-1)}$$

Wobei d_i den Rangdifferenzen $rg(x_i)-rg(y_i)$ entspricht.

13.0 Regressionsanalyse

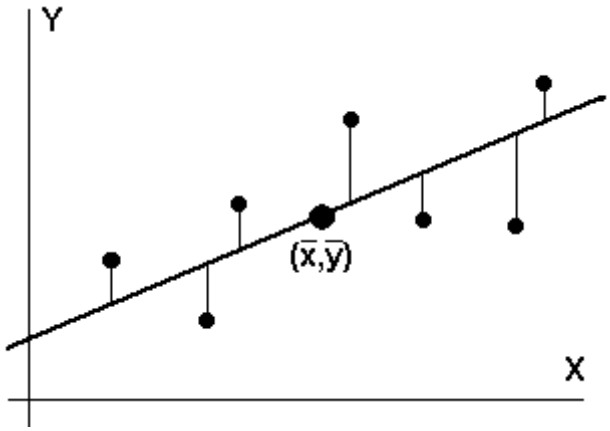
Lineare Kleinste Quadrate Regression

Ich werde in diesem Kapitel zunächst die Lineare Kleinste Quadrate Regression kurz wiederholen und dabei die wichtigsten Formeln auflisten.

Jetzt zur Materie:

Bei der linearen Kleinste Quadrate Regression liegt eine Datenmenge vor und es ist das Ziel eine Gerade durch diese Datenmenge zu ermitteln, die möglichst nahe an allen Daten liegt. Optimal wäre es, wenn alle Punkte der Datenmenge auf der Geraden liegen würden. Mit Hilfe dieser Gerade werden dann Y-Werte zu X-Werten, die nicht in der Stichprobe lagen, prognostiziert

Grafisch sieht das so aus:



Ein oft verwendetes Beispiel ist die Entwicklung des BSP mit der Zeit. Mit Hilfe der Regressionsgerade wird dann das zukünftige BSP geschätzt.

Für den relativ einfachen Fall der linearen Kleinste Quadrate Regression müssen die Koeffizienten der Gleichung

$$f(x) = a + bx$$

geschätzt werden. Im Folgenden werde ich die folgende Notation für die Regressionsgleichung verwenden:

$$\hat{y} = a + bx$$

Das Dach über dem y zeigt an, dass es sich um eine Schätzung handelt und nicht um eine tatsächliche Ausprägung.

Die Schätzung der Koeffizienten geschieht folgendermaßen:

Die Summe der quadrierten Abweichungen zwischen der Regressionsfunktion und den Ausprägungen beträgt

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

(n steht immer für den Stichprobenumfang)

Diese quadrierten Abweichungen zwischen Ausprägungen und der Regressionsgerade gilt es zu minimieren. Diese werden minimal, wenn die ersten Ableitungen nach a und b gleich Null sind.

Auch wenn die Herleitung der Regressionskoeffizienten im Skript nicht angegeben ist, wurde sie schon in Klausuren geprüft. Daher solltest du diese beherrschen.

Indem man die ersten Ableitungen nach a und b gleich Null setzt, erhält man ein Gleichungssystem mit zwei Gleichungen und zwei Unbekannten.

$$\frac{df(a,b)}{da} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = 0$$

$$\frac{df(a,b)}{db} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) = 0$$

Nach einigen einfachen Umformungen erhält man:

Gleichung 1:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

Gleichung 2:

$$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

Dieses Gleichungssystem wird nach a und b aufgelöst:

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Man erhält also:

$$a = \bar{y} - b\bar{x}$$

(erhält man, indem man Gleichung 1 durch n teilt)

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Für b kann man mathematisch auch schreiben:

$$b = R_{xy}\frac{S_y}{S_x}$$

,wobei R_{xy} der Korrelationskoeffizient zwischen X und Y ist.

Beispiel:

Zu den folgenden Wertepaaren sollen die Regressionskoeffizienten der Regressionsfunktion

$$\hat{y} = a + bx$$

geschätzt werden.

x	2	2	6	6
y	2	6	4	8

Es müssen zunächst die folgenden Werte ermittelt werden:

$$\bar{x} = 4$$

$$\bar{y} = 5$$

$$\sum x_i y_i = 2 * 2 + 2 * 6 + 6 * 4 + 6 * 8 = 88$$

$$\sum x_i^2 = 2 * 2 + 2 * 2 + 6 * 6 + 6 * 6 = 80$$

Eingesetzt in $a = \bar{y} - b\bar{x}$ und $b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$ erhält man:

$$a = 5 - 4b$$

$$b = \frac{88 - 4 * 4 * 5}{80 - 4 * 4 * 4} = \frac{8}{16} = 0{,}5$$

$$a = 3$$

Interpretation der Regressionskoeffizienten: Wie bei jeder Funktion gibt das Absolutglied a der Funktion

$$\hat{y} = a + bx$$

die Lage des Funktionsgraphen an, und der Koeffizient b gibt die Steigung des Funktionsgraphen an.

Kleinste Quadrate Schätzung

Hat man die Regressionskoeffizienten a und b bestimmt, so kann man eine Schätzfunktion aufstellen:

$$\widehat{Y_n} = \hat{a} + \hat{b}X_n$$

Die Differenz aus den tatsächlichen Werten und den geschätzten Werten nennt man Residuen und schreibt:

$$\widehat{\epsilon_n} = Y_n - \widehat{Y_n}$$

Da die Gerade genau in der Mitte der Werte liegt, addieren sich die Residuen zu Null (die Abweichungen nach oben und unter gleichen einander aus).

Die Varianz um die Regressionsgerade herum ist dann:

$$\widehat{\sigma^2} = \frac{1}{N-2}\sum_n \widehat{\epsilon_n}^2$$

Merke: Die Regressionsgerade geht immer durch die Mittelwerte $\bar{x}$ und $\bar{y}$.

Beispiel: Es seien die folgenden Daten gegeben:

x	2	4	6	8
y	2	5	7	10

Die Regressionsgerade lautet:

$$\hat{y} = -0{,}5 + 1{,}3x$$

Die Residuen $Y_n - \widehat{Y_n}$ ermittelt man aus folgender Tabelle:

X_n	2	4	6	8
Y_n	2	5	7	10
$\widehat{Y_n}$	-0,2	2,4	5	7,6
$Y_n - \widehat{Y_n}$	2,2	2,6	2	2,4

Aus den Residuen kann man nun die Varianz um die Regressionsgerade herum ermitteln.

X_n	2	4	6	8
Y_n	2	5	7	10
$\widehat{Y_n}$	2,1	4,7	7,3	9,9
$Y_n - \widehat{Y_n}$	-0,1	0,3	-0,3	0,1
$\left(Y_n - \widehat{Y_n}\right)^2$	0,01	0,09	0,09	0,01

Die letzte Zeile wird nun aufaddiert und durch N-2 geteilt. Man erhält:

$$\widehat{\sigma^2} = \frac{1}{N-2}\sum_n \widehat{\epsilon_n}^2 = 0{,}1$$

Annahmen des klassischen Regressionsmodells

Es gelten folgende Annahmen:

-die Residuen sind normalverteilt mit Erwartungswert Null und untereinander unkorreliert.

-die Varianz der Residuen ist im Zeitablauf konstant.

- $X_n$ist deterministisch, also extern vorgegeben.

Die Annahmen der Normalverteilung oder konstanten Varianz kann man oft grafisch erkennen. Steigt oder fällt die Streuung der Werte im Zeitablauf, so ist die Varianz nicht konstant. Sind nur wenige Werte in der Nähe des Mittelwertes und viele weit entfernt, so ist die Annahme der Normalverteilung verletzt.

Erwartungstreue der KQ-Schätzer a und b

Die Schätzer für a, b und σ^2sind erwartungstreu. Den mathematischen Beweis findest du im Fernuni-Skript. Es ist nicht auszuschließen, dass dieser geprüft wird. Dasselbe gilt für die Varianzen der KQ-Schätzer. Ich verzichte an dieser Stelle aber darauf die Herleitungen abzutippen, da es nichts weiter zu erklären oder verstehen gibt. Sollte die Herleitung geprüft werden, so kannst Du sie einfach aus dem Skript abschreiben.

Konfidenzintervalle und Tests

Für die Regressionskoeffizienten gilt

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

(siehe letztes Kapitel).

Da a und b Schätzwerte sind, lassen sich nun Konfidenzintervalle ableiten, genau wie für andere Parameter auch.

Dabei gilt:

a und b sind normalverteilt mit unbekannter Varianz. Diese muss also aus der Stichprobe geschätzt werden und es muss die t-Verteilung genutzt werden. Die Freiheitsgrade entsprechen in diesem Fall N-2.

Es gilt:

$$\widehat{\sigma_a} = \hat{\sigma}\sqrt{\frac{\sum_n X_n^2}{N\sum_n (X_n - \bar{X})^2}}$$

$$\widehat{\sigma_b} = \frac{\hat{\sigma}}{\sqrt{\sum_n (X_n - \bar{X})^2}}$$

Die Konfidenzintervalle sind entsprechend:

$$\hat{a} \pm t\left(1 - \frac{a}{2}, N - 2\right)\widehat{\sigma_a}$$

für a und

$$\hat{b} \pm t\left(1 - \frac{a}{2}, N - 2\right)\widehat{\sigma_b}$$

für b. Hier ist es wichtig, das a des Signifikanzniveaus und das a der Regressionageraden nicht zu verwechseln!

Beispiel zum Konfidenzintervall:

Wir bleiben bei dem Beispiel aus Seite 159. Es soll ein 95% Konfidenzintervall für $a = -0{,}8$ angegeben werden.

Man benötigt folgende Werte:

$$\sum_n X_n^2 = 120$$

$$\sum_n (X_n - \bar{X})^2 = 20$$

$\widehat{\sigma^2}$ wurde bereits berechnet:

$$\widehat{\sigma^2} = \frac{1}{N-2}\sum_n \widehat{\epsilon_n}^2 = 0{,}1$$

$$\hat{\sigma} = 0{,}32$$

Man erhält für $\widehat{\sigma_a}$:

$$\widehat{\sigma_a} = \hat{\sigma}\sqrt{\frac{\sum_n X_n^2}{N\sum_n (X_n - \bar{X})^2}} = 0{,}32\sqrt{\frac{120}{4 * 20}} = 0{,}39$$

Für das Konfidenzintervall gilt nun:

$$\hat{a} \pm t\left(1 - \frac{a}{2}, N-2\right)\widehat{\sigma_a}$$

$\hat{a}$ wurde schon auf -0,8 geschätzt. Nun kann man einsetzen:

$$-0{,}8 \pm 4{,}3 * 0{,}39$$

$$-2{,}48 \leq a \leq 0{,}88$$

Der geringe Stichprobenumfang und die hohe Varianz führen zu einem sehr großen Intervall.

Tests für a und b

Es gilt für die Prüfgrößen:

$$T_a = \frac{\hat{a} - a_0}{\widehat{\sigma_a}} \sim t(N-2)$$

$$T_b = \frac{\hat{b} - b_0}{\widehat{\sigma_b}} \sim t(N-2)$$

Beispiel:

Angenommen es soll die Nullhypothese $a = 0$ mit $a = 5\%$ getestet werden (Achtung, die "a"s bezeichnen verschiedene Dinge). Wieder muss nur in die richtige Formel eingesetzt werden:

$$T_a = \frac{\hat{a} - a_0}{\widehat{\sigma_a}} \sim t(N-2)$$

Als Prüfgröße erhält man:

$$T_a = \frac{-0{,}8 - 0}{0{,}39} = -2{,}1$$

Dies liegt innerhalb der kritischen Werte der t-Verteilung für $N = 4$, nämlich $(-4{,}3\ und\ +4{,}3)$.

Die Nullhypothese kann also nicht abgelehnt werden.

Konfidenzintervalle für die prognostizierten Werte

Hat man erst einmal die Regressionsgleichung aufgestellt, so kann man für beliebige X-Werte die zugehörigen Y-Werte schätzen. Für diese Schätzungen können dann wieder Konfidenzintervalle angegeben werden.

Bei prognostizierten Werten gibt es zwei Fehlerquellen:

1) die Werte liegen eh nicht alle direkt auf der Regressionsgeraden, sondern schwanken darum herum. Dies wird durch die Residuen erfasst.

2) die Regressionsgerade ist selber nur eine Schätzung und beruht auf den geschätzten Parametern a und b.

Beide Fehlerquellen müssen bei der Konstruktion eines Konfidenzintervalles berücksichtigt werden.

Man unterscheidet zwei verschiedene Fälle:

1) Ein Konfidenzintervall für die Regressionsgerade

2) Ein Prognoseintervall für einen bestimmten Y-Wert.

Zu Fall 1):

Es gilt

$$\frac{\hat{E} - E}{\sqrt{\widehat{Var}\left(\hat{E}\right)}} \sim t(N-2)$$

,wobei E die hypothetische Regressionsgerade ist und $\hat{E}$ die aus der Stichprobe geschätzte.

Für $\widehat{Var}\left(\hat{E}\right)$ gilt außerdem:

$$\widehat{Var}\left(\hat{E}\right) = \sigma^2\left(\frac{1}{N} + \frac{(X-\bar{X})^2}{\sum_n (X_n - \bar{X})^2}\right)$$

Das entsprechende Konfidenzintervall beträgt:

$$\hat{E} \pm t(1-\frac{a}{2}, N-2)\hat{\sigma}\sqrt{\frac{1}{N} + \frac{(X-\bar{X})^2}{\sum_n (X_n - \bar{X})^2}}$$

Das Konfidenzintervall ist also von x abhängig (und zwar die Lage und die Breite des Konfidenzintervalls).

Beispiel:

Ein Konfidenzintervall zu $a = 5\%$ für

$$a = -2{,}8$$

$$b = 1{,}3$$

$$\widehat{\sigma^2} = 10{,}68$$

$$\bar{x} = 5$$

$$\sum_n (X_n - \bar{X})^2 = 20$$

beträgt:

$$(-2{,}8 + 1{,}3x) \pm 4{,}3 * 3{,}27 \sqrt{\frac{1}{4} + \frac{(x-5)^2}{20}}$$

Zu Fall 2):

Das gesuchte Prognoseintervall für einen konkreten Wert Y_o beträgt:

$$\widehat{Y_0} \pm t(1 - \frac{a}{2}, N-2)\hat{\sigma} \sqrt{1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum_n (X_n - \bar{X})^2}}$$

Im Gegensatz zu Fall 1) enthält dieses Konfidenzintervall also einen zusätzlichen Unsicherheitsfaktor $\hat{\sigma}$, was der Streuung der einzelnen Werte um die Regressionsgerade Rechnung trägt. Das Konfidenzintervall ist entsprechend breiter.

Bestimmtheitsmaß

Das Bestimmtheitsmaß ist das Quadrat des Pearsonschen Korrelationskoeffizienten. Wie man mathematisch zeigen kann, entspricht das Bestimmtheitsmaß dem Quotienten aus der Varianz der durch lineare Regression geschätzten Y-Werte und der Varianz der beobachteten Y-Werte.

$$R^2 = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2}$$

Interpretation des Bestimmtheitsmaßes

Das Bestimmtheitsmaß gibt an, welcher Anteil an Streuung (Varianz) der beobachteten Y-Werte durch die Varianz der geschätzten Y-Werte erklärt wird. Der Ausdruck „erklärt“ ist etwas hier etwas missverständlich. Es soll bedeuten, dass dieser Anteil an Varianz durch den Einfluss des Merkmals X herbeigeführt wird. Angenommen das Merkmal Y hängt zu 50% von dem Merkmal X ab und zu 50% von einem dritten unbekannten Merkmal. In diesem Fall würde das Bestimmtheitsmaß bezüglich X und Y 50% (0,5) betragen.

Hängt Y ausschließlich von X ab, so beträgt das Bestimmtheitsmaß 1. Da die gesamte Änderung des Merkmals Y von der Änderung des Merkmals X abhängt, liegen alle tatsächlich beobachteten Werte auch auf der Regressionsgerade.

Varianzzerlegung

Wie schon beim Bestimmtheitsmaß angesprochen, kann man die Varianz in „erklärte Varianz“ bzw. erklärte Streuung und Rest-Varianz unterteilen. Beide zusammen bilden die gesamte Varianz der Daten. Anschaulich läßt sich das so erklären: Die Varianz wird gemessen als die Summe der quadrierten Abweichungen vom Mittelwert. Der Mittelwert kann grafisch durch eine Horizontale dargestellt werden. Es wird also die Streuung um diese Horizontale herum gemessen.

Durch eine Drehung der Geraden kann es sein, dass sich die Streuung um diese herum verringert. Diese Differenz an Streuung ist die durch die Regressionsgerade erklärte Varianz.

Mathematisch schreibt man:

Für die totale Streuung/Varianz:

$$SQT = \sum_n (Y_n - \bar{Y})^2$$

Für die erklärte Streuung:

$$SQE = \sum_n \left(\hat{Y}_n - \bar{Y}\right)^2$$

Für die Rest-Streuung:

$$SQR = \sum_n \left(Y_n - \hat{Y}_n\right)^2$$

Und natürlich gilt:

$$SQT = SQE + SQR$$

Globaler F-Test

Wie oben schon gesagt, ist eine Regressionsgerade, die parallel zur X-Achse verläuft, dasselbe, wie eine Horizontale durch den Mittelwert von Y. In diesem Fall würde also kein Unterschied bestehen zwischen totaler und Rest-Streuung. In diesem Fall würde die Regressionsgerade keine Erklärungskraft haben. In diesem Fall wäre $b = 0$. Ob dies der Fall ist, kann man mit dem globalen F-Test testen:

Die Teststatistik dazu ist:

$$F = \frac{SQE}{SQR/(N-2)} \sim F(1, N-2)$$

,wobei man F auch ausschließlich mit Hilfe des Korrelationskoeffizienten schreiben kann:

$$F = \frac{R_{xy}^2}{1 - R_{xy}^2}(N-2) \sim F(1, N-2)$$

Hierzu muss nun also die Tabelle der F-Verteilung genutzt werden.

Achtung: Normalerweise mußt du zu den einzelnen Tests nichts wissen, da man alle Tests gleich macht und nur Teststatistik und Verteilung aus der Formelsammlung abliest. Der globale F-Test hat aber eine Besonderheit: Beim Test der Hypothese $b = 0$ wird kein zweiseitiger Test durchgeführt, sondern ein einseitiger Test. Der Annahmebereich ist $F < F(1, N-2)$.

Beispiel:

Gegeben sind:

$$SQE = 36$$

$$SQR = 24$$

$$N = 36$$

Die Nullhypothese sei $b = 0$ und soll zu $a = 5\%$ getestet werden.

Für die Teststatistik folgt nun:

$$F = \frac{SQE}{\frac{SQR}{N-2}} = \frac{36}{24/34} = 51$$

Der Annahmebereich der F-Verteilung für 0,95 und 1 Zählerfreiheitsgrad und 34 Nennerfreiheitsgrade ist 4,13.

Die Nullhypothese muss also verworfen werden.

ANOVA Tafel

Die ANOVA Tafel ist nur eine Darstellungsform von verschiedenen Varianztermen. Für die Klausur solltest du wissen welcher Varianzterm wo steht.

Die ANOVA Tafel für den globalen F-Test ist zum Beispiel:

Varianz-Quelle	SQ		Freiheitsgrade df	
Hypothese	SQE	$\sum_n \left(\hat{Y}_n - \bar{Y}\right)^2$	1	
Residuen	SQR	$\sum_n \left(Y_n - \hat{Y}_n\right)^2$	$N-2$	$F = \frac{SQE}{SQR/(N-2)}$
Total	SQT	$\sum_n (Y_n - \bar{Y})^2$	$N-1$	$\sim F(1, N-2)$

Eine ANOVA Tafel für das letzte Beispiel würde wie folgt aussehen:

Varianz-Quelle	SQ		Freiheitsgrade df	
Hypothese	SQE	36	1	
Residuen	SQR	24	34	51
Total	SQT	60	35	$4{,}13$

Einfaktorielle Varianzanalyse

Zum Modell:

Bei der einfaktoriellen Varianzanalyse bestehen folgende Modellannahmen:

Es liegen mehrere Gruppen vor, in denen jeweils wieder mehrere Personen enthalten sind. Es wird angenommen, dass die Personen in den Gruppen voneinander unabhängig und identisch verteilt sind.

Beispiel: Die Klausur „Vertiefung Mathematik und Statistik“ wird an 10 Klausurorten geschrieben. Die Studenten, die an den Klausuren teilnehmen, sind identisch verteilt und unabhängig voneinander. Es ist also nicht zu erwarten, dass die Teilnehmer in einer Stadt eine bessere Note schreiben, als die in einer anderen Stadt. Außerdem wird angenommen, dass die Klausurergebnisse voneinander unabhängig sind (nur weil die Studenten in Hamburg einen guten Tag haben, heißt das nicht, dass deshalb auch die Studenten in München einen guten Tag haben).

Die Gruppen werden mit i bezeichnet und die Personen in diesen Gruppen mit j.

Da die Personen die gleiche Verteilung haben, haben sie auch denselben Mittelwert.

Mathematisch schreibt man:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Für die Fehler wird angenommen, dass sie Erwartungswert 0 und Varianz σ^2 haben.

Für den Mittelwert der i-ten Gruppe gilt:

$$\hat{\mu_i} = \frac{1}{J}\sum_j Y_{ij}$$

Für den Mittelwert werden wir im folgenden $\overline{Y_i}$ schreiben, sodass gilt:

$$\hat{\mu_i} = \frac{1}{J}\sum_j Y_{ij} \coloneqq \bar{Y_i}$$

Für den Gesamt-Mittelwert gilt dann:

$$\bar{Y} = 1/IL \sum_{ij} Y_{ij}$$

Für die Varianz-Zerlegung gilt dann:

$$SQR = \sum_{ij} (Y_{ij} - \bar{Y_i})^2$$

$$SQE = \sum_{ij} (\bar{Y_i} - \bar{Y})^2$$

Für SQE kann man vereinfacht auch schreiben:

$$SQE = J \sum_i \bar{Y}_i^2 - IJ * \bar{Y}^2$$

Für SQT kann man vereinfacht auch schreiben:

$$SQT = \sum_{ij} Y_{ij}^2 - IJ * \bar{Y}^2$$

Für die mittlere Residual-Quadratsumme innerhalb der Gruppen gilt:

$$\sigma^2 = \frac{SQR}{I(J-1)} = MQR$$

Nun ist alles definiert, was man für die ANOVA Tafel benötigt.

Die Hypothese lautet

$$H_0: \mu_i = \mu$$

,also Gleichheit aller Mittelwerte.

Die ANOVA Tafel für diese Hypothese lautet:

Varianz-Quelle	SQ		Freiheitsgrade df	
Hypothese	SQE	$\sum_{ij}(\bar{Y}_\iota - \bar{Y})^2$	$I-1$	
Residuen	SQR	$\sum_{ij}(Y_{ij} - \bar{Y}_\iota)^2$	$I(J-1)$	$F = \frac{\frac{SQE}{I-1}}{SQR/(I(J-1))}$
Total	SQT	$\sum_{ij}(Y_{ij} - \bar{Y})^2$	$IJ-1$	$\sim F(I-1, J-2)$

Beispiel:

Es werden in 4 Städten von je 20 Studenten eine Klausur geschrieben. Der Notendurchschnitt ist der einzelnen Städte ist folgender Tabelle zu entnehmen:

Stadt	A	B	C	D
Notenschnitt	2,5	2,7	3	2,3

Weiter sei die Quadratsumme der Einzelwerte mit 300 gegeben.

Teste auf Gleichheit der Notendurchschnitte zu $a = 5\%$.

Antwort:

Für die Prüfgröße $F = \frac{\frac{SQE}{I-1}}{SQR/(I(J-1))}$ benötigt man SQE und SQR:

$$SQE = J\sum_i \bar{Y}_i^2 - IJ * \bar{Y}^2$$

$$SQR = \sum_{ij}(Y_{ij} - \bar{Y}_i)^2$$

Man muss also den Gesamtnotenschnitt $\bar{Y}$ ermitteln:

$$\bar{Y} = \frac{1}{4}(2{,}5 + 2{,}7 + 3 + 2{,}3) = 2{,}625$$

I ist die Anzahl an Gruppen, also in diesem Fall 4. J ist die Anzahl an Gruppenmitgliedern, also in diesem Fall 20.

Für SQR sind nun auch alle Werte ermittelt und man muss nur noch einsetzen:

$$SQE = 20\sum_i 2{,}5^2 + 2{,}7^2 + 3^2 + 2{,}3^2 - 4 * 20 * 2{,}625^2 = 5{,}35$$

SQR ist mir zu aufwendig zu berechnen, daher berechne ich lieber SQT und ziehe SQE davon ab:

$$SQT = \sum_{ij} Y_{ij}^2 - IJ * \bar{Y}^2 = 300 - 4 * 20 * 2{,}625^2 = -251{,}25$$

$$SQR = SQT - SQE = 256{,}6$$

$$F = \frac{\frac{SQE}{I-1}}{SQR/(I(J-1))} = \frac{\frac{5{,}35}{4-1}}{256{,}6/(4(20-1))} = 0{,}53$$

Der kritische Wert der F-Verteilung ist

$$F(4{,}36) = 2{,}65$$

Die Nullhypothese kann also nicht abgelehnt werden.

Wie du siehst, ist hier nichts gefragt, außer Zahlen in Formeln einzusetzen. Es ist daher sehr wichtig, dass du alle Variablen in der Formel (mit den ganzen Indizierungen und Akzenten) zuordnen kannst.

Post hoc Varianzanalyse

Sollte die Nullhypothese abgelehnt werden, so ist nicht klar, welche Mittelwertsunterschiede zur Ablehnung der Nullhypothese geführt haben. Es wird also nach der Varianzanalyse noch für jedes Mittelwertepaar ein T-Vergleich durchgeführt.

Da nun aber mehrere Tests durchgeführt werden, und bei jedem Test die Fehlerwahrscheinlichkeit von 5% besteht, muss diese Fehlerwahrscheinlichkeit nun angepasst werden, damit das Niveau des globalen F-Tests eingehalten wird. Die Anpassung erfolgt folgendermaßen:

$$a^* = a/k$$

k bezeichnet hier die Anzahl an Tests, die durchgeführt werden müssen und es gilt:

$$k = \binom{I}{2} = \frac{I!}{2! * (I-2)!}$$

(gesprochen I über 2)

Die Teststatistik ist nun:

$$T_{ii'} = \frac{\bar{Y}_i - \bar{Y}_{i'}}{s\sqrt{2/J}} \sim t(IJ - I)$$

,wobei hier zum Niveau a^* geprüft wird.

Lösungen zu den Aufgaben der Statistik KE I

Im Folgenden findest du Lösungsvorschläge für die Aufgaben im Anhang des Fernuni-Skriptes. Da die Fernuni keine Lösungen veröffentlicht und laut Lehrstuhl eine ausreichende Prüfungsvorbereitung darin besteht, die alten Klausuren und die Übungsaufgaben zu lösen(neben dem Studium des Lehrmaterials natürlich), halte ich diese Aufgaben für sehr klausurrelevant.

Lösung zu Aufgabe 1

a) Alle Wahlberechtigten mit Telefonanschluss.

b) Die Parameter sind:

$$\pi_{CDUCSU} = 39\%$$

$$\pi_{SPD} = 30\%$$

$$\pi_{FDP} = 10\%$$

$$\pi_{Linke} = 8\%$$

$$\pi_{Grüne} = 9\%$$

Die Schätzungen entsprechen den Werten der Stichprobe.

c) Die Regierungsparteien kommen demnach auf 69%. Es liegt nun eine dichotome Grundgesamtheit vor. Das Konfidenzintervall für den Anteilsert lautet:

$$\hat{\pi} - z_{1-\frac{a}{2}} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \leq \pi \leq \hat{\pi} + z_{1-\frac{a}{2}} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

Es gilt:

$$\hat{\pi} = 0{,}69$$

Man muß nur noch einsetzen:

$$0{,}69 - 1{,}96 * \sqrt{\frac{0{,}69(0{,}31)}{1.343}} \leq \pi \leq 0{,}69 + 1{,}96 * \sqrt{\frac{0{,}69(0{,}31)}{1.343}}$$

$$0{,}69 - 1{,}96 * \sqrt{\frac{0{,}69(0{,}31)}{1.343}} \leq \pi \leq 0{,}69 + 1{,}96 * \sqrt{\frac{0{,}69(0{,}31)}{1.343}}$$

$$66{,}53\% \leq \pi \leq 71{,}47$$

d) A ist eigentlich binomialverteilt, aber für große N ist es näherungsweise normalverteilt.

e) Es muss die Standardabweichung der verschiedenen Kategorien ermittelt werden (Der Fehlerbereich für $a = 5\%$ ist 2σ). Es gilt:

$$Var(\hat{\pi}_j) = \frac{1}{N}\pi_j(1 - \pi_j)$$

Entsprechend erhält man:

$$Var(\hat{\pi}_{CDUCSU}) = \frac{1}{1.343}0{,}39(1 - 0{,}39) = 0{,}000177$$

$$Var(\hat{\pi}_{SPD}) = 0{,}000156$$

$$Var(\hat{\pi}_{FDP}) = 0{,}000067$$

$$Var(\hat{\pi}_{Linke}) = 0{,}000055$$

$$Var(\hat{\pi}_{Grüne}) = 0{,}000061$$

Jetzt muss man noch die Wurzel ziehen und mal zwei nehmen und erhält als Fehlerbereich:

CDU/CSU: 2,66%

SPD:2,5%

FDP:1,64%

Linke:1,48%

Grüne:1,56%

Gerundet sind die Fehlerbereiche also korrekt.

Lösung zu Aufgabe 2

Ich bin der Meinung, daß diese Aufgabe mit Hilfe der Wahrscheinlichkeitsrechnung/Kombinatorik gelöst werden muss. Diese ist aber nicht Bestandteil des Kurses. Es könnte also sein, dass der Lehrstuhl einen anderen Lösungsweg erwartet, bei dem die Hilfsmittel des Kurses genutzt werden- ich wüßte nur nicht welcher das sein soll.

Zur Lösung:

2.1) $N = 2$

Hier gibt es 4 mögliche Ereignisse: $AA, \bar{A}\bar{A}, A\bar{A}, \bar{A}A$. In zwei von 4 Möglichkeiten entspricht die relative Häufigkeit der Stichprobe der relativen Häufigkeit der Grundgesamtheit.

2.2) $N = 10$

Ähnlich wie 2.1, nur daß es nun sehr aufwendig ist, alle Möglichkeiten aufzuschreiben. Aus der Kombinatorik weiß man aber: Es gibt „10 über 5“ Möglichkeiten, daß 5 A und 5 $\bar{A}$ im Ergebnis vorkommen. Und es gibt insgesamt 2^{10} mögliche Ergebnisse.

Die gesuchte Wahrscheinlichkeit ist also:

$$P(f_{10}(A) = 0{,}5) = \frac{\binom{10}{5}}{2^{10}} = \frac{(5*4*3*2*1)(5*4*3*2*1)}{5*4*3*2*1} = \frac{120}{1.024} = 11{,}72\%$$

2.3) N=15

Bei N=15 ist es nicht möglich genauso viele Ereignisse A, wie $\bar{A}$ zu erhalten. Die Wahrscheinlichkeit ist also 0.

Lösung zu Aufgabe 3

Es handelt sich um ein binomialverteiltes Merkmal. Ein Schätzer für die Erfolgswahrscheinlichkeit ist der Mittelwert der Stichprobe, also $210/300 = 0{,}7$.

Ein 95% Konfidenzintervall erhält man nach der Formel:

$$\hat{\pi} - z_{1-\frac{a}{2}} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \leq \pi \leq \hat{\pi} + z_{1-\frac{a}{2}} * \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

Man muss nur noch einsetzen:

$$0{,}7 - 1{,}96 * \sqrt{\frac{0{,}7(1-0{,}7)}{300}} \leq \pi \leq 0{,}7 + 1{,}96 * \sqrt{\frac{0{,}7(1-0{,}7)}{300}}$$

$$0{,}65 \leq \pi \leq 0{,}75$$

Lösung zu Aufgabe 4

a1) Modell: Verteilung bekannt, Varianz bekannt.

Das Merkmal ist normalverteilt und entsprechend gilt für das Konfidenzintervall:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Man muss nur noch einsetzen:

$$105 - 1{,}96\frac{20}{\sqrt{10}} \leq \mu \leq 105 + 1{,}96\frac{20}{\sqrt{10}}$$

$$92{,}6 \leq \mu \leq 117{,}4$$

a2) Modell: Verteilung bekannt, Varianz unbekannt.

Da die Varianz aus der Stichprobe geschätzt wurde muss folgende Formel verwendet werden:

$$\bar{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{N}}$$

Man muss nur noch einsetzen:

$$105 - 2{,}262\frac{20}{\sqrt{10}} \leq \mu \leq 105 + 2{,}262\frac{20}{\sqrt{10}}$$

$$90{,}69 \leq \mu \leq 119{,}30$$

Das Konfidenzintervall bei unbekannter Varianz der Grundgesamtheit ist größer, da dem Schätzfehler der Varianzschätzung aus der Stichprobe Rechnung getragen wird.

b1) Modell: Verteilung unbekannt, Varianz bekannt Stichprobenumfang groß.

Im Gegensatz zu der Aufgabenstellung in a) gibt es nun keine Aussage über die Verteilung der Grundgesamtheit. Es liegt aber ein großer Stichprobenumfang vor, daher wird die Aufgabe wie die Teilaufgabe a1) gelöst. Für das Konfidenzintervall gilt:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Einsetzen:

$$105 - 1{,}96\frac{17{,}32}{\sqrt{50}} \leq \mu \leq 105 + 1{,}96\frac{17{,}32}{\sqrt{50}}$$

$$100{,}20 \leq \mu \leq 109{,}80$$

b2)Modell: Verteilung unbekannt, Varianz unbekannt, Stichprobenumfang groß.

Wegen des großen Stichprobenumfangs berechnet sich das Konfidenzintervall wie in a2)

$$\bar{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{N}}$$

Einsetzen:

$$105 - 2{,}01\frac{17{,}32}{\sqrt{50}} \leq \mu \leq 105 + 2{,}01\frac{17{,}32}{\sqrt{50}}$$

$$100{,}08 \leq \mu \leq 109{,}92$$

Lösung zu Aufgabe 5

a) Modell: Verteilung unbekannt, Varianz bekannt.

Die Verteilung der Grundgesamtheit ist zwar nicht gegeben, aber wegen des großen Stichprobenumfangs ist diese asymptotisch normalverteilt.

Für das Konfidenzintervall gilt:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Einsetzen:

$$590 - 1{,}96\frac{50}{\sqrt{250}} \leq \mu \leq 590 + 1{,}96\frac{50}{\sqrt{250}}$$

$$583{,}80 \leq \mu \leq 596{,}20$$

b) Modell: Verteilung unbekannt, Varianz bekannt.

Die Verteilung der Grundgesamtheit ist zwar nicht gegeben, aber wegen des großen Stichprobenumfangs ist diese asymptotisch normalverteilt.

Für das Konfidenzintervall gilt:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

In diesem Fall gilt $\alpha = 1\%$. Für $z_{1-\alpha/2}$ folgt also

$$z_{1-\alpha/2} = 2{,}58$$

Es muss also gelten:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} = 583{,}8$$

Einsetzen:

$$590 - 2{,}58\frac{50}{\sqrt{N}} = 583{,}8$$

Auflösen nach N:

$$N = 432{,}91$$

Es müssen 433 Personen befragt werden.

c) Ich bin mir nicht sicher, worauf diese Frage abzielt. Man könnte argumentieren, dass die Umfrage nicht repräsentativ ist, da Studenten, die arbeitstätig sind und teilzeit studieren, mehr verdienen und seltener am Haupteingang angetroffen werden.

Ansonsten würde ich sagen, dass für das Konfidenzintervall die Annahme der Normalverteilung getroffen wurde, was auf die Grundgesamtheit nicht zutreffen muss. Dennoch kann die Normalverteilung genutzt werden, da Summen von beliebig verteilten Zufallsvariablen approximativ normalverteilt sind.

Lösung zu Aufgabe 6

Modell: Unbekannte Verteilung, kleiner Stichprobenumfang, unbekannte Varianz.

In dieser Aufgabenstellung ist die Verteilung unbekannt und der Stichprobenumfang ist klein. Ich bin der Meinung, dass zur Lösung der Aufgabe die Annahme der Normalverteilung der Grundgesamtheit getroffen werden muss.

a) Punktschätzer für den Mittelwert:

Ein Punktschätzer für den Mittelwert der Grundgesamtheit ist der Mittelwert der Stichprobe. Für das Merkmal Kundenzufriedenheit gilt:

$$\bar{X} = \frac{1}{N}\sum X_n = \frac{1}{6}(46 + 48 + 31 + 32 + 43 + 68) = 44{,}67$$

Für das Merkmal Loyalität gilt:

$$\bar{X} = \frac{1}{N}\sum X_n = \frac{1}{6}(102 + 106 + 88 + 90 + 100 + 122) = 101{,}33$$

Ein Punktschätzer für die Varianz ist:

$$S^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X})^2$$

Für das Merkmal Kundenzufriedenheit erhält man:

$$S^2 = \frac{1}{6-1}((46 - 44{,}67)^2 + (48 - 44{,}67)^2 + (31 - 46{,}67)^2 + (32 - 44{,}67)^2 + (43 - 44{,}67)^2 + (68 - 44{,}67)^2) = 181{,}47$$

Für das Merkmal Loyalität erhält man:

$$S^2 = \frac{1}{6-1}((102 - 101{,}33)^2 + (106 - 101{,}33)^2 + (88 - 101{,}33)^2 + (90 - 101{,}33)^2 + (100 - 101{,}33)^2 + (122 - 101{,}33)^2) = 151{,}47$$

b) Das Konfidenzintervall für p lautet:

$$\frac{e^A - 1}{e^A + 1} \leq p \leq \frac{e^B - 1}{e^B + 1}$$

,wobei gilt:

$$A = \ln\frac{1+R}{1-R} - \frac{2z}{\sqrt{N-3}}$$

$$B = \ln\frac{1+R}{1-R} + \frac{2z}{\sqrt{N-3}}$$

R ist wiederum:

$$R = \frac{S(X,Y)}{S(X)S(Y)} = \frac{\sum_{n=1}^{N}(X_n - \bar{X})(Y_n - \bar{Y})}{\sqrt{\sum_{n=1}^{N}(X_n - \bar{X})^2 \sum_{n=1}^{N}(Y_n - \bar{Y})^2}}$$

Man benötigt noch R. $S(x)$ und $S(Y)$ wurden schon in a) berechnet, es fehlt also noch $S(X,Y)$.

$$S(X,Y) = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X})(Y_n - \bar{Y}) = 165{,}33$$

Es folgt:

$$R = \frac{S(X,Y)}{S(X)S(Y)} = \frac{165{,}33}{12{,}3 * 13{,}47} = 99{,}72\%$$

Wie schon in a) erwähnt, fehlt eine Aussage über die Verteilung der Grundgesamtheit in der Aufgabenstellung. Der Stichprobenumfang ist auch nicht groß, daher ist die Annahme nicht gerechtfertigt (Selbst bei großem Stichprobenumfang wäre die Annahme nicht gerechtfertigt, aber man könnte sie nutzen, da die Summe mehrerer beliebig verteilter Zufallsvariablen bei großem Stichprobenumfang approximativ normalverteilt ist).

c) Ich weiß nicht genau, worauf diese Frage abzielt. Die Punktschätzer sind die Parameter der Stichprobe und die Populationsparameter sind die Parameter der Population.

Die Verteilungen der Punktschätzer lassen sich wie folgt skizzieren:

Erwartungswert:

$$E(\hat{\Theta}) = \Theta$$

Varianz:

$$S^2(\hat{\Theta}) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\Theta}_m - \bar{\hat{\Theta}}\right)^2$$

, wobei Θ für den jeweils zu schätzenden Parameter steht und $M = 10$ gilt.

Lösung zu Aufgabe 7

a) Es gilt:

$$E(\bar{X}) = \mu$$

und

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N}$$

So kann der Erwartungswert mit

$$E(\bar{X}) = 100$$

Und die Varianz mit:

$$\sigma_{\bar{X}}^2 = \frac{100}{5} = 20$$

bestimmt werden.

b) Die Dichtefunktion der Zufallsvariablen $X_n$besitzt eine ähnliche Dichtefunktion, wie die Standardnormalverteilung. Statt des Mittelwertes von Null hat sie einen Mittelwert von 100 und aufgrund der hohen Varianz (Standardnormalverteilung hat Varianz 1) ist sie deutlich breiter und flacher.

Die Dichtefunktion der Zufallsvariablen $\bar{X}$ hat denselben Erwartungswert nur eine deutlich kleinere Varianz. Damit ist die Dichtefunktion deutlich schmaler und höher als die von X_n aber immernoch breiter und flacher, als die der Standardnormalverteilung.

c) Für sehr große Stichprobenumfänge strebt die Varianz des Mittelwertes gegen Null. Dies sieht man an der Formel

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N}$$

Lösung zu Aufgabe 8

a) Diese Begriffe sind hinten im Skript definiert.

b) Beispiel Mittelwert: Der Schätzwert des Mittelwertes ist eine Zufallsvariable, die sich aus der Summe vieler Zufallsvariablen ergibt. Der Schätzwert ist also selber auch eine Zufallsvariable (Zieht man auch einer Grundgesamtheit von 100 nacheinander mehrmals eine Stichprobe vom Umfang 10, so werden sich die Stichproben sehr wahrscheinlich unterscheiden-sie sind zufällig zusammengesetzt). Für diese Zufallsvariable kann man nach mehrmaligem Ziehen wieder die bekannten Parameter Mittelwert und Varianz berechnen.

Dasselbe gilt für den Schätzer für die Varianz.

Lösung zu Aufgabe 9

Die Werte können aus den Tabellen abgelesen werden. Dies musst du sicher und schnell beherrschen. Am besten kennst du die Werte für 5% und 1,5% der Standardnormalverteilung auswendig!

Achte auch immer darauf, ob es sich um einseitige oder zweiseitige Tests handelt.

a) 1,64

b) 1,70

c) 7,81

d) 3,63

Lösung zu Aufgabe 10

a) Der Anteilswert ist hier „Person hat das alkoholfreie Bier identifiziert“. Angenommen beide Biere schmecken gleich, müßte dieser Anteilswert 0,5 sein. Ich wähle

$$H_0: \pi_0 = 0{,}5$$

gegen

$$H_1: \pi_0 \neq 0{,}5$$

Der Annahmebereich beträgt:

$$\pi_0 - z_{1-\alpha/2}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{N}} \leq \hat{\pi} \leq \pi_0 + z_{1-\alpha/2}\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{N}}$$

Unabhängig von a wird die Nullhypothese also abgelehnt, wenn $\hat{\pi}$ von 0,5 verschieden ist. Aus der Aufgabenstellung ist $\hat{\pi} = 98/150$ bekannt. Daher wird die Nullhypothese angelehnt.

b) Der Annahmebereich aus a) ist unabhängig von N. Da auch bei b) $\hat{\pi} \neq 0{,}5$ gilt, wird auch hier die Hypothese

$$H_0: \pi_0 = 0{,}5$$

abgelehnt.

Lösung zu Aufgabe 11

a) Zunächst muss die Hypothese aufgestellt werden. Da die Aussage ist, dass ca. 100 Tage im Jahr regnet, wähle ich

$$H_0: \pi_0 = 100/365$$

gegen

$$H_1: \pi_0 \neq 100/365$$

Die Prüfgröße ist

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1-\pi_0)}}$$

Es muss nur noch eingesetzt werden:

$$Z = \frac{\frac{80}{365} - \frac{100}{365}}{\sqrt{\frac{\frac{1}{365}100}{365\left(1-\frac{100}{365}\right)}}} = -2{,}35$$

Die kritischen Werte der Standardnormalverteilung zu $a = 0{,}05$ sind -1,96 und 1,96. Die Prüfgröße liegt außerhalb dieses Bereiches und somit muss die Nullhypothese abgelehnt werden.

b) Bei solchen Fragen ist es immer schwierig zu erahnen, worauf die Frage genau abzielt. Ich würde hier argumentieren, dass „es hat an x Tagen im Jahr geregnet“ zwar binomialverteilt ist, aber alleine keine Auassagekraft über die Änderung des Klimas hat. Die Regenmenge wäre beispielsweise auch wichtig.

Rein statistisch kann man sagen, dass π binomialverteilt ist und aufgrund des hohen Stichprobenumfangs durch die Normalverteilung approximiert werden kann.

Lösung zu Aufgabe 12

Zunächst muss die Hypothese aufgestellt werden. Da die Aussage ist, dass das Medikament in 90% der Fälle wirkt und nicht in mindestens 90% der Fälle, wähle ich

$$H_0: \pi_0 = 0{,}9$$

gegen

$$H_1: \pi_0 \neq 0{,}9$$

Die Prüfgröße ist

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1-\pi_0)}}$$

Es muss nur noch eingesetzt werden:

$$Z = \frac{0{,}8 - 0{,}9}{\sqrt{\frac{1}{200}0{,}9(1-0{,}9)}} = -4{,}71$$

Die kritischen Werte der Standardnormalverteilung zu $\alpha = 0{,}01$ sind -2,58 und 2,58. Die Prüfgröße liegt außerhalb dieses Bereiches und somit muss die Nullhypothese abgelehnt werden.

Lösung zu 13

Zunächst muss die Hypothese aufgestellt werden. Da die Aussage ist, dass die Münze symmetrisch sei, wähle ich

$$H_0: \pi_0 = 0{,}5$$

gegen

$$H_1: \pi_0 \neq 0{,}5$$

Die Prüfgröße ist

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1-\pi_0)}}$$

Es muss nur noch eingesetzt werden:

$$Z = \frac{\frac{8}{12} - \frac{6}{12}}{\sqrt{\frac{\frac{1}{12}6}{12\left(1-\frac{6}{12}\right)}}} = 1{,}55$$

Die kritischen Werte der Standardnormalverteilung zu $a = 0{,}05$ sind -1,96 und 1,96. Die Prüfgröße liegt innerhalb dieses Bereiches und somit kann die Nullhypothese nicht abgelehnt werden.

Lösung zu Aufgabe 14

Bei einer 9:3:3:1 Verteilung müssten die 556 Erbsen wie folgt verteilt sein:

Runde gelbe Erbsen: 312,75

Runde grüne Erbsen: 104,25

Kantige gelbe Erbsen: 104,25

Kantige grüne Erbsen: 34,75

Zur Lösung nutzt man den Chi Quadrat Anpassungstest.

Als Prüfgröße nutzt man:

$$\chi_*^2 = \sum_{i=1}^{m} \frac{(h_{oi} - h_{ei})^2}{h_{ei}}$$

Dabei ist m die Anzahl an Merkmalsausprägungen (bzw. Merkmalsklassen bei stetigen Merkmalen). Es muss gelten $h_{ei} > 5$, ansonsten kann die Verteilung nicht verwendet werden.

Dies ist ein einseitiger Test und die obere Annahmegrenze ist der Wert der Chi Quadrat Verteilung mit $m - 1$ Freiheitsgraden und Signifikanzniveau $1 - a$.

Man erhält

$$\chi_*^2 = 0{,}0162 + 0{,}135 + 0{,}0101 + 0{,}218 = 0{,}47$$

Die Hypothese der Gleichheit kann nicht abgelehnt werden.

Lösung zu Aufgabe 15

a)

Ein Schätzer für die Wahrscheinlichkeit ist die Häufigkeit der einzelnen Note in der Subpopulation geteilt durch die Anzahl an Klausuren in der Subpopulation.

Note	1	2	3	4	5
Weiblich	23,33%	33,33%	33,33%	6,67%	3,33%
Männlich	15,56%	44,44%	24,44%	4,44%	11,11%

Deskriptiver Vergleich: Hier kann man zu jeder Note sagen, ob ein Mann oder eine Frau eine höhere Wahrscheinlichkeit hat diese Note zu bekommen, wenn er/sie an der Klausur teilnimmt.

b) Hier kann man den Chi-Quadrat Unabhängigkeitstest verwenden (siehe Aufgabe 16). Da aber kein Signifikanzniveau gegeben ist und die Daten offensichtlich abhängig sind, reicht es in meinen Augen zu sagen, dass bei Unabhängigkeit die bedingten Wahrscheinlichkeiten der Männer und der Frauen ungefähr gleich sein müßten. Dies ist nicht so, daher besteht ein Zusammenhang.

Lösung zu Aufgabe 16

Hier ist der Chi-Quadrat Anpassungstest zu verwenden.

Die Nullhypothese ist (immer) die Unabhängigkeit der Merkmale.

Die Testgröße ist:

$$\chi_*^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(h_{oij} - h_{eij}\right)^2}{h_{eij}}$$

Man stellt zunächst die Tabelle der beobachteten Werte mit ihren Randverteilungen auf und dann die Tabelle der hypothetischen Verteilung (unter Annahme der Unabhängigkeit) mit ihren Randverteilungen.

Ballbesuch	Viele	Wenig	Gar keinen	Summe
Studenten	25	15	10	50
Nicht-Studenten	15	20	15	50
Summe	40	35	25	100

Die hypothetische Verteilung ist:

Ballbesuch	Viele	Wenig	Gar keinen	Summe
Studenten	20	17,5	12,5	50
Nicht-Studenten	20	17,5	12,5	50
Summe	40	35	25	100

Ich werde nun nicht die gesamte Rechnung hinschreiben. Für i=1, j=1 gilt zum Beispiel:

$$\frac{(25-20)^2}{20}$$

Im Ergebnis erhält man:

$$\chi_*^2 = 4{,}21$$

Der kritische Wert der Chi-Quadrat Verteilung mit $v = (r-1)*(s-1) = 2$ Freiheitsgraden zu $a = 10\%$ ist 4,61. Somit kann die Nullhypothese nicht abgelehnt werden.

Lösung zu Aufgabe 17

a) Die kritische Region der Standardnormalverteilung für 1% ist> 2,33. Bei einem Mittelwert von 1.000, Standardabweichung von 5 und $N = 64$ berechnet sich der kritische Wert für den Stichprobenmittelwert zu:

$$\mu_0 - z_{1-a}\frac{\sigma}{\sqrt{N}} = 1.000 - 2{,}33 * \frac{5}{8} = 998{,}54$$

Unterhalb dieses Wertes muss die Nullhypothese abgelehnt werden.

b) Die Wahrscheinlichkeit, dass bei einer Verteilung mit Mittelwert 998 und Varianz 25 bei N=25 ein Wert über 998,54 angenommen wird, berechnet man wie folgt:

Zunächst muss standardisiert werden:

$$\frac{998{,}54 - 998}{\frac{\sigma}{\sqrt{N}}} = \frac{0{,}54}{5/8} = 0{,}86$$

Gesucht ist also die Wahrscheinlichkeit, mit der eine standardnormalverteilte Zufallsvariable 1-0,86 von Ihrem Mittelwert abweicht.

Aus der Tabelle erhält man den Wert von 55,56%.

c) Der Wert der Standardnormalverteilung für 5% ist 1,65. Es muss also folgende Formel aufgelöst werden:

$$\frac{998{,}54 - 998}{\frac{5}{\sqrt{N}}} = 1{,}65$$

Daraus folgt N=233,41. Der Stichprobenumfang müßte also mindestens 234 betragen.

Lösung zu Aufgabe 18

a) Die Tabelle lautet:

	A kompetent	B kompetent	Summe
A kompetent	25	35	60
B kompetent	35	5	40
Summe	60	40	100

b) Für den Unabhängigkeitstests benötigt man die hypothetische Verteilung bei Unabhängigkeit:

	A kompetent	B kompetent	Summe
A kompetent	36	24	60
B kompetent	24	16	40
Summe	60	40	100

Die Testgröße ist dann:

$$\chi_*^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(h_{oij} - h_{eij}\right)^2}{h_{eij}}$$

wobei r die Anzahl an Ausprägungen von X und s die Anzahl an Ausprägungen von Y bezeichnet.

Die Nullhypothese ist immer die Unabhängigkeit der beiden Merkmale. Der Ablehnungsbereich für die Nullhypothese ist

$$\chi_*^2 > \chi_{(1-a,v)}^2$$

wobei v die Anzahl an Freiheitsgeraden bezeichnet. Diese berechnen sich zu $v = (r-1) * (s-1)$.

H_0 wird also abgelehnt, wenn die Testgröße größer ausfällt als $\chi_{(1-a,v)}^2$.

Man erhält:

$$\chi_*^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(h_{oij} - h_{eij}\right)^2}{h_{eij}} = 21$$

Der kritische Wert der Chi-Quadrat Verteilung mit 1 Freiheitsgrad zu $a = 5\%$ ist 3,84. Die Nullhypothese kann also abgelehnt werden.

c) Der entsprechende Teil der Tabelle ist dieser:

	A kompetent	B kompetent	Summe
A kompetent	25	35	60

In 35 von 60 Fällen wird B für kompetent gehalten. In 25 von 60 Fällen nicht.

Lösung zu Aufgabe 19

a) Der Erwartungswert ist der Mittelwert der Daten:

$$E(X) = \frac{50 + 82 + 73 + 65 + 64 + 59 + 72 + 84 + 69 + 75}{10} = 69{,}3$$

Die Varianz ist

$$S^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X})^2 = 106{,}23$$

b) Als Nullhypothese wähle ich:

$$H_0: \mu_0 \leq 50$$

Nun muss ein Test auf den Mittelwert bei normalverteilter Grundgesamtheit mit bekannter Varianz durchgeführt werden. Die Testgröße ist:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{N}}}$$

$$Z = \frac{69{,}3 - 50}{\frac{9}{\sqrt{10}}} = 6{,}78$$

Der Wert liegt deutlich außerhalb des Annahmebereiches für die Nullhypothese (kritischer Wert der Standardnormalverteilung ist 1,65).

Daher kann die Nullhypothese abgelehnt werden.

c) Jetzt muss die Varianz aus der Stichprobe geschätzt werden und es muss die t-Verteilung genutzt werden.

Die Prüfgröße ist:

$$T = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{N}}}$$

T ist t-verteilt mit N-1 Freiheitsgraden und es gelten die entsprechenden kritischen Werte.

$$T = \frac{69{,}3 - 50}{\frac{10{,}31}{\sqrt{10}}} = 5{,}92$$

Der Wert liegt deutlich außerhalb des Annahmebereiches für die Nullhypothese (kritischer Wert der t-Verteilung ist 1,833).

Daher kann die Nullhypothese abgelehnt werden.

d1) Vorweg eine Anmerkung zu dieser Aufgabe: ich habe im Fernuni nichts zu Konfidenzintervallen für Erwartungswerte gefunden und löse diese Aufgabe etwas frei.

Das Konfidenzintervall für den Mittelwert einer normalverteilten Grundgesamtheit erhälst du nach folgender Formel:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Achtung: In diesem Fall soll ein Konfidenzintervall für den Erwartungswert berechnet werden, nicht für den Mittelwert. Es müssen also die Parameter der Verteilung des Erwartungswertes berechnet werden.

Für den Erwartungswert gilt folgendes Konfidenzintervall:

$$E(X) - z_{1-\alpha/2}\frac{\sigma_{\bar{x}}}{\sqrt{N}} \leq E(X) \leq E(X) + z_{1-\alpha/2}\frac{\sigma_{\bar{x}}}{\sqrt{N}}$$

N ist gleich 1, da nur eine Stichprobe durchgeführt wurde.

Die Varianz des Erwartungswertes beträgt:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N} = \frac{81}{10} = 8{,}1$$

Der Erwartungswert von $E(X)$ ist $E(X)$.

Einsetzen:

$$69{,}3 - 1{,}65\frac{\sqrt{8{,}1}}{\sqrt{1}} \leq E(X) \leq 69{,}3 + 1{,}65\frac{\sqrt{8{,}1}}{\sqrt{1}}$$

$$64{,}60 \leq E(X) \leq 74{,}00$$

d2) Bei unbekannter Varianz muss diese aus der Stichprobe geschätzt werden. Ansonsten alles wie bei d1):

$$E(X) - z_{1-\alpha/2}\frac{s_{\bar{x}}}{\sqrt{N}} \leq E(X) \leq E(X) + z_{1-\alpha/2}\frac{s_{\bar{x}}}{\sqrt{N}}$$

Einsetzen:

$$69{,}3 - 1{,}65\frac{\sqrt{10{,}63}}{\sqrt{1}} \leq E(X) \leq 69{,}3 + 1{,}65\frac{\sqrt{10{,}63}}{\sqrt{1}}$$

$$63{,}92 \leq E(X) \leq 74{,}68$$

Dieses Konfidenzintervall ist größer als das aus d1). Der Grund liegt in der Schätzung der Varianz der Grundgesamtheit.

Lösung zu Aufgabe 20

a) Als Nullhypothese verwende ich die Hypothese, dass die durchschnittliche Körpergröße kleiner oder gleich 180 cm ist. Die Prüfgröße ist in der Aufgabenstellung gegeben. Der kritische Wert für einen Test auf den Mittelwert bei normalverteilter Grundgesamtheit mit unbekannter Varianz ist 1,708 (t-Verteilung, $a = 5\%$, 25 Freiheitsgrade). Die Prüfgröße liegt oberhalb des kritischen Wertes. Damit kann die Nullhypothese abgelehnt werden.

b)Eine Verringerung des Signifikanzniveaus bewirkt, dass der kritische Wert steigt. Damit steigt auch die Wahrscheinlichkeit eines Fehlers 2ter Art (dieser ist die Wahrscheinlichkeit, dass der Mittelwert unter dem kritischen Wert liegt, wenn die wahre Verteilung angenommen wird).

Eine Erhöhung des Stichprobenumfangs verringert die Fehlerwahrscheinlichkeit 2ter Art.
Begründung: Wählt man N maximal, so gäbe es keine Chance auf einen Fehler 2ter Art.

Lösung zu Aufgabe 21

Es ist zwar etwas gewagt von mir das zu sagen, aber ich denke, dass diese Aufgabe nicht zu lösen ist. Ohne Angaben über die wahre Verteilung kann man auch keine Aussagen über die Wahrscheinlichkeiten einer richtigen oder falschen Entscheidung treffen. Angenommen der wahre Mittelwert ist 0: Ein Stichprobenmittelwert über 205 wäre dann sehr unwahrscheinlich und der Entscheider würde fast immer eine falsche Entscheidung treffen. Ist der wahre Mittelwert 200, so ergeben sich schon ganz andere Wahrscheinlichkeiten. Man könnte auch argumentieren, dass der Entscheider immer die falsche Entscheidung trifft, da das Merkmal stetig ist.

Lösung zu Aufgabe 22

a) Hier geht es um den Vergleich zweier Mittelwerte zweier beliebig verteilter Grundgesamtheiten.

In diesem Fall gilt für die Varianz S:

$$S = \sqrt{\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}}$$

Die Prüfgröße ist

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{S}$$

mit $\delta_0 = 0$.

T ist approximativ standardnormalverteilt für $N_x, N_y > 30$.

Die Prüfgröße ist in der Aufgabenstellung mit 3,85 gegeben. Der kritische Wert der Standardnormalverteilung für $a = 5\%$ (zweiseitiger Test) ist 1,96. Daher kann die Nullhypothese der Gleichheit der Mittelwerte nicht abgelehnt werden.

b) Die Prüfgröße ist

$$T = \frac{\bar{X} - \bar{Y} - \delta_0}{S}$$

mit $\delta_0 = 1$

Die Prüfgröße ist in der Aufgabenstellung mit 0,96 gegeben, daher kann die Nullhypothese abgelehnt werden (alles andere ist wie in a)).

c)Ja, da die Stichprobenumfänge jeweils größer als 30 sind, sind die Prüfgrößen approximativ normalverteilt.

Lösung zu Aufgabe 23

a) Für den Test müssen die Werte zunächst zu einer Menge zusammengefasst werden:

5,6,7,8,8,9,9,10,11,12,12,13,13,14,15,16,16,17,17,18,19,21

Mit den entsprechenden Rängen:

1	2	3	4,5	4,5	6,5	6,5	8	9	10,5	10,5	12,5
5	6	7	8	8	9	9	10	11	12	12	13

12,5	14	15	16,5	16,5	18,5	18,5	20	21	22
13	14	15	16	16	17	17	18	19	21

Die Rangsumme aus der zweiten Gruppe ist:

$$T_W = \sum_n rg(X_i) = 1 + 2 + 3 + 4{,}5 + 4{,}5 + 6{,}5 + 8 + 9 + 12{,}5 + 14 + 15 + 16{,}5 = 96{,}5$$

Die Prüfgröße ist nun:

$$Z = \frac{T_W - E[T_W]}{\sqrt{Var[T_W]}}$$

,wobei gilt:

$$E[T_W] = \frac{N(N+M+1)}{2} = \frac{12(12+10+1)}{2} = 138$$

$$\sqrt{Var[T_W]} = \sqrt{\frac{12 * 10(10+12+1)}{12}} = \sqrt{230}$$

Man erhält:

$$Z = \frac{T_W - E[T_W]}{\sqrt{Var[T_W]}} = \frac{96{,}5 - 138}{\sqrt{230}} = -2{,}74$$

Der kritische Wert der Standardnormalverteilung ist -1,96 (und 1,96). Die Nullhypothese der Unabhängigkeit kann abgelehnt werden.

b) Der Wilcoxon-Rangsummen-Test ist unabhängig von Verteilungsannahmen.

Lösung zu Aufgabe 24

Dies ist eine sehr allgemeine Fragestellung. Für den Vergleich zweier Verteilungen fällt mir nur der Chi-Quadrat Anpassungstest ein. Andere Tests vergleichen nicht die gesamte Verteilung sondern nur Parameter, wie Mittelwert oder Varianz. Andere Tests testen auf Unabhängigkeit. Beispiele findet man im Skript bei dem jeweiligen Test.

Lösung zu Aufgabe 25

a) Zunächst müssen die Werte in einer Menge zusammengefasst werden:

5,28	5,30	5,48	5,56	5,57	5,60	5,62	5,7	5,75	5,8
1	2	3	4	5	6	7	8	9	10

5,91	5,92	6,01	6,04	6,10	6,11	6,17	6,20	6,24	6,29
11	12	13	14	15	16	17	18	19	20

Die Ränge der Gruppe vor dem Training sind:

19,10,5,14,4,12,17,3,16,18

Die Rangsumme ist 118.

Die Prüfgröße ist nun:

$$Z = \frac{T_W - E[T_W]}{\sqrt{Var[T_W]}}$$

,wobei gilt:

$$E[T_W] = \frac{N(N+M+1)}{2} = \frac{10(10+10+1)}{2} = 105$$

$$\sqrt{Var[T_W]} = \sqrt{\frac{10 * 10(10+10+1)}{12}} = \sqrt{175}$$

Man erhält:

$$Z = \frac{T_W - E[T_W]}{\sqrt{Var[T_W]}} = \frac{118 - 105}{\sqrt{175}} = 0{,}98$$

Der kritische Wert der Standardnormalverteilung ist 1,96 bzw. -1,96. Damit kann die Nullhypothese nicht abgelehnt werden.

b) Ich bin mir nicht sicher, welcher Test mit „t-Test“ gemeint ist und halte den im Skript als preTest/Post Test bezeichneten Test für angebracht.

Man bildet zunächst die paarweisen Differenzen:

Davor	Danach	Differenz
6,24	6,1	0,14
5,8	5,75	0,05
5,57	5,6	-0,03
6,04	5,91	0,13
5,56	5,3	0,26
5,92	5,62	0,3
6,17	6,29	-0,12
5,48	5,28	0,2
6,11	5,7	0,41
6,2	6,01	0,19

Mittelwert der Differenz ist 0,153.

Varianz der Differenz ist 0,022.

Die Varianz könnte man auch umständlich berechnen (für den Fall, dass in der Aufgabenstellung keine Daten gegeben sind, sondern nur Parameter.

Es gilt:

$$S_D = \sqrt{S_x^2 + S_Y^2 - 2S_x S_y R_{xy}}$$

R_{xy} ist die Korrelation zwischen X und Y.

δ_0 ist die hypothetische Differenz).

Die Prüfgröße ist dann:

$$T = \frac{\overline{D} - \delta_0}{S_D/\sqrt{N}}$$

Die Prüfgröße T ist t-verteilt mit N-1 Freiheitsgraden und es gelten die entsprechenden kritischen Werte.

Einsetzen:

$$T = \frac{0{,}153 - 0}{\frac{0{,}149}{\sqrt{10}}} = 3{,}25$$

Der kritische Wert der t-Verteilung 2,262. Damit muss die Nullhypothese abgelehnt werden. Das Training hat also einen Einfluss auf die Lösungszeit.

Lösung zu Aufgabe 26

a) Die Kontingenztabelle lautet:

		Danach	
		dafür	dagegen
Davor	dafür	36	4
	dagegen	20	10

Achte immer darauf, daß die Summen der Randverteilungen mit dem Umfrageumfang übereinstimmen.

b) Ich würde hier den Chi-Quadrat Anpassungstest verwenden. Die hypothetische Verteilung ist

		Danach	
		dafür	dagegen
Davor	dafür	40	0
	dagegen	0	30

Diese Verteilung kann aber nicht genutzt werden, da nicht alle $h_{ei} \geq 5$ sind.

Man kann aber die folgenden Verteilungen nutzen:

Beobachtete Verteilung:

	Davor	Danach
Dafür	40	56
Dagegen	30	14

Hypothetische Verteilung:

	Davor	Danach
Dafür	40	40
Dagegen	30	30

Als Prüfgröße nutzt man:

$$\chi_*^2 = \sum_{i=1}^{m} \frac{(h_{oi} - h_{ei})^2}{h_{ei}}$$

Dabei ist m die Anzahl an Merkmalsausprägungen (bzw. Merkmalsklassen bei stetigen Merkmalen). Es muss gelten $h_{ei} \geq 5$, ansonsten kann die Verteilung nicht verwendet werden.

Dies ist ein einseitiger Test und die obere Annahmegrenze ist der Wert der Chi Quadrat Verteilung mit $m-1$ Freiheitsgraden und Signifikanzniveau $1-a$.

Nach Einsetzen erhält man 14,93 für di Prüfgröße. Der kritische Wert der Chi-Quadrat Verteilung ist 3,84. Die Nullhypothese der Gleichheit der Verteilungen muss abgelehnt werden.

Lösung zu Aufgabe 27

a) Für den Test

$$H_0: p = 0$$

$$H_1: p \neq 0$$

nutzt man die Prüfgröße:

$$T = \sqrt{N-2}\frac{R}{\sqrt{1-R^2}}$$

T ist t-verteilt mit N-2 Freiheitsgraden.

Für R erhält man

$$r = \frac{Cov(X,Y)}{\tilde{s}_x * \tilde{s}_y} = \frac{-1{,}62}{16{,}85 * 11{,}88} = 0$$

Die Nullhypothese kann nicht abgelehnt werden (Die Werte der t-Verteilung müssen bei einer Prüfgröße von 0 nicht berechnet werden).

b) Man erhält folgende Ränge:

X	124	79	118	102	86	89	109	128	114	95
Rang(X)	9	1	8	5	2	3	6	10	7	4
Y	100	94	101	112	76	98	91	73	90	84
Rang(Y)	8	6	9	10	2	7	5	1	4	3

Nun wird der Korrelationskoeffizient für die Ränge berechnet:

$$r = \frac{Cov(X,Y)}{\tilde{s}_x * \tilde{s}_y} = \frac{2{,}5}{3{,}03 * 3{,}03} = 0{,}27$$

Für die Prüfgröße gilt dann:

$$T = \sqrt{10-2}\frac{0{,}27}{\sqrt{1-0{,}27^2}} = 0{,}79$$

Auch hier wird der kritische Wert der t-Verteilung nicht überschritten. Die Nullhypothese kann nicht abgelehnt werden.

Lösung zu Aufgabe 28

Nullhypothese des Chi-Quadrat-Unabhängigkeitstests ist die Unabhängigkeit der beiden Merkmale. Die Alternativhypothese ist die Abhängigkeit.

Bei Test des Korrelationskoeffizienten gilt:

Für den Test

$$H_0: p = 0$$

$$H_1: p \neq 0$$

a) Es kann sein, dass die Merkmale abhängig sind, obwohl der Korrelationskoeffizient Null ist. Dieser misst nur die lineare Abhängigkeit, aber nicht die nicht-lineare. Existiert also keine nicht-lineare Abhängigkeit, so sind die Nullhypothesen äquivalent.

b) Wie schon in a) gesagt, misst der Korrelationskoeffizient nur die lineare Abhängigkeit. Beim Chi-Quadrat-Unabhängigkeitstest werden alle Abhängigkeiten getestet und er gilt für beliebiges Meßniveau. Der Korrelationskoeffizient gilt nur für metrisch messbare Merkmale.

Beispiel: Ein Datensatz mit einem nicht metrisch messbaren Merkmal kann nicht mit dem Test über den Korrelationskoefizienten behandelt werden. Als Beispiel kann das Beispiel aus dem Fernuni-Skript zum Chi-Quadrat-Unabhängigkeitstest gewählt werden.

c) Eine asymptotische Verteilung nähert sich für große Stichprobenumfänge einer Verteilung an. Ein t-verteiltes Merkmal ist beispielsweise asymptotisch normalverteilt, da sich die t-Verteilung für große N der Normalverteilung annähert.

Lösung zu Aufgabe 29

Ein signifikanter Korrelationskoeffizient kann zufällig zustandekommen. Angenommen auf der Suche nach einem profitablen Aktienhandelssystem teste ich die Wertentwicklung einer bestimmten Aktie im Vergleich zu den Fussballergebnissen einer Fussballmannschaft. Offensichtlich gibt es hier keinen Kausalzusammenhang und es sollte auch keine signifikanten Korrelationskoeffizienten geben. Teste ich nun aber sehr viele Fussballmanschaften gegen sehr viele Aktien, so wird es zufällig einen signifikant von 0 verschiedenen Korralationskoeffizienten geben. Die Aussage „Immer wenn Mannschaft X gewinnt, steigt Aktie Y“ wäre dennoch Unsinn.

Lösung zu Aufgabe 30

a) Die Kontingenztafel lautet:

	Dafür	Dagegen	Keine Meinung	Summe
Parteigebunden	50	130	20	200
Parteilos	350	1270	180	1800
Summe	400	1400	200	2000

b) Die Nullhypothese ist die Unabhängigkeit der Merkmale. Die Prüfgröße ist Chi-Quadrat-verteilt und steigt mit der Abhängigkeit.

c) Die hypothetische Verteilung unter Unabhängigkeit erhält man jeweils aus dem Produkt der Randverteilungen geteilt durch den Stichprobenumfang.

	Dafür	Dagegen	Keine Meinung	Summe
Parteigebunden	40	140	20	200
Parteilos	360	1260	180	1800
Summe	400	1400	200	2000

Die 40 erhält man beispielsweise aus 400*200/2000

Die Testgröße ist dann:

$$\chi_*^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(h_{oij} - h_{eij}\right)^2}{h_{eij}}$$

wobei r die Anzahl an Ausprägungen von X und s die Anzahl an Ausprägungen von Y bezeichnet.

Man erhält

$$\chi_*^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(h_{oij} - h_{eij}\right)^2}{h_{eij}} = 3{,}57$$

Der kritische Wert der Chi-Quadrat-Verteilung mit 2 Freiheitsgraden und $a = 1\%$ ist 9,21. Die Nullhypothese kann nicht abgelehnt werden.

Lösung zu Aufgabe 31

a) Es gilt:

$$r = \frac{Cov(X,Y)}{\tilde{s}_x * \tilde{s}_y} = \frac{10}{5 * 4} = 0{,}5$$

b) Möchte man den Korrelationskoeffizienten nicht aus Null, sondern auf einen konkreten Wert p_0 testen, so nutzt man als Prüfgröße:

$$Z = \frac{\sqrt{N-3}}{2}\left(\ln\frac{1+R}{1-R} - \ln\frac{1+p_0}{1-p_0}\right)$$

Z ist für große Stichprobenumfänge approximativ standardnormalverteilt.

Einsetzen:

$$Z = \frac{\sqrt{500-3}}{2}\left(\ln\frac{1+0{,}5}{1-0{,}5} - \ln\frac{1+0{,}6}{1-0{,}6}\right) = -3{,}21$$

Der kritische Bereich für die Nullhypothese ist <1,65. Damit kann die Nullhypothese abgelehnt werden.

c) Es gilt:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = R_{xy}\frac{S_y}{S_x}$$

Einsetzen;

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = 0{,}5\frac{4}{5} = 0{,}4$$

$$a = \bar{y} - b\bar{x} = 14$$

d)Es gilt für die Prüfgröße:

$$T_b = \frac{\hat{b} - b_0}{\widehat{\sigma_b}} \sim t(N-2)$$

Einsetzen:

$$T_b = \frac{0{,}4}{1{,}75} = 0{,}23$$

Damit kann die Nullhypothese nicht abgelehnt werden. Der kritische Bereich der t-Verteilung wäre $< -1{,}96$ und $> 1{,}96$.

e) Es muss folgende Formel gelöst werden:

$$a + bx = 48$$

Einsetzen:

$$14 + 0{,}4x = 48$$

$$x = 85$$

f)

Das gesuchte Prognoseintervall für einen konkreten Wert Y_o beträgt:

$$\widehat{Y_0} \pm t(1 - \frac{a}{2}, N - 2)\hat{\sigma}\sqrt{1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sum_n (X_n - \bar{X})^2}}$$

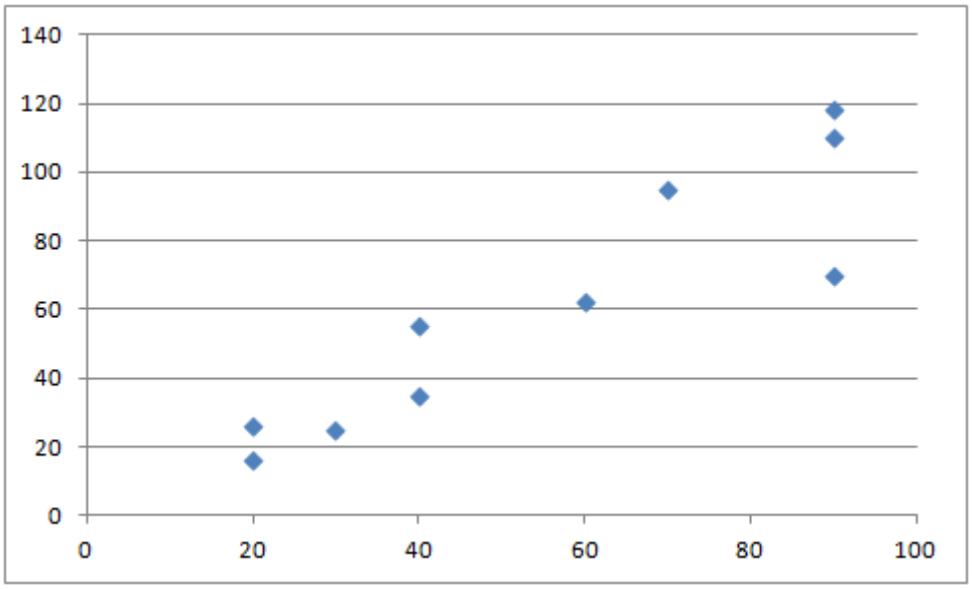
Einsetzen:

$$48 \pm t\left(1 - \frac{a}{2}, N - 2\right) 1{,}75\sqrt{1 + \frac{1}{500} + \frac{(85 - 40)^2}{12475}} = 48 \pm 3{,}44$$

Das Intervall beträgt 44,56 bis 51,44.

Lösung Aufgabe 32

a)



b) Es gilt folgende Formel für die Berechnung der Parameter:

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

Zur Berechnung müssen noch die Mittelwerte berechnet werden- die Summen sind schon in der Aufgabenstellung gegeben.

Man erhält:

$$\bar{x} = 55$$

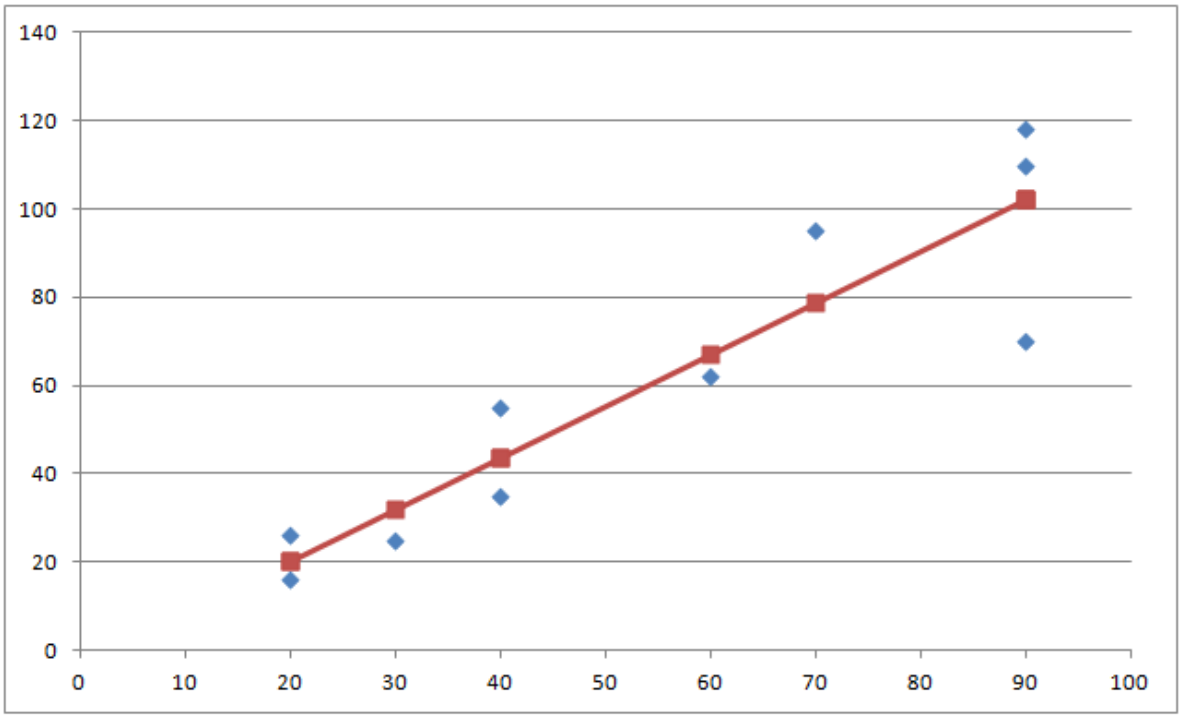
$$\bar{y} = 61{,}2$$

Einsetzen:

$$b = \frac{42380 - 10 * 55 * 61{,}2}{37700 - 10 * 55^2} = 1{,}17$$

$$a = 61{,}2 - 1{,}17 * 55 = -3{,}18$$

Grafisch sieht das dann so aus:



c) In diesem Fall ist die Grundgesamtheit unendlich groß, nämlich alle Entfernungen, die geschätzt werden könnten. Aus dieser Grundgesamtheit werden dann 10 verschiedene Werte gewählt, die die Stichprobe bilden.

d) Nein, mit steigender Entfernung, steigt auch die Abweichung der Schätzung zum wahren Wert. Das ist auch grafisch gut erkennbar.

e)Die Hypothese würde bedeuten, dass die Schätzungen die wahren Werte perfekt treffen.

Lösung zu Aufgabe 33

a) Es gilt folgende Formel für die Berechnung der Parameter:

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

Zur Berechnung müssen noch die Mittelwerte berechnet werden- die Summenterme sind schon in der Aufgabenstellung gegeben.

Man erhält:

$$\bar{x} = 2{,}5$$

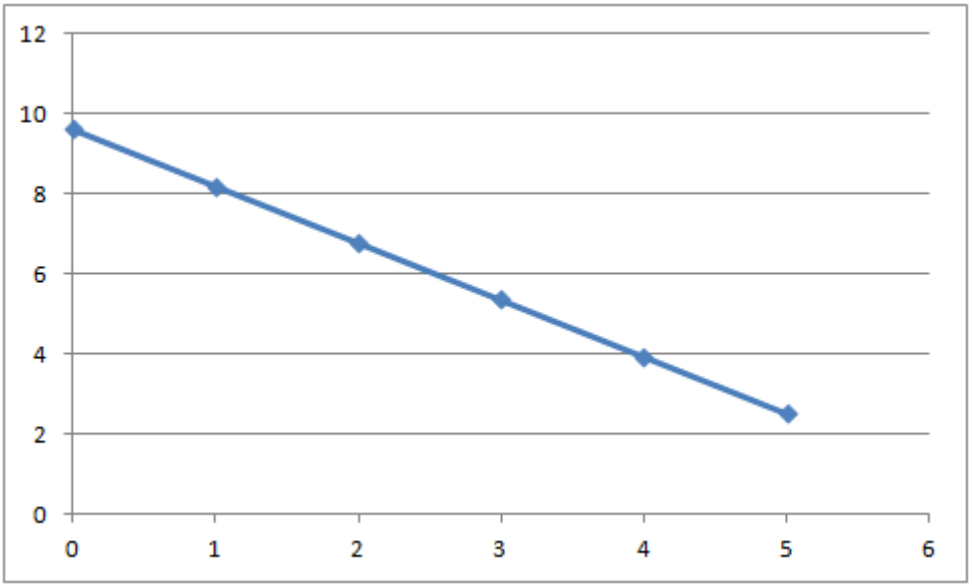
$$\bar{y} = 6{,}07$$

Einsetzen:

$$b = \frac{331 - 30 * 2{,}5 * 6{,}07}{275 - 30 * 2{,}5^2} = -1{,}42$$

$$a = \bar{y} - b\bar{x} = 6{,}07 - (-1{,}42) * 2{,}5 = 9{,}62$$

b)



c) Für das Konfidenzintervall gilt:

$$\hat{b} \pm t\left(1-\frac{a}{2}, N-2\right)\widehat{\sigma_b}$$

Und für $\widehat{\sigma_b}$ gilt:

$$\widehat{\sigma_b} = \frac{\hat{\sigma}}{\sqrt{\sum_n (X_n - \bar{X})^2}}$$

Einsetzen:

$$\widehat{\sigma_b} = \frac{1{,}23}{9{,}35} = 0{,}13$$

$$\hat{b} \pm t\left(1-\frac{a}{2}, N-2\right)\widehat{\sigma_b} = \hat{b} \pm 2{,}048 * 0{,}13$$

Das Konfidenzintervall beträgt: -1,69 bis -1,15.

Damit ist b signifikant von Null verschieden. Die Null liegt nicht innerhalb des Konfidenzintervalls.

d) Das gesuchte Prognoseintervall für einen konkreten Wert Y_o beträgt:

$$\widehat{Y_0} \pm t(1-\frac{a}{2}, N-2)\hat{\sigma}\sqrt{1+\frac{1}{N}+\frac{(X_0-\bar{X})^2}{\sum_n (X_n-\bar{X})^2}}$$

Einsetzen:

Das gesuchte Prognoseintervall für einen konkreten Wert Y_o beträgt:

$$\widehat{Y_0} \pm t\left(1-\frac{a}{2}, N-2\right)1{,}23\sqrt{1+\frac{1}{30}+\frac{(2{,}5-2{,}5)^2}{\sum_n (X_n-\bar{X})^2}}$$

$$\widehat{Y_0} \pm t\left(1-\frac{a}{2}, N-2\right)1{,}23\sqrt{1+\frac{1}{30}} = \widehat{Y_0} \pm 2{,}56$$

Für $\widehat{Y_0}$ gilt:

$$\widehat{Y_0} = a - bx_0 = 9{,}62 - 1{,}42 * 2{,}5 = 6{,}07$$

(dies ist der Mittelwert, da für x auch der Mittelwert gewählt wurde).

Für das Intervall erhält man 3,51 bis 8,63.

Lösung zu Aufgabe 34

a) Es gilt folgende Formel für die Berechnung der Parameter:

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

Zur Berechnung müssen noch die Mittelwerte und die Summenterme berechnet werden. Man erhält:

$$\bar{x} = 20$$

$$\bar{y} = 15$$

$$\sum x_i y_i = 2028$$

$$\sum x_i^2 = 2662$$

Einsetzen:

$$b = \frac{2028 - 6 * 20 * 15}{2662 - 6 * 400} = 0{,}87$$

$$a = \bar{y} - b\bar{x} = 15 - 0{,}87 * 15 = -2{,}4$$

Die Zeichnung überlasse ich Dir.

b) Es gilt für die Prüfgröße:

$$T_b = \frac{\hat{b} - b_0}{\widehat{\sigma_b}} \sim t(N-2)$$

und

$$\widehat{\sigma_b} = \frac{\hat{\sigma}}{\sqrt{\sum_n (X_n - \bar{X})^2}}$$

Einsetzen:

$$\widehat{\sigma_b} = \frac{1{,}55}{16{,}19} = 0{,}1$$

$$T_b = \frac{0{,}87 - 1}{0{,}1} = -1{,}3$$

Der kritische Bereich der t-Verteilung mit 4 Freiheitsgraden ist <-2,77 und >2,77. Daher kann die Nullhypothese nicht abgelehnt werden.

c) Die Korrelation kann man direkt aus den Daten berechnen, oder aus den Korrelationskoeffizienten (wie hier gefordert).

Es gilt:

$$b = R_{xy}\frac{S_y}{S_x}$$

Es müssen noch die Standardabweichungen berechnet werden.

$$S_y = 6{,}45$$

$$S_x = 7{,}24$$

Einsetzen:

$$0{,}87 = R_{xy}\frac{6{,}45}{7{,}24}$$

Daraus folgt

$$R_{xy} = 0{,}98$$

Hier kann man gut die Probe über die Berechnung der Korrelation aus den Daten machen, aber das überlasse ich Dir.

d) Es gilt:

$$R^2 = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2}$$

$\tilde{s}_{\hat{y}}^2$ muss noch berechnet werden. Man nutzt folgende Schätzwerte, die sich aus a) ergeben:

(Achtung, ich habe die Reihenfolge der X-Werte geändert)

X	10	15	19	20	26	30
Y	6,3	10,65	14,13	15	20,22	23,7

Man erhält:

$$R^2 = \frac{39{,}66}{41{,}6} = 0{,}95$$

e)

$$-2{,}4 + 0{,}87 * 25 = 19{,}35$$

f) Das gesuchte Prognoseintervall für einen konkreten Wert Y_o beträgt:

$$\widehat{Y_0} \pm t(1-\frac{a}{2}, N-2)\hat{\sigma}\sqrt{1+\frac{1}{N}+\frac{(X_0-\bar{X})^2}{\sum_n (X_n-\bar{X})^2}}$$

Es muss die Varianz um die Regressionsgerade herum berechnet werden:

$$\widehat{\sigma^2} = \frac{1}{N-2}\sum_n \widehat{\epsilon_n}^2$$

Dazu nutzt man folgende Tabelle

X	10	15	19	20	26	30
Y	8	10	12	15	20	25
$\widehat{Y}$	6,3	10,65	14,13	15	20,22	23,7

Daraus erhält man:

$$\widehat{\sigma^2} = \frac{1}{6-2}(8-6{,}2)^2 + \cdots + (25-23{,}7)^2 = 2{,}4$$

Einsetzen:

$$\widehat{Y_0} \pm t\left(1-\frac{a}{2}, N-2\right) 1{,}55\sqrt{1+\frac{1}{6}+\frac{(25-20)^2}{262}}$$

Das Konfidenzintervall beträgt

$$19{,}35 \pm 4{,}83$$

Lösung zu Aufgabe 35

Für die Prüfgröße $F = \frac{\frac{SQE}{I-1}}{SQR/(I(J-1))}$ benötigt man SQE und SQR:

$$SQE = J\sum_i \bar{Y}_i^2 - IJ * \bar{Y}^2$$

I ist die Anzahl an Gruppen, also in diesem Fall 3. J ist die Anzahl an Gruppenmitgliedern, also in diesem Fall 8.

Man muss also den Gesamtdurchschnitt $\bar{Y}$ ermitteln:

$$\bar{Y} = \frac{1}{3}(46{,}38 + 36{,}75 + 20{,}38) = 34{,}5$$

Und den quadrierten Durchschnitt der einzelnen Gruppen $\bar{Y}_i^2$:

$$\bar{Y}_i^2 = 46{,}38^2 + 36{,}75^2 + 20{,}38^2 = 3.916{,}34$$

Einsetzen:

$$SQE = J\sum_i \bar{Y}_i^2 - IJ * \bar{Y}^2$$

$$SQE = 8 * 3.916{,}34 - 3 * 8 * 1.190{,}25 = 2.764{,}72$$

Für SQR gilt:

$$SQR = \sum_{ij}(Y_{ij} - \bar{Y}_i)^2 = 359{,}25$$

Nun sind alle Werte für die Prüfgröße bekannt. Einsetzen:

$$F = \frac{\frac{2.764{,}72}{3-1}}{359{,}25/(3(8-1))} = 80{,}81$$

Der Wert ist deutlich größer als der kritische Wert der F-Verteilung (das kann man sagen, ohne den genauen Wert der F-Verteilung abzulesen).

Die Nullhypothese kann also abgelehnt werden. Die Gruppen haben nicht die gleichen Mittelwerte und damit über der Faktor Lärm einen Einfluss auf die Arbeitsleistung aus.

Lösung zu Aufgabe 36

a) Man könnte hier auch die Gleichheit der Mittelwerte mit einem Test aus dem vorherigen Kapitel durchführen, aber es ist ausdrücklich nach einem einseitigen Test gefragt. Daher muss man die Varianzzerlegung nutzen.

Als Nullhypothese wählt man die Gleichheit der Mittelwerte.

Die Prüfgröße ist

$$F = \frac{\frac{SQE}{I-1}}{SQR/(I(J-1))}$$

SQE und SQR werden wie in Aufgabe 35 berechnet.

Es folgt:

$$SQE = J\sum_i \bar{Y}_i^2 - IJ * \bar{Y}^2 = 0{,}31$$

Für SQR gilt:

$$SQR = \sum_{ij}(Y_{ij} - \bar{Y}_i)^2 = 0{,}66$$

Einsetzen:

$$F = \frac{\frac{0{,}31}{2-1}}{0{,}66/(2(10-1))} = 8{,}45$$

Der kritische Wert der F-Verteilung für $F(I-1, J-2)$ ist 5,32. Die Nullhypothese muss abgelehnt werden.

b) Die Lösung erfolgt äquivalent zu a)

Es folgt:

$$SQE = J\sum_i \bar{Y}_i^2 - IJ * \bar{Y}^2 = 0{,}37$$

Für SQR gilt:

$$SQR = \sum_{ij}(Y_{ij} - \bar{Y}_i)^2 = 1{,}05$$

Einsetzen:

$$F = \frac{\frac{0{,}37}{3-1}}{1{,}05/\big(3(10-1)\big)} = 4{,}76$$

Der kritische Wert der F-Verteilung für $F(I-1, J-2)$ ist 4,46. Die Nullhypothese muss abgelehnt werden.

Lösung zu Aufgabe 37

a) Dies ist wieder die Standardaufgabe, nämlich die Gleichheit der Mittelwerte zu prüfen.

Die Prüfgröße ist

$$F = \frac{\frac{SQE}{I-1}}{SQR/(I(J-1))}$$

SQE und SQR werden wie in Aufgabe 35 berechnet.

Es folgt:

$$SQE = J\sum_i \bar{Y}_i^2 - IJ * \bar{Y}^2 = 120$$

Für SQR gilt:

$$SQR = \sum_{ij} (Y_{ij} - \bar{Y}_i)^2 = 78$$

Einsetzen:

$$F = \frac{\frac{120}{3-1}}{78/(3(5-1))} = 9{,}23$$

Der kritische Wert der F-Verteilung für $F(I-1, J-2)$ ist 3,89. Die Nullhypothese muss abgelehnt werden.

b) Nun gilt es zu testen, welche der Mittelwerte gleich und welche ungleich sind. Für mich geht aus der Aufgabenstellung nicht klar hervor, ob a^* oder a als Fehlerwahrscheinlichkeit genutzt werden soll. Ich entscheide mich hier für a^*.

Die Fehlerwahrscheinlichkeit berechnet man nach

$$a^* = a/k$$

k bezeichnet hier die Anzahl an Tests, die durchgeführt werden müssen und es gilt:

$$k = \binom{I}{2} = \frac{I!}{2! * (I-2)!}$$

In diesem Fall also $k = 3$.

Man erhält

$$a^* = 0{,}0167$$

Die Teststatistik ist nun:

$$T_{ii'} = \frac{\bar{Y}_i - \bar{Y}_{i'}}{S\sqrt{2/J}} \sim t(IJ - I)$$

,wobei hier zum Niveau a^* geprüft wird.

Es muss also s ermittelt werden. Es gilt:

$$\sigma^2 = \frac{SQR}{I(J-1)} = MQR$$

Für die geschätzte Standardabweichung erhält man:

$$\sigma^2 = \frac{78}{3(5-1)} = 6{,}5$$

$$s = \hat{\sigma} = 2{,}55$$

Einsetzen:

Für den Vergleich von Klima I und Klima II gilt:

$$T_{ii'} = \frac{0}{2{,}55\sqrt{2/5}}$$

Offensichtlich kann diese Nullhypothese nicht abgelehnt werden.

Für den Vergleich von Klima I und Klima III gilt:

$$T_{ii'} = \frac{-6}{2{,}55\sqrt{2/5}} = -3{,}72$$

Der kritische Bereich der t-Verteilung für 12 Freiheitsgrade und $a^* = 1{,}67\%$ ist ca. <3 und >3. Die Nullhypothese muss also abgelehnt werden.

Da Klima II und Klima I denselben Mittelwert haben, wird auch hier die Nullhypothese abgelehnt werden.

c) Man kann für „unter 12“ die Null wählen und für „12 oder mehr“ die 1. Dann wird mit den entsprechenden Daten das Testverfahren aus a) und b) durchgeführt. Dabei gehen die Daten über das Ausmaß der Abweichung von der 12 verloren.

14.0 Statistik Kurseinheit II

Diese Kurseinheit wurde bisher deutlich seltener getestet als die erste und auch der Aufgabenteil ist deutlich kleiner. Während in Kurseinheit I sehr viele Formeln zu nutzen waren, sind diese nun eher selten. Daher lässt sich die Kurseinheit auch deutlich angenehmer lesen, was Du auch auf jeden Fall einmal tun solltest. Für die Klausur musst du viele Fachbegriffe beherrschen, die in MC-Form abgeprüft werden. Ich werde mich auf einige wenige Begriffserklärungen beschränken, da es keinen Mehrwert für Dich darstellt, wenn ich einen Skriptabschnitt abschreibe. Wird in der Klausur nach einem Begriff gefragt, so brauchst du diesen nur nachzuschlagen. Für besonders hilfreich halte ich die Beantwortung der Fragen am Ende des Kapitels.

Falsifikation: Falsifikation ist der Nachweis der Ungültigkeit einer Aussage. Je konkreter eine Aussage und je mehr Bedingungen (Falsifikatoren), desto eher kann sie wiederlegt werden. Eine Aussage wie „Wenn der Dax über 7000 steigt, dann könnte er steigen, aber es besteht auch das Risiko von fallenden oder stagnierenden Kursen“ ist also nicht falsifizierbar.

Verifikation: die Überprüfbarkeit einer Aussage. Eine Aussage kann falsifizierbar sein, aber dies ist evtl. mit den gegebenen technischen Möglichkeiten nicht überprüfbar.

Tautologie: Eine Aussage, die immer stimmt. Siehe Beispiel Falsifikation.

Kontradiktion: Bei einer Kontradiktion kann man von der Wahrheit der einen Aussage auf die Falschheit der anderen schliessen. Beispiel: „Die Welt ist eine Kugel“ und „die Welt ist eine Scheibe“.

Reliabilität: Zuverlässigkeit/Genauigkeit der erhobenen Daten. In der Testtheorie setzt sich der beobachtete Wert aus einem wahren Wert und einem Meßfehler zusammen. Die Reliabilität ist das Verhältnis der Varianz des wahren Wertes zur gesamten Varianz. Gute Reliabilität ist eine Voraussetzung für Validität, wobei zu hohe Reliabilität zu Lasten der Validität geht. Hohe Reliabilität ist Voraussetzung für die Wiederholbarkeit von Messungen.

Validität: Die Validität, ist ein Maß dafür, ob die bei der Messung erzeugten Daten wie beabsichtigt die zu messende Größe repräsentieren

Exhaustion: Modifikation oder Erweiterung einer Theorie aufgrund von Untersuchungsergebnissen, die die ursprüngliche Form der Theorie falsifizieren.

Hypothesenprüfung mit Prädiktionsregeln

Hypothesen werden teils unscharf formuliert und bieten einen relativ hohen Interpretationsspielraum. Um die Hypothese möglichst genau zu formulieren, nutzt man sogenannte Wahrheitstafeln.

Wahrheitstafeln

Am besten erklärt man dies an einem **Beispiel:**

Die Hypothese: „Rauchen verursacht Lungenkrebs“ kann auf verschiedene Arten interpretiert werden. Es ist zum Beispiel fraglich:

-Bekommt jeder Raucher Lungenkrebs?

-Wenn man Lungenkrebs bekommt, ist man dann immer Raucher?

In einer Wahrheitstafel trägt man nun die Merkmale „Raucher“ und „Lungenkrebs in einer Kreuztabelle ein und für jede Kombination, die sich mit der Hypothese deckt, trägt man eine 1 ein und für jede, die sich nicht deckt, trägt man eine 0 ein.

Möchte man beispielsweise aussagen, dass nur Raucher Lungenkrebs bekommen, aber nicht alle, so lautet die Wahrheitstafel:

	Raucher	Nichtraucher
Lungenkrebs	1	0
Kein Lungenkrebs	1	1

Möchte man dagegen aussagen, dass sowohl Raucher als auch Nichtraucher Lungenkrebs bekommen können, aber Raucher immer Lungenkrebs bekommen, so lautet die Wahrheitstafel:

	Raucher	Nichtraucher
Lungenkrebs	1	1
Kein Lungenkrebs	0	1

A priori Regeln-Cohens Kappa

A-priori Regeln, sind Prädiktionsregeln, die aus inhaltlichen Theorien abgeleitet werden. Im Gegensatz dazu werden a-posteriori Regeln (später dazu mehr) anhand von Daten ermittelt.

Ich muß sagen, daß es mir nicht leicht fiel, die Erklärung von Cohens Kappa im Fernuni-Skript zu verstehen. Ich empfehle jedem Studenten kurz einmal das Kapitel über Cohens Kappa im Fernuni-Skript zu lesen. Am besten bevor und nachdem du meine Erklärung gelesen hast. Solltest du trotz aller Hilfe die Erklärung im Skript nicht verstehen, so würde ich auch keine weitere Zeit damit verschwenden- es mag aber hilfreich sein, die Formulierungen des Lehrstuhls einmal gelesen zu haben.

Cohens Kappa möchte ich auch wieder an einem **Beispiel** beschreiben: Zwei Kritiker bewerten einen Untersuchungsgegenstand. Dieser kann positiv, neutral oder negativ bewertet werden.

Insgesamt haben beide 100 Bewertungen abgegeben. Ihre Bewertungen sind in folgender Tabelle dargestellt:

Kritiker A/B	Positiv	Neutral	Negativ	Summe
Positiv	23	12	4	39
Neutral	3	25	5	33
Negativ	1	10	17	28
Summe	27	47	26	100

Die Hypothese ist, dass die Kritiker die Untersuchungsgegenstände gleich bewerten.

Aus der Tabelle berechnet man nun folgende Werte:

-Die Wahrscheinlichkeit von übereinstimmenden Urteilen. Dies ist die Summe der Werte der Hauptdiagonalen geteilt durch die Gesamtsumme. Diese nenne ich G(+).

-Die Wahrscheinlichkeit, mit der zwei zufällig urteilende Kritiker übereinstimmen (diese sind unabhängig voneinander und entsprechen Hauptdiagonale der Tabelle bei Unabhängigkeit geteilt durch N). Diese nenne ich G(-).

Cohens Kappa berechnet sich nun nach folgender Formel:

$$\mathrm{K} = \frac{G(+) - G(-)}{1 - G(-)}$$

,wobei gilt:

$$G(+) = \sum_{i=1}^{n} f_{ii}$$

$$G(-) = \sum_{i=1}^{n} f_i * f_i$$

G(-) ist die Summe der Produkte der Randverteilungen (Solltest du hier Verständnisprobleme haben, wiederhole bitte das Kapitel über statistische Unabhängigkeit aus Grundlagen der Wirtschaftsmathematik und Statistik).

In meinem Beispiel war also die Wahrheitstabelle

Kritiker A/B	Positiv	Neutral	Negativ
Positiv	1	0	0
Neutral	0	1	0
Negativ	0	0	1

Als Beispiel werde ich einmal das Kappa für die Beispielwerte berechnen:

$$G(+) = \sum_{i=1}^{n} f_{ii} = 0{,}23 + 0{,}25 + 0{,}17 = 0{,}65$$

$$G(-) = \sum_{i=1}^{n} f_i * f_i = 0{,}27 * 0{,}39 + 0{,}47 * 0{,}33 + 0{,}26 * 0{,}28 = 0{,}33$$

$$\mathrm{K} = \frac{G(+) - G(-)}{1 - G(-)} = \frac{0{,}65 - 0{,}33}{1 - 0{,}33} = 0{,}48$$

Je nachdem welche Hypothese getestet werden soll, kann die Wahrheitstabelle auch anders aussehen und die Berechnung von G(-) muß entsprechend angepasst werden. Es muss also nicht auf Unabhängigkeit geprüft werden, sondern es kann auf jede beliebige hypothetische Verteilung geprüft werden.

A posteriori Regeln

Für die A posteriori Regeln werde ich die Variablenbezeichnung des Fernuni Skriptes nutzen.

Ich werde zunächst $\lambda_{X\to Y}$ kurz erläutern:

Man geht von einer relativen Häufigkeitstabelle zweier Merkmale X und Y aus und berechnet die Modalwerte der bedingten Häufigkeiten von Y gegeben X. 1 minus die Summe dieser Modalwerte nennt man F(+). Außerdem berechnet man den Modalwert von Y ohne X. 1- diesen Modalwert nennt man F(-).

Das Ganze an einem **Beispiel:**

	Y1	Y2	Summe
X1	0,4	0,2	0,6
X2	0,3	0,1	0,4
Summe	0,7	0,3	1

Der Modalwert von Y gegeben X1 ist nun der größte Wert aus 0,4 und 0,2. Der Modalwert für Y gegeben X2 ist der größte Wert aus 0,3 und 0,1. Der Modalwert von Y ohne X ist der größte Wert aus 0,7 und 0,3.

F(+) ist also $1-(0{,}4+0{,}3)=0{,}3$

F(-) ist $1-0{,}7=0{,}3$.

Das Prädiktionsmaß Lambda X->Y berechnet man nun wie folgt:

$$\lambda_{X\to Y}=\frac{F(-)-F(+)}{F(-)}$$

Analog kann auch das Prädiktionsmaß Lambda Y->X (dann mit X gegeben Yi) berechnet werden und Lambda X<->Y. Bei letzerem ist F(+) die Summe aus F(+) für $\lambda_{X\to Y}$ und F(+) für $\lambda_{Y\to X}$. Dasselbe gilt für F(-).

Lösungen zu den Aufgaben der Kurseinheit II

Lösung zu Aufgabe 1

a) Es sind folgende Kombinationen möglich:

ABC

ACB

BAC

BCA

CAB

CBA

b)Nein, die Gruppen sind „vergleichbar“ und nicht identisch.

c) Antworten sind oft Kontextabhängig. So ist es möglich, dass bei drei items A,B,C verschiedene Mittelwerte in Abhängigkeit der Reihenfolge auftreten. Dies ist auch in der Quantenmechanik zu beobachten, bei der Orts- und Impulsmessungen nicht verträglich sind.

Es tritt auch oft das Problem der Reaktivität auf. Eine Person, die weiß, dass sie an einer statistischen Untersuchung teilnimmt, wird nicht natürlich reagieren.

Anmerkung: Natürlich wird vom Studenten nicht erwartet sich mit Quantenmechanik auszukennen- der Trick ist hier das Thema der Frage im Skript aufzuschlagen und entsprechend abzuschreiben.

Lösung zu Aufgabe 2

a) Die Antwort findest du auf S.189:

Erkundungsphase

Theoretische Phase

Planungsphase

Untersuchungsphase

Auswertungsphase

Entscheidungsphase

b)

Indikator: Ein beobachtbarer Sachverhalt.

Operationalisierung: Die Art, wie den Begriffen die Indikatoren zugeordnet werden. (S.191)

Falsifikatoren: Bedingungen, mit denen eine Aussage als falsch beurteilt werden können.

Tautologie: Eine Aussage, die immer stimmt.

Bei einer Kontradiktion kann man von der Wahrheit der einen Aussage auf die Falschheit der anderen schliessen. Beispiel: „Die Welt ist eine Kugel“ und „die Welt ist eine Scheibe“.

c) Im Labor kann die Anzahl an situativen Störvariablen kontrolliert werden, im Experiment die Anzahl an personellen Störvariablen. Bei Survey-Studien sind auch die unabhängigen Variablen zufällig.

d)Störvariablen sind Variablen, die einen Einfluss auf das Ergebnis haben, aber nicht mit erhoben werden. Sie werden meist als stochastische Größen modelliert.

e) Nein. Möchte ich zum Beispiel mit einer falsch geeichten Waage 1 kg Fleisch wieder, so wird auch eine empirische Untersuchung immer das falsche Ergebnis bestätigen. Somit kann ich keine Aussage über die Validität machen.

Lösung zu Aufgabe 3

a)Falsch. Der Informationsgehalt hängt von der Falsifizierbarkeit ab.

b)Richtig.

c)Falsch. Siehe Lösung zu 2b.

d)Richtig.

Lösung zu Aufgabe 4

a) In Worten: Da aus A die Gültigkeit von B folgt, folgt auch aus $\bar{B}$ die Gültigkeit von $\bar{A}$ und andersherum.

b) Hier kann man nichts berechnen. Der Ausdruck bezeichnet das Komplement der Schnittmenge aus A und B.

c) Siehe S.211. Es folgt, dass wenn O nicht gilt entweder T oder A oder R falsch sind.

d)Nein, die absolute Richtigkeit einer Theorie ist durch ein signifikantes Ergebnis nicht bewiesen (Siehe S.211).

e) $A \cup B$

Die linke Gleichungsseite bezeichnet umgangssprachlich die Schnittmenge aus „der Vereinigungsmenge von A und B“ und „A“. Dies ist wieder A. Die rechte Gleichungsseite bezeichnet die Vereinigungsmenge aus „A“ und “der Schnittmenge aus A und B“. Das ist auch A.

Lösung zu Aufgabe 5

C wird widerlegt.

Lösung zu Aufgabe 6

A: Falsch. Sie müssen in die WENN Komponente aufgenommen werden.

B:Richtig.

C,D,E: Aussage III besitzt den höchsten Informationsgehalt, da hier ein klarer Zusammenhang zwischen Schulabschluss, Engagement und Einkommen, Firmenposition ermittelt und quantifiziert werden kann.

Lösung zu Aufgabe 7

A: Falsch. Die Reliabilität ist das Verhältnis der Varianz des wahren Wertes zur gesamten Varianz.

B: Falsch. Siehe A.

C: Richtig. Siehe Erklärung zum Bestimmtheitsmaß: r^2 entspricht dem Anteil der erklärten Varianz an der Gesamtvarianz.

D:Richtig.

Lösung zu Aufgabe 8

A: Falsch. Statt Repräsentativität müßte es Objektivität heißen.

B: Falsch. Gute Reliabilität ist eine Voraussetzung für Validität, wobei zu hohe Reliabilität zu Lasten der Validität geht.

C:Richtig.

D: Nein, es ist ein Anteilswert, der Werte zwischen 0 und 1 annimmt.

Lösung zu Aufgabe 9

A: Richtig. Siehe Definition.

B:Falsch. Siehe Definition.

C: Falsch. Reliabilität misst den Anteil an Zufallsfehlern. Systematische Fehler beeinträchtigen die Reliabilität nicht.

D:Falsch. Der Meßfehler bezieht sich auf die Validität.

Lösung zu Aufgabe 10

A: Falsch. Beim Kohen Kappa werden die Beurteiler verglichen.

B: Richtig. Siehe Beschreibung von Cohens Kappa.

C:Richtig. Diese Darstellung hätte der Professor auch gerne in das Skript schreiben können.

D:Nein. Es kann hohe negative Werte annehmen (Falls G(+) sehr klein und G(-) relativ groß ist).

15.0 Statistik Kurseinheit III

In der Kurseinheit III werden kaum neue Themen eingeführt. Es geht eher um die Wiederholung und Vertiefung des bisher Gelernten. Ich muss zugeben, dass ich mich oft über mehrere Seiten gefragt habe, was hier zu lernen oder zu verstehen ist, bis ich gemerkt habe, dass es nichts zu lernen oder zu verstehen gab. Ich werde mich im folgenden auf das beschränken, was ich für prüfungsrelevant halte.

Nominale Skalenniveaus

Hier solltest du nochmal die Messbarkeiten von Variablen aus Grundlagen der Wirtschaftsmathematik und Statistik wiederholen:

Messbarkeitseigenschaften von Merkmalen

Es wird generell zwischen den folgenden Merkmalsarten unterschieden:

- **Qualitative Merkmale**: Die Merkmale können nur nach dem Kriterium „gleich“ oder „ungleich“ gemessen werden. Eine Rangfolge ist nicht möglich. Man sagt: Das Merkmal ist nur **nominal** messbar. **Beispiel:** Staatsangehörigkeit.

- **Intensitätsmäßige Merkmale**: Die Merkmale können nach dem Kriterium „gleich“ oder „ungleich“ gemessen werden und eine Rangfolge ist möglich (Der Abstand zwischen den einzelnen Ausprägungen besitzt aber keine Aussagekraft). Das Merkmal ist **ordinal** messbar.

Beispiel: Die Studenten der Fernuni werden nach Notendurchschnitt in 3 Klassen eingeteilt.

Klasse 1: bis 2.0

Klasse 2: bis 4.0

Klasse 3 : bis 6.0

(Immer wenn ein Merkmal in Klassen eingeteilt wird, handelt es sich um ein ordinal messbares Merkmal.)

- Quantitative Merkmale: Dem Merkmal wird eine Zahl zugewiesen und es ist **metrisch** bzw. **kardinal** messbar. Der Abstand zwischen den einzelnen Ausprägungen besitzt also Aussagekraft. Die Skala nennt man auch Kardinalskala oder Intervallskala. Existiert ein absoluter Nullpunkt (Temperatur in Kelvin), so spricht man von einer Verhältnisskala, existiert keiner, so spricht man von einer Intervallskala (IQ-Skala).

Es wird weiter zwischen diskreten und stetigen Merkmalen unterschieden:

- Diskretes Merkmal: Das Merkmal kann endlich viele, oder abzählbar unendlich viele Ausprägungen annehmen.

Beispiel: Schulnoten, Kinderzahl, Familienstand.

- Stetiges Merkmal: Das Merkmal kann mehr als abzählbar unendlich viele Ausprägungen annehmen.

Beispiel: Körpergröße, Zeitdauer.

Achte bei der Bestimmung des Merkmalstyp auf die Begrenzung der Messgenauigkeit. Ist diese unbegrenzt, so handelt es sich um ein stetiges Merkmal. Ist sie begrenzt, der minimale Abstand zwischen zwei Merkmalen also bestimmbar, so handelt es sich um ein diskretes Merkmal.

Skalentransformation

Ähnlich wie bei Funktionen, wird durch eine Skalentransformation jedem Merkmal einer Skala ein Wert zugeordnet. Die Eigenschaften der Skala müssen dabei unverändert werden.

Beispiel: Bei einer Umfrage nach der beliebtesten Automarke könnte man

BMW $= 1$

Audi $= 2$

VW $= 3$

…

zuordnen.

Die Transformation muss eindeutig sein. Es können also nicht 2 Merkmale dieselbe Zuordnung bekommen (Audi und VW beide die 2).

Die Transformation ist monoton, wenn die Rangfolge erhalten bleibt.

Beispiel: Schulnoten

Sehr gut $= 1$

Gut $= 2$

Befriedigend $= 3$

Ausreichend $= 4$

Mangelhaft $= 5$

Transformationen können immer von einer Skala mit hohem Informationsniveau auf eine Skala mit niedrigerem Informationsniveau vorgenommen werden. Dabei sind folgende Regeln zu beachten:

- Die Transformation einer Rangskala muss mindestens monoton sein.

- Die Transformation einer Kardinalskala muss linear sein.

Achte bei Skalentransformationen immer darauf, dass diese Sinn ergeben müssen. In Klausuren wurde z.B. oft gefragt ob eine lineare Transformation der Form $y = ax + b$ zulässig sei. Solch eine Transformation macht für nominale Merkmale keinen Sinn (Aus „VW“ $\text{wird } a * \text{VW} + b$).

Numerische Codierung von nominalen Merkmalen

Ordnet man nominalen Merkmalen Zahlenwerte zu, so kann man aus diesen Zahlenwerten statistische Kennzahlen wie Varianz und Mittelwerte berechnen. Diese haben aber keine Aussagekraft, da die Zuordnung willkürlich geschieht. Da nominale Merkmale auch keine Rangordnung haben, kann man keine Quantile oder Summenhäufigkeiten berechnen.

Beispiel: Bei einer Wahlumfrage werden dern Parteien CDU, SPD und Grüne jeweils die Zahlenwerte 1,2 und 3 zugeordnet. Man kann nun aber nicht sagen, dass Grüne einen höheren Wert hätte, als die SPD, nur weil ihr der Zahlenwert 3 zugeordnet wurde. Die 3 ist in diesem Fall als Wort zu behandeln.

Sinnvoll wäre zum Beispiel eine Codierung von 0 und 1 bei dichotomen Merkmalen. Der Mittelwert wäre dann die relative Häufigkeit der Ausprägung, der die 1 zugewiesen wurde.

Beispiel: Wahlumfrage zwischen „CDU“ und „nicht CDU“ mit der Codierung „CDU“=0 und „nicht CDU“ =1. Der Mittelwert gleicht in diesem Fall der relativen Häufigkeit der „nicht-CDU“ Ausprägung.

0-1 codierste Variablen haben folgende Eigenschaften, die nominale Merkmale eigentlich nicht haben:

-Mittelwerte und Streuungen können interpretiert werden.
-In Regressionsgleichungen kann $\hat{\beta}$als Mittelwertsunterschied der durch X definierten Gruppen interpretiert werden und die Korrelation ist ein Maß für den standardisierten Mittelwertsunterschied.
-In Regressionsgleichungen können Interaktionseffekte der Form $\beta_{12}X_1X_2$ definiert werden.
-Es lassen sich gruppenspezifische Steigungen für X definieren.
-Produkt-Moment-Korrelationen von X mit anderen Indikatoren Y können als $\emptyset$ −Koeffizient oder biseriale Korrelation gelesen werden.

Ansonsten sind 0-1 codierte Variablen als Bernoulli-verteilt anzusehen.

Ordinale Zusammenhangsmaße

Bei ordinal skalierbaren Merkmalen kann der Abstand zwischen zwei Ausprägungen nicht bestimmt werden, es gibt aber eine klare Rangfolge. Als Zusammenhangsmaße werden der Gamma-Koeffizient (Goodman-Kruskal) und der Tau-Koeffizient (Kendall) behandelt. Ich werde den „Formelsalat“ aus dem Fernuni-Skript hier jetzt nicht wiederholen und kurz umgangssprachlich erläutern, was hier berechnet wird. Parallel solltest du dir die Formeln anschauen.

Gamma-Koeffizient

Da im Skript kein Beispiel aufgeführt ist, werde ich den Gamma Koeffizienten an einem Beispiel erklären. Es liegen folgende Daten vor:

X	1	3	4
Y	2	1	4

Nun wird für jedes Wertepaar eine <> Beziehung ermittelt. Ich nenne dazu die Ausprägungen X1,X2,X3 und Y1,Y2,Y3.

Zunächst betrachtet man das Wertepaar X1, X2 und Y1, Y2. Ist X1>X2 und auch Y1>Y2 oder aber X1<X2 und Y1<Y2, bestehen also dieselben Rangbeziehungen, so spricht man von Konkordanz.

Ist dagegen X1<X2 und Y1<Y2 oder aber X1>X2 und Y1>Y2, bestehen also entgegengesetzte Rangbeziehungen,so spricht man von Diskordanz.

Zu jedem möglichen Wertepaar X!,X2 und dem dazugehörigen Wertepaar Y1,Y2 wird nun ermittelt ob Konkordanz oder Diskordanz besteht. Gilt entweder X1=X2 oder Y1=Y2, also Gleichheit bei irgendeinem Wertepaar, so besteht weder Konkordanz noch Diskordanz. Diese Fälle sollen uns nicht weiter interessieren.

Es wird nun die Anzahl an Wertepaaren mit Konkordanzn_{konk} und die Anzahl an Wertepaaren mit Diskordanz n_{disk} ermittelt. Der Gamma Koeffizient ist dann

$$\gamma = \frac{n_{konk} - n_{disk}}{n_{konk} + n_{disk}}$$

Für das Beispiel gilt:

Zunächst wähle ich das Wertepaar X1,X2 mit dem dazugehörigen Wertepaar Y1,Y2:

Da 3>1 und 1<2 gilt, besteht eine Diskordanz. Das wiederhole ich mit den Wertepaaren (1,4;2,4) und (3,4;2,4).

Für (1,4;2,4) besteht Konkordanz und für (3,4;2,4) ebenfalls. Wir haben also $n_{konk} = 2$ und $n_{disk} = 1$.

Es folgt:

$$\gamma = \frac{2-1}{2+1} = \frac{1}{3}$$

Tau Koeffizient

Beim Tau Koeffizienten nutzt man wieder n_{konk} und n_{disk}. Es wird zwischen τ_a und τ_b unterschieden:

$$\tau_a = \frac{n_{konk} - n_{disk}}{N(N-1)/2}$$

$$\tau_b = \frac{n_{konk} - n_{disk}}{\sqrt{(n_{konk} + n_{disk} + n_{>=}) * (n_{konk} + n_{disk} + n_{=>})}}$$

,wobei $n_{>=}$ für die Anzahl der Fälle mit X1>X2 und Y1=Y2 steht. $n_{=>}$ entsprechend für X1=X2 und Y1>Y2.

Zusammenhänge in Kreuztabellen

Symmetrische Maße

Markiere Dir die Seite mit einem Postit, damit du die Formeln schnell nachschlagen kannst. Ansonsten gibt es hier nichts zu sagen.

Prädiktionsregeln

Cohens Kappa ist aus Kurseinheit II bekannt. Das Lambda von Goodman Kruskal ist aus den a-posteriori-Regeln aus Kurseinheit 2 bekannt.

Scheinkorrelation

Scheinkorrelation entsteht, wenn 2 Merkmale von einem dritten Merkmal abhängig sind, das in der Analyse nicht erfasst wird. Zum Beispiel leben Vegetarier länger. Dabei wirkt sich der Fleischkonsum garnicht auf die Gesundheit aus, sondern Vegetarier achten einfach mehr auf ihre Ernährung und essen deshalb weniger ungesunde Dinge.

Dieses Thema wird im Fernuni Skript mit gefühlten 100 Formeln und 1000 Indizierungen behandelt. Ich halte diese Formeln für nicht klausurrelevant und glaube auch nicht, dass sie dem Verständnis helfen.

Partielle Korrelation

Ich möchte versuchen kurz und verständlich zu erklären, worum es bei der partiellen Korrelation geht. Ich werde es aber komplett anders erklären, als der Lehrstuhl und bin mir nicht sicher, ob Du die Erklärung im Skript verstehst, wenn du die partielle Korrelation verstanden hast.

Mathematisch ermittelt man die partielle Korrelation wie folgt:

Man hat im Grundmodell zwei Variablen X und Y und nimmt nun die Variable Z hinzu, die mit X und Y korreliert ist. Um die tatsächliche Korrelation zwischen X und Y (die partielle Korrelation) zu ermitteln, muss die Korrelation zwischen X und Y um den Einfluss von Z bereinigt werden.

Bereinigung von X um den Einfluss von Z:

Dazu schätzt man X in Abhängigkeit von Z durch lineare Regression:

$$\hat{x} = a + bz$$

Die Änderung von X, die nicht durch Z beeinflusst wird, ist nun

$$x_i^* = x_i - \hat{x_i}$$

Dasselbe macht man für Y und erhält entsprechend:

$$y_i^* = y_i - \hat{y_i}$$

Die Korrelation zwischen x_i^* und y_i^* wird nicht durch Z beeinflusst. Dies ist die partielle Korrelation.

In Formeln:

$$r_{x^*y^*} = \frac{Cov(x^*y^*)}{s_{x^*}s_{y^*}}$$

Anmerkung: Dies entspricht der im Skript angegebenen Formel:

$r_{x^*y^*} = \frac{SQR_{XY}}{\sqrt{SQR_{XX} * SQR_{YY}}}$

Sind die Korrelationen zwischen den drei Variablen gegeben, so berechnet man die partielle Korrelation nach der Formel:

$$r_{x^*y^*} = \frac{r_{XY} - r_{XZ} * r_{YZ}}{\sqrt{(1 - r_{XZ}^2) * (1 - r_{YZ}^2)}}$$

Optimale Prognose bei Kenntnis von Z

Ist Z bekannt, so kann die abhängige Variable Y in Abhängigkeit von Z bestimmt werden. Die Prognose ist optimal, wenn sie dem Erwartungswert von Y gegeben Z entspricht. Hierzu nutzt man das lineare Regressionsmodell der Form:

$$y = a + bz + \epsilon$$

,wobei gilt:

$$a = E(y) - bE(z)$$

$$b = Cov(y, z) * Cov(z, z)^{-1}$$

Reliabilität (interne Konsistenz)

Was die Reliabilität ist, haben wir schon gesagt. Es ist der Varianzanteil der wahren Werte an der Gesamtvarianz. Das Problem ist natürlich, dass die wahren Werte nicht bekannt sind, sonst müßte man ja keine Stichproben durchführen. Ein Lösungsversuch für dieses Problem ist der Parallel-Test/Re-Test.

Parallel-Test / Re-Test

Beim Parallel-Test wird derselbe Test noch einmal durchgeführt. Es wird davon ausgegangen, dass das Merkmal X aus dem wahren Wert T und einer zufälligen Störgröße(Fehler) ϵ besteht. Es wird angenommen, dass die Fehler unkorreliert sind. Man erhält bei zweimaliger Testdurchführung mathematisch:

$$X_1 = T + \epsilon_1$$

$$X_2 = T + \epsilon_2$$

Da die Fehler unkorreliert sind gilt:

$$Cov(X_1, X_2) = Cov(T, T)$$

Für die Reliabilität gilt dann:

$$rel = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1) * Var(X_2)}}$$

Ändert sich im Zeitablauf der Wert T, so berechnet man mit der Methode die Autokovarianz von T.

Test-Halbierung

Anstatt den Test zu wiederholen, kann man ihn auch halbieren und beide Hälften betrachten. Die Reliabilität ergibt sich dann zu:

$$rel = \frac{2p_{1,2}}{1 + p_{1,2}}$$

Allgemeine Spearman-Brown-Formel

Halbiert man den Test nicht, sondern teilt ihn in k Teile, so erhält man für die Reliabilität:

$$rel = \frac{kp_{i,j}}{1 + (k-1)\bar{p}}$$

mit

$$\bar{p} = \frac{1}{k(k-1)} \sum_{i \neq j} p_{i,j}$$

mit K als Anzahl an Items.

Cronbach's Alpha

Bei Kronbach's Alpha werden die einzelnen Korrelationen der Allgemeinen Spearman-Brown-Formel durch $\bar{p}$ ersetzt:

$$rel = \frac{k\bar{p}}{1 + (k-1)\bar{p}}$$

Mit K als Anzahl an Items.

Abschwächungs-Korrektur

Die Abschwächungs-Korrektur bezeichnet die Tatsache, dass aufgrund von Meßfehlern die gemessene Korrelation schwächer ist als die tatsächliche.

Es wird folgendes Modell angenommen:

$$X_1 = T_1 + \epsilon_1$$

$$X_2 = T_2 + \epsilon_2$$

Die wahre Korrelation wird aus den gemessenen Daten wie folgt ermittelt:

$$Corr(T_1, T_2) = Corr(X_1, X_2) / \sqrt{rel_1 * rel_2}$$

, wobei Corr hier für Korrelation steht.

Theorem von Bayes / Totale Wahrscheinlichkeit

Wir betrachten zunächst die Wahrscheinlichkeit eines Ereignisses, das von mehreren sich gegenseitig ausschließenden Ereignissen eingeschlossen wird. Um das auch zu verstehen, hier ein Beispiel:

Es werden Daten über die Essgewohnheiten und die Erkrankung an einer Krebsart erhoben

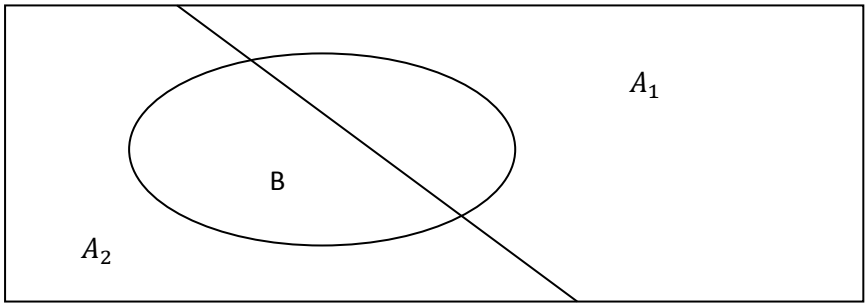
Ereignis *B*: Krebserkrankung positiv

Ereignis A_1:Vegetarier

Ereignis A_2:Fleisch-essender Mensch

Das Ereignis B wird von den Ereignissen $A_1 \text{ und } A_2$ eingeschlossen.

Graphisch sieht das dann so aus:



Die „totale Wahrscheinlichkeit“ des Ereignisses B läßt sich nun wie folgt darstellen:

$$P(B) = P(A_1) * P(B|A_1) + P(A_2) * P(B|A_2)$$

Interessiert man sich nur für die Frage:

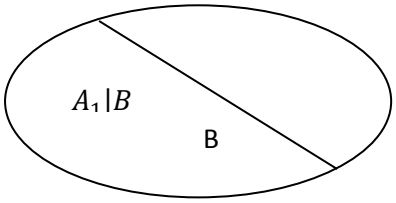
Wie groß ist die Wahrscheinlichkeit, daß ein zufällig ausgewählter Krebskranker Vegetarier ist, so kann dies nach dem Satz von Bayes gelöst werden:

Demnach gilt formal:

$$P(A_1|B) = \frac{P(A_1) * P(B|A_1)}{\sum_{i=1}^{n} P(B|A_i) * P(A_i)} = \frac{P(A_1) * P(B|A_1)}{P(B)}$$

Anschaulich bedeutet das, dass $P(A_1|B)$ dem Verhältnis von A_1**innerhalb von B** zu B entspricht.

Es ist wichtig, dass du diesen Satz verstanden hast. Deshalb nochmal eine Erklärung anhand der grafischen Darstellung: Durch die Bedingung, dass B eingetreten ist verringert sich die „Anzahl aller möglichen Fälle“ auf B Und die „für A_1günstigen Fälle“ auf $A_1|B$:



Die Bedingung, dass B eingetreten ist beschränkt die Grundgesamtheit auf B. Deshalb steht P(B) ja auch im Nenner. Die Wahrscheinlichkeit aus B nun A_1 zu wählen entspricht dem Anteil von B, der zu A_1 gehört.

Bespiel:

10% aller Menschen sind Vegetarier. 1% aller Vegetarier erkranken irgendwann an Krebs. 10% aller Fleisch-essenden Menschen erkranken irgendwann an Krebs. Wähle man zufällig einen krebskranken Menschen aus, was ist die Wahrscheinlichkeit, dass dies ein Vegetarier ist?

Antwort: Zunächst berechnet man wie viele krebskranke Vegetarier und wie viele krebskranke Fleich-essende Menschen es gibt:

Krebskranke Vegetarier: $10\% * 1\% = 0{,}1\%$

Krebskranke Fleisch-essende Menschen: $90\% * 10\% = 9\%$

Krebskranke insgesamt: 9,1%

Es gibt also 90 mal mehr krebskranke fleisch-essende Menschen als krebskranke Vegetarier. Aus einer Stichprobe, in der das Verhältnis krebskranker Vegetarier zu krebskranken Fleisch-essenden Menschen 1 zu 90 ist, ist die Wahrscheinlichkeit, gerade den krebskranken Vegetarier zu ziehen, 1 zu 91.

Item-Analyse

Als Item wird eines von mehreren Merkmalen einer Untersuchungseinheit bezeichnet. Im Fernuni Skript werden mit Items zu erfüllende Aufgaben bezeichnet. Ein Beispiel wäre zum Beispiel ein Intelligenztest, bei dem mehrere Aufgaben gelöst werden müssen. Bei der Item-Analyse werden folgende Eigenschaften analysiert:

1) Item-Schwierigkeit: Auf einer Skala von 0 bis 1 wird die Schwierigkeit , die Aufgabe zu lösen, angegeben. Der Wert entspricht dem Anteil der Personen, die die Aufgabe richtig gelöst haben. Oft sind die Daten auf einer Skala von 1 bis x und der Mittelwert der Ergebnisse gegeben. Die Item-Schwierigkeit ist dann der Anteil der Skala, der unterhalb des Mittelwertes liegt.
   Beispiel:
   Eine Skala von 1-10 hat einen item-Mittelwert von 6. Die Skala geht von 1-10, also 9 Punkte Spannweite. Unterhalb der 6 liegt ein Bereich von 6-1=5. Die Schwierigkeit ist 5/9.
2) Trennschärfe: Der Trennschärfe eines Items ist zu entnehmen, wie gut das Testergebnis eines einzelnen Items das Gesamttestergebnis prognostizieren kann. Items mit sehr hohem oder sehr niedrigem Schwierigkeitsgrad haben eine schlechte Trennschärfe und werden aus einem eindimensionalen Test entfernt. Die Korrelation eines Items mit sich selbst wird bei der Berechnung weggelassen.
3) Homogenität: Die Homogenität bezeichnet die durchschnittliche Korrelation der Items untereinander. Items mit geringer Homogenität sollten aus dem Test entfernt werden.
4) Dimensionalität: Mit der Faktoren- oder Clusteranalyse kann die Dimensionalität einer „Item-Batterie“ untersucht werden. Zum Beispiel kann die Korrelationsmatrix der Items umgeordent werden, bis mehrere Blöcke mit homogener Korrelation entstehen.

Faktorenanalyse

In diesem Kapitel werde ich mich ebensowenig am Skript orientieren, wie bei der partiellen Korrelation. Ich werde mich darauf beschränken, die Faktoren- und Hauptkomponentenanalyse zu erklären. Ob Du das Kapitel im Skript noch lesen möchtest oder nicht, musst Du dann selber entscheiden. Generell kann man sagen: Es gibt sehr dicke Bücher über die Faktorenanalyse und da der Lehrstuhl dieses sehr schwierige und komplexe Thema auf wenigen Seiten abhandelt sind komplexe Fragen zu diesem Thema sehr unwahrscheinlich. Ich gehe davon aus, dass entweder das grundsätzliche Verständnis abgeprüft wird, oder aber die Lösung direkt aus dem Skript abgelesen werden kann (ohne das man verstehen muss, was da steht oder warum).

Bei der partiellen Korrelation haben wir schon besprochen, dass die Korrelation zwischen zwei Variablen eine Scheinkorrelation sein kann und auf eine dritte Variable zurückzuführen sein kann. Bei der Faktorenanalyse versucht man die wichtigen Faktoren, also die, die einen Einfluss auf die abhängige Variable haben, zu bestimmen.

Das Grundmodell ist der Regressionsanalyse dabei sehr ähnlich. Der wichtige Unterschied ist der, dass die Parameter nicht beobachtbar (latent) sind, sondern geschätzt werden müssen. Ich betrachte beispielsweise den Absatz von Limousinen von Mercedes, BMW und Audi. Ich vermute, dass es latente Faktoren gibt, die alle drei Werte beeinflussen (welche das sind ist dabei unbekannt). Aus der Korrelationsmatrix heraus können diese aber geschätzt werden.

Das Grundmodell der Faktorenanalyse lautet:

$$x = \Lambda\xi + \epsilon$$

Wobei gilt:

Λ ist die Faktor-Ladungs-Matrix.

ξ ist der Vektor von Faktoren.

ϵ ist der Fehlerterm.

Im Skript ist auch eine Formel für die Schätzung der Faktoren angegeben. Ich erspare mir jetzt aber diese abzuschreiben.

Die durch die Faktoren erklärte Streuung nennt man Kommunalitäten. Diese berechnet man nach der Formel:

$$h_i^2 = (\Lambda\Lambda')_{ii}$$

Die Faktorenanalyse verfolgt drei Ziele:

1. Reduktion der Variablenanzahl: Die Faktorenanalyse erkennt Variablengruppen, in denen jeweils alle Variablen ähnliche Informationen erfassen. Diese Variablen können zusammengefasst werden.

2. Ermittlung verlässlicher Messgrößen: Werden die Variablen zu einem Faktor zusammengefasst, so besitzt dieser Faktor günstigere messtechnische Eigenschaften als die einzelnen Variablen.
3. Analytische Zielsetzung: Die Faktorenanalyse ermöglicht es, von den manifesten Variablen (den Indikatorvariablen) auf übergeordnete latente Variablen zu schließen.

Hauptkomponentenanalyse

Bei der Hauptkomponentenanalyse werden aus einer gegebenen Korrelationsmatrix die Faktoren herausgefiltert, die überflüssig sind, weil sie stark korreliert sind.

Beispiel: Man untersucht den Einfluß der Preisentwicklung der Aktien Shell, Gasprom und Total fina elf auf den Dax. Angenommen, man stellt fest, dass jede dieser Aktien eine negative Korrelation zum Dax hat und die Aktien untereinander stark positiv korreliert sind. Man könnte also statt drei stark korrelierter Aktien auch nur eine Aktie verwenden.

Formal läßt sich die Hauptkomponentenanalyse wie folgt darstellen:

Für eine Datenmenge vom Umfang n und p Merkmalen, die durch die p x n Matrix $X' = (x_1, x_2, \dots x_p)$ dargestellt wird, entsprechen die m Hauptkomponenten den m unkorrelierten Linearkombinationen der Vektoren $x_1, x_2, \dots x_m$ mit maximaler Varianz. Gesucht ist also ein p-dimensionaler Vektor a_1 für den gilt: $a_1'x$ hat maximale Varianz. Als nächstes wird eine Linearkombination $a_2'x$ gesucht, die zu $a_1'x$ unkorrelliert ist und maximale Varianz besitzt.

Das Maximum von

$$\mathrm{Var}(a_1'x) = a_1'\Sigma a_1$$

kann nur ermittelt werden, wenn $a_1'a_1$ durch eine Bedingung begrenzt wird. In dieser Herleitung wird als Bedingung $a_1'a_1 = 1$ gewählt. Zur Maximierung von $a_1'\Sigma a_1$ unter der Bedingung $a_1'a_1 = 1$ wird der Lagrange Multiplikator λ genutzt: Maximiere

$$a_1'\Sigma a_1 - \lambda(a_1'a_1 - 1)$$

Ableitung nach a_1 ergibt

$$\Sigma a_1 - \lambda a_1 = 0$$

Schreibt man diesen Ausdruck als

$$(\Sigma - \lambda I)a_1 = 0$$

,so erkennt man, dass es sich hier um ein Eigenwert-Problem handelt. λ ist ein Eigenwert und a_1 der dazugehörige Eigenvektor. Es gilt den Eigenwert zu finden, für den $a_1'\Sigma a_1$maximal wird.

Da gilt:

$$a_1'\Sigma a_1 = a_1'\lambda a_1 = \lambda a_1'a_1 = \lambda$$

muss der größte Eigenwert gesucht werden. Ist der Eigenvektor zum größten Eigenwert gefunden, so muss der zum ersten Eigenvektor unkorrelierte Eigenvektor zum zweitgrößten Eigenwert gefunden werden.

Es muss also gelten:

$$a_1' a_2 = 0$$

(Die beiden Vektoren sind linear unabhängig, wenn sie senkrecht aufeinander stehen). Es gilt also, den Vektor a_2 zu finden, für den folgende Lagrange-Multiplikation maximal wird:

$$a_2' \Sigma a_2 - \lambda(a_2' a_2 - 1) - \phi a_1 a_2'$$

(λ und ϕ sind die Lagrange Multiplikatoren).

Ableitung nach a_2' und Nullsetzen ergibt:

$$\Sigma a_2 - \lambda a_2 - \phi a_1 = 0$$

Multiplikation dieser Gleichung mit a_1' von links ergibt:

$$a_1' \Sigma a_2 - \lambda a_1' a_2 - \phi a_1' a_1 = 0$$

Da die ersten beiden Terme für unkorrellierte Vektoren Null ergeben, folgt $\phi = 0$. Equivalent zum ersten Eigenwert folgt erneut die Gleichung

$$(\Sigma - \lambda I) a_2 = 0$$

und damit die Suche nach dem zweitgrößten Eigenwert. Nach demselben Schema erfolgt die Ermittlung der dritten, vierten usw. Linearkombinationen $a_i x$ und damit die Ermittlung der weiteren Eigenwerte (Eigenwertzerlegung).

Anleitung zur Hauptkomponentenanalyse:

Ausgegangen wird von einer Korrelationsmatrix.

1) Ermittle die Eigenwerte der Korrelationsmatrix. Das Verhältnis der Eigenwerte entspricht dem Verhältnis der erklärten Varianz durch die zugehörigen Eigenvektoren. Hat eine Korrelationsmatrix beispielsweise drei Eigenwerte 0,7; 0,2 und 0,1 so erklärt der erste Eigenvektor 70% der Varianz.
2) Ermittle zunächst den Eigenvektor zum größten Eigenwert
3) Ermittle nun den Eigenvektor zum zweitgrößten Eigenwert, der orthogornal zum ersten Eigenvektor ist.
4) Führe das fort, bis die Eigenvektoren einen vorgegebenen Anteil der Varianz erklären.

Maximum-Likelihood Analyse

Die ML-Faktorenanalyse geht vom selben Grundmodell aus, wie die Hauptkomponentenanalyse. Bei der ML-Faktorenanalyse werden aber Annahmen über die Verteilung der Parameter getroffen. Auf dieser Basis wird die Likelihoodfunktion aufgestellt, die dann maximiert wird.

Identifikation

Es ist möglicherweise nicht möglich bei der ML-Faktorenanalyse alle Parameter eindeutig zu identifizieren. Es darf daher nicht mehr manivestierte Variablen geben als latente Faktoren. Ist das Modell nicht identifizierbar, so müssen Restriktionen eingeführt werden oder weitere Indikatoren gemessen werden.

Lösungen zu den Aufgaben Kurseinheit III

Aufgabe 1 und 2: Die Antworten auf diese Fragen kannst du dem SPSS Programm entnehmen.

Lösung zu Aufgabe 3

a) Laut wikipedia sind dies: Histogramm, Boxplot, QQ-Plot, Scatterplot, Mosaikplot.

b) Ein Histogramm ist ein Flächendiagramm bei dem die Fläche einer Ausprägung die relative Häufigkeit darstellt. Beim Stabdiagramm ist dies nicht der Fall.

c) Durch die Codierung können beispielsweise Mittelwerte sinnvoll berechnet werden.

d) Mittelwert ist relative Häufigkeit für 0/1 codierte Ausprägung.

Lösung zu Aufgabe 4

a) Chi-Quadrat-Koeffizient, Kontingenz-Koeffizient, normierter Kontingenz-Koeffizient, Phi-Koeffizient

b) Man geht von einer Häufigkeitstabelle zweier Merkmale X und Y aus und berechnet die Modalwerte der bedingten Häufigkeiten von Y gegeben X. Die Summe dieser Modalwerte nennt man F(+). Außerdem berechnet man den Modalwert von Y ohne X. Diesen Modalwert nennt man F(-).

Das Prädiktionsmaß Lambda X->Y berechnet man nun wie folgt:

$$\lambda_{X\to Y} = \frac{F(-) - F(+)}{F(-)}$$

Es ist ein a-posteriori Maß, da es eine Prädiktionsregel ist, die aus den Daten abgeleitet wird.

c) A-priori Maße werden aus inhaltlichen Theorien abgeleitet und können anhand von Daten überprüft werden.

d) Man benötigt die Randverteilung um daraus die hypothetische Verteilung unter Unabhängigkeit zu berechnen:

	A	B	C	Summe
A	0,3	0,1	0	0,4
B	0,2	0,2	0	0,4
C	0	0	0,1	0,1
Summe	0,5	0,3	0,1	0,9

	A	B	C	Summe
A	0,22	0,13	0,04	0,4
B	0,22	0,13	0,04	0,4
C	0,06	0,03	0,01	0,1
Summe	0,5	0,3	0,1	0,9

Es gilt:

$$G(+) = \sum_{i=1}^{n} f_{ii} = \frac{0{,}3 + 0{,}2 + 0{,}1}{0{,}9} = 0{,}67$$

$$G(-) = \sum_{i=1}^{n} f_i * f_i = \frac{0{,}22 + 0{,}13 + 0{,}01}{0{,}9} = 0{,}4$$

$$\text{K} = \frac{G(+) - G(-)}{1 - G(-)} = \frac{0{,}67 - 0{,}4}{1 - 0{,}4} = 0{,}45$$

e) Es werden wieder die Randverteilungen benötigt. Ich benenne die Spalten mit X und die Zeilen mit Y.

	X1=A	X2=B	X3=C	Summe
Y1=A	0,3	0,1	0	0,4
Y2=B	0,2	0,2	0	0,4
Y3=C	0	0	0,1	0,1
Summe	0,5	0,3	0,1	0,9

Man könnte nun noch die relativen Häufigkeiten berechnen, dies würde aber am Ergebnis nichts ändern. Ich rechne daher mit den gegebenen absoluten Häufigkeiten.

Nun müssen die Modalwerte der bedingten Verteilungen bestimmt werden.

Für Y gegeben x=A ist der Modalwert 0,3. Für Y gegeben x=B ist er 0,2 und für Y gegeben x=C ist er 0,4. Es bleiben F(+)=0,2+0,1+0=0,3.

Der Modalwert der Randverteilung von Y ist 0,4. Es folgt also F(-)=0,1.

Einsetzen:

$$\lambda_{X \to Y} = \frac{0{,}1 - 0{,}3}{0{,}1} = -2$$

Für $\lambda_{Y \to X}$ erhält man entsprechend:

$$\lambda_{Y \to X} = \frac{0{,}4 - 0{,}1}{0{,}4} = \frac{3}{4}$$

Für $\lambda_{Y \leftrightarrow X}$ erhält man:

$$\lambda_{Y\leftrightarrow X} = \frac{(0{,}1 + 0{,}4) - (0{,}3 + 0{,}1)}{0{,}1 + 0{,}4} = 0{,}2$$

Lösung zu Aufgabe 5

a) Der Parameter α bezeichnet den Wert, den Y unabhängig von X annehmen kann. Der Parameter β bezeichnet die Abweichung nach oben in Abhängigkeit von X. Der letzte Term ist der Fehlerterm. Die Funktion nimmt die beiden Werte α +Fehlerterm und $\alpha + \beta$+ Fehlerterm an – je nachdem ob X den Wert 0 oder 1 annimmt.

b) Nimmt X1n den Wert 1 an, so ensteht einerseits ein konstanter Einfluss in Höhe von $\beta 1$ und ein von $\beta 12 * X2n$ abhängiger Einfluss.

c) Mir ist nicht klar, auf welches Problem sich die Aufgabenstellung bezieht. Wenn mit „Problem" der Test auf Gleichheit der Mittelwerte gemeint ist, so ist ein F-Test zu verwenden (siehe S.327 Fernuni-Skript).

Lösung zu Aufgabe 6

a) Nominales Skalenniveau: Automarken mit den Codierungen: Audi=1 ,VW=2, BMW=3

Ordinales Skalenniveau: Einkommen eingestuft in hoch, mittel, niedrig

Intervallskalenniveau: IQ Skala.

Verhältnisskala: Körpergröße

b) Lineare Transformationen.

c)Nur Skalentransformationen, die das Informationsniveau unverändert lassen oder verringern (Man kann nicht von einer Skala mit niedrigem Informationsniveau auf eine höhere transformieren).,

Lösung zu Aufgabe 7

a) Siehe Beschreibung zum Beginn der KE 3 in diesem Skript.

b) Auftrittshäufigkeiten und Modalwert

c) Eine wichtige Voraussetzung ist die Transitivität der Schulnoten. Eine 1 ist also besser als eine 2 usw. Außerdem gibt es zu beachten, dass Noten schon Kategorien sind (es wird z.B. nicht zwischen einer guten und einer schlechten 2 unterschieden) und das Merkmal metrisch sein muss.

Lösung zu Aufgabe 8

a) Kurz zum Kontrapositionsgesetz: Wenn A eintritt, dann tritt auch B ein besagt die erste Klammer. Daraus soll folgen, dass wenn B nicht eintritt, auch A nicht eintreten kann. Mir bleibt nicht mehr zu sagen, als dass das selbstverständlich ist, da ansonsten ja der Fall „A tritt ein, aber B nicht" eingetreten wäre, der dem ersten Klammerausdruck widerspricht. Davon unabhängig sind korrelierte Variablen natürlich abhängig, da die Korrelation die lineare Abhängigkeit bezeichnet. Unabhängige Variablen sind stets unkorreliert.

b) Nicht unbedingt. Es gibt auch nichtlineare Abhängigkeiten, die von der Korrelation nicht erfaßt werden. Dies ist übrigens ein Lieblingsthema des Lehrstihls, das in sämtlichen quantitativen Modulen in Klausuren abgeprüft wird.

c) Die Frage ist so allgemein gestellt, dass ich nicht weiß, worauf sie abzielt. Ich habe sie in mehreren Mathe-Foren gestellt und den Lehrstuhl um eine Antwort gebeten. Sobald ich eine sinnvolle Antwort erhalte, werde ich diese in die Korrekturhinweise des Skriptes stellen. Ich kann nur vermuten, dass hier erwartet wird, dass gilt: Sind X und Y unkorreliert, dann sind sie auch stochastisch unabhängig und andersrum.

Lösung zu Aufgabe 9

a)

X\Y	Diskret	metrisch
Diskret	Kreuztabellen	Varianz-Analyse
Metrisch	Kategoriale Regression	Regression

b) Das Bestimmtheitsmaß ist das Quadrat des Pearsonschen Korrelationskoeffitionen:

$$r = \frac{Cov(X,Y)}{\tilde{s}_x * \tilde{s}_y}$$

Interpretation des Bestimmtheitsmaßes

Das Bestimmtheitsmaß gibt an, welcher Anteil an Streuung (Varianz) der beobachteten Y-Werte durch die Varianz der geschätzten Y-Werte erklärt wird. Der Ausdruck „erklärt" ist etwas hier etwas missverständlich. Es soll bedeuten, dass dieser Anteil an Varianz durch den Einfluss des Merkmals X herbeigeführt wird. Angenommen das Merkmal Y hängt zu 50% von dem Merkmal X ab und zu 50% von einem dritten unbekannten Merkmal. In diesem Fall würde das Bestimmtheitsmaß bezüglich X und Y 50% (0,5) betragen.

Hängt Y ausschließlich von X ab, so beträgt das Bestimmtheitsmaß 1. Da die gesamte Änderung des Merkmals Y von der Änderung des Merkmals X abhängt, liegen alle tatsächlich beobachteten Werte auch auf der Regressionsgerade.

c) Der globale F-Test testet die Gleichheit der Gruppenmittel.

Lösung zu Aufgabe 10

a) Scheinkorrelation entsteht, wenn 2 Merkmale von einem dritten Merkmal abhängig sind, das in der Analyse nicht erfasst wird. Zum Beispiel leben Vegetarier länger. Dabei wirkt sich der Fleischkonsum garnicht auf die Gesundheit aus, sondern Vegetarier achten einfach mehr auf ihre Ernährung und essen deshalb weniger ungesunde Dinge.

b) Der Einfluss von dritten Variablen muss eliminiert werden.

c) Die dritte Variable, die evtl. einen Einfluss auf die beiden Variablen X und Y hat, nennt man Kontrollvariable.

d) Es gilt:

$$r_{x^*y^*} = \frac{r_{XY} - r_{XZ} * r_{YZ}}{\sqrt{(1 - r_{XZ}^2) * (1 - r_{YZ}^2)}}$$

Einsetzen:

$$r_{x^*y^*} = \frac{0{,}7 - 0{,}6 * 0{,}3}{\sqrt{(1 - 0{,}36) * (1 - 0{,}09)}} = 0{,}68$$

e) Ist Z bekannt, so kann die abhängige Variable Y in Abhängigkeit von Z bestimmt werden. Hierzu nutzt man das lineare Regressionsmodell der Form:

$$y = a + bz + \epsilon$$

,wobei gilt:

$$a = E(y) - bE(z)$$

$$b = Cov(y, z) * Cov(z, z)^{-1}$$

f) i) Die Prognose ist optimal wenn X und Y gemeinsam normalverteilt sind.

ii) Einsetzen in die Formel aus e):

$$y = a + bx + \epsilon$$

,wobei gilt:

$$a = E(y) - bE(x)$$

$$b = Cov(y, x) * Cov(x, x)^{-1}$$

Einsetzen:

$$b = Cov(y, x) * Cov(x, x)^{-1} = 0{,}7$$

$$a = E(y) - bE(x) = 1 - 0{,}7 = 0{,}3$$

$$y = 0{,}3 + 0{,}7x$$

Lösung zu Aufgabe 11

a) Bei Kronbach's Alpha werden die einzelnen Korrelationen der Allgemeinen Spearman-Brown-Formel durch $\bar{p}$ ersetzt

b) Es gilt:

$$rel = \frac{k\bar{p}}{1 + (k-1)\bar{p}}$$

Einsetzen:

$$rel = \frac{5 * 0{,}5}{1 + (5-1)5} = 0{,}12$$

c) Ändert sich im Zeitablauf der Wert T, so berechnet man mit der Methode die Autokovarianz von T.

d) Die Abschwächungs-Korrektur bezeichnet die Tatsache, dass aufgrund von Meßfehlern die gemessene Korrelation schwächer ist als die tatsächliche.

Es gilt:

$$Corr(T_1, T_2) = Corr(X_1, X_2)/\sqrt{rel_1 * rel_2}$$

Einsetzen:

$$Corr(T_1, T_2) = \frac{0{,}3}{0{,}7} = 0{,}43$$

Lösung zu Aufgabe 12

a) Die Skala geht von 1-5. Die Schwierigkeit berechnet sich also zu $\frac{4{,}7-1}{5-1} = 0{,}925\%$.

b) Die Trennschärfe beträgt 0,71.

c) Items mit sehr hohem oder sehr niedrigem Schwierigkeitsgrad haben eine schlechte Trennschärfe und werden aus einem eindimensionalen Test entfernt.

d) Mit der Faktoren- oder Clusteranalyse kann die Dimensionalität (ein oder mehrdimensional) einer „Item-Batterie" untersucht werden. Zum Beispiel kann die Korrelationsmatrix der Items umgeordent werden, bis mehrere Blöcke mit homogener Korrelation entstehen.

Lösung zu Aufgabe 13

a) Die klassische Testtheorie ist eine einfache Version der Faktorenanalyse.

b) Das Grundmodell der Faktorenanalyse lautet:

$$x = \Lambda\xi + \epsilon$$

Wobei gilt:

Λ ist die Faktor-Ladungs-Matrix.

ξ ist der Vektor von Faktoren.

ϵ ist der Fehlerterm.

c) Die Faktoren werden durch eine optimale Prognose geschätzt (siehe Formel S.375).

d) Es gilt:

$$h_i^2 = (\Lambda\Lambda')_{ii}$$

Einsetzen:

$$h_i^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 5 & 1 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 5 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 5 \\ 0 & 1 & 1 \\ 5 & 1 & 26 \end{pmatrix}$$

e) Für eine Datenmenge vom Umfang n und p Merkmalen, die durch die p x n Matrix $\mathrm{X}' = (\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_\mathrm{p})$ dargestellt wird, entsprechen die m Hauptkomponenten den m unkorrelierten Linearkombinationen der Vektoren $\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_\mathrm{m}$ mit maximaler Varianz. Gesucht ist also ein p-dimensionaler Vektor a_1 für den gilt: $\mathrm{a}_1'\mathrm{x}$ hat maximale Varianz. Als nächstes wird eine Linearkombination $\mathrm{a}_2'\mathrm{x}$ gesucht, die zu $\mathrm{a}_1'\mathrm{x}$ unkorrelliert ist und maximale Varianz besitzt.

Das Maximum von

$$\mathrm{Var}(\mathrm{a}_1'\mathrm{x}) = \mathrm{a}_1'\Sigma\mathrm{a}_1$$

kann nur ermittelt werden, wenn $\mathrm{a}_1'\mathrm{a}_1$ durch eine Bedingung begrenzt wird. In dieser Herleitung wird als Bedingung $\mathrm{a}_1'\mathrm{a}_1 = 1$ gewählt. Zur Maximierung von $\mathrm{a}_1'\Sigma\mathrm{a}_1$ unter der Bedingung $\mathrm{a}_1'\mathrm{a}_1 = 1$ wird der Lagrange Multiplikator λ genutzt: Maximiere

$$\mathrm{a}_1'\Sigma\mathrm{a}_1 - \lambda(\mathrm{a}_1'\mathrm{a}_1 - 1)$$

Ableitung nach a_1 ergibt

$$\Sigma\mathrm{a}_1 - \lambda\mathrm{a}_1 = 0$$

Schreibt man diesen Ausdruck als

$$(\Sigma - \lambda\mathrm{I})\mathrm{a}_1 = 0$$

,so erkennt man, dass es sich hier um ein Eigenwert-Problem handelt. λ ist ein Eigenwert und a_1 der dazugehörige Eigenvektor. Es gilt den Eigenwert zu finden, für den $a_1' \Sigma a_1$ maximal wird.

Da gilt:

$$a_1' \Sigma a_1 = a_1' \lambda a_1 = \lambda a_1' a_1 = \lambda$$

muss der größte Eigenwert gesucht werden. Ist der Eigenvektor zum größten Eigenwert gefunden, so muss der zum ersten Eigenvektor unkorrelierte Eigenvektor zum zweitgrößten Eigenwert gefunden werden.

Es muss also gelten:

$$a_1' a_2 = 0$$

(Die beiden Vektoren sind linear unabhängig, wenn sie senkrecht aufeinander stehen). Es gilt also, den Vektor a_2 zu finden, für den folgende Lagrange-Multiplikation maximal wird:

$$a_2' \Sigma a_2 - \lambda(a_2' a_2 - 1) - \phi a_1 a_2'$$

(λ und ϕ sind die Lagrange Multiplikatoren).

Ableitung nach a_2' und Nullsetzen ergibt:

$$\Sigma a_2 - \lambda a_2 - \phi a_1 = 0$$

Multiplikation dieser Gleichung mit a_1' von links ergibt:

$$a_1' \Sigma a_2 - \lambda a_1' a_2 - \phi a_1' a_1 = 0$$

Da die ersten beiden Terme für unkorrellierte Vektoren Null ergeben, folgt $\phi = 0$. Equivalent zum ersten Eigenwert folgt erneut die Gleichung

$$(\Sigma - \lambda I) a_2 = 0$$

und damit die Suche nach dem zweitgrößten Eigenwert. Nach demselben Schema erfolgt die Ermittlung der dritten, vierten usw. Linearkombinationen $a_i x$ und damit die Ermittlung der weiteren Eigenwerte (Eigenwertzerlegung).

f) Die Komponenten eines Zufallsvektors sind unabhängig. Hauptkomponenten sind also Null. Sollten doch Korrelationen auftreten, so sind diese zufällig und sollten ignoriert werden.

g)Beide gehen vom selben Grundmodell aus. Bei der ML-Faktorenanalyse werden Annahmen über die Verteilung der Parameter getroffen. Auf dieser Basis wird die Likelihoodfunktion aufgestellt, die dann maximiert wird.

h)Es ist möglicherweise nicht möglich bei der ML-Faktorenanalyse alle Parameter eindeutig zu identifizieren. Es darf daher nicht mehr manivestierte Variablen geben als latente Faktoren. Ist das

Modell nicht identifizierbar, so müssen Restriktionen eingeführt werden oder weitere Indikatoren gemessen werden.

Lösung zu Aufgabe 14

a) Mit einem Mosaikdiagramm können Kontingenztafeln (d.h. zwei- oder mehrdimensionale Häufigkeitstabellen) dargestellt werden. Für das erste Attribut wird die horizontale Richtung wie in einem Streifendiagramm unterteilt. Jeder Abschnitt wird dann gemäß der Anteile des zweiten Attributes vertikal aufgeteilt, und zwar wieder wie in einem Streifendiagramm.

b) Ich zitiere hier mal aus Statwiki: Ein Stem-and-Leaf-Plot ist eine halbgraphische Darstellung der Beobachtungswerte einer metrisch skalierten Variablen. Der Plot dient dem Auffinden von extremen Beobachtungswerten bzw. Ausreißern in der Datenreihe und der Einschätzung der Verteilungsform der Variablen.

Für die Erstellung eines Stem-and-Leaf-Plots gibt es verschiedene Varianten, z.B.:

Stem-and-Leaf-Plot für monatliches persönliches Nettoeinkommen (Euro)

```
Frequency    Stem &  Leaf
  17,00       0 .  011111
  72,00       0 .  222222233333333333333333
  55,00       0 .  4444444444455555555
  82,00       0 .  66666666666666777777777777
  87,00       0 .  88888888999999999999999999999
  96,00       1 .  000000001111111111111111111111111
 106,00       1 .  22222222222222222222222223333333333
  64,00       1 .  444444445555555555555
  37,00       1 .  6666677777777
  33,00       1 .  88888888999
  21,00       2 .  0001111
  13,00       2 .  22233
  15,00       2 .  44445
   1,00       2 .  &
  17,00 Extremes    (>=2700)
Stem width:   1000,00
Each leaf:      3 case(s)
& denotes fractional leaves.
```

Unterhalb des Diagramms wird die Stamm-Einheit (stem width) angegeben. Ist diese wie im Beispiel 1000, so nimmt der Stamm die Tausender-Ziffern auf. Die Blätter (leafs) entsprechen dann den Hunderter-Ziffern. Alle weiteren Ziffern der Beobachtungswerte werden vernachlässigt. Die letzte Angabe unterhalb des Diagramms, bezeichnet mit each leaf, gibt an, wieviele Fälle durch eine Blatt-Ziffer repräsentiert werden. In einer Vorspalte zum Stem-and-Leaf-Plot steht unter der Überschrift Frequency die absolute Häufigkeit der Fälle der jeweiligen Zeile.

Mit der Erstellung des Stem-and-Leaf-Plots geht eine Klassenbildung der Beobachtungswerte einher. So bedeutet z.B. die 2. Zeile mit dem Stamm 2, dass 9 Fälle (Personen) ein monatliches Nettoeinkommen von 2200 bis unter 2300 Euro und 6 Fälle ein monatliches Nettoeinkommen von 2300 bis unter 2400 Euro haben.

Desweiteren ist ersichtlich, dass die Variable monatliches persönliches Nettoeinkommen eine rechtsschiefe Verteilung aufweist.

c) Mit einem gruppierten Balkendiagramm oder einem Mosaik-Plot.

Lösung zu Aufgabe 15

a) Im Skript sind dazu folgende Gründe aufgeführt:

-Mittelwerte und Streuungen können interpretiert werden.

-In Regressionsgleichungen kann $\hat{\beta}$ als Mittelwertsunterschied der durch X definierten Gruppen interpretiert werden und die Korrelation ist ein Maß für den standardisierten Mittelwertsunterschied.

-In Regressionsgleichungen können Interaktionseffekte der Form $\beta_{12}X_1X_2$ definiert werden.

-Es lassen sich gruppenspezifische Steigungen für X definieren.

-Produkt-Moment-Korrelationen von X mit anderen Indikatoren Y können als $\emptyset$ −Koeffizient oder biseriale Korrelation gelesen werden.

b) Die Korrelation zwischen X und Y bezeichnet die standardisierte Mittelwertdifferenz der Y-Variable multipliziert mit der Standardabweichung der 0-1-Variable X.

c) Ja, so geschehen z.B. in Aufgabe 5.

d) Bei mehr als 2 Ausprägungen gelten die Aussagen aus a) nicht mehr.