

Wiederholung, Ergänzung, Erklärung & Intuition:
Statistik I & II für Studierende der Wirtschaftswissenschaften
(Ludwig-Maximilians-Universität München)

Autoren:

Matthias Aßenmacher^{*},

Ann-Kathrin Köpple[†],

Christoph Luther[‡],

Patricia Haro[§],

Maximilian Mandl[¶]

Stand: August 26, 2020

Dieses Dokument wurde aus verschiedenen Quellen erstellt. Es soll als kleine Verständnishilfe für Studierende angesehen werden, wobei auf mathematische Genauigkeit und Vollständigkeit explizit verzichtet wird. Außerdem wird jedes Thema durch einen Block an Multiple-Choice Aufgaben und Hinweisen auf die passenden R-Funktionen für die behandelten Methoden ergänzt. Für Fehler wird keine Haftung übernommen.

Der erste Teil dieses Dokuments (\approx Statistik I) wurde von Ann-Kathrin in Zusammenarbeit mit Matthias im Sommer 2020 verfasst. Ann-Kathrin verantwortete das Schreiben des Erstentwurfs, Matthias war verantwortlich für intensives Korrekturlesen, Anpassungen und Erweiterungen.

Der zweite Teil (\approx Statistik II) basiert auf Vorlesungszusammenfassungen von Max aus dem Sommer 2019, welche in Zusammenarbeit mit Christoph, Patricia & Matthias in dieses Format gegossen und detaillierter ausgearbeitet wurden. Besonderer Dank für diesen zweiten Teil gilt Herrn Dr. Alexander Engelhardt, der freundlicherweise einen Teil seines Materials zur Verfügung gestellt hat (siehe auch: <https://www.crashkurs-statistik.de>), sodass wir uns hiervon inspirieren lassen konnten. Aufgrund des Fehlens der Themenbereiche Kombinatorik, Wahrscheinlichkeitsrechnung & Multiple Regression wurden diese im Sommer 2020 von Ann-Kathrin & Matthias ergänzt.

^{*}Institut für Statistik, LMU München; Kontakt bei Fragen & Anregungen: matthias@stat.uni-muenchen.de

[†]Studentische Hilfskraft (SoSe20), Institut für Statistik, LMU München

[‡]Studentische Hilfskraft (WiSe 18/19 - SoSe 20), Institut für Statistik, LMU München

[§]Studentische Hilfskraft (WiSe 18/19 - WiSe 19/20), Institut für Statistik, LMU München

[¶]Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie, LMU München

Contents

1	Grundbegriffe, Skalenniveaus, Datenerhebung	9
1.1	Was ist die Statistik?	9
1.2	Grundbegriffe	9
1.2.1	Untersuchungseinheit und Grundgesamtheit	9
1.2.2	Bestandsmasse und Bewegungsmasse	9
1.2.3	Merkmale und Merkmalsausprägungen	10
1.2.4	Skalentypen	11
1.3	Datenerhebung	11
1.3.1	Erhebungsarten	12
1.3.2	Umfang	12
1.3.3	Datenform	12
1.3.4	Erhebungsmethode	12
1.4	Datenaufbereitung	13
1.4.1	Datenstruktur	13
1.4.2	Kodierung	13
1.4.3	Transformation	14
1.4.4	Statistik-Software	14
2	Häufigkeitsverteilungen, (univariate) grafische Darstellung	15
2.1	Berechnung von Häufigkeiten	15
2.1.1	absolute Häufigkeit	15
2.1.2	Klassenbildung	15
2.1.3	relative Häufigkeit	16
2.1.4	Häufigkeitstabelle	16
2.2	Graphische Darstellung von Häufigkeiten	17
2.2.1	Balken- und Säulendiagramm	17
2.2.2	Kreisdiagramm	17
2.2.3	Histogramme	18
2.3	Ordnungsstatistik	18
2.4	Empirische Verteilungsfunktion	19
2.4.1	Vorgehensweise bei ordinalen und diskreten Merkmalen	19
2.4.2	Vorgehensweise bei stetigen Merkmalen	19
3	Lagemaße	21
3.1	Modus	21
3.2	Median/Zentralwert	21
3.3	Quantile	22
3.4	Boxplots	23

3.5	Mittelungen	23
3.5.1	arithmetisches Mittel	24
3.5.2	harmonisches Mittel	24
3.5.3	geometrische Mittel	24
3.5.4	Vergleich von Mittelwert & Median	25
3.6	Aufgaben	26
4	Streuungsmaße	27
4.1	Spannweite	27
4.2	Quartilsabstand	27
4.3	Mittlere absolute Abweichung (MAD)	27
4.4	Varianz	28
4.5	Standardabweichung	29
4.6	Variationskoeffizient	29
4.7	Aufgaben	30
5	Konzentrationsmaße	31
5.1	Absolute Konzentrationsmaße	31
5.1.1	Konzentrationsrate	31
5.1.2	Konzentrationskurve	31
5.1.3	Herfindahl-Index	32
5.2	Relative Konzentrationsmaße	33
5.2.1	Lorenzkurve	33
5.2.2	Gini-Koeffizient	37
5.3	Aufgaben	39
6	Zusammenhangsmaße	40
6.1	Die Kontingenztafel	40
6.2	Unabhängigkeit	42
6.3	Zusammenhangsmaße für nominale Merkmale	43
6.4	Odds Ratio	45
6.5	Zusammenhangsmaße für ordinale Merkmale	47
6.5.1	Gamma nach Goodman and Kruskal	49
6.5.2	Tau-Maße	50
6.5.3	Rangkorrelationskoeffizient nach Spearman	50
6.6	Zusammenhangsmaße für metrische Merkmale	52
6.6.1	Kovarianz	52
6.6.2	Korrelationskoeffizient nach Bravais-Pearson	52
6.7	Wrap-Up	53
6.8	Aufgaben	54

7	Lineare Einfachregression	56
7.1	Einführung	56
7.2	Plots und Annahmen	56
7.3	Kleinste-Quadrate-Schätzer	57
7.4	Besonderheiten der Regressionsgerade	58
7.5	Güte der Anpassung	58
7.6	Kategoriale Regression	59
7.7	Aufgaben	64
8	Indizes	66
8.1	Verhältniszahlen	66
8.1.1	Gliederungszahlen	66
8.1.2	Beziehungszahlen	66
8.1.3	Indexzahlen	66
8.2	Preisindizes	67
8.2.1	nach Laspeyres	67
8.2.2	nach Paasche	67
8.3	Mengenindizes	68
8.3.1	nach Laspeyres	68
8.3.2	nach Paasche	68
8.4	Umsatzindex	68
8.5	Spezielle Probleme	68
8.5.1	Erweiterung des Warenkorbs	68
8.5.2	Substitution einer Ware	69
8.5.3	Subindizes	69
8.6	Aufgaben	70
9	Zeitreihen	71
9.1	Zerlegung von Zeitreihen, Komponentenmodell	71
9.2	Gleitende Durchschnitte	72
9.3	Saisonale Komponente, konstante Saisonfigur	73
9.4	Zerlegung in Trend und Saison	73
9.4.1	Trend und Saisonkomponente mit Regression	74
9.5	Alternative Ansätze	74
9.6	Aufgaben	75
10	R-Einführung Teil I	76
11	Kombinatorik	78
11.1	Permutation	78
11.1.1	Anzahl Permutationen ohne Wiederholungen	78

11.1.2	Permutation mit Wiederholungen	78
11.2	Kombination	79
11.2.1	Kombination ohne Wiederholung und ohne Reihenfolge	80
11.2.2	Kombination ohne Wiederholung und mit Reihenfolge	80
11.2.3	Kombination mit Wiederholung und ohne Reihenfolge	80
11.2.4	Kombination mit Wiederholung und mit Reihenfolge	80
11.3	Aufgaben	81
12	Wahrscheinlichkeitsrechnung	83
12.1	Grundlagen & -begriffe	83
12.2	Relative Häufigkeit	83
12.3	Laplacesche Wahrscheinlichkeit	84
12.4	Axiome	84
12.5	Bedingte Wahrscheinlichkeit	84
12.5.1	Satz von der totalen Wahrscheinlichkeit	85
12.5.2	Der Satz von Bayes	85
12.6	Stochastische Unabhängigkeit	86
12.7	Aufgaben	87
13	Zufallsvariablen	89
13.1	Diskrete Zufallsvariablen	89
13.2	Stetige Zufallsvariablen	89
13.3	Träger einer Zufallsvariablen	90
13.4	Verteilungsfunktion	90
13.5	Erwartungswert & Varianz	90
13.6	Zweidimensionale Zufallsvariablen	90
13.7	Aufgaben	91
14	Spezielle Verteilungen	92
14.1	Diskrete Verteilungen	92
14.1.1	Diskrete Gleichverteilung	92
14.1.2	Bernoulliverteilung	92
14.1.3	Binomialverteilung	93
14.1.4	Geometrische Verteilung	94
14.1.5	Hypergeometrische Verteilung	94
14.1.6	Poissonverteilung	95
14.1.7	Multinomialverteilung	96
14.1.8	Aufgaben	98
14.2	Stetige Verteilungen	100
14.2.1	Stetige Gleichverteilung	100

14.2.2	Exponentialverteilung	101
14.2.3	Normalverteilung (aka Gauß'sche Glockenkurve)	101
14.2.4	Chi-Quadrat-Verteilung	102
14.2.5	t-Verteilung	102
14.2.6	F-Verteilung	103
14.2.7	Aufgaben	104
14.3	Wichtige Schlüsselbegriffe und "Konzepte" anhand der diskreten Gleichverteilung	105
14.3.1	Parameter von Verteilungen	105
14.3.2	Träger einer Verteilung	105
14.3.3	Verteilungsfunktion, Wahrscheinlichkeitsfunktion und Dichtefunktion	105
14.3.4	Erwartungswert und Varianz	106
14.3.5	Aufgaben	107
15	Grenzwertsätze und Approximationen von Verteilungen	108
15.1	Grenzwertsätze	108
15.1.1	Gesetz der großen Zahlen	108
15.1.2	Zentraler Grenzwertsatz (ZGS)	108
15.2	Approximationen	108
15.2.1	Approximation der Binomial- durch die Normalverteilung	109
15.2.2	Approximation der Binomial- durch die Poissonverteilung	109
15.2.3	Approximation der Poisson- durch die Normalverteilung	110
15.2.4	Approximation der hypergeometrischen durch die Binomialverteilung	110
15.3	Aufgaben	111
16	Schätzen	112
16.1	Die Maximum Likelihood Schätzung	112
16.2	Konfidenzintervalle (KI)	113
16.3	Aufgabe	114
17	Testtheorie	115
17.1	Der p-Wert	115
17.1.1	Der p-Wert beim einseitigen t-Test	115
17.1.2	Der p-Wert beim zweiseitigen t-Test	116
17.2	Hypothesentests	117
17.2.1	Einfacher Gauss-Test	117
17.2.2	Einfacher t-Test	117
17.2.3	Approximativer Einfacher Binomialtest	117
17.2.4	Chi-Quadrat-Anpassungstest	117
17.2.5	F-Test	117
17.2.6	Doppelter Gauss-Test	117

17.2.7	Doppelter t-Test	118
17.2.8	Welch-Test	118
17.2.9	Paired t-Test	118
17.2.10	Approximativer Doppelter Binomialtest	118
17.2.11	Mann-Whitney-U-Test für zwei unabhängige Stichproben	118
17.2.12	Kolmogorov-Smirnov-Anpassungstest	118
17.2.13	Chi-Quadrat-Unabhängigkeitstest	119
17.2.14	Odds-Ratio-Test	119
17.3	Unterschied zwischen Gauss-Tests und t-Tests	119
17.3.1	Unterschied zwischen doppeltem t-Test und Welch-Test	120
17.4	Aufgaben	121
18	Lineare Regression	123
18.1	Kleinste-Quadrate-Schätzer	124
18.2	Multiple lineare Regression	124
18.3	Signifikanztests und Konfidenzintervalle	124
18.4	Bestimmtheitsmaß und Overall F-Test	125
18.4.1	Bestimmtheitsmaß	125
18.4.2	Overall F-Test	125
18.5	Aufgaben	127
19	R-Einführung Teil II	129

Statistik I

1 Grundbegriffe, Skalenniveaus, Datenerhebung

1.1 Was ist die Statistik?

Die Statistik kann man in drei verschiedene Grundaufgaben einteilen. Die deskriptive, explorative und induktive Statistik. In Statistik I wird hauptsächlich die deskriptive und die explorative Statistik thematisiert.

- Deskriptive Statistik: Das Ziel der deskriptiven Statistik ist es, umfangreiches Datenmaterial in Tabellen, Graphiken und Kennzahlen übersichtlich darzustellen. Viele dieser Methoden sind bereits aus der Schule bekannt (z.B. Kreis- und Balkendiagramm oder arithmetisches Mittel a.k.a. der Durchschnitt) und werden in dieser Veranstaltung durch weitere Maßzahlen und Darstellungsweisen ergänzt.
- Explorative Statistik: Hierbei wird das aufbereitete Datenmaterial auf Strukturen und Muster untersucht um mögliche Hypothesen aufzustellen.

Der Begriff *Daten* mag für den ein oder anderen etwas neu sein, bezeichnet jedoch im Grunde genommen nichts anderes als eine Messung oder Erhebung von Werten. Vereinfachend kann man sich einen *Datensatz* z.B. einfach als die Messung der Größe aller Personen im Hörsaal vorstellen.

1.2 Grundbegriffe

1.2.1 Untersuchungseinheit und Grundgesamtheit

Die **Untersuchungseinheit** ist ein einzelnes zu untersuchendes Objekt, welches durch das Symbol ω dargestellt wird.

Die **Grundgesamtheit** ist die Menge an Objekten, über die man etwas sagen möchte. Das Symbol der Grundgesamtheit ist Ω . Somit sind alle Untersuchungseinheiten zusammen die Grundgesamtheit. Diese Beziehung lässt sich wie folgt umschreiben: $\omega \in \Omega$

Beispiel: Die gesamten Studenten im Hörsaal sind die Grundgesamtheit. Ein einzelner Student ist die Untersuchungseinheit.

1.2.2 Bestandsmasse und Bewegungsmasse

Wenn man die Grundgesamtheit Ω zu einem bestimmten Zeitpunkt einmal misst, dann spricht man von einer **Bestandsmasse**.

Beispiel hierfür ist die Messung, bei der festgestellt wird wie viele Studenten am Semesteranfang (10. Oktober) immatrikuliert sind.

Im Gegensatz dazu spricht man von einer **Bewegungsmasse** wenn Ereignisse gemessen werden, die über einen bestimmten Zeitraum eintreten können. Das wären zum Beispiel die Studenten, die während des Wintersemesters das Studium abbrechen.

1.2.3 Merkmale und Merkmalsausprägungen

Wenn wir von einer bestimmten Eigenschaft oder einem Aspekt der Untersuchungseinheit sprechen, nennt man dies **Merkmal** oder **statistische Variable**.

Wenn man sich nun für einen gemessenen/konkreten Wert eines Merkmals interessiert, dann nennt man das **Merkmalsausprägung**. In unserem **Beispiel** könnte man sich für die Leistung in Statistik I der Studenten interessieren, somit sind die **Merkmalesträger** *Studenten, die Statistik I belegt haben*. Eine konkrete **Merkmalsausprägung** mit dem **Merkmal**: "*Leistung in Statistik I*" wäre dann auf der Notenskala 1,0 bis 5,0 beispielsweise die Note 2,3.

Es gibt zwei Arten von Merkmalsausprägungen:

- **Qualitative:** Merkmalsausprägungen, sind Ausprägungen, die keinen mathematischen Wert annehmen, also nicht aus Zahlen bestehen. In unserem Beispiel wäre das z.B. die Einteilung der Leistungen in *bestanden* und *nicht bestanden*.
- **Quantitativ:** Merkmalsausprägungen sind *messbar* und werden somit mit Zahlen angegeben. Beispielsweise in der Klausur 40 von 60 Punkten erreicht. Quantitative Merkmalsausprägungen kann man weiter unterscheiden in diskret, stetig und quasistetig.
 - **Diskrete** Merkmale haben abzählbar viele mögliche Merkmalsausprägungen, das heißt nicht quasi unendlich viele Ausprägungen wie z.B. Sandkörner am Meer (mehr dazu s.u. bei *quasistetig*), sondern man kann die möglichen Merkmalsausprägungen mit nicht allzu großem Aufwand abzählen. (Bsp.: Geschlecht (da gibt es nur männlich oder weiblich) oder die Platzierung beim Schönheitswettbewerb (bei einer Teilnahme von 5 Personen, kann ich nur Platz 1, 2, 3, 4 oder 5 bekommen), Studiendauer in Semestern)
 - **Stetige** Merkmale können unendlich viele verschiedene Merkmalsausprägungen haben. (Das Alter, da zwischen bspw. 18 und 19 unendlich viele Nachkommastellen vorhanden sein können, oder aber auch Anteilswerte sind hierfür ein gutes Beispiel)
 - **Quasistetigen** Daten sind theoretisch stetig (bspw. Körpergröße, Gewicht, o.ä.) Daten, werden aber nur auf einer diskreten Skala (in sehr kleinen Einheiten) gemessen. Da solche Daten *praktisch* in den meisten Fällen auch wie stetige Daten behandelt werden, spricht man hier von *quasistetigen* Daten.
 - Werden (quasi)stetige Daten in Klassen eingeteilt (bspw. Abfrage von Gehalt in Fragebögen), so spricht man von **klassierten** oder **klassiert-stetigen** Daten. Diese Klassenbildung hat weitreichende Implikationen für die anwendbaren Methoden (vgl. folgende Kapitel).

Der **Merkmalsraum** oder **Zustandsraum** ist die Menge aller möglichen Merkmalsausprägungen. Hier in unserem Beispiel hat der Merkmalsraum des Merkmals Notenleistung gerundet auf ganze Noten eine Mächtigkeit von 5 (Die Noten: 1, 2, 3, 4, 5)

1.2.4 Skalentypen

Merkmale besitzen aufgrund der Eigenschaften ihrer möglichen Merkmalsausprägungen bestimmte Skalierungen. Diese richtig zuordnen zu können ist sehr wichtig, denn je nach Skalentyp kann man in den folgenden Kapiteln unterschiedliche Maßzahlen bestimmen.

Nominalskala: Die Ausprägungen bei einer Nominalskala können nur voneinander unterschieden werden, jedoch nicht geordnet oder ins Verhältnis gesetzt werden. Daher kann man die Merkmalsausprägungen nicht werten, in dem man bspw. sagt, blaue Autos seien besser als (oder doppelt so gut wie) rote Auto. Die einzige Aussage die getroffen werden kann ist, ob die Ausprägungen *gleich oder ungleich* sind.

Ordinalskala: Bei der Ordinalskala können wir nicht nur eine Aussage über gleich/ungleich (wie bei der Nominalskala) treffen, sondern zusätzlich auch die Ordnung, d.h. über *kleiner und größer*. Somit können ordinale Merkmale in eine natürliche Rangfolge/Ordnung gebracht werden. Diese Ordnung kann interpretiert werden, jedoch nicht die Abstände. Beispiele hierfür wären die Schweregrade eines Computerspiels oder die Güteklassen eines Hotels, da man zwar sagen kann, dass das 4-Sterne-Hotel besser als das 2-Sterne-Hotel ist, nicht jedoch, dass es doppelt so gut ist o.ä.

Metrische Skala: Die metrische Skala besitzt den höchsten Informationsgehalt. Denn hier kann man zusätzlich zu den Aussagen gleich/ungleich und größer/kleiner auch Aussagen über Abstände zwischen den Merkmalsausprägungen treffen. Somit kann man die Merkmalsausprägungen in eine Ordnung bringen und die Abstände zwischen den Merkmalsausprägungen messen und interpretieren. Die metrische Skala kann man weiter auf splitten in Intervall- und Verhältnisskala: Bei der **Intervallskala** können Differenzen gebildet werden um eine Aussage über den Abstand zu machen, jedoch keine Quotienten, da es kein (natürlichen) Nullpunkt gibt um zwei Werte in Relation zu einander zu setzen. (**Beispiel** Temperatur: Man kann sagen, dass es heute 10 Grad kälter als gestern ist, jedoch nicht , dass es heute halb so warm ist wie gestern. Bei der **Verhältnisskala** gibt es diesen natürlichen Nullpunkt. Deshalb kann man Quotienten bilden und Verhältnisse sinnvoll interpretieren. (**Beispiel** Größe: Man kann sagen, dass Person A doppelt so groß ist wie Person B.) Ein Spezialfall der Verhältnisskala ist die Absolutskala, da nur natürliche Einheiten vorkommen (keine physikalischen Größen). *Natürliche Einheiten* sind bspw. Anzahlen, im Sinne von 10 Äpfel oder 5 Blumen.

1.3 Datenerhebung

Um mit Daten arbeiten zu können müssen diese erst einmal "entstehen oder hergestellt" werden. Dafür gibt es die Datenerhebung. Diese *beschafft* Informationen bzw. *gewinnt* Daten.

1.3.1 Erhebungsarten

Primärerhebung: Wenn ich selbst eine Erhebung (Befragung, Beobachtung, Experiment) starte ohne auf vorhandenes Material zurück zu greifen, dann wird es als Primärerhebung bezeichnet.

Sekundärerhebung: Wenn ich auf bereits vorhandenes Material zurückgreife (z.B. Daten/Statistiken aus dem Internet) dann handelt es sich um eine Sekundärerhebung. Das Material existierte schon vor meiner Recherche.

1.3.2 Umfang

Bei Erhebungen, kann man entweder alle Untersuchungseinheiten einer Grundgesamtheit miteinbeziehen, dann spricht man von einer **Voll-/Totalerhebung** oder nur eine Teilmenge der Grundgesamtheit miteinbeziehen. Dann spricht man von einer **Teilerhebung (Stichprobe)**. Ein Beispiel für die Totalerhebung wäre eine Volkszählung oder eine Evaluation, bei der alle Studenten befragt werden.

Teilerhebungen sind zum Beispiel Qualitätsprüfungen von Produkten, bei denen einzelne Produkte überprüft werden oder die Sonntagsumfrage bei der einzelne, zufällig ausgewählte Bürger aus der Bevölkerung abstimmen können, wen sie wählen würden wenn aktuell Bundestagswahl wäre.

1.3.3 Datenform

Je nachdem wie oft und über welchen Zeitraum eine Erhebung gemacht wird, gibt es bestimmte Datenformen. Bei **Querschnittsdaten** wird der Ist-Zustand zu einem bestimmten Zeitpunkt aufgenommen. An mehreren Untersuchungseinheiten werden ein oder mehrere Merkmale nur *einmal* erhoben. **Beispiele** hierfür sind Lehrerevaluationen oder der Mietspiegel.

Eine Erweiterung hiervon sind die **Longitudinal-, Längsschnitt-, oder Paneldaten**. Hierbei werden ein oder mehrere Merkmale an mehreren Untersuchungseinheiten zu *verschiedenen* Zeitpunkten wiederholt erhoben. Somit interessiert uns hier u.a. auch die Entwicklung der Merkmale im Zeitverlauf.

Beispiele: SOEP (Wiederholungsbefragung von privaten Haushalten in Deutschland), Deutsches Mobilitätspanel (Befragung von Haushalten nach ihrem Mobilitätsverhalten und ihrer PKW-Nutzung). Bei einer **Zeitreihe** wird *ein Merkmal* an aufeinander folgenden Zeitpunkten beobachtet. Hierbei wird die Entwicklung eines Merkmals im Zeitverlauf beobachtet. **Beispiele** hierfür sind Aktienkurse oder (Preis-/Mengen-)Indizes.

1.3.4 Erhebungsmethode

Man unterscheidet bei den Erhebungsmethoden grundsätzlich die **Beobachtung**, die **Befragung** und das **Experiment**. Im Gegensatz zur Befragung ist die Beobachtung, wenn sie verdeckt ausgeführt werden kann, weitgehend unverfälscht, da die Merkmalsträger nicht mit in die Erhebung einbezogen werden. Bei Befragungen sind Personen direkt mit einbezogen und können teils eher schwerlich unverfälschte Aussagen über das eigene Verhalten machen. Jedoch kann bei der Beobachtung

nur das äußerliche Verhalten ermittelt werden, wohingegen man bei einer Befragung auch innere Einstellungen und gedankliche Prozesse durch gezielte Fragen messen kann. Das Experiment ermöglicht Ursachenforschung, ein Nachteil ist jedoch (genauso wie bei der Beobachtung), dass es sehr zeit- und kostenaufwändig sein kann.

1.4 Datenaufbereitung

Nachdem Daten erhoben wurden, müssen diese nun aufbereitet werden, in einem sinnvollen Format abgespeichert und ausgewertet werden können.

1.4.1 Datenstruktur

Um Daten abspeichern zu können, werden diese üblicherweise in Datenmatrix dargestellt. Dies ist wie eine Tabelle, in der

- jede Zeile die Information über eine **Untersuchungseinheit** enthält
- jede Spalte einem **Merkmal** entspricht
- jedes Element der Matrix einer **Merkmalsausprägung** entspricht

Nr	nm	nmqm	wfl	rooms	bj	bez
1	608.40	12.67	48	2	1957	Untergiesing
2	780.00	13.00	60	2	1983	Bogenhausen
3	822.60	7.48	110	5	1957	Obergiesing
4	500.00	8.62	58	2	1957	Schwanthh
5	595.00	8.50	70	3	1972	Aubing
6	960.00	11.85	81	3	2006	Schwanthh

Table 1: *Mietspiegel Beispiel für eine Datenmatrix*

R-Befehl für die Erstellung von Datensätzen: <code>> data.frame()</code>	Dokumentation
---	-------------------------------

⚠ Meist werden Datensätze in der Praxis nicht in R erstellt, sondern aus einer externen Quelle (z.B. einer .csv-Datei oder einer .txt-Datei) importiert.

R-Befehl für den Import von Datensätzen: <code>> read.table()</code>	Dokumentation
---	-------------------------------

1.4.2 Kodierung

Da man mit Zeichenketten (Wörtern) nicht rechnen kann, müssen diese aufbereitet werden. Der Vorgang, bei dem Zeichenkette/Merkmalsausprägungen Zahlen zugeordnet werden nennt man Kodierung (Vergleiche Tabelle 1: Hier würden die Verschiedenen Stadtteile mit Nummern kodiert).

1.4.3 Transformation

Bei der Daten Aufbereitung kann es sinnvoll sein Daten zu transformieren, also mathematisch verändern. Je nach Skalentyp sind verschiedene Arten von Transformationen zulässig.

Nominalskala: Alle *eindeutigen* Transformationen (z.B. vgl. Tabelle 1: Transformation von Stadtteilbezeichnungen in Zahlencodes)

Ordinalskala: Alle Transformationen, welche die *vorliegende Ordnung erhalten*

Intervallskala: Alle Transformationen der Form $g(x) = a + bx, b > 0$ (z.B. Temperaturumrechnung von °F in °C)

Verhältnisskala: Alle Transformationen der Form $g(x) = bx, b > 0$ (z.B. Umrechnung von Minuten in Stunden)

1.4.4 Statistik-Software

Es gibt verschiedene Software-Programme um statistische Analysen durchzuführen. In Statistik-Veranstaltungen oder studentischen Projekten vereinfacht es die Analyse enorm, wenn man nicht alles per Hand rechnen muss, sondern die Software das für einen mit ein paar Befehlen macht. In der praktischen Arbeit mit Daten ist dies ebenfalls der Standard, auch aus Gründen der Reproduzierbarkeit, etc. Hier in der Vorlesung behandeln wir hauptsächlich die Programmiersprache R. Am Ende jeweils von Statistik I und II gibt es in der Veranstaltung eine genauere Einführung dazu, jedoch stolpert man im Verlauf des Skripts auch immer wieder über Befehle und Outputs aus R. Mit R kann man so gut wie alles aus der Vorlesung berechnen. Hat man einmal einen Datensatz importiert, kann man die verschiedensten Sachen damit berechnen ohne den Original-Datensatz selbst zu verändern. Dies ist einer der vielen Vorteile gegenüber Programmen wie bspw. Excel.

Dieses Skript liefert, ergänzend zu den intuitiven und eher nicht-technischen Erklärungen der verschiedenen Themen, zu jeder Methode/Maßzahl eine kurze Info darüber, wie diese in R angewendet bzw. berechenbar ist. Dies war bereits weiter oben in Kapitel 1.4.1 zu sehen und wird sich wie ein roter Faden durch dieses Manuskript ziehen.

R selbst ist kostenlos beziehbar unter

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

Der Editor R-Studio kann hier heruntergeladen werden:

<https://www.rstudio.com/>

Es finden sich auch zahlreiche Hilfe-Seiten, wie z.B.

<https://www.rdocumentation.org/>

oder

<https://stat.ethz.ch/R-manual/>

2 Häufigkeitsverteilungen, (univariate) grafische Darstellung

Häufigkeiten, deren Berechnung und Darstellung in Diagrammen, sind sicher schon weitgehend aus der Schule bekannt. Im Folgenden wird dies wiederholt und weiter ergänzt.

2.1 Berechnung von Häufigkeiten

2.1.1 absolute Häufigkeit

Um die Anzahl der Untersuchungseinheiten zu erfassen, welche eine bestimmte Merkmalsausprägung aufweisen, verwendet man die absolute Häufigkeit.

Diese kann man bei Nominal-, Ordinal- und der metrischen Skala anwenden. Die **allgemeine Formel für die absolute Wahrscheinlichkeit** lautet:

$$n_j = \sum_{i=1}^n I_{a_j}(x_i), \quad j = 1, \dots, k$$

Hierbei muss die folgende Bedingung erfüllt sein:

$$I_{a_j} = \begin{cases} 1 & \text{falls } x_i = a_j \\ 0 & \text{sonst} \end{cases}$$

Wenn eine Untersuchungseinheit I_{a_j} die entsprechende Merkmalsausprägung hat, dann bekommt sie den Wert 1 und wird in die Summe mit einbezogen. Fällt eine Untersuchungseinheit nicht in die zu erfassende Menge, also hat eine Untersuchungseinheit nicht die entsprechende Merkmalsausprägung, dann tritt der 2. Fall ein, nämlich diese erhält den Wert 0. Zum Schluss werden alle Untersuchungseinheiten, die diese gewünschte Ausprägung aufweisen aufsummiert.

Die Berechnung der absoluten Häufigkeit macht vor allem für Merkmale Sinn, bei denen nicht allzu viele verschiedene Merkmalsausprägungen beobachtet werden (z.B. Noten, Lieblingsfarbe, o.ä.).

2.1.2 Klassenbildung

Um die Übersicht bei stetigen und diskreten Merkmalen mit vielen Ausprägungen (=quasistetig) zu behalten, macht man sich die Klassenbildung zu Nutze. Um eine sinnvolle und brauchbare Verteilung bei der Klassifizierung zu bekommen, bietet es sich an, die Grundgesamtheit in \sqrt{n} Klassen zu teilen (grobe, sehr allgemeine Faustregel!)

Allgemein gibt es zwei Möglichkeiten zur Wahl der Klassen

1. nach sachologischen Gegebenheiten
2. nach willkürlichen Kriterien

da man jedoch mit den willkürlichen Kriterien Strukturen verfälschen kann, sollten diese eher vermieden werden.

Einschub: Mathematische Notation bei Klassenbildung

k	Anzahl der Klassen
e_{j-1}	untere Klassengrenze der j-ten Klasse
e_j	obere Klassengrenze der j-ten Klasse
$d_j = e_j - e_{j-1}$	Klassenbreite der j-ten Klasse
$a_j = \frac{1}{2}(e_j + e_{j-1})$	Klassenmitte der j-ten Klasse
n_j	Anzahl der Beobachtungen in der j-ten Klasse

2.1.3 relative Häufigkeit

Die absolute Häufigkeit ist bei unterschiedlichen Stichprobenumfängen nicht vergleichbar. Um dieses Problem zu umgehen kann man die relative Häufigkeitsverteilung verwenden, die die gesamte Verteilung auf 1 normiert. Somit sind (relative) Häufigkeiten nun auch für unterschiedliche Stichprobenumfänge vergleichbar.

Die relativen Häufigkeiten f_j sind *die Anteile* die auf jede Ausprägung entfallen. Man berechnet sie durch den Quotienten aus der absoluten Häufigkeit und dem Stichprobenumfang.

$$\frac{\text{absolute Häufigkeit}}{\text{Stichprobenumfang}} = \frac{n_j}{n} = f_j$$

2.1.4 Häufigkeitstabelle

Die Häufigkeitstabelle umfasst alle möglichen Ausprägungen eines Merkmals (bzw. alle gebildeten Klassen) und deren (relative & absolute) Häufigkeiten. Man kann sie bei *diskreten* und bei *gruppierten stetigen* Merkmalen (vgl. Tabelle 2) verwenden, jedoch nicht bei stetigen, da jede Beobachtung einen anderen Wert hat und somit die Tabelle "unendlich lang" werden würde. Für gruppierte stetige Merkmale kommen zusätzlich zu den Spalten *Merkmalsausprägung* a_j , *absolute Häufigkeit* n_j , und *relative Häufigkeit* f_j , noch die Spalten *Klassengrenzen* $[e_{j-1}; e_j[$ und die *Klassenbreite* d_j hinzu.

j	$[e_{j-1}; e_j[$	d_j	n_j	f_j
1	$[e_0; e_1[$	d_1	n_1	f_1
:	:	:	:	:
:	:	:	:	:
k	$[e_{k-1}; e_k[$	d_k	n_k	f_k
Σ			n	1

Table 2: Allgemeine Form bei gruppierten (quasi-)stetigen Merkmalen

2.2 Graphische Darstellung von Häufigkeiten

Da graphische Darstellungen leichter verständlich und übersichtlicher sind, werden die Daten meist ergänzend zu den Häufigkeitstabellen auf diese Art und Weise dargestellt.

2.2.1 Balken- und Säulendiagramm

Gestaltung von Säulendiagrammen

- Auf der x-Achse (Abszisse) sind die verschiedenen Merkmalsausprägungen abgetragen, darüber entstehen die Säulen. Jede Säule entspricht einer Merkmalsausprägung
- Auf der y-Achse (Ordinate) wird die Skala abgetragen, um ablesen zu können, wie groß die Anzahl oder der Anteil einer Merkmalsausprägung ist.
- Die Höhe der Säule kann die absoluten oder die relative Häufigkeit darstellen.

Gestaltung von Balkendiagrammen

Das Balkendiagramm ist identisch zum Säulendiagramm, jedoch um 90 Grad gedreht.

- Auf der x-Achse (Abszisse) ist die Skala abgetragen
- Auf der y-Achse (Ordinate) sind die Merkmalsausprägungen abgetragen.

Gestapeltes Balkendiagramm

Bei einem gestapelten Balkendiagramm nutzt man die Tatsache, dass sich die relativen Häufigkeiten zu 1 aufsummieren. Hat man beispielsweise in verschiedenen Jahren unterschiedliche Zusammensetzungen der relativen Anteile, kann man diese in einem gestapelten Säulendiagramm gut vergleichen.

R-Befehl für Balken-/Säulendiagramme: <code>> barplot(data)</code>	Dokumentation
---	-------------------------------

⚠ Für gestapelte Balkendiagramme muss der Funktion eine Tabelle übergeben werden.

2.2.2 Kreisdiagramm

Gestaltung von Kreisdiagrammen

Jede Merkmalsausprägung erhält einen Sektor des Kreises. Man berechnet den Winkel durch die Multiplikation der relativen Häufigkeit mit 360° .

Das Kreisdiagramm kann bei allen Skalen verwendet werden, jedoch kann die Ordnung von Ausprägungen nicht wiedergegeben werden. Somit würde bei Verwendung einer Ordinalskala der Informationsgehalt über die Ordnung verloren gehen.

R-Befehl für Kreisdiagramme: <code>> pie(data)</code>	Dokumentation
--	-------------------------------

2.2.3 Histogramme

Das Histogramm ist, im Gegensatz zu den bisher vorgestellten Diagramm-Typen, nur bei metrischer Skala anwendbar.

Gestaltung von Histogrammen

- Da hier ein metrisches Merkmal vorliegt, muss dieses zunächst einmal in Klassen eingeteilt werden.
- Auf der x-Achse (Abszisse) ist die Skala des Merkmals abgebildet.
- Da die Fläche der einzelnen Balken den relativen Häufigkeiten entspricht (bzw. Histogrammfläche und relative Häufigkeit sind proportional zueinander), lässt sich die Höhe eines Balkens (Ordinate) wie folgt berechnen:

$$h_j = \frac{f_j}{d_j} \quad (<=> h_j \cdot d_j = f_j)$$

- Werden Klassen gleicher Breite verwendet, so wird das Öfteren auch die relative/absolute Häufigkeit auf der Ordinate abgetragen, da auch hierdurch die Forderung der Proportionalität (siehe oben) gewahrt wird.

Probleme

Da das Aussehen von der gewählten Klassengröße abhängt, sollte man, wie schon in (vgl. *Klassensbildung* Kapitel 2.1.1) erwähnt, sachlogische Gegebenheiten bei der Wahl der Klassenbreiten heranziehen.

Offene Klassen, die gegen unendlich gehen, sind nicht abbildbar. Eine Möglichkeit ist, die Klasse so zu wählen, dass darin schon die Mehrheit der Merkmalsausprägungen enthalten sind. (vgl. Induktive Statistik im zweiten Semester)

R-Befehl für Histogramme: <code>> hist(data)</code>	Dokumentation
--	-------------------------------

⚠ Per default nutzt R gleiche Klassenbreiten, deren Anzahl mit Hilfe der [Formel von Sturges](#) berechnet wird. Selbst gewählte Klassen (auch ungleicher Breite) können mit dem `breaks`-Argument übergeben werden.

2.3 Ordnungsstatistik

Für die Berechnungen einige Maße ist es wichtig, dass die Merkmalsausprägungen geordnet sind. Die *Ordnungsstatistik* ist nur bei ordinaler und metrischer Skala verwendbar, da man nominale Merkmale nicht ordnen kann (vgl. Kapitel 1.2.4). Dabei werden die Ausprägungen der Urliste in eine aufsteigende Ordnung gebracht. Damit man erkennt, ob eine Urliste oder eine Ordnungsstatistik vorliegt, werden die tiefgestellten Indizes bei der Ordnungsstatistik in Klammern $x_{(i)}$ gesetzt. Die tiefgestellte Zahl in Klammern gibt den Rang an. Gibt es zwei Merkmalsausprägungen mit der gleichen Ausprägung, dann nennt man das *Bindung* (Tie). Wenn man Bindungen in einer Ordnungsstatistik berücksichtigt, dann erhalten diese den gemittelten Wert ihrer bisherigen Position

in der Ordnungsstatistik.

Beispiel: Urliste: $x_1 = 8$; $x_2 = 4$; $x_3 = 5$; $x_4 = 1$; $x_5 = 4$

Ordnungsstatistik (mit Bindungen): $x_{(1)} = 1$; $x_{(2,5)} = 4$; $x_{(2,5)} = 4$; $x_{(4)} = 5$; $x_{(5)} = 8$

Rangabfrage: $Rg(4) = 2, 5$

R-Befehl für die Ordnungsstatistik: <code>> sort(data)</code>
--

Dokumentation

2.4 Empirische Verteilungsfunktion

Bei der empirischen Verteilungsfunktion benötigt man die Ordnungsstatistik, daher ist sie nur für Merkmale mit ordinaler und metrischer Skala anwendbar. Hier werden nicht die einzelnen Merkmalsausprägungen in ihrer Häufigkeit einzeln dargestellt, sondern die Häufigkeiten werden *kumuliert*, d.h. aufsummiert.

$$F(x) = \sum_{a_j \leq x} f(a_j)$$

$F(x)$ ist die kumulierte relative Häufigkeit an der Stelle x , das bedeutet, dass alle Wahrscheinlichkeiten $f(a_j)$ der Merkmalsausprägungen a_j kleiner gleich x aufsummiert werden.

Die relativen Häufigkeiten werden immer weiter aufsummiert, weshalb die Funktion monoton wachsend ist und dann schließlich bei 1 stagniert, da die kumulierte relative Häufigkeit nicht höher als 1 sein kann. Somit ist der Wertebereich von $F(x)$ von 0 und 1.

Es gibt verschiedene Vorgehensweisen bei diskreten und stetigen Merkmalen, deshalb betrachten wir im Folgenden die Verteilungsfunktion für die beiden Skalen separat.

2.4.1 Vorgehensweise bei ordinalen und diskreten Merkmalen

1. Ordnungsstatistik bilden
2. Relativen Häufigkeiten berechnen
3. Kumulierte Häufigkeiten $F(x)$ für jede *unterschiedliche* Merkmalsausprägung berechnen
4. Graph: Trage die kumulierten Häufigkeiten als $(x_i; F(x_i))$ in ein Diagramm ein und verlängere die Punkte mit einem *horizontalen* Strich, bis zum Abszissenwert der nächsten Merkmalsausprägung. Somit entsteht eine Treppenfunktion, welche von 0 bis 1 geht.

Die Rechenregeln für ordinale und diskrete Merkmale: Skript (vgl. Slide 2.39)

2.4.2 Vorgehensweise bei stetigen Merkmalen

1. (Geordnete) Klassen bilden (vgl. Kapitel 2.1.2)
2. Relativen Häufigkeiten der Klassen berechnen

3. Kumulierten Häufigkeiten $F(e_j)$ für jede Klasse berechnen
4. Graph: Trage die kumulierten Häufigkeiten als $(e_j; F(e_j))$ in das Diagramm ein und verbinde die Punkte (da man eine Gleichverteilung innerhalb der Klassen annimmt).

Berechnung der empirischen Verteilungsfunktion von klassierten Daten:

Die folgende Formel verwendet man, um eine kumulierte Häufigkeit einer bestimmten Merkmalsausprägung zu bekommen.

$$F(x) = \begin{cases} 0 & x < e_0 \\ F(e_{j-1}) + \frac{f_j}{d_j}(x - e_{j-1}) & x \in [e_{j-1}] \\ 1 & x > e_k \end{cases}$$

Die Schwäche ist, dass man hierfür von einer Gleichverteilung innerhalb der Klassen ausgehen muss, was eine sehr starke Annahme ist und nicht immer unbedingt realistisch ist.

R-Befehl für die emp. Verteilungsfunktion: `> ecdf(data)`

[Dokumentation](#)

3 Lagemaße

Grob gesagt beschreiben (zentrale) Lageparameter, wo sich der Schwerpunkt der Daten auf einer Skala befindet. Manche Lageparameter machen nur bei bestimmten Skalen Sinn. In der folgenden Tabelle ist markiert, bei welchen Skalen die einzelnen Lageparameter jeweils Sinn machen.

	Nominalskala	Ordinalskala	metrische Skala
Modus	x	x	x
Median		x	x
Quantile		x	x
Box-Plots		x	x
Mittelungen			x

3.1 Modus

Die Merkmalsausprägung, die am häufigsten auftritt, nennt man Modus. Diese Maßzahl macht sowohl bei diskreten oder bei (quasi)stetigen Merkmalen Sinn, solange es eine überschaubare Anzahl an verschiedenen Merkmalsausprägungen gibt. Dies ist bei diskreten Merkmalen logischerweise öfter der Fall als bei (quasi)stetigen Merkmalen. Bei klassiert-stetigen Merkmalen nimmt man oft die Klassenmitte der Klasse mit der höchsten absoluten Häufigkeit.

R-Befehl für den Modus: `> Mod(data)`

[Dokumentation](#)

⚠ Die R-Funktion ist **nicht** Teil von **base-R** sondern Teil des Paketes **DescTools**. Dieses muss zunächst installiert (`install.packages("DescTools")`) und anschließend importiert werden (`library(DescTools)`).

3.2 Median/Zentralwert

Um den Median bei diskreten Merkmalen anwenden zu können, bringt man die Urliste zuerst einmal in eine Ordnungsstatistik. Sind die Merkmalsausprägungen aufsteigend geordnet, teilt der Median die Grundgesamtheit in zwei gleich große Teile. Unterhalb des Medians liegen die Hälfte (50%) der Werte die *kleiner oder gleichen* dem Median sind und im zweiten Teil befinden sich die andere Hälfte der Merkmalsausprägungen die *größer oder gleich* dem Median sind. Der Median wird mit $\tilde{x}_{0,5}$ bezeichnet. Alternativ kann man den Median auch mit der Verteilungsfunktion berechnen, indem man den Wert bestimmt, bei dem die kumulierte relative Häufigkeit 0,5 beträgt. Dies würde dann so aussehen: $F(\tilde{x}_{0,5}) = 0,5$.

Die Stärke des Medians ist, dass dieser relativ unempfindlich gegenüber Ausreißern und Extremwerten ist. Das heißt, wenn bspw. 10 Merkmalsausprägungen einer Grundgesamtheit im Bereich zwischen 0 und 10 haben, macht es keinen Unterschied, ob der größte Wert auch innerhalb dieses Bereichs liegt (also z.B. den Wert 10 hat) oder weit drüber hinaus geht (z.B. 60), da für den Median lediglich die *Anzahl* der Beobachtungen über-/unterhalb herangezogen werden, nicht jedoch deren konkreter Wert.

Für den Median gibt es eine Fallunterscheidung bzgl. der Berechnung zwischen gerader und ungerader Anzahl der Beobachtungen. Für die ungerade Anzahl an Beobachtungen n ist es einfach der mittlere Wert der Ordnungsstatistik. Da es jedoch bei gerade Anzahl von Beobachtungen keine Mitte gibt, ist es hier ein bisschen aufwändiger. Man bildet das arithmetische Mittel der beiden Beobachtungen, zwischen denen die Mitte der Ordnungsstatistik wäre.

$$\tilde{x}_{0,5} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{falls } n \text{ gerade} \end{cases}$$

Bei klassierten metrischen Merkmalen kann man nicht so wie oben beschrieben vorgehen, da man die exakten Merkmalsausprägungen nicht kennt.

Deshalb greift man hier wieder auf die Annahme der Gleichverteilung innerhalb der Klassen zurück. Man bestimmt die Klasse, in der der Median liegt, d.h. man schaut in welcher Klasse die kumulierte relative Häufigkeit 0,5 liegt. Unter Verwendung der Formel für klassierten Daten (siehe Formelsammlung) kann man nun den Median berechnen.

Man subtrahiert von 0,5 die kumulierte relative Häufigkeit der unteren Klassengrenze der Klasse, die den Median enthält. Das Ergebnis dividiert man durch die relative Häufigkeit der Klasse in der der Median enthalten ist. Den Quotienten multipliziert man mit der Klassenbreite. Zum Schluss addiert man die untere Klassengrenze, der Klasse, in der der Median liegt dazu.

R-Befehl für Median: <code>> median(data, ...)</code>
--

Dokumentation

3.3 Quantile

Die Quantile sind wie der Median bei ordinalen und metrischen Daten anwendbar (Um genau zu sein ist der Median lediglich ein bestimmtes Quantil). Uns interessiert möglicherweise nicht nur der Median, der uns durch seine Lage sagt, an welcher Stelle 50% der Werte kleiner oder gleich diesem Wert sind, sondern potentiell die gleiche Aussage auch für andere Prozentwerte (bspw. 25%). Somit können wir die Grundgesamtheit durch ein Quantil in zwei beliebig großen Teilbereich aufteilen, welche sich zu 100% aufsummieren. Man wählt einen Wert α zwischen 0 und 1 auswählen und kann somit die Lage des Quantils zu bestimmen. Um die Realisation des Quantils zu bestimmen, multipliziert man die Anzahl der Beobachtungen n mit α .

- Wenn das Produkt $n\alpha$ keine ganze Zahl ist, dann ist die nächst größere ganze Zahl die Realisation des Quantils.
- Wenn das Produkt $n\alpha$ eine ganze Zahl ist, dann addiert man diese mit dem nächst größeren Wert und bildet daraus den Mittelwert.

Wie der Median sind auch die Quantile (relativ) unempfindlich gegenüber Ausreißern, da sie in erster Linie nur auf der Lage der Ausprägungen basiert und nicht auf den konkreten Werten.

R-Befehl für Quantile: <code>> quantile(data, probs, ...)</code>

Dokumentation

Besondere Quantile Zusätzlich zum Median, der dem 50%-Quantil entspricht, gibt es noch zwei weitere besondere Quantile. Zum einen das **untere Quartil**, das dem 25%-Quantil entspricht, und zum anderen das **oberen Quartil**, welches dem 75%-Quantil entspricht. Diese beiden sind u.a. im folgenden Kapitel 3.4 für die Box-Plots relevant.

3.4 Boxplots

Ein Boxplot ist eine grafische Darstellung, mit der man sich schnell einen Überblick über die Verteilung der Ausprägungen eines Merkmals schaffen kann und sehr schnell die Unterschiede zwischen Verteilungen von verschiedenen Merkmalen erkennen. Um den *einfachen* Box-Plot zu zeichnen, benötigt man den Median und das obere und untere Quartil, sowie das Minimum und das Maximum.

Vorgehensweise:

1. Zuerst zeichnet man die Box, die durch das untere und obere Quartil begrenzt ist.
2. Danach zeichnet man in die Box den Median als dicke Linie ein.
3. Die Striche die von der Box weggehen sind die sog. *Whiskers*. Sie gehen bis zum Minimum und Maximum.

Zeichnet man den modifizierten Boxplot, so kommen drei Schritte hinzu. Die Modifikation liegt hierbei in der Kennzeichnung von Ausreißern. Ausreißer sind hierbei (für gewöhnlich) definiert als Werte, die *"mehr als 1,5-mal die Boxlänge von einem der beiden Quartile entfernt sind"*.

1. Man berechnet die Länge der Box, indem man das untere vom oberen Quartil abzieht. Anschließend bestimmt man damit die "Grenze", ab wo ein Wert ein Ausreißer nach oben oder unten wäre.
2. Die Whiskers gehen nun nicht mehr bis zum Minimum/Maximum, sondern lediglich bis zum kleinsten/größten Wert innerhalb der berechneten Grenzen.
3. Alle Werte, außerhalb dieser Grenzen zeichnet man als Kreis ein. Dies sind Ausreißer.
4. *Anmerkung:* Von Extremwerten spricht man, wenn ein Wert mehr als 3 Boxlängen entfernt von dem oberen bzw. unteren Boxenrand entfernt liegt. Diese werden manchmal separat mit einem Sternchen eingezeichnet.

R-Befehl für den Boxplot: <code>> boxplot(data, ...)</code>
--

Dokumentation

3.5 Mittelungen

Im Folgenden werden drei verschiedene Mittelungen aufgeführt. Je nach dem ob die Daten in gleicher/unterschiedlicher Gewichtung in einen Mittelwert einfließen sollen oder ob es sich um eine multiplikative Verknüpfung zwischen den Werten handelt werden arithmetisches, harmonisches oder geometrisches Mittel verwendet.

3.5.1 arithmetisches Mittel

Das arithmetische Mittel ist den meisten sicherlich als "*Durchschnitt*" bekannt. Hierbei gehen alle Daten mit *gleicher* Gewichtung in die Berechnung ein. Diese erfolgt durch Aufsummieren aller Merkmalsausprägungen und Teilen durch die Anzahl n der Merkmalsausprägungen.

Liegen klassierte/gruppierte Daten mit unterschiedlichen Klassengrößen n_j vor, so macht man Gebrauch vom gewichteten arithmetischen Mittel. Hierbei werden die Merkmalsausprägungen beim Aufsummieren mit ihrer Klassengröße n_j gewichtet und diese Summe anschließend durch n geteilt. Eine Schwäche des Mittelwertes ist die Empfindlichkeit gegenüber Ausreißern/Extremwerten, welche dadurch zustande kommt, dass die Abweichungen aller Ausprägungen in Summe Null ergeben. Somit verschiebt sich der Mittelwert durch extreme Werte sehr schnell in deren Richtung.

R-Befehl für das arithm. Mittel: <code>> mean(data, ...)</code>
--

Dokumentation

3.5.2 harmonisches Mittel

Das harmonische Mittel wird im Gegensatz zum arithmetischen Mittel verwendet, wenn die Merkmalsausprägungen unterschiedlich gewichtet werden sollen. Ein Hinweis darauf, dass man das harmonische Mittel verwenden muss, sind Verhältniszahlen, bspw. $\frac{km}{h}$ oder $\frac{EUR}{h}$.

Man berechnet das harmonische Mittel durch den Quotienten aus den aufsummierten Anteilen und der Summe der Quotienten aus der Gewichtung und der Merkmalsausprägung x_i . Wenn man das harmonische Mittel aus einer Häufigkeitstabelle berechnet, ergibt sich ein Sonderfall. Denn der Zähler (die aufsummierten Anteile), ist n (bzw. 1 bei relativen Häufigkeiten) und der Nenner ergibt sich aus dem Quotient der absoluten Häufigkeit (bzw. relativen Häufigkeit) der Merkmalsausprägung durch die Merkmalsausprägung selbst.

Kein R-Befehl für das harmonische Mittel in base-R verfügbar.

3.5.3 geometrische Mittel

Das geometrische Mittel verwendet man bei relativen Merkmalsausprägungen (z.B. Wachstumsfaktoren), die sich auf einen bestimmten Ausgangswert beziehen. Bei solchen Werten kann man Aussagen treffen, wie sich Werte zwischen zwei Zeitpunkten verändert haben. Mit dem geometrischen Mittel berechnet man dann die durchschnittliche Veränderung eines Wertes im Zeitverlauf. Da es sich um eine multiplikative Verknüpfung handelt, multipliziert man alle relativen Veränderungen und zieht dann die n -te Wurzel aus der Anzahl der n relativen Veränderungen. Alternativ kann man auch anstelle des Aufmultiplizierens auch einfach den Quotient des n -ten Ausgangswert durch den ersten Ausgangswert teilen und anschließend die n -te Wurzel ziehen.

Kein R-Befehl für das geometrische Mittel in base-R verfügbar.
--

3.5.4 Vergleich von Mittelwert & Median

Ein Unterschied zwischen arithmetischem Mittel und Median ist die Empfindlichkeit bzw. Robustheit gegenüber den Ausreißern. Der Median ist relativ robust ggü. Ausreißern, während das arithmetische Mittel eher empfindlich ist. Basierend auf diesen Eigenschaften können durch deren Vergleich Rückschlüsse auf die Verteilung der Daten gezogen werden.

Symmetrische Verteilung Fallen Median und arithmetisches Mittel zusammen (d.h. sind in etwa gleich), dann spricht man von einer symmetrischen Verteilung, da hierdurch der Schluss gezogen werden kann, dass entweder (i) keine Ausreißer vorliegen oder (ii) sich die Ausreißer auf beiden Seiten (d.h. nach oben und unten) die Waage halten.

Asymmetrische Verteilung Fallen Median und arithmetisches Mittel auseinander, dann spricht man von einer asymmetrischen Verteilung. Wenn das arithmetische Mittel größer ist als der Median, kann man daraus schließen, dass es tendenziell eher Ausreißer nach oben gibt. Eine solche Verteilung wird als linkssteil bzw. rechtsschief bezeichnet. Linkssteil, da sich die untere Hälfte der Daten (links vom Median) eher nah am Median befindet und das Histogramm somit eher steil ansteigend aussieht. Rechtssteil, da das arithmetische Mittel durch die potenziellen Ausreißer weiter nach "rechts gezogen" (rechtsschief) wird und das Histogramm eher flach abfallend aussieht. Deshalb ist das arithmetische Mittel in diesem, Fall größer als der Median.

Die Verteilung heißt im Gegensatz dazu rechtssteil bzw. linksschief, wenn der Median größer ist als das arithmetische Mittel. D.h. es kann genau dieselbe Intuition wiederverwendet werden, nur diesmal in die andere Richtung.

3.6 Aufgaben

1. Unterschiede zwischen Mittelwert und Median?

- a) Mittelwert ist robuster ggü. Ausreißern ☐
- b) Median ist robuster ggü. Ausreißern ☐
- c) Keine ☐

2. Der Median ..

- a) .. liegt immer genau in der Mitte der Box. ☐
- b) .. entspricht dem 50%-Quantil. ☐
- c) .. entspricht dem 2. Quartil. ☐
- d) .. ist wichtig dafür, zu berechnen wann ein Wert ein Ausreißer ist. ☐

3. Welche Mittelung ist geeignet, um den durchschnittlichen Anstieg der Transferausgaben in der Fußballbundesliga zu ermitteln?

- a) Arithmetisches Mittel ☐
- b) Geometrisches Mittel ☐
- c) Harmonisches Mittel ☐
- d) Alle drei machen Sinn ☐

4 Streuungsmaße

Bis jetzt haben wir uns Lagemaße angeschaut, welche nur etwas über die (zentrale) Lage der Daten aussagen. Um zu quantifizieren, wie stark die Daten schwanken/streuen (Hier: Wie stark die Daten um einen Mittelwert schwanken). Weil man zur Berechnung von Streuungsmaßen Differenzen benötigt, ist auch hier wieder nur für bestimmte Skalenniveaus die Berechnung der vorgestellten Streuungsmaße möglich.

	Nominalskala	Ordinalskala	metrische Skala
Spannweite		x	x
Quartilsabstand		x	x
Mittlere absolute Abweichung (MAD)			x
Varianz			x
Standardabweichung			x
Variationskoeffizient			x

4.1 Spannweite

Als **Streubereich** bezeichnet man den Bereich in dem die gesamten Merkmalsausprägungen liegen. Dessen Breite bezeichnet man als **Spannweite**. Dabei subtrahiert man den kleinsten Wert (*Minimum*) vom größten Wert (*Maximum*). Da es nur auf diesen beiden Werten basiert, ist es anfällig gegenüber Ausreißern, welche zu sehr großen Spannweiten führen können.

R-Befehl für die Spannweite: `> max(data) - min(data)`

[Dokumentation](#)

4.2 Quartilsabstand

Im Gegensatz zur Spannweite ist der (Inter-)Quartilsabstand *robust* gegenüber Ausreißern (d.h. er wird nicht von ihnen beeinflusst). Im vorigen Kapitel wurden bei den Quantilen (vgl. Kap. 3.3) die beiden besonderen Quantile, *oberes und unteres Quartil*, vorgestellt. Aus deren Abstand ergibt sich der (Inter-)Quartilsabstand. Grafisch kann man sich dies als die Länge der Box im Boxplot veranschaulichen (vgl. Kap. 3.4). Im (Inter-)Quartilsabstand liegen somit die mittleren/zentralen 50% der Werte, darunter logischerweise auch der Median (vgl. Kap. 3.2).

$$d_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$$

R-Befehl für den Quartilsabstand: `> IQR(data)`

[Dokumentation](#)

4.3 Mittlere absolute Abweichung (MAD)

Die *mittlere absolute Abweichung* (Englisch: mean absolute deviation) gibt die durchschnittliche Abweichung der Merkmalsausprägungen um einen bestimmten (zentralen) Wert A an. A kann beispielsweise der Median oder der Mittelwert sein. Bei der Berechnung werden die betragsmäßigen

Differenzen aus den einzelnen Datenpunkten und A aufsummiert und durch die Beobachtungszahl n dividiert.

$$D_A = \frac{1}{n} \sum_{i=1}^n |x_i - A|$$

Die Betragsstriche verhindern dabei, dass sich die positiven und negativen Abweichungen "aufheben". Eine weitere Möglichkeit dies zu vermeiden, wäre die Differenzen zu quadrieren (vgl. Kap. 4.4).

R-Befehl für den MAD: `> mad(data, center = median(data), ...)` [Dokumentation](#)

4.4 Varianz

Die Varianz s^2 ist die mittlere *quadratische* Abweichung zum arithmetischen Mittel. Die Varianz ist dabei das gängigste Maß für die Streuung von Merkmalsausprägungen um das arithmetische Mittel. Wie oben bereits erwähnt, wird durch die Quadrierung verhindert, dass sich die positiven & negativen Abweichungen "aufheben" können.

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

Verschiebungssatz Jede einzelne Abweichung auszurechnen und zu quadrieren kann bei großer Anzahl n von Datenpunkten sehr umständlich sein. Deshalb kann durch die Umformung mittels des Verschiebungssatzes eine handrechnerisch leichtere Form erreicht werden:

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

Anmerkung: Neben dem arithm. Mittel (welches auch für die "normale" Varianz-Formel benötigt wird) muss hier nur noch das arithm. Mittel der *quadratierten* Daten berechnet werden, um schließlich die Varianz berechnen zu können.

Varianz aus klassierten/gruppierten Daten Liegen nun keine Einzeldaten vor sondern gruppierte Daten von denen die Varianz bestimmt werden soll, geht man folgendermaßen vor: Die Streuung in zwei Teile zerlegt ($s_{zwischen}^2$ und $s_{innerhalb}^2$), diese separat berechnet und anschließend addiert werden.

Anmerkung: Dies mag auf den ersten Blick etwas kontra-intuitiv erscheinen, jedoch ist diese Berechnung auch ohne Kenntnis der Einzeldaten (d.h. mit de facto weniger Information) möglich. Aufgrund dessen ist diese Zerlegung in manchen Fällen hilfreich.

$s_{zwischen}^2$: Bei der Streuung zwischen den Klassen wird die durchschnittliche quadratische Abweichung der Mittelwerte der Klassen (\bar{x}_j) vom Mittelwert aller Daten (\bar{x}) berechnet.

$$s_{zwischen}^2 = \frac{1}{n} \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2$$

$s_{innerhalb}^2$: Bei der Streuung innerhalb der Klassen wird zuerst die Streuung jeder einzelnen Gruppe ($s_j^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$) berechnet. Von diesen gruppenspezifischen Varianzen wird anschließend das gewichtete arithmetische Mittel (vgl. Kap. 3.5.1) gebildet:

$$s_{innerhalb}^2 = \frac{1}{n} \sum_{j=1}^k n_j s_j^2$$

Somit kann man die gesamte Varianz durch die Summe aus $s_{zwischen}^2$ und $s_{innerhalb}^2$ berechnen.

R-Befehl für die Varianz: `> var(data)`

[Dokumentation](#)

⚠ Die R-Funktion teilt bei der Berechnung der Varianz nicht durch n sondern durch $n - 1$. Die Hintergründe dafür werden in Statistik II (vgl. Kap. 16.1) erläutert. Um dies zu umgehen und die (empirische) Varianz zu berechnen, sollte das Ergebnis dieses R-Befehls mit $(n - 1) / n$ multipliziert werden.

4.5 Standardabweichung

Die Standardabweichung erhält man, indem man die positive Wurzel der Varianz (vgl. Kap. 4.4) zieht. Der Vorteil der Standardabweichung ist, dass diese wieder in der gleichen Einheit wie die Beobachtungswerte vorliegt, da wir das Quadrieren aus der Formel für die Varianz durch das Wurzelziehen wieder auflösen.

R-Befehl für die Standardabweichung: `> sd(data)`

[Dokumentation](#)

⚠ Wie bei der Varianz wird auch beim R-Befehl für die Standardabweichung nicht durch n sondern durch $n - 1$ geteilt.

4.6 Variationskoeffizient

Beim Variationskoeffizient wird die Standardabweichung in Beziehung zum arithm. Mittel gesetzt, damit die Streuungen von Datensätzen mit unterschiedlichen Mittelwerten miteinander verglichen werden können. Die Berechnung erfolgt durch den Quotienten aus Standardabweichung und arithm. Mittel, dadurch wird der Variationskoeffizient dimensionslos.

R-Befehl(e) für den Variationskoeffizient: `> sd(data) / mean(data)`

4.7 Aufgaben

1. Bei welcher Maßzahl werden hohe Abweichungen vom Mittelwert stärker gewichtet?

- a) MAD ☐
- b) Varianz ☐
- c) Bei beiden gleich stark ☐

2. Welche Aussagen zur Streuungszerlegung sind wahr?

- a) Die Varianz innerhalb der Gruppen ist immer größer als zwischen den Gruppen. ☐
- b) Man kann die Varianz innerhalb und zwischen den Gruppen einfach addieren um die Gesamtvarianz zu erhalten. ☐
- c) Es gibt Sonderfälle, bei denen die Streuung zwischen den Gruppen der Gesamtstreuung entspricht. ☐
- d) Es muss immer eine Streuung innerhalb der Gruppen vorliegen. ☐

3. Der Verschiebungssatz ..

- a) .. erleichtert die Berechnung des arithmetischen Mittels. ☐
- b) .. kann auch bei gruppierten Daten verwendet werden. ☐
- c) .. dient zur Berechnung des arithmetischen Mittels der quadrierten Daten. ☐
- d) .. benötigt das arithmetische Mittel der quadrierten Daten. ☐

4. Welche der folgenden Aussagen zum Variationskoeffizienten sind wahr?

- a) Der Variationskoeffizient ermöglicht den Vergleich von Streuungen von Merkmalen, die in verschiedenen Einheiten gemessen werden. ☐
- b) Der Variationskoeffizient ermöglicht den Vergleich von Streuungen von Merkmalen, die in verschiedenen Größenordnungen liegen. ☐
- c) Für die Berechnung des Variationskoeffizienten müssen beide Merkmale in der gleichen Einheit vorliegen. ☐
- d) Zur Berechnung des Variationskoeffizienten benötigt man den Median. ☐

5 Konzentrationsmaße

Bis jetzt können wir bei einem Datensatz Aussagen über die (zentrale) Lage der Daten und das Ausmaß der Streuung treffen. Im Folgenden werden auch Aussagen über die *Konzentration* der Daten von Interesse sein, sowie deren graphische Darstellung. Somit kann bspw. ausgesagt werden, ob eine eher gleiche ($\hat{=}$ faire) Verteilung oder möglicherweise ein Monopol vorliegt.

Da man die Daten ins Verhältnis zueinander setzt, sind diese Maßzahlen nur noch Merkmale mit metrischem Skalenniveau möglich. Konzentrationsmaße werden im weiteren Verlauf in absolute und relative Konzentrationsmaße geteilt.

5.1 Absolute Konzentrationsmaße

5.1.1 Konzentrationsrate

Die Konzentrationsrate ist ein eher simples Maß, mit dem man Aussagen à la "*Die drei größten Marktteilnehmer machen 60% des Umsatzes.*" treffen kann. Hierfür addiert man einfach die (Markt-)Anteile der g Merkmalsträger mit den größten Anteilen zusammen, wobei g je nach Kontext/Interesse vorab gewählt werden muss:

$$CR_g = \sum_{i=n-g+1}^n p_i = \sum_{i=n-g+1}^n \left(\frac{x_i}{\sum_{i=1}^n x_i} \right)$$

Würde ich beispielsweise bei 10 Merkmalsträgern die Konzentrationsrate der $g = 2$ größten Merkmalsträger berechnen wollen, dann würde ich die Summe der Anteile p_i von $i = 10 - 2 + 1 = 9$ bis $n = 10$ berechnen. Somit also die Summe der Anteile des neunten und zehnten Merkmalsträgers. Der Wertebereich der Konzentrationsrate kann von $\frac{g}{n}$, bei Gleichverteilung, wenn die Anzahl der g größten Merkmalsträger auch dem Anteil an den gesamten Merkmalsträgern entspricht, bis hin zur 1 gehen. Bei einer Konzentrationsrate von 1 liegt ein Monopol (bei $g = 1$) oder ein Oligopol (bei $g \geq 2$) vor.

Einschub: Praktische Relevanz // Wahl von g

Das Gesetz gegen Wettbewerbsbeschränkung (§18 GWB) legt fest, wann man bei Unternehmen von marktbeherrschend sprechen kann. Um bei *einem* Unternehmen von marktbeherrschend sprechen kann, benötigt dieses ein Marktanteil von mind. 40%. Eine Gesamtheit von *drei oder weniger* Unternehmen muss ein Marktanteil von mindestens 50% erreichen und eine Gesamtheit von *fünf oder weniger* Unternehmen muss einen Marktanteil von mindestens zwei Drittel erreichen, damit von Marktbeherrschung gesprochen werden kann.

5.1.2 Konzentrationskurve

Die Konzentrationsrate wird durch die Konzentrationskurve graphisch dargestellt. Bei der Konzentrationskurve werden zuerst die Merkmalsträger absteigend nach ihrer Größe abgetragen. Auf der

x-Achse ist somit die kumulierte Anzahl der Merkmalsträger mit den größten Ausprägungen abgetragen (also $1, 2, \dots, n$) und auf der y-Achse die kumulierten relativen Marktanteile. Je flacher der Graph ist, desto ähnlicher sind die Anteile verteilt. Entspricht der Graph einer Geraden, so liegt eine Gleichverteilung vor.

5.1.3 Herfindahl-Index

Ein weiteres absolutes Konzentrationsmaß ist der Herfindahl-Index. Dieser bezieht sich nicht wie die Konzentrationsrate (vgl. Kap. 5.1.1) nur auf die g größten Merkmalsträger, sondern liefert somit eine allgemeinere Aussage über alle Merkmalsträger. Den Herfindahl-Index berechnet man als Quotienten aus der Summe der quadrierten Beobachtungen $\sum_{i=1}^n x_i^2$ und der quadrierten Summe aller Beobachtungen $(\sum_{i=1}^n x_i)^2$. Da im Zähler zuerst quadriert und anschließend aufsummiert wird, im Nenner hingegen zuerst aufsummiert und dann quadriert wird, ist klar, dass der Zähler stets kleiner oder gleich dem Nenner sein wird. Somit ergibt sich folgender Wertebereich: Liegt ein Monopol vor (ein Merkmalsträger besitzt die gesamte Merkmalssumme), so sind Zähler & Nenner identisch, was zu einem Wert von $H = 1$ führt. Bei einer Gleichverteilung (alle Merkmalsträger besitzen die gleiche Merkmalsausprägung a), erhalten wir den Wert $H = \frac{1}{n}$.

Einschub I: Beweis für die untere Grenze:

$$H = \frac{\sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i)^2} = \frac{n \cdot a^2}{(n \cdot a)^2} = \frac{n \cdot a^2}{n^2 \cdot a^2} = \frac{1}{n}$$

Einschub II: Aussagen über Veränderungen

Beim Herfindahl-Index kann man relativ einfach pauschale Aussagen über Veränderungen treffen, solange die Merkmalssumme gleich bleibt.

- *Beispiel 1: Fusion zweier Marktteilnehmer*

In diesem Fall würde sich der Nenner nicht ändern, da sich die gesamte Merkmalssumme nicht ändert. Der Zähler würde jedoch größer werden, da $(a+b)^2 > a^2 + b^2$. Wichtig ist aber hierbei auch, die damit einhergehende Veränderung des Wertebereichs zu beachten.

- *Beispiel 2: Transfer von einem großen Merkmalsträger zu einem kleineren*

Wird ein Teil der Merkmalssumme von einem größeren zu einem kleineren Merkmalsträger transferiert (und bleibt der kleinere dadurch weiterhin kleiner), so wird der Herfindahl-Index ebenfalls sinken. Der Wertebereich ändert sich dabei nicht. Umgekehrt (Transfer von klein zu groß) gilt dieselbe Intuition.

R-Befehl für den Herfindahl-Index: `> Herfindahl(data)` [Dokumentation](#)

⚠ Die R-Funktion ist **nicht** Teil von **base-R** sondern Teil des Paketes **DescTools**. Dieses muss zunächst installiert (`install.packages("DescTools")`) und anschließend importiert werden (`library(DescTools)`).

5.2 Relative Konzentrationsmaße

5.2.1 Lorenzkurve

Die Lorenzkurve ist eine grafische Methode um Konzentration auf eine *relative* Art und Weise darzustellen. Sie ist bereits auf den ersten Blick von der Konzentrationskurve zu unterscheiden, da sowohl auf der x- als auch auf der y-Achse ausschließlich *relative* Werte abgetragen sind. Somit gehen beide Achsen von 0 bis 1. Anhand der Lorenzkurve kann man Aussagen à la "*Die ärmsten X% der Merkmalsträger besitzen einen Anteil von Y% der Merkmalssumme.*" treffen. Der Begriff "arm" und "besitzen" soll hierbei aber noch signalisieren, dass dieses Konzept lediglich auf Einkommen, o.ä. anwendbar ist, sondern dient hier einfach als plastische Beispielformulierung. Man kann diese Aussage auch invertieren und zu folgendem Schluss kommen: "*Die reichsten (100 - X) % der Merkmalsträger besitzen einen Anteil von (100 - Y) % der Merkmalssumme.*"

All diese beispielhaften Formulierungen zeigen, dass es essentiell ist die Daten geordnet vorliegen zu haben, bevor man die Punkte zur Erstellung der Lorenzkurve berechnen kann. Bei der Berechnung werden im Folgenden zwei Fälle unterschieden: Individualdaten (der "normale" Fall) und gruppierte Daten (der etwas "kompliziertere" Fall).

Berechnung Lorenzkurve bei Individualdaten Da jeder Merkmalsträger jeweils einen Anteil von $\frac{1}{n}$ an der Gesamtheit der Merkmalsträger ausmacht, teilt man die x-Achse in n gleichgroße Abschnitte. Liegen z.B. 5 Beobachtungen vor, so wird die x-Achse durch Markierungen bei $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{5}$ & $\frac{4}{5}$ unterteilt. Der zugehörige Wert der Ordinate gibt für jeden dieser x-Werte den kumulierten Anteil der Merkmalssumme dieses Anteils der Merkmalsträger an. Nehmen wir also an, unsere 5 Beobachtungen hätten folgende Ausprägungen:

$$x_1 = 4; \quad x_2 = 1; \quad x_3 = 7; \quad x_4 = 5; \quad x_5 = 3$$

Die geordneten Daten sähen in diesem Fall so aus:

$$x_{(1)} = 1; \quad x_{(2)} = 3; \quad x_{(3)} = 4; \quad x_{(4)} = 5; \quad x_{(5)} = 7$$

Da wir eine gesamte Merkmalssumme von 20 hätten, hätte die erste Beobachtung daran einen kumulierten Anteil von $\frac{1}{20}$, die ersten beiden hätten einen kumulierten Anteil von $\frac{1+3}{20} = \frac{4}{20}$, die ersten drei hätten einen kumulierten Anteil von $\frac{1+3+4}{20} = \frac{8}{20}$ und die ersten vier hätten einen kumulierten Anteil von $\frac{1+3+4+5}{20} = \frac{13}{20}$.

Somit ergeben sich folgende Punkte für die Lorenzkurve.

$$\left(\frac{1}{5} \middle| \frac{1}{20}\right); \quad \left(\frac{2}{5} \middle| \frac{4}{20}\right); \quad \left(\frac{3}{5} \middle| \frac{8}{20}\right); \quad \left(\frac{4}{5} \middle| \frac{13}{20}\right)$$

Außerdem gehören zu **jeder** Lorenzkurve die Punkte (0|0) und (1|1), da logischerweise 0% der Merkmalsträger auch 0% der Merkmalsumme besitzen und 100% der Merkmalsträger auch 100% der Merkmalsumme besitzen.

⚠ *Notation:* Die Werte auf der x-Achse werden oft als u_i und die Werte auf der y-Achse als v_i bezeichnet. Dies ist für die Formeln in Kapitel 5.2.2 wichtig.

Für die Interpretation ist es wichtig sich klar zu machen, wie eine perfekt Gleichverteilung der Merkmalssumme sich in der Lorenzkurve widerspiegeln würde. In unserem Beispiel wäre in diesem Fall jedes $x_i = 4$ (gesamte Merkmalssumme von 20 aufgeteilt auf 5 Merkmalsträger) und die Punkte für die Lorenzkurve wären:

$$(0|0); \quad \left(\frac{1}{5}|\frac{1}{5}\right); \quad \left(\frac{2}{5}|\frac{2}{5}\right); \quad \left(\frac{3}{5}|\frac{3}{5}\right); \quad \left(\frac{4}{5}|\frac{4}{5}\right); \quad (1|1)$$

In diesem Fall würde die Lorenzkurve perfekt mit der Winkelhalbierenden übereinstimmen, was auch der Grund ist, warum die Winkelhalbierende oft mit in die Grafik eingezeichnet ist. Sie dient quasi als Referenz, wie es im Fall der Gleichverteilung aussähe um abschätzen zu können wie stark die Lorenzkurve davon abweicht.

Das andere Extrem, ein Monopol, läge vor falls die gesamte Merkmalssumme einem Merkmalsträger zugeordnet würde. In unserem Beispiel wären dann $x_{(1)} = \dots = x_{(4)} = 0$ und $x_{(5)} = 20$, was zu folgenden Punkten für die Lorenzkurve führen würde:

$$(0|0); \quad \left(\frac{1}{5}|0\right); \quad \left(\frac{2}{5}|0\right); \quad \left(\frac{3}{5}|0\right); \quad \left(\frac{4}{5}|0\right); \quad (1|1)$$

In Abbildung 1 sind die Lorenzkurven für die drei Beispielszenarien dargestellt.

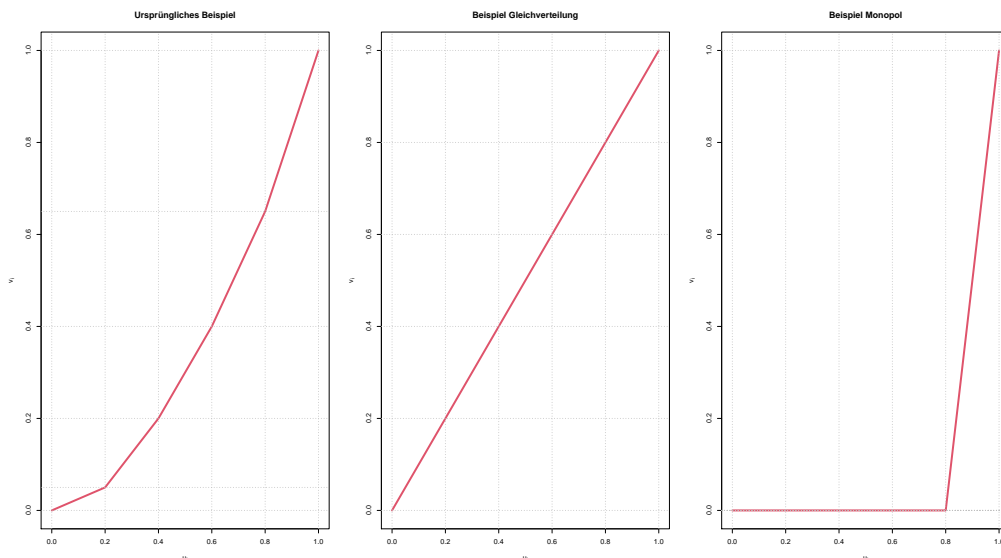


Figure 1: Lorenzkurven zum Beispiel für Individualdaten

Zwischen diesen beiden Extrema gibt es sehr viele Abstufungen von (Un)Gleichverteilung. Generell lässt sich festhalten: Je weiter der Graph der Lorenzkurve (nach links unten) von der Winkelhalbierenden entfernt ist, also je größer die Fläche dazwischen ist, desto ungleicher verteilt, also desto konzentrierter ist ein Merkmal. Vice versa, je näher an der Winkelhalbierenden, desto gleichmäßiger verteilt, also desto weniger konzentriert ist ein Merkmal. Diese angesprochene Fläche spielt

auch beim Gini-Koeffizienten (vgl. Kap. 5.2.2) eine entscheidende Rolle.

Weitere Eigenschaften der Lorenzkurve sind, dass sie **immer** unterhalb der Winkelhalbierenden und niemals darüber verlaufen muss. Da die Merkmalsausprägungen kumuliert (also aufsummiert) werden, kann der Graph nur monoton steigend sein. Zudem muss die Steigung eines Kurvensegments immer größer oder gleich dem vorigen Segment sein, da die Merkmalsausprägungen bei der Lorenzkurve nach Größe geordnet wurden.

Berechnung Lorenzkurve bei gruppierten Daten Hat man obige Erklärungen für Individualdaten verstanden, so wird auch das Verständnis für das Vorgehen bei gruppierten Daten nicht schwer fallen. Der erste wichtige, und visuell auffälligste, Unterschied besteht darin, dass die Abstände auf der x-Achse nicht mehr identisch ist. Ansonsten sind die Berechnungen weitestgehend ähnlich zum dem Fall für Individualdaten.

Nehmen wir an wir hätten im obigen Beispiel nun nicht mehr 5 Beobachtungen sondern 100. Dabei haben 10 Beobachtungen eine Merkmalsausprägung von 1, 40 haben eine Merkmalsausprägung von 5, 20 eine Merkmalsausprägung von 7 und 30 eine Merkmalsausprägung von 15. Insgesamt entspricht dies einer Merkmalssumme von 800. Die Gruppe mit der geringsten Merkmalsausprägung hätte damit einen Anteil von $\frac{10}{100} = 10\%$ an den Merkmalsträgern und einen Anteil von $\frac{10 \cdot 1}{800} = 0,0125$ an der Merkmalssumme, die beiden Gruppen mit den geringsten Merkmalsausprägungen einen Anteil von $\frac{10+40}{100} = 50\%$ an den Merkmalsträgern und einen Anteil von $\frac{10 \cdot 1 + 40 \cdot 5}{800} = 0,2625$ an der Merkmalssumme, usw.

Dies führt zu folgenden Punkten für die Lorenzkurve:

$$(0|0); \quad (0,1|0,0125); \quad (0,5|0,2625); \quad (0,7|0,4375); \quad (1|1)$$

Abbildung 2 zeigt die Lorenzkurve für dieses Beispielszenario.

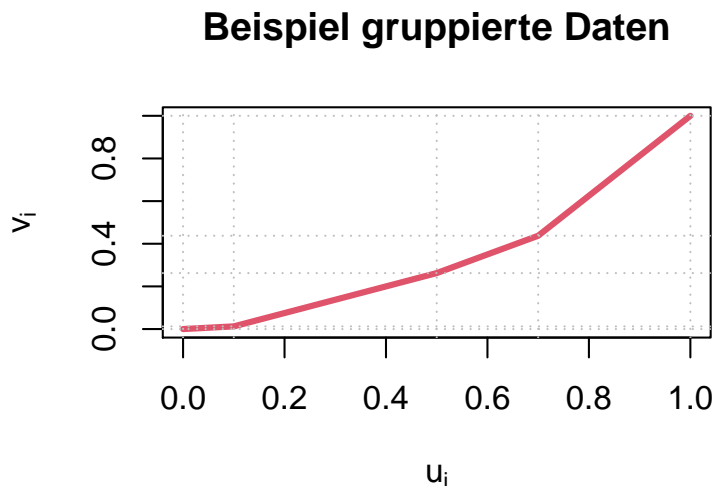


Figure 2: Lorenzkurve zum Beispiel für gruppierte Daten

⚠ *Notation:* Um Formeln für gruppierte von Formeln für Individualdaten abzuheben wird eine Tilde verwendet, d.h. x-Werte als \tilde{u}_i und y-Werte als \tilde{v}_i bezeichnet.

⚠ *Anmerkung I:* Die gleiche Lorenzkurve wie in Abbildung 2 hätte man auch für die Individualdaten zeichnen können, jedoch wäre dies ein um einiges höherer Aufwand gewesen. In diesem Fall wäre es jedoch möglich gewesen, da wir tatsächlich für jedes Individuum dessen genaue Merkmalsausprägung kennen.

⚠ *Anmerkung II:* Kennen wir **nicht** für jedes Individuum dessen genaue Merkmalsausprägung, sondern lediglich einen Gruppenmittelwert, so wird jedem Individuum in einer Gruppe dieser Gruppenmittelwert als Merkmalsausprägung zugeordnet. Dadurch kann man ganz normal, wie oben für gruppierte Daten gezeigt, vorgehen. Wichtig ist dabei jedoch im Hinterkopf zu behalten, dass damit implizit die Annahme einhergeht, dass innerhalb der Gruppen Gleichverteilung herrscht, da wir jedem Individuum einer Gruppe denselben Wert zuordnen. Diese Annahme muss nicht immer realistisch sein und sollte stets kritisch hinterfragt werden.

R-Befehl für die Lorenzkurve: <code>> Lc(data)</code>
--

Dokumentation

⚠ Die R-Funktion ist **nicht** Teil von **base-R** sondern Teil des Paketes **ineq**. Um diese Funktion verwenden zu können, muss das Paket zunächst installiert werden (`install.packages("ineq")`) und anschließend importiert werden (`library(ineq)`).

5.2.2 Gini-Koeffizient

Die im vorigen Kapitel erwähnte Fläche F , die zwischen dem Graph und der Winkelhalbierenden liegt, stellt die Basis für den Gini-Koeffizienten dar, welcher ein Maß für die relative Konzentration ist. Der Gini ist definiert als "*Zweimal die Fläche zwischen Winkelhalbierender und Lorenzkurve.*"

Kennt man bereits die kumulierten Anteile an der Merkmalssumme (y-Werte der Punkte auf der Lorenzkurve), so ist der Gini recht einfach zu berechnen:

- Zunächst addiert man jeweils zu jedem Anteilswert den Anteilswert der vorherigen Punktes (angefangen bei 0 bis zur 1): $(v_{i-1} + v_i)$
- Diese Summen werden anschließend addiert: $\sum_{i=1}^n (v_{i-1} + v_i)$
- und mit $\frac{1}{n}$ multipliziert: $\frac{1}{n} \sum_{i=1}^n (v_{i-1} + v_i)$
- Dieses Produkt wird zum Schluss von 1 abgezogen.

$$G = 1 - \frac{1}{n} \sum_{i=1}^n n_j \cdot (v_{i-1} + v_i)$$

Für gruppierte Daten ändern sich Vorgehen und Formel nicht dramatisch. Der einzige Unterschied zur obigen Formel besteht darin, dass man mit den Gruppengrößen n_j gewichten muss:

$$G = 1 - \frac{1}{n} \sum_{i=1}^n (\tilde{v}_{i-1} + \tilde{v}_i)$$

Man kann sich das ein bisschen wie bei gewichteten arithmetischen Mittel in Kapitel 3.5.1 vorstellen, da auch dort Unterschiedlichen Gruppengrößen auf ähnliche Art & Weise Rechnung getragen wird.

Der Wertebereich des Gini-Koeffizient beginnt bei 0, was für eine absolute Gleichverteilung spricht. Dies macht intuitiv Sinn, da bei der absoluten Gleichverteilung die Fläche zwischen Winkelhalbierender und Lorenzkurve nicht existiert. Dass die obere Grenze nicht bei Eins, sondern bei $(\frac{n-1}{n})$, ist auf den ersten Blick vielleicht etwas weniger intuitiv. Bei einem Blick auf das Monopol-Szenario in Abbildung 1 sollte jedoch klar werden, dass die Fläche zwischen Winkelhalbierender und Lorenzkurve nicht den Wert 0,5 erreichen kann & somit der Gini (also das Doppelte dieser Fläche) nicht 1 werden kann. Grund dafür ist das Dreieck, welches unten rechts stets per Konstruktion ausgespart wird. Die Größe dieses ausgesparten Dreieckes hängt von der Anzahl der Merkmalsträger n ab und damit auch der Wertebereich.

Da mit einem variierendem Wertebereich (je nach Anzahl der Merkmalsträger) schwierig Konzentrationen für verschiedene Merkmale verglichen werden können, berechnet man den **normierten Gini-Koeffizienten** G^+ .

Der normierte Gini-Koeffizient erhält man, indem man den berechneten Gini-Koeffizient mit $\frac{n}{n-1}$ multipliziert:

$$G^+ = \frac{n}{n-1}G$$

Der Wertebereich von G^+ geht dann noch von 0 bis 1 und ist unabhängig von n . Die Konzentration 0 steht dabei für absolute Gleichverteilung, also dafür, dass es *keine* Konzentration gibt, während 1 für eine *vollständige* Konzentration, also für ein Monopol, steht. Somit ist es nun kein Problem mehr verschiedene Merkmale mit unterschiedlichem n in Bezug auf ihre Konzentration zu vergleichen.

⚠ Bei der Normierung des Gini für gruppierte Daten entspricht n weiterhin der Anzahl der Beobachtungen und **nicht** der Anzahl der Gruppen.

Einschub: Aussagen über Veränderungen

Beim Gini kann man relativ einfach pauschale Aussagen über Veränderungen treffen, solange die Anzahl der Merkmalsträger gleich bleibt.

- *Beispiel 1: Alle Merkmalsträger erfahren dieselbe relative Steigerung*

Alle Merkmalsträger steigern ihre Merkmalssumme um 10%. In diesem Fall würde sich Gini nicht verändern, da sich an den Relationen nichts ändert hat.

- *Beispiel 2: Alle Merkmalsträger erfahren dieselbe absolute Steigerung*

Alle Merkmalsträger steigern ihre Merkmalssumme um 10 Einheiten. In diesem Fall würde Gini nicht sinken, da in Relationen zueinander nun alle etwas gleichere Anteile besitzen. Man kann sich das gut an einem Extremfall veranschaulichen: Angenommen jeder Merkmalsträger würde seine Merkmalssumme um das 100-fache der bisher größten Ausprägung steigern. Dadurch wurden alle bisher dagewesenen Unterschiede quasi irrelevant werden und jeder hätte nahezu gleich viel.

R-Befehl für den Gini: <code>> Gini(data)</code>

Dokumentation

⚠ Die R-Funktion ist **nicht** Teil von **base-R** sondern Teil des Paketes **ineq**. Um diese Funktion verwenden zu können, muss das Paket zunächst installiert werden (`install.packages("ineq")`) und anschließend importiert werden (`library(ineq)`).

5.3 Aufgaben

1. Welche Aussagen bzgl. Gini & Lorenzkurve sind wahr?

- a) Die absolute Merkmalssumme ist unerheblich für den Gini. ☐
- b) Höherer Gini bedeutet (global) steilere Lorenzkurve. ☐
- c) Der Gini ist uneingeschränkt geeignet um die Konzentration in zwei Gruppen zu vergleichen. ☐
- d) Erhalten alle Merkmalsträger dieselbe prozentuale Steigerung ihres (absoluten) Teils der Merkmalssumme, so verändert sich der Gini nicht. ☐

2. Welche Aussagen bzgl. des Herfindahl-Index sind wahr?

- a) Der Herfindahl-Index ist uneingeschränkt geeignet um die Konzentration in zwei Gruppen zu vergleichen. ☐
- b) Falls sich die Merkmalssumme ändert, können definitive Aussagen über der Änderung des Herfindahl-Index getroffen werden. ☐
- c) Falls sich die Verteilung Merkmalssumme ändert, können definitive Aussagen über der Änderung des Herfindahl-Index getroffen werden. ☐
- d) Höherer Herfindahl-Index bedeutet ungleichere Verteilung. ☐

3. Der Gini für gruppierte Daten ist nur identisch zum "normalen" Gini, falls ..

- a) .. alle Gruppen gleich groß sind. ☐
- b) .. absolute Gleichverteilung herrscht. ☐
- c) .. Gleichverteilung innerhalb der Gruppen herrscht. ☐
- d) .. die Anzahl der Gruppen kleiner als 10 ist. ☐

6 Zusammenhangsmaße

Bis jetzt haben wir stets lediglich eine Variable (bzw. Merkmal) X und dessen Ausprägungen x_1, x_2, \dots, x_i betrachtet. Auf diese Art und Weise war es uns möglich, Aussagen über dessen Lage, Streuung und Konzentration zu treffen.

Da in diesem Kapitel Aussagen über Zusammenhänge getroffen werden sollen bzw. Zusammenhänge quantifiziert werden sollen, werden nun stets **zwei** Variablen/Merkmale X & Y gleichzeitig betrachtet. Dies führt dazu, dass die beobachteten Ausprägungen der Merkmale nun folgendermaßen vorliegen: $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$

Beispiel: Wir betrachten nun bei Personen gleichzeitig ihre Körper- (X) und ihre Schuhgröße (Y). Dies könnte bspw. zu folgender Stichprobe führen: $(185\text{cm}; 44), (170\text{cm}; 39), \dots, (195\text{cm}; 47)$

Von zentraler Bedeutung für die Quantifizierung von Zusammenhängen ist es, sich im Vorhinein darüber klar zu werden, auf welchen Skalenniveaus die beobachteten Merkmale gemessen werden. Dies hat erhebliche Auswirkungen darauf, welche Zusammenhangsmaße berechnet werden können. Als Richtlinie gilt hier stets: Es kann lediglich ein Zusammenhangsmaß berechnet werden, welches für die Skala desjenigen Merkmals geeignet ist, welches auf der "niedrigeren" Skala gemessen wird.

Beispiel: Betrachten wir bei Personen ihr Geschlecht (X) und ihr Einkommen (Y), so sind nur Zusammenhangsmaße anwendbar, die für nominale Merkmale geeignet sind, da das Merkmal Geschlecht lediglich nominal skaliert ist. Berechnet man stattdessen Betriebszugehörigkeit (X) und Einkommen (Y), so sind Zusammenhangsmaße für metrische Merkmale anwendbar.

6.1 Die Kontingenztafel

Eines der grundlegendsten Instrumente zur gemeinsamen Darstellung zweier Merkmale ist die Kontingenztafel. Hierbei wird pro unterschiedliche Merkmalsausprägung von X eine Zeile, für jede untersch. Merkmalsausprägung von Y eine Spalte belegt. Dies führt bei der Betrachtung der Merkmale X : "Parteilichkeit im Bundestag" (Ausprägungen: *Union, SPD, Grüne, FDP, Linke, AfD*) und Y : Geschlecht (Ausprägungen: *weiblich, männlich, divers*) zu einer Kontingenztafel mit 6 Zeilen und 3 Spalten, d.h. einer (6×3) -Kontingenztafel.

Die Entscheidung, welche Variable in den Zeilen und welche Variable in den Spalten abgetragen wird, kann mehr oder minder frei getroffen werden. Oftmals wird die Auswahl jedoch nach zwei Kriterien getroffen:

- Kann ein betrachtetes Merkmal als "Schichtungsmerkmal" betrachtet werden, so steht es für gewöhnlich in den Zeilen. Ein Schichtungsmerkmal ist eine Variable, die die Stichprobe in verschiedene Gruppen einteilt, wie z.B. "Geschlecht", "Bildungsabschluss" oder "Raucher/Nicht-Raucher".

- Vermutet man in den Daten eine kausale Wirkungsstruktur (d.h. ein Merkmal hat einen Einfluss auf das Andere), so steht meist dasjenige Merkmal in den Zeilen, welches als Ursache angesehen wird, z.B. bei der Betrachtung von "Raucher/Nicht-Raucher" und "Auftreten von Lungenkrebs ja/nein" würde man das Merkmal "Raucher/Nicht-Raucher" in den Zeilen abtragen.

Da in der Kontingenztafel für jede mögliche Ausprägung eines Merkmals eine zusätzliche Zeile bzw. Spalte in der Darstellung benötigt wird, ist sie lediglich für Merkmale mit nicht allzu vielen verschiedenen Ausprägungen sinnvoll. Für metrische Merkmale bedeutet dies, dass diese nur sinnvoll in Kontingenztafeln dargestellt werden können, falls sie vorher in Klassen eingeteilt wurden (bspw. Altersgruppen oder Einkommensklassen).

Die inneren Zellen der Kontingenztafeln werden in der Regel mit den absoluten Häufigkeiten n_{ij} ("Wie oft beobachte ich die Ausprägung in der i-ten Zeile von Merkmal X gemeinsam mit der Ausprägung in der j-ten Spalte vom Merkmal Y ?") oder den relativen Häufigkeiten $f_{ij} = \frac{n_{ij}}{n}$ befüllt. Die (absoluten) Randhäufigkeiten werden mit $n_{i\bullet}$ ("Wie oft beobachte ich die Ausprägung in der i-ten Zeile von Merkmal X ?") und $n_{\bullet j}$ ("Wie oft beobachte ich die Ausprägung in der j-ten Spalte von Merkmal Y ?") bezeichnet.

Teilt man alle Häufigkeiten der inneren Zellen in der i-ten Zeile durch die Randhäufigkeit der i-ten Zeile, so erhält man die bedingten relativen Häufigkeiten $f_{j|i} = \frac{n_{ij}}{n_{i\bullet}}$. Analog lassen sich auch die bedingten relativen Häufigkeiten $f_{i|j} = \frac{n_{ij}}{n_{\bullet j}}$ für die j-te Spalte bestimmen.

Beispiel: *Lungenkrebs bei weiblichen Rauchern*¹

X/Y	Lungenkrebs	kein Lungenkrebs	Σ
Nicht-Raucher	458	1598	2056
Raucher	877	412	1289
Σ	1335	2010	3345

Table 3: *Absolute Häufigkeiten (Lungenkrebsinzidenz bei weiblichen Rauchern)*

X/Y	Lungenkrebs	kein Lungenkrebs	Σ
Nicht-Raucher	$\frac{458}{3345}$	$\frac{1598}{3345}$	$\frac{2056}{3345}$
Raucher	$\frac{877}{3345}$	$\frac{412}{3345}$	$\frac{1289}{3345}$
Σ	$\frac{1335}{3345}$	$\frac{2010}{3345}$	1

Table 4: *Relative Häufigkeiten (Lungenkrebsinzidenz bei weiblichen Rauchern)*

R-Befehl für Kontingenztafeln: `> table(data$x, data$y)`

[Dokumentation](#)

¹Quelle: <https://www.aerzteblatt.de/archiv/8369/Lungenkrebsrisiko-hoeher-als-bisher-angenommen-Europaweite-Studie-vorgestellt>

X/Y	Lungenkrebs	kein Lungenkrebs	Σ
Nicht-Raucher	$\frac{458}{2056}$	$\frac{1598}{2056}$	1
Raucher	$\frac{877}{1289}$	$\frac{412}{1289}$	1

Table 5: *Bedingt auf Raucher/Nicht-Raucher (Lungenkrebsinzidenz bei weiblichen Rauchern)*

X/Y	Lungenkrebs	kein Lungenkrebs	
Nicht-Raucher	$\frac{458}{1335}$	$\frac{1598}{2010}$	
Raucher	$\frac{877}{1335}$	$\frac{412}{2010}$	
Σ	1	1	

Table 6: *Bedingt auf Lungenkrebs-Status (Lungenkrebsinzidenz bei weiblichen Rauchern)*

R-Befehl für (bedingte) Kontingenztafeln: <code>> prop.table(data)</code>	Dokumentation
--	-------------------------------

6.2 Unabhängigkeit

Ein zentrales Konzept bei der Betrachtung von Zusammenhängen ist die *Unabhängigkeit*.

Betrachtet man in Tabelle 3 die Randhäufigkeiten $n_{\bullet j}$ der Spalten, so erkennt man das Verhältnis von Lungenkrebs zu Nicht-Lungenkrebs, unabhängig davon ob jemand geraucht hat oder nicht. Dieses Verhältnis liegt bei etwas mehr als 2 zu 3. Betrachtet man nun lediglich die Häufigkeiten in der ersten Zeile (d.h. nur die Nicht-Raucher) n_{11} und n_{12} , so erkennt man, dass unter diesen Beobachtungen das Verhältnis von Lungenkrebs zu Nicht-Lungenkrebs bei ca. 1 zu 4 liegt (bei den Nicht-Lungenkrebs-Fällen liegt es bei etwas mehr als 2 zu 1). Somit scheint es einen Zusammenhang zwischen Rauchen und dem Auftreten von Lungenkrebs zu geben, da man unter Unabhängigkeit erwarten würde, in beiden Gruppen (Nicht-Raucher und Raucher) dasselbe Verhältnis, nämlich das der Randhäufigkeiten zu beobachten. Berechenbar sind die *unter Unabhängigkeit erwarteten absoluten Häufigkeiten* über die Formel $\hat{n}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$.

Fortsetzung Beispiel: *Lungenkrebs bei weiblichen Rauchern*

X/Y	Lungenkrebs	kein Lungenkrebs	Σ
Nicht-Raucher	$\frac{2056 \cdot 1335}{3345} = 820,5561$	$\frac{2056 \cdot 2010}{3345} = 1235,4439$	2056
Raucher	$\frac{1289 \cdot 1335}{3345} = 514,4439$	$\frac{1289 \cdot 2010}{3345} = 774,5561$	1289
Σ	1335	2010	3345

Table 7: *Unter Unabhängigkeit erwartete absolute Häufigkeiten*

Betrachtet man in dieser Häufigkeitstabelle nun die Verhältnisse von Lungenkrebs zu Nicht-Lungenkrebs bei den Rauchern bzw. bei den Nicht-Rauchern, so erkennt man, dass das Verhältnis in beiden Gruppen jeweils bei ca. 2 zu 3 liegt und somit dem Verhältnis der Randhäufigkeiten der Spalten entspricht.

6.3 Zusammenhangsmaße für nominale Merkmale

Dadurch, dass einfach berechenbar ist, wie eine Häufigkeitstabelle unter Unabhängigkeit der beiden Merkmale aussieht (vgl. Tabelle 7), kann man nun messen wie stark die tatsächlich beobachtete Häufigkeitstabelle davon abweicht. Um diese beobachteten Abweichungen in eine einzige Maßzahl zusammenzufassen, geht man wie folgt vor:

- Man bildet für jede Zelle die Differenz aus beobachteter absoluter Häufigkeit und unter Unabhängigkeit erwarteter absoluter Häufigkeit: $n_{ij} - \hat{n}_{ij}$
- Man quadriert diese Differenzen, damit sich negative und positive Abweichungen später beim Aufsummieren nicht gegenseitig aufheben: $(n_{ij} - \hat{n}_{ij})^2$
- Man teilt diese quadrierten Differenzen durch die \hat{n}_{ij} , d.h. man setzt sie ins Verhältnis zu dem Wert, den man eigentlich erwartet hätte: $\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$
- Man summiert diese Werte der einzelnen Zellen auf: $\sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$

Die Maßzahl, welche durch diese Vorgehensweise berechnet wird, nennt sich χ^2 -Koeffizient und nimmt bei Unabhängigkeit (d.h. unter Unabhängigkeit erwartete sind gleich den beobachteten absoluten Häufigkeiten) den Wert 0 an, bei Abhängigkeit hingegen sehr hohe Werte. Diese Maßzahl besitzt jedoch noch einige gravierende Nachteile:

- Je mehr Beobachtungen wir in unserer Stichprobe haben, desto leichter sind hohe Abweichungen zu beobachten, d.h. desto größer wird *tendenziell* χ^2
- Je mehr Zellen die Kontingenztafel besitzt, desto mehr Abweichungen sind zu beobachten, d.h. desto größer wird *tendenziell* χ^2

Um diesen Nachteilen entgegenzutreten existieren verschiedene Erweiterungen, welche eine (oder beide) dieser Einschränkungen beheben.

R-Befehl für χ^2 : `> chisq.test(data$x, data$y)`

[Dokumentation](#)

⚠ Diese R-Funktion führt neben der Berechnung von χ^2 auch den zugehörigen χ^2 -Unabhängigkeitstest durch. Dies ist erst Teil des Stoffs von Statistik II (vgl. Kap. 17.2.13) und kann vorläufig ignoriert werden. Der Wert `X-squared` in der Ausgabe entspricht dem χ^2 -Koeffizienten.

Loslösung vom Stichprobenumfang Der Φ -Koeffizient und der Kontingenzkoeffizient C sind zwei Möglichkeiten um den Wertebereich von χ^2 vom Stichprobenumfang zu lösen. Sie berechnen sich auf folgende, leicht unterschiedliche Weisen


$$\Phi = \sqrt{\frac{\chi^2}{n}}, \quad 0 \leq \Phi \leq \sqrt{\min(k, l) - 1}$$
$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \quad 0 \leq C \leq \sqrt{\frac{\min(k, l) - 1}{\min(k, l)}}$$

 **Wichtig:** Die Zeilen- und Spaltenzahl werden in diesem Kontext mit k & l bezeichnet.

Eine Berechnung diese beiden Maßzahlen macht Sinn, wenn wir die Zusammenhänge aus zwei Kontingenztafeln vergleichen wollen, bei denen zwar der Stichprobenumfang, nicht aber jedoch die Dimensionen der Kontingenztafel unterschiedlich sind. Beispielsweise könnte man den berechneten Zusammenhang aus Tabelle 3 mit einer Gruppe von männlichen Rauchern vergleichen. Hier wären dann die Dimensionen der beiden Tafeln identisch, die Stichprobenumfänge könnten sich jedoch unterscheiden.

R-Befehl für den Φ -Koeffizient: <code>> Phi(data\$x, data\$y)</code>	Dokumentation
---	-------------------------------

R-Befehl für den C : <code>> ContCoef(data\$x, data\$y)</code>	Dokumentation
---	-------------------------------

 Die R-Funktionen sind **nicht** Teil von base-R sondern Teil des Paketes DescTools. Dieses muss zunächst installiert (`install.packages("DescTools")`) und anschließend importiert werden (`library(DescTools)`).


Loslösung von der Dimension der Kontingenztafel Möchte man diese beiden Maße nun auch noch von der Dimension der Kontingenztafel loslösen, so kann man dies durch Berechnung von *Cramers V* oder des korrigierten Kontingenzkoeffizienten C_{korrr} tun:

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(k, l) - 1)}}, \quad 0 \leq V \leq 1$$
$$C_{korrr} = \sqrt{\frac{\min(k, l)}{\min(k, l) - 1} \cdot \underbrace{\sqrt{\frac{\chi^2}{\chi^2 + n}}}_C}, \quad 0 \leq C_{korrr} \leq 1$$

Eine Berechnung diese beiden Maßzahlen macht Sinn, wenn wir die Zusammenhänge aus zwei Kontingenztafeln vergleichen wollen, bei denen sowohl der Stichprobenumfang, als auch die Dimensionen der Kontingenztafel unterschiedlich sind. Beispielsweise könnte man den berechneten Zusammenhang aus Tabelle 3 mit einer Gruppe von männlichen Rauchern vergleichen, bei denen nicht nur die Lungenkrebsinzidenz sondern auch das Auftreten anderer Krebsarten betrachtet wird. Hier wären dann auch die Dimensionen der beiden Tafeln unterschiedliche, da die Kontingenztafel der Männer mehr Spalten besäße als die der Frauen.

R-Befehl für den Cramers V: <code>> CramerV(data\$x, data\$y)</code>	Dokumentation
---	-------------------------------

R-Befehl für C_{korrr} : <code>> ContCoef(.., correct = TRUE)</code>	Dokumentation
---	-------------------------------

 Die R-Funktionen sind **nicht** Teil von base-R sondern Teil des Paketes DescTools. Dieses muss zunächst installiert (`install.packages("DescTools")`) und anschließend importiert werden (`library(DescTools)`).

6.4 Odds Ratio

Der Odds-Ratio ist ein Maß, welche speziell für Vier-Felder-Tafeln definiert ist. Durch Zwischenschritte über das (relative) Risiko bis letztendlich hin zum Odds Ratio lässt er sich gut nachvollziehbar erklären.

	y_1	y_2	Σ
x_1	a	b	a+b
x_2	c	d	c+d
Σ	a+c	b+d	n

Table 8: Allgemeine Darstellung einer 2×2 -Kontingenztafel (Vier-Felder-Tafel)

Risiko Betrachtet man zunächst ein *Risikomerkm*al, so kann man für dieses Risikomerkm al die Wahrscheinlichkeit bestimmen, dass es eintritt. Diese Wahrscheinlichkeit wird als Risiko bezeichnet. So könnte man bspw. für Tabelle 3 das Risiko bestimmen, dass eine Frau an Lungenkrebs erkrankt: Entweder für die gesamte Population ($\frac{1335}{3345}$) oder für eine bestimmte "Schicht" (z.B. für die Raucher: $\frac{877}{1289}$). Im Folgenden werden wir vom Geschlecht als *Schichtungsmerkmal* und vom Lungenkrebs als *Risikomerkm*al sprechen.

Relatives Risiko Möchte man nun die Risiken in zwei verschiedenen Schichten miteinander vergleichen, so bietet sich das relative Risiko als Maßzahl an. Hierbei dividiert man die Risiken zweier Schichten durcheinander, z.B. das Risiko für Lungenkrebs bei den Rauchern und das Risiko für Lungenkrebs bei den Nicht-Rauchern:

$$RR_{Lungenkrebs} = \frac{\frac{877}{1289}}{\frac{458}{2056}} = \frac{0,68}{0,22} = 3,09$$

Dieser Wert bedeutet, dass das Risiko für Lungenkrebs bei den Rauchern ca. dreimal so hoch ist wie bei den Nicht-Rauchern. Welches Schichtungsmerkmal dabei im Zähler steht ist nicht fix vorgegeben, ist jedoch für die Interpretation des Ergebnisses im Nachhinein wichtig. Man könnte also auch die Nicht-Raucher in den Zähler packen:

$$RR_{Lungenkrebs} = \frac{\frac{458}{2056}}{\frac{877}{1289}} = \frac{0,22}{0,68} = 0,32$$

Die Interpretation wäre hier, dass das Risiko für Lungenkrebs bei den Nicht-Rauchern lediglich einem Drittel des Risikos bei den Rauchern entspricht. Letztlich die gleiche Aussage wie zuvor, nur auf andere Art und Weise ausgedrückt.

Chance Vergleicht man die Risiken von zwei konkurrierenden Risikomerkm alen für ein Schichtungsmerkmal, so spricht man von der Chance oder auf Englisch *Odds*. Man spricht dabei immer von der Chance auf dasjenige Merkmal, dessen Risiko im Zähler steht, d.h.

$$\frac{\text{Risiko für "Lungenkrebs" bei Rauchern}}{\text{Risiko für "kein Lungenkrebs" bei Rauchern}} = \frac{\frac{877}{1289}}{\frac{412}{1289}} = \frac{877}{412} = 2,13$$

beschreibt die Chance für Lungenkrebs (bei Rauchern), während

$$\frac{\text{Risiko für "kein Lungenkrebs" bei Rauchern}}{\text{Risiko für "Lungenkrebs" bei Rauchern}} = \frac{\frac{412}{1289}}{\frac{877}{1289}} = \frac{877}{412} = 0,47$$

die Chance für kein Lungenkrebs (bei Rauchern) beschreibt. Letztendliche liefern beide Berechnungen jedoch wieder dieselbe Aussage.

Odds Ratio Ähnlich wie beim relativen Risiko möchte man auch beim Odds Ratio zwei verschiedene Schichten miteinander vergleichen, aber nicht hinsichtlich der Risiken sondern hinsichtlich der Chancen. Hierbei dividiert man also die Chancen zweier Schichten durcheinander, z.B. die Chance für Lungenkrebs bei den Nicht-Rauchern und die Chance für Lungenkrebs bei den Rauchern:

$$OR = \frac{\frac{458}{1598}}{\frac{877}{412}} = 0,02$$

Dieser Wert bedeutet, dass die Chance für Lungenkrebs bei den Nicht-Rauchern ca. 0,02 mal so hoch ist wie bei den Rauchern. Welches Schichtungsmerkmal dabei im Zähler steht ist nicht fix vorgegeben, ist jedoch für die Interpretation des Ergebnisses im Nachhinein wichtig. Man könnte also auch die Raucher in den Zähler packen:

$$OR = \frac{\frac{877}{412}}{\frac{458}{1598}} = 58,65$$

Die Interpretation wäre hier, dass die Chance für Lungenkrebs bei den Rauchern 58,65 mal dem der Nicht-Raucher entspricht. Letztlich die gleiche Aussage wie zuvor, nur auf andere Art und Weise ausgedrückt.

⚠ Bei der generellen Formel für den Odds Ratio steht (sofern nicht explizit etwas anderes gesagt wird) das Schichtungsmerkmal aus der ersten Zeile im Zähler, was man an der Formel (bezogen auf Tabelle 8) erkennen kann:

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

R-Befehl für den Odds Ratio: <code>> OddsRatio(data)</code>
--

Dokumentation

⚠ Die R-Funktion ist **nicht** Teil von **base-R** sondern Teil des Paketes **DescTools**. Dieses muss zunächst installiert (`install.packages("DescTools")`) und anschließend importiert werden (`library(DescTools)`).

6.5 Zusammenhangsmaße für ordinale Merkmale

Bei ordinalen Daten haben wir, durch die Möglichkeit die Daten zu ordnen, mehr Informationsgehalt. Deshalb macht es hier Sinn, andere Zusammenhangsmaße zu benutzen, damit dieser Informationsgehalt nicht verloren geht. Die Idee hinter den ordinalen Zusammenhangsmaßen ist die, dass man jede Beobachtung mit jeder anderen Beobachtung vergleicht und sich dabei anschaut, wie diese zueinander stehen. Dabei ordnet man solche Beobachtungspaare als *konkordant*, *diskordant* oder als *Bindung* ein.

Konkordanz liegt vor, wenn sich X und Y in die gleiche Richtung bewegen, also die Beobachtung mit einem größeren x-Wert auch einen größeren y-Wert aufweist.

Beispiel: $(x_1, y_1) = (2, 3)$ ist konkordant zu $(x_2, y_2) = (4, 8)$, denn beim Übergang von der ersten zur zweiten Beobachtung steigt X von 2 auf 4 und Y von 3 auf 8. Konkordanz läge ebenfalls vor, wenn X fallen & Y ebenfalls fallen würde. Somit bewegen sich beide Werte in die gleiche Richtung. "Je größer X, desto größer Y" spricht für einen positiven/gleichgerichteten Zusammenhang.

Diskordanz ist das Gegenteil, d.h. wenn X sich in eine andere Richtung als Y bewegt.

Beispiel hierfür ist $(x_1, y_1) = (2, 3)$ und $(x_2, y_2) = (5, 1)$. Beim Übergang von der ersten zur zweiten Beobachtung steigt X von 2 auf 5, Y sinkt jedoch von 3 auf 1. Diskordanz läge ebenfalls vor, wenn X fallen & Y steigen würde. Es geht somit nur um die Entwicklung in unterschiedliche Richtungen, also um einen negativen Zusammenhang.

Bindungen (Ties) liegen vor wenn sich ein Wert verändert, der andere jedoch gleich bleibt.

Beispiel für eine Bindung in X wäre hier $(x_1, y_1) = (3, 5)$ und $(x_2, y_2) = (3, 8)$ bzw. für eine Bindung in Y $(x_1, y_1) = (4, 7)$ und $(x_2, y_2) = (2, 7)$. Bindungen enthalten keine Aussagekraft für mögliche Zusammenhänge.

Um basierend auf einer Kontingenztafel die Anzahlen der konkordanten/diskordanten Paare und der Bindungen zu berechnen, ist es zunächst einmal wichtig, dass die Merkmalsausprägungen in den Zeilen & Spalten geordnet sind. Ist dies nicht der Fall, ist ein strukturiertes Vorgehen unmöglich. Ist dies gewährleistet, gibt es eine bestimmte, strukturierte Vorgehensweise:

Konkordante Paare:

- Man beginnt **links** oben (Zelle für die kleinste Ausprägung sowohl von X, als auch von Y).
- Man multipliziert die Beobachtungszahl in dieser Zelle mit allen Beobachtungszahlen, die in Zellen *unter* UND *rechts* von dieser Zelle liegen.
Grund: Diese Beobachtungen haben sowohl größere x-Werte, als auch größere y-Werte.
- Man wendet dieses Prinzip auf **jede** Zelle in der Tabelle an und summiert die Werte für die einzelnen Zellen am Schluss auf, um die Gesamtzahl konkordanter Paare zu erhalten.

Diskordante Paare


- Man beginnt **rechts** oben (Zelle für die kleinste Ausprägung von X und die größte von Y).
- Anders als bei der Konkordanz multipliziert man die Beobachtungszahl in dieser Zelle mit allen Beobachtungszahlen, die in Zellen *unter* UND *links* von dieser Zelle liegen.
Grund: Diese Beobachtungen haben sowohl größere x-Werte, jedoch kleinere y-Werte.
- Man wendet dieses Prinzip auf **jede** Zelle in der Tabelle an und summiert die Werte für die einzelnen Zellen am Schluss auf, um die Gesamtzahl diskordanter Paare zu erhalten.

Bindungen (Ties) in X

- Man beginnt **links** oben (Zelle für die kleinste Ausprägung sowohl von X, als auch von Y).
- Man multipliziert die Beobachtungszahl in dieser Zelle mit allen Beobachtungszahlen, die in Zellen *in derselben Zeile* wie diese Zelle liegen.
Grund: Diese Beobachtungen haben den gleichen x-Wert, jedoch andere y-Werte.
- Man wendet dieses Prinzip auf **jede** Zelle in der Tabelle an und summiert die Werte für die einzelnen Zellen am Schluss auf, um die Gesamtzahl an Bindungen in X zu erhalten.

Bindungen (Ties) in Y

- Man beginnt **links** oben (Zelle für die kleinste Ausprägung sowohl von X, als auch von Y).
- Man multipliziert die Beobachtungszahl in dieser Zelle mit allen Beobachtungszahlen, die in Zellen *in derselben Spalte* wie diese Zelle liegen.
Grund: Diese Beobachtungen haben andere x-Werte, jedoch den gleichen y-Wert.
- Man wendet dieses Prinzip auf **jede** Zelle in der Tabelle an und summiert die Werte für die einzelnen Zellen am Schluss auf, um die Gesamtzahl an Bindungen in Y zu erhalten.

 *Notation:* Für alle nachfolgend aufgeführten Formeln werden die Anzahlen der konkordanten & diskordanten Paare, sowie der Bindungen in X bzw. Y, mit K , D , T_x & T_y bezeichnet.

Ein schönes Beispiel für die Berechnung von K und D findet sich hier:

[Beispiel](#)

Mit der Kenntnis über die Anzahlen der konkordanten & diskordanten Paare, sowie der Bindungen in X bzw. Y, kann man nun einige Zusammenhangsmaße berechnen.

6.5.1 Gamma nach Goodman and Kruskal

Das γ nach Goodman and Kruskal quantifiziert den Zusammenhang zwischen zwei Merkmalen allein auf Grundlage der konkordanten und diskordanten Paare. Im Zähler wird die Differenz der beiden Werte berechnet und im Nenner werden die beiden addiert.

$$\gamma = \frac{K - D}{K + D}$$

Aufgrund der Differenzenbildung im Zähler dann der Bruch sowohl positive, als auch negative Werte annehmen. Liegen nur diskordante Paare vor (d.h. $K = 0$), besteht ein perfekter negativer Zusammenhang und γ nimmt den Wert -1 an. Liegen ausschließlich konkordante Paare vor (d.h. $D = 0$), so spricht man von einem perfekten positiven Zusammenhang und γ beträgt +1.

Abstufungen im negativen & positiven Bereich deuten auf *tendenziell* negative bzw. positive Zusammenhänge hin und je näher der Wert betragsmäßig an der 1 liegt, als desto stärker wird der Zusammenhang bewertet. Liegen (annähernd) gleich viele konkordante und diskordante Paar vor, so ist wird γ (annähernd) einen Wert von 0 annehmen und es ist kein eindeutiger Zusammenhang erkennbar.

⚠ Im Gegensatz zu den den Maßen für nominale Merkmale beinhaltet γ nicht nur eine Aussage über die **Stärke**, sondern auch über die **Richtung** des Zusammenhangs. Dies spiegelt sich im Wertebereich wieder, der in Kapitel 6.3 bei allen Maßen nur im positiven lag und bei den Maßen in diesem Kapitel sowohl im positiven als auch im negativen liegt.

⚠ Bei der Berechnung von γ werden die Bindungen komplett außen vor gelassen werden. Dies führt (tendenziell) zu einer Überschätzung des Zusammenhangs, da Beobachtungspaare die gegen einen Zusammenhang sprechen (die Bindungen) ignoriert werden. Die τ -Maße beheben dieses Problem.

R-Befehl für γ : <code>> GoodmanKruskalGamma(data)</code>

Dokumentation

⚠ Die R-Funktion ist **nicht** Teil von **base-R** sondern Teil des Paketes **DescTools**. Dieses muss zunächst installiert (`install.packages("DescTools")`) und anschließend importiert werden (`library(DescTools)`).

6.5.2 Tau-Maße

Kendalls τ_b Hier werden die T_x und T_y Bindungen berücksichtigt. Der Zähler ist identisch zu γ , jedoch werden im Nenner die Bindungen mit eingerechnet:

$$\tau_b = \frac{K - D}{\sqrt{(K + D + T_x)(K + D + T_y)}}$$

⚠ Liegen keine Bindungen vor, so kollabiert der Nenner zu

$$\sqrt{(K + D + 0)(K + D + 0)} = \sqrt{(K + D)^2} = K + D,$$

sodass τ_b identisch zu γ ist.

R-Befehl für τ_b : <code>> cor(data\$x, data\$y, method = "kendall")</code>

Dokumentation

Kendalls/Stuarts τ_c Bei Kendalls/Stuarts τ_c werden nicht die Bindungen, sondern das Minimum aus der Spaltenzahl und der Zeilenanzahl in der Berechnung berücksichtigt. Ausgangspunkt ist jedoch weiterhin die Differenz aus den Anzahlen der konkordanten und diskordanten Paare im Zähler.

$$\tau_c = \frac{2\min(k, l)(K - D)}{n^2(\min(k, l) - 1)}$$

⚠ **Remember:** Die Zeilen- und Spaltenzahl werden in diesem Kontext mit k & l bezeichnet.

R-Befehl für τ_c : <code>> StuartTauC(data)</code>
--

Dokumentation

⚠ Die R-Funktion ist **nicht** Teil von **base-R** sondern Teil des Paketes **DescTools**. Dieses muss zunächst installiert (`install.packages("DescTools")`) und anschließend importiert werden (`library(DescTools)`).

⚠ Kendalls τ_b ist das wohl gebräuchlichste dieser drei Maße, was man auch daran erkennen kann, dass es in *base-R* verfügbar ist während die anderen beiden nur über ein spezielles Paket verfügbar sind.

6.5.3 Rangkorrelationskoeffizient nach Spearman

Der Spearman Korrelations-Koeffizient (r_{SP}) wird auch Rangkorrelationskoeffizient genannt, da man nicht die Abstände der echten Datenpunkte zueinander in Beziehung setzt, sondern nur die *Rangabstände* betrachtet. Dies ist auch der große Unterschied zum Korrelationskoeffizient nach Bravais-Pearson (vgl. Kap. 6.6.2).

Berechnung: Die Basis bildet eine Auflistung der Beobachtungen und die Zuordnung der passende Ränge zu den einzelnen Merkmalsausprägungen. Anschließend bildet man *für jeder Beobachtung* die Differenz des Ranges von x_i und des Ranges von y_i . Auf Basis dieser Rangdifferenzen d_i wird r_{SP} schließlich berechnet:

$$r_{SP} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

⚠ Diese Formel funktioniert nur wenn **keine** Bindungen vorliegen. Ist dies doch der Fall, so muss man dafür korrigieren (siehe unten).

Korrigierter Rangkorrelationskoeffizient nach Spearman Liegen *Bindungen* innerhalb eines Merkmals oder beider Merkmale vor, wird die Bestimmung der Ränge etwas komplexer. Zusätzlich zu den d_i kommen hier noch b_j und c_k hinzu.

⚠ b_j und c_k haben einen anderen Index als die d_i . Während i der Laufindex für die Beobachtungen ist, beziehen sich j und k auf die verschiedenen Merkmalsausprägungen von X bzw. Y.

Die b_j sind die Anzahlen, die angeben wie oft die unterschiedlichen Merkmalsausprägungen beim Merkmal X jeweils auftreten. Die c_k bedeuten analog dazu dasselbe für Y.

Liegen keine Bindungen vor, so passieren zwei Dinge: Alle b_j und alle c_k nehmen den Wert 1 an, da jede Merkmalsausprägung nur einmal vorkommt. Die Teile der Formel mit b_j und c_k werden dadurch gleich Null und fallen also weg. Passiert dies, so reduziert sich die Formel auf die obige Formel, welche also nur einen Spezialfall der nachfolgenden Formel darstellt:

$$r_{SP} = \frac{n(n^2 - 1) - \frac{1}{2} \sum_j b_j(b_j^2 - 1) - \frac{1}{2} \sum_k c_k(c_k^2 - 1) - 6 \sum_i d_i^2}{\sqrt{n(n^2 - 1) - \sum_j b_j(b_j^2 - 1)} \sqrt{n(n^2 - 1) - \sum_k c_k(c_k^2 - 1)}}$$

Der Wertebereich von r_{SP} geht von -1 bis +1 und auch die Interpretation ist analog zu den γ - & τ -Maßen: Bei $r_{SP} < 0$ liegt ein negativer Zusammenhang zwischen den beiden Merkmalen vor, bei $r_{SP} > 0$ liegt ein positiver Zusammenhang vor und bei $r_{SP} = 0$ liegt kein Zusammenhang vor. Zudem ist r_{SP} dimensionslos und symmetrisch. Symmetrisch bedeutet, dass $r_{SP}(X, Y) = r_{SP}(Y, X)$.

R-Befehl für r_{SP} : `> cor(data$x, data$y, method = "spearman")` [Dokumentation](#)

⚠ Auch wenn alle Maße aus diesem Kapitel für ordinale Merkmale gedacht sind, kann man sie auch für metrische Daten verwenden. Dabei entsteht allerdings Informationsverlust, da diese Maße nicht die volle Information der metrischen Skala (Interpretierbarkeit von Abständen) ausnutzen.

6.6 Zusammenhangsmaße für metrische Merkmale

6.6.1 Kovarianz

Die Kovarianz ist ein Maß für den Zusammenhang zweier metrischer Merkmale. Sie misst die gemeinsame Streuung zweier Merkmale, weshalb die Varianz lediglich ein Spezialfall der Kovarianz ist, nämlich die gemeinsame Streuung einer Variable mit sich selbst.

Die Berechnung kann als Verallgemeinerung der Varianz angesehen werden:

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

zieht man den Vergleich zur Formel der Varianz aus Kapitel 4.4, so fällt auf, dass $\text{Cov}(X, X) = \text{Var}(X)$. Des weiteren ist die Kovarianz symmetrisch, d.h. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$, was man eigentlich recht gut an obiger Formel sieht.

Ist die Kovarianz positiv bzw. negativ, so liegt ein positiver bzw. negativer Zusammenhang zwischen den beiden Merkmalen vor. Die Kovarianz hat einen Wertebereich von $-\infty$ bis ∞ und hängt von zwei Faktoren ab:

- Zusammenhang der beiden Merkmale
- Streuung der beiden Merkmale

wobei der Einfluss von Letzterem die Quantifizierung des reinen Zusammenhangs behindert.

⚠ Aufpassen: Es kann passieren, dass ein Merkmal X (mit hoher Streuung) eine höhere Kovarianz mit Merkmal Y besitzt als das Merkmal Z (mit niedriger Streuung), obwohl Z stärker mit Y zusammenhängt als X .

R-Befehl für $\text{Cov}(X, Y)$: <code>> cov(data\$x, data\$y)</code>

Dokumentation

6.6.2 Korrelationskoeffizient nach Bravais-Pearson

Der Korrelationskoeffizient nach Bravais-Pearson bereinigt die Kovarianz um den Einfluss der Streuung und kann somit zur reinen Quantifizierung des Zusammenhangs verwendet werden. Dies erreicht man, indem man die Kovarianz durch die Wurzel aus dem Produkt der beiden Varianzen dividiert.

$$r_{BP} = \frac{\text{Cov}(X; Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

bzw. mit $\frac{1}{n}$ rausgekürzt:
$$r_{BP} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)(\sum_{i=1}^n y_i^2 - n\bar{y}^2)}}$$

Dies ist auch wieder ein symmetrisches Maß, allerdings mit einem Wertebereich von -1 bis 1, da r_{BP} standardisiert & dimensionslos ist. Ist $r_{BP} > 0$, so spricht man von einem positiven linearen Zusammenhang, ist $r_{BP} < 0$ so hängen die beiden Merkmale negativ linear zusammen. Bei $r_{BP} = 0$ sind die beiden Merkmale unkorreliert.

R-Befehl für r_{BP} : `> cor(data$x, data$y, method = "pearson")` [Dokumentation](#)

6.7 Wrap-Up

Der Übersichtlichkeit halber hier nochmal alle Skalenniveaus und die passenden Zusammenhangsmaße auf einen Blick.

	Nominal	Ordinal	metrisch
Nominal	χ^2 -basiert		
Ordinal	χ^2 -basiert	τ -und γ -Maße/ Rangkorrelation	
Metrisch	χ^2 -basiert	Rangkorrelation	Korrelation (Kovarianz)

Table 9: Übersicht über verschiedene Zusammenhangsmaße

Diese Tabelle verdeutlicht drei zentrale Punkte:

- Interessiert man sich für den Zusammenhang von zwei Merkmalen, die auf *unterschiedlichen* Skalenniveaus gemessen werden, kann man nur Zusammenhangsmaße verwenden, die dem *niedrigeren* Skalenniveau genügen.

Beispiel: Den Zusammenhang zwischen Studienfach (*nominal*) & Statistik-Note (*ordinal*) kann man lediglich mit χ^2 -basierten Maßen quantifizieren.

- Es ist immer *möglich* Zusammenhangsmaße zu verwenden, die für niedrigere Skalenniveaus geeignet sind. Ob das *sinnvoll* ist, ist eine andere Frage, da man dadurch nicht den vollen Informationsgehalt des vorliegenden Skalenniveaus ausschöpft.

Beispiel: Den Zusammenhang zwischen Größe (*metrisch*) & Gewicht (*metrisch*) könnte man auch (unter Inkaufnahme von Informationsverlust) mit r_{SP} quantifizieren.

- Es ist *nicht möglich* Zusammenhangsmaße zu verwenden, die ausschließlich für höhere Skalenniveaus geeignet sind.

Beispiel: Den Zusammenhang zwischen Studienfach (*nominal*) & Statistik-Note (*ordinal*) könnte man nicht mit r_{SP} oder gar r_{BP} quantifizieren.

6.8 Aufgaben

1. Welche Aussagen bzgl. relativem Risiko & Odds Ratio sind wahr?

- a) Bei beiden Maßzahlen werden zwei Gruppen verglichen. ☐
- b) Der Odds Ratio kann auf Basis von relativen Risiken berechnet werden. ☐
- c) Beim relativen Risiko werden zwei Risikomerkmale verglichen. ☐
- d) Ein Odds Ratio < 0 bedeutet eine geringere Chance in der ersten Gruppe. ☐

2. Die unter Unabhängigkeit erwarteten absoluten Häufigkeiten ..

- a) .. müssen stets ganzzahlig sein. ☐
- b) .. können ohne Kenntnis der gemeinsamen Verteilung berechnet werden. ☐
- c) .. sind identisch zu der bedingten Verteilung. ☐
- d) .. sind maximal so hoch wie die tatsächlich beobachteten Häufigkeiten. ☐

3. Welche der folgenden Aussagen über Zusammenhangsmaße für nominale Merkmale sind wahr?

- a) Mit Cramers V sind Zusammenhänge für Kontingenztafeln von verschiedener Dimension und mit unterschiedlichem n vergleichbar. ☐
- b) Φ besitzt einen kleineren Wertebereich als χ^2 . ☐
- c) Um Kontingenztafeln mit dem korrigierten Kontingenzkoeffizienten vergleichen zu können muss deren Dimension gleich sein. ☐
- d) Kein Zusammenhangsmaß für nominale Merkmale kann negative Werte annehmen. ☐

4. Welche der folgenden Aussagen sind wahr?

- a) Rang-basierte Zusammenhangsmaße sind bei metrischen Merkmalen nicht anwendbar. ☐
- b) Das Prinzip der Kon-/Diskordanz kann auch bei nominalen Merkmalen angewendet werden. ☐
- c) Zusammenhangsmaße für nominale Merkmale können nicht bei ordinalen oder metrischen Merkmalen verwendet werden. ☐
- d) Zusammenhangsmaße für nominale Merkmale können keine Richtung des Zusammenhangs angeben. ☐

5. Bindungen in Y ..

- a) .. sprechen für einen negativen Zusammenhang. ☐
- b) .. haben keinen Einfluss auf den Wert von γ . ☐
- c) .. erhöhen den Wert von Kendalls τ_b . ☐
- d) .. haben einen Einfluss auf den Wert von Kendalls/Stuarts τ_c . ☐

7 Lineare Einfachregression

7.1 Einführung

Das Ziel der Regression ist es herauszufinden, den Einfluss von einem Merkmal X (Einflussgröße) auf ein Merkmal Y (Zielgröße) zu quantifizieren. Die Einflussgröße X wird auch Regressor oder unabhängige Variable genannt. Das Merkmal Y wird daher abhängige Variable, da sie abhängig von X ist, Response oder Regressand genannt. Um eine Regressionsanalyse sinnvoll durchführen zu können, benötigt man Beobachtungspaare (wie bereits in Kap. 6), mit unterschiedlichen Merkmalsausprägungen von X und Y . Damit kann man dann die Merkmale auf einen (linearen) Zusammenhang prüfen und versuchen, diesen mit einem Modell zu schätzen.

⚠ Zusammenhangsmaße geben (im best case) lediglich Stärke & Richtung des Zusammenhangs an. Die lineare Regression geht *zwei* Schritte weiter: Es wird (1) eine Wirkungskette angenommen (X beeinflusst Y und nicht andersherum) & (2) der Zusammenhang quantifiziert ("Wenn X um eine Einheit steigt, erwarten wir eine Änderung von Y um ...")

Der einfachste Zusammenhang zwischen zwei Merkmalen ist der lineare Zusammenhang, welcher durch den Ansatz $Y = a + b \cdot X$ angegeben wird. Bildet man den Zusammenhang zwischen X und Y durch dieses Modell ab, so spricht man von einer *linearen (Einfach-)Regression*.

Da nicht alle Beobachtungspaare zwingend auf einer Geraden liegen, wird der obige Ansatz durch ein *Fehlerglied* (bzw. *-term*) oder *Residuum* e ergänzt. Dadurch werden zufällige Abweichungen von der Geraden mit in das Modell einbezogen und es wird erhalten $Y = a + b \cdot X + e$ als Modellgleichung.

7.2 Plots und Annahmen

Da man bei zwei Merkmalen nicht automatisch von einem linearen Zusammenhang bzw. einer Ursachen-Wirkung-Beziehung ausgehen kann, ist es ratsam, zuerst einmal durch eine *graphische Darstellung* zu überprüfen, ob die Annahme eines linearen Zusammenhang überhaupt zu rechtfertigen ist. Hierbei ist ein *Streudiagramm* oder *Scatter-Plot* hilfreich.

Hierbei trägt man auf die x-Achse die Einflussgröße X und auf der y-Achse die Zielgröße Y ab und jedes Beobachtungspaar $(x_i; y_i)$ wird als ein Punkt in das Koordinatensystem eingezeichnet. Liegt ein Zusammenhang zwischen den Merkmalen vor, kann man das oft schon hier erkennen. Es gibt jedoch nicht nur lineare, sondern u.a. auch zyklische, exponentiell wachsende/fallende oder logarithmische Zusammenhänge. Wir fokussieren uns hier ausschließlich auf lineare Zusammenhänge. Oft kommt es vor, dass zwar ein grober Zusammenhang zu erkennen ist, dieser jedoch durch nicht dazu passende Beobachtungspaare, sogenannte *Ausreißer*, gestört wird. Diese müssen gesondert betrachtet werden und gegebenenfalls (mit guter Rechtfertigung) aus dem Datensatz entfernt werden, da diese (teils großen) Einfluss auf die Schätzung (vgl Kap. 7.3) haben können.

R-Befehl für Scatter-Plots: <code>> plot(data\$x, data\$y, ...)</code>

Dokumentation

7.3 Kleinste-Quadrate-Schätzer

Motivation Angenommen die Modellparameter wären mit $\hat{a} = 4$ und $\hat{b} = 2$ bestimmt und wir betrachten das Beobachtungspaar $(x_1; y_1) = (5; 10)$ aus dem Datensatz. Für dieses Beobachtungspaar wäre der (durch das Modell) *vorhergesagte Wert*

$$\hat{y}_1 = a + b \cdot x_1 = 4 + 2 \cdot 5 = 14,$$

was keine besonders akkurate Vorhersage wäre (das wahre y_1 ist 10). Den daraus resultierenden Fehler \hat{e}_1 könnte man wie folgt berechnen:

$$y_1 = \underbrace{a + b \cdot x_1}_{\hat{y}_1} + e_1 \quad \Leftrightarrow \quad \hat{e}_1 = y_1 - \hat{y}_1 = 5$$

Genauso könnte man für das zweite, dritte, ..., n -te Beobachtungspaar den Fehler bestimmen. Je besser das Modell, desto kleiner wäre die Summe der *Beträge der Fehler*. Das Wort *Betrag* ist hier wichtig, da sowohl Abweichungen nach oben als auch nach unten schlecht sind.

 **Notation:** *Geschätzte* Werte werden stets mit einem *Dach* versehen (z.B. \hat{a} oder \hat{y}).

Diese Fehler spielen eine zentrale Rolle in dem Optimierungsproblem zur Bestimmung der Werte für a und b . Bei diesem Optimierungsproblem wird die **Summe der quadrierten Fehler** minimiert:

$$\text{Fehler für ein Beobachtungspaar: } e_i = y_i - \hat{y}_i = y_i - a + b \cdot x_i$$

$$\text{Quadrierter Fehler für ein Beobachtungspaar: } e_i^2 = (y_i - \hat{y}_i)^2 = (y_i - a + b \cdot x_i)^2$$

$$\text{Summe der quadrierten Fehler: } \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a + b \cdot x_i)^2$$

Man verwendet die quadrierten statt betragsmäßigen Fehler, da sich diese besser optimieren lassen.


Vorgehensweise der Schätzung: Um diesen Ausdruck zu minimieren, also $\min_{a,b} \sum_{i=1}^n e_i^2$, bildet man die partiellen Ableitungen nach a und b ab und setzt diese jeweils gleich Null. Durch Auflösen der zweiten Gleichung erhält man:

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n \bar{y} \bar{x}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

Eingesetzt in die erste Gleichung ergibt sich:

$$\hat{a} = \bar{y} - \hat{b} \bar{x}$$

Würde man den Schätzer für b mit $\frac{1}{n}$ erweitern, so stünde im Zähler die Kovarianz und im Nenner die Varianz der Einflussgröße, d.h. $\hat{b} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$.

 \hat{b} muss immer vor \hat{a} berechnet werden, da es Teil der Formel zur Berechnung von \hat{a} ist.

7.4 Besonderheiten der Regressionsgerade

Sinnvoller Wertebereich Da nur die Beobachtungspaar $(x_1; y_1), \dots, (x_n; y_n)$ in die Schätzung einfließen, können wir nur über deren Wertebereich eine sinnvolle Aussage treffen.

Beispiel: Schätzt man ein Regressionsmodell mit Gewicht als Ziel- und Körpergröße als Einflussgröße basierend auf Daten von Erwachsenen, so sollten basierend darauf keine Aussagen/Prognosen für Kleinkinder getroffen werden.

Arithmetisches Mittel Da im vorigen Kapitel zur Berechnung von \hat{a} die Mittelwerte verwendet wurden, liegt der Punkt mit den Mittelwerten (\bar{x}, \bar{y}) auch auf der Regressionsgeraden.

Kleiner Tipp: Wenn man die Regressionsgerade zeichnen muss, dann kann man diesen Punkt immer verwenden, da die arithmetischen Mittel meist schon bekannt sind.

Bedeutung des Korrelationskoeffizient Für die Berechnung von \hat{b} spielt die Kovarianz eine entscheidende Rolle. Daher gibt das Vorzeichen von $Cov(X; Y)$ (bzw. r_{BP}) Auskunft über das Vorzeichen von \hat{b} . Eine höheren Kovarianz/Korrelation bedeutet jedoch nicht automatisch größeres \hat{b} , da nicht nur diese eine Rolle bei der Berechnung von \hat{b} spielt, sondern auch die Streuung von X .

7.5 Güte der Anpassung

Die Frage "Wie gut repräsentiert unser Modell die Originaldaten?" lässt sich durch eine Varianzanalyse beantworten. Dabei wird die Gesamtstreuung der Daten zerlegt in die Streuung der vorhergesagten Werte und die Streuung der Fehler:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\sum_{i=1}^n e_i^2}$$
$$SQ_{Total} = SQ_{Regression} + SQ_{Residual}$$

Die erste der drei Summen (SQ_{Total}) ist die Gesamtstreuung von Y . Die zweite ($SQ_{Regression}$) ist die Streuung der *vorhergesagten* y -Werte. Daher spricht man hier von der Streuung, die durch das Regressionsmodell erklärt wird. Das heißt je größer $SQ_{Regression}$ im Verhältnis zu SQ_{Total} ist, desto besser passt das Regressionsmodell zu den Originaldaten. Diese dritte Summe ($SQ_{Residual}$) entspricht den quadrierten Fehlern und sollte daher im Verhältnis zu SQ_{Total} möglichst klein sein.

⚠ Da die \hat{y}_i basierend auf den x_i vorhergesagt werden, kennen wir den Grund (nämlich die unterschiedlichen x -Werte) für die Abweichung von \bar{y} . Daher spricht man vom der *erklärten Streuung*.

Bestimmtheitsmaß R^2 Dieses Maß gibt den Anteil der erklärten Streuung ($SQ_{Regression}$) an der gesamten Streuung (SQ_{Total}) an. Analog könnte man auch den Anteil der nicht erklärten Streuung ($SQ_{Residual}$) von 1 subtrahieren:

$$R^2 = \frac{SQ_{Regression}}{SQ_{Total}} = 1 - \frac{SQ_{Residual}}{SQ_{Total}}$$

Da es sich hier um einen Anteilswert handelt, geht der Wertebereich von R^2 von 0 bis 1. Liegen sämtliche Beobachtungspaare exakt auf der Regressionsgerade, so ist $SQ_{Regression} = SQ_{Total}$ und R^2 nimmt folglich den Wert 1 an. Man spricht hier von *perfekter Anpassung*.

Wenn jedoch das Bestimmtheitsmaß 0 beträgt, dann wird kein Teil der Streuung in den Originaldaten durch das Modell erklärt. Dies erkennt man grafisch an einer Regressionsgerade die parallel zur x-Achse verläuft, da $\hat{b} = 0$ und somit für jedes x_i das gleiche $\hat{y}_i = \bar{y}$ vorhergesagt wird. Man spricht hier von *Nullanpassung*.

⚠ Für die lineare Regression mit einer Einflussgröße ergibt sich ein Spezialfall: Hier kann R^2 auch durch Quadrierung von r_{BP} berechnet werden.

7.6 Kategoriale Regression

Alle bisherigen Ausführungen bezogen sich auf den Fall eines metrischen Merkmals als Einflussgröße. Besitzt X jedoch nur kategoriales bzw. nominales Skalenniveau, erfordert dies eine andere Vorgehensweise. Um auch solche Merkmale als Einflussgröße verwenden zu können, müssen diese entsprechend *umzukodiert* werden. Dies kann mit Hilfe der **Dummykodierung** oder der **Effektkodierung** geschehen, welche im Folgenden anhand eines Beispiels erläutert werden:

Hierfür betrachten wir das Merkmal X : "*Parteizugehörigkeit eines Bundestagsabgeordneten*" mit den möglichen Ausprägungen $\{Union, SPD, Grüne, Linke, AfD, FDP\}$ als Einflussgröße. Die Zielgröße Y ist die "*Höhe der Nebeneinkünfte des Abgeordneten*".

Dummy- und Effektkodierung haben folgende Schritte gemeinsam:

Schritt 1: Auswahl einer Referenzkategorie

Eine der möglichen Merkmalsausprägungen muss als Referenz ausgewählt werden. Für diese Kategorie wird keine Dummyvariable gebildet. Was als Referenz gewählt wird ist prinzipiell egal, muss allerdings bei der anschließenden Interpretation berücksichtigt werden.

Beispiel: Wir entscheiden uns für die *Union* als Referenzkategorie.

Schritt 2: Bildung von $k-1$ Dummyvariablen

Anschließend wird die *Anzahl* der Dummies bestimmt. Wir benötigen eine Dummyvariable weniger als es mögliche Merkmalsausprägungen gibt. Es wird für jede mögliche Merkmalsausprägung eine Dummyvariable kreiert.

Beispiel: Wir benötigen hier $6 - 1 = 5$ Dummyvariablen, konkret:

x_{SPD} , $x_{Grüne}$, x_{Linke} , x_{AfD} , x_{FDP}

Schritt 3: Aufstellen der Regressionsgleichung

Da wir hier nun technisch gesehen mehrere Einflussgrößen haben, sieht auch die Gleichung anders aus.

Beispiel: $y_i = a + b_1 \cdot x_{SPD} + b_2 \cdot x_{Grüne} + b_3 \cdot x_{Linke} + b_4 \cdot x_{AfD} + b_5 \cdot x_{FDP} + e_i$

Sie unterscheiden sich in folgenden Schritten:

Schritt 4 [Dummykodierung]: *Definition der Ausprägungen der Dummyvariablen*

Die i -te Dummyvariable nimmt den Wert 1 an, falls die i -te Kategorie vorliegt und Null sonst.

$$\textbf{Beispiel: } x_{SPD} = \begin{cases} 1, & \text{falls } X = SPD \\ 0, & \text{sonst} \end{cases} ; \quad x_{Grüne} = \begin{cases} 1, & \text{falls } X = Grüne \\ 0, & \text{sonst} \end{cases} ; \text{ usw.}$$

Schritt 4 [Effektkodierung]: *Definition der Ausprägungen der Dummyvariablen*

Die i -te Dummyvariable nimmt weiterhin den Wert 1 an, falls die i -te Kategorie vorliegt. Sie nimmt den Wert -1 an, falls die Referenz vorliegt und Null bei jeder anderen Kategorie.

$$\textbf{Beispiel: } x_{SPD} = \begin{cases} 1, & \text{falls } X = SPD \\ -1, & \text{falls } X = Union \\ 0, & \text{sonst} \end{cases} ; \quad x_{Grüne} = \begin{cases} 1, & \text{falls } X = Grüne \\ -1, & \text{falls } X = Union \\ 0, & \text{sonst} \end{cases} ; \text{ usw.}$$

Schritt 5 [Dummykodierung]: Berechnung der Parameter

1. Berechnung der Mittelwerte von Y für jede Kategorie
2. Der Mittelwert der Referenz ist der Intercept \hat{a}
3. Die Differenz des Mittelwerts der i -ten Kategorien zu \hat{a} ergeben die \hat{b}_i

Beispiel: $\hat{a} = \bar{y}_{Union}$; $\hat{b}_1 = \bar{y}_{SPD} - \hat{a} = \bar{y}_{SPD} - \bar{y}_{Union}$, usw.

Schritt 5 [Effektkodierung]: Berechnung der Parameter

1. Berechnung der Mittelwerte von Y für jede Kategorie
2. Der *Mittelwert aller Mittelwerte* ist der Intercept \hat{a}
3. Die Differenz des Mittelwerts der i -ten Kategorien zu \hat{a} ergeben die \hat{b}_i

Beispiel: $\hat{a} = \frac{1}{6} \cdot (\bar{y}_{Union} + \bar{y}_{SPD} + \dots + \bar{y}_{FDP})$; $\hat{b}_1 = \bar{y}_{SPD} - \hat{a}$, usw.

Der Unterschied besteht als allein darin, welche Ausprägung die Dummies annehmen, falls die Referenzkategorie vorliegt. Dies hat jedoch weitreichende Implikationen für die Interpretation:

Interpretation Dummykodierung: \hat{a} ist der Mittelwert der Referenzkategorie und $\hat{b}_1, \dots, \hat{b}_{k-1}$ entsprechen jeweils den Abweichungen der Mittelwerte der anderen Kategorien zur Referenz.

Beispiel: Für einen SPD-Agebordneten erwarten man um \hat{b}_1 höhere/niedrigere Nebeneinkünfte als für einen Abgeordneten der Union.

Interpretation Effektkodierung : Hier bezieht sich \hat{a} nicht auf die Referenz, sondern auf eine fiktive "*durchschnittliche Kategorie*". Die $\hat{b}_1, \dots, \hat{b}_{k-1}$ entsprechen jeweils den Abweichungen der Mittelwerte der anderen Kategorien zu dieser "*Durchschnittskategorie*". Für die Abweichung der Referenz zum Intercept multipliziert man jedes \hat{b}_i mit -1 und summiert die Produkte dann auf.

Beispiel: Für einen SPD-Agebordneten erwarten man um \hat{b}_1 höhere/niedrigere Nebeneinkünfte als für den fiktiven "*Durchschnittsabgeordneten*".

R-Befehl für lineare Regression: `> lm(formula = .., data = ..)` [Dokumentation](#)

Für die folgenden Erläuterungen wird der Datensatz `mtcars` aus dem `datasets`-Paket verwendet. Für das Verständnis der Beispiele sollte man sich die Dokumentation kurz ansehen.

⚠ In dem `formula`-Argument wird die Modellgleichung übergeben. Dabei wird der Intercept **nicht** mit angegeben, da diese per default mitgeschätzt wird. Anstatt dem Ist-Gleich-Zeichen muss hier eine Tilde (`~`) eingegeben werden.

Eine typische Eingabe sähe hier in etwa so aus: `formula = mpg ~ wt`

⚠ In dem `data`-Argument wird der Datensatz übergeben, in dem die Variablen aus der Modellgleichung zu finden sind. Der vollständige Befehl sähe hier in etwa so aus:

```
lm(formula = mpg ~ wt, data = mtcars)
```

⚠ Mit dem `summary`-Befehl kann der Output eines geschätzten Modell angezeigt werden. Man weist dem Modell typischerweise einen Namen zu und gibt diesen dann als Argument in den `summary`-Befehl (s.u.). Ein typischer Output sähe wie folgt aus:

```
modell <- lm(formula = mpg ~ wt, data = mtcars)
summary(modell)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min       1Q   Median       3Q      Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt          -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

In der `Coefficients`-Tabelle kann man in der ersten Spalte ("Estimate") \hat{a} in der Zeile `Intercept` und \hat{b} in der Zeile `wt` ablesen. Unter `Multiple R-squared` findet man den Wert für R^2 .

⚠ Wurde bereits ein Scatter-Plot der Daten mit `plot(data$x, data$y)` erstellt, so kann mit `abline(modell)` relativ unkompliziert die geschätzte Regressionsgerade eingezeichnet werden:

R-Befehl fürs Einzeichnen von Geraden: `> abline(modell)`

[Dokumentation](#)

⚠ Kategoriale Variablen erkennt man in R daran, dass sie den Variablentyp `factor` aufweisen. Den Variablentyp kann man mit dem `str()`-Befehl abfragen. Übergibt man dem `lm()`-Befehl im `formula`-Argument auf der rechten Seite der Tilde eine Variable von Typ `factor`, so bildet R automatisch $k - 1$ Dummyvariablen und nimmt diese mit ins Modell auf. Ist eine Variable nicht vom Typ `factor`, so kann man sie mit dem `factor()`-Befehl umwandeln (so wie hier notwendig, s.u.). Ein typischer Output sähe wie folgt aus:

```
modell2 <- lm(formula = mpg ~ factor(cyl), data = mtcars)
summary(modell2)

Call:
lm(formula = mpg ~ factor(cyl), data = mtcars)

Residuals:
    Min       1Q   Median       3Q      Max
-5.2636 -1.8357  0.0286  1.3893  7.2364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.6636     0.9718  27.437  < 2e-16 ***
factor(cyl)6  -6.9208     1.5583  -4.441  0.000119 ***
factor(cyl)8 -11.5636     1.2986  -8.905  8.57e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.223 on 29 degrees of freedom
Multiple R-squared:  0.7325,    Adjusted R-squared:  0.714
F-statistic: 39.7 on 2 and 29 DF,  p-value: 4.979e-09
```

Nötiges Hintergrundwissen zu den Daten: Die Variable `cyl` steht für die Anzahl der Zylinder und nimmt lediglich die Ausprägungen 4, 6 und 8 an. Daher machte es auch Sinn, sie hier als kategoriale Regressor in das Modell mit aufzunehmen.

Wieder liegt unser Fokus auf der ersten Spalte ("Estimate") der `Coefficients`-Tabelle: Zunächst fällt auf, dass es zwei Dummyvariablen für die Ausprägungen 6 und 8 gibt. Daran erkennen wir, dass R die Ausprägung 4 als Referenz gewählt hat. \hat{a} in der Zeile `Intercept` ist somit der Mittelwert dieser Referenz. Die Parameter \hat{b}_1 & \hat{b}_2 in den Zeilen `factor(cyl)6` und `factor(cyl)8` geben

die Abweichungen der Mittelwerte für die jeweiligen Kategorien an. Unter **Multiple R-squared** findet man erneut den Wert für R^2 .

⚠ Per default verwendet R stets die Dummy-Kodierung. Um auf die Effektkodierung umzustellen muss man auf das **contrasts**-Argument zurückgreifen. Um das adäquat tun zu können, sollte man sich **cyl** zunächst als **factor**-Variable definieren, und dies nicht erst in der **lm**-Funktion tun:

```
mtcars$fac_cyl <- factor(mtcars$cyl)
modell13 <- lm(mpg ~ fac_cyl, data = mtcars, contrasts = list(fac_cyl = "contr.sum"))
summary(modell13)
```

Call:

```
lm(formula = mpg ~ fac_cyl, data = mtcars, contrasts = list(fac_cyl = "contr.sum"))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-5.2636	-1.8357	0.0286	1.3893	7.2364

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	20.5022	0.5935	34.543	< 2e-16 ***
fac_cyl1	6.1615	0.8167	7.544	2.57e-08 ***
fac_cyl2	-0.7593	0.9203	-0.825	0.416

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.223 on 29 degrees of freedom

Multiple R-squared: 0.7325, Adjusted R-squared: 0.714

F-statistic: 39.7 on 2 and 29 DF, p-value: 4.979e-09

Hier ist zunächst ein bisschen undurchsichtig, was passiert ist. R benennt nun die Dummyvariablen anders, wodurch auf den ersten Blick nicht mehr klar ersichtlich ist, was als Referenz gewählt wurde und welcher Dummy zu welcher Kategorie gehört. Rechnet man jedoch $\hat{a} + \hat{b}_1 = 20,5022 + 6,1615 = 26,6637$ so wird (durch Vergleich mit dem dummykodierte Modell) klar, dass **fac_cyl1** sich auf die Kategorie *4 Zylinder* bezieht. Rechnet man $\hat{a} + \hat{b}_2 = 20,5022 - 0,7593 = 19,7429$ so erhält man dasselbe Ergebnis wie bei $\hat{a} + \hat{b}_1 = 26,6636 - 6,9208 = 19,7428$ aus dem dummykodierte Modell. Somit ist klar, dass sich **fac_cyl2** auf die Kategorie *6 Zylinder* bezieht & die Kategorie *8 Zylinder* damit die Referenz sein muss.

Zur Kontrolle: $\hat{a} + \hat{b}_1 \cdot (-1) + \hat{b}_2 \cdot (-1) = 20,5022 - 6,1615 + 0,7593 = 15,1$ ergibt im effektkodierten Modell dasselbe Ergebnis für die Kategorie *8 Zylinder* wie $\hat{a} + \hat{b}_2 = 26,6636 - 11,5636 = 15,1$ im dummykodierte Modell.

7.7 Aufgaben

1. Welche der folgenden Aussagen sind wahr?

- a) Regressionskoeffizient & Korrelationskoeffizient haben die gleiche Aussagekraft. ☐
- b) Man kann bereits aus dem Korrelationskoeffizienten auf das Vorzeichen des Regressionskoeffizienten schließen. ☐
- c) Höherer Korrelationskoeffizient, bedeutet automatisch auch höherer Regressionskoeffizient. ☐
- d) Sowohl Korrelation als auch Regressionskoeffizient haben einen Wertebereich von -1 bis 1. ☐

2. R^2 bei der linearen Einfachregression ..

- a) .. ist immer kleiner/gleich dem Korrelationskoeffizienten. ☐
- b) .. kann Werte von -1 bis 1 annehmen. ☐
- c) .. beschreibt den Anteil der erklärten Streuung der Zielgröße durch die Einflussgröße. ☐
- d) .. deutet bei kleinen Werten auf eine eher schlechte Modellgüte hin. ☐
- e) .. beschreibt den Anteil der erklärten Streuung des Modells. ☐

3. Was ist bei der Interpretations der geschätzten Koeffizienten im Regressionsmodell wichtig?

- a) Stets nur den Absolutbetrag interpretieren. ☐
- b) Der Intercept ist der erwartete Wert der Zielgröße wenn die Einflussgröße ihren Durchschnittswert annimmt. ☐
- c) Interpretation des Steigungsparameters pro Einheit. ☐
- d) Der Intercept ist nicht immer sinnvoll interpretierbar. ☐
- e) Vorhersagen sollte man nur für Werte der Einflussgröße durchführen, die auch so (ähnlich) in der Stichprobe vorkommen. ☐

4. Welche Aussagen bzgl. kategorialen Regressoren sind korrekt?

- a) Sowohl Dummy- als auch Effekt-Kodierung führen zu gleichen Zahl an Dummy-Variablen. ☐
- b) Die Interpretation der geschätzten Koeffizienten ist bei Dummy- & Effekt-Kodierung identisch. ☐
- c) Der Intercept ist weder bei Dummy- noch bei Effekt-Kodierung interpretierbar. ☐
- d) Bei einer höheren Anzahl an verschiedenen Kategorien ist die Dummy-Kodierung sinnvoller. ☐

5. Interpretation des Intercepts bei Dummy-Kodierung:

- a) Der Intercept entspricht dem erwarteten Wert der Zielgröße bei Vorliegen der Referenzkategorie. ☐
- b) Eine Änderung der Referenzkategorie hat eine Änderung der geschätzten Koeffizienten für alle Dummy-Variablen zur Folge. ☐
- c) Eine Änderung der Referenzkategorie hat (potenziell) eine Änderung der Anpassungsgüte (R^2) zur Folge. ☐
- d) Die Referenzkategorie ist immer auf natürliche Art & Weise vorgegeben. ☐

8 Indizes

8.1 Verhältniszahlen

Verhältniszahlen (Indizes) sind Quotienten aus zwei Maßzahlen, welche grob in Gliederungszahlen, Beziehungszahlen und einfache Index-/Messzahlen zu unterteilen sind.

8.1.1 Gliederungszahlen

Gliederungszahlen sind Quotienten, bei denen der Zähler eine Teilmenge des Nenners ist, z.B. die Erwerbs- oder die Arbeitslosenquote.

8.1.2 Beziehungszahlen

Bei *Beziehungszahlen* ist dies nicht der Fall, jedoch stehen Zähler & Nenner in sachlich sinnvollem Zusammenhang. Ein Beispiel ist die Bevölkerungsdichte (Quotient aus Einwohnerzahl & Fläche).

8.1.3 Indexzahlen

Einfache Indexzahlen beschreiben den Zusammenhang zwischen einer Maßzahl, die zu verschiedenen Zeitpunkten gemessen wurde. Somit sieht man die zeitliche Entwicklung einer Größe, bezogen auf einen Basiszeitpunkt. Dabei bezeichnet x_0 den Wert der Maßzahl in der Basisperiode, x_t den Wert derselben Maßzahl in einer *Berichtsperiode* t . Die Indexzahl ist der Quotient aus dem Wert der Maßzahl in der Bericht- und der Basisperiode.

$$I_{0t} = \frac{x_t}{x_0}$$

Da Zähler und Nenner die gleiche Einheit haben, kürzt sich diese raus und der Index hat keine Einheit. Einfache Indexzahlen werden häufig für Preise (Preismesszahl/Preisindex) oder Mengen (Mengenmesszahl/Mengenindex) berechnet, um deren zeitliche Entwicklung nachzuvollziehen:

$$P_{0t} = \frac{p_t}{p_0}(\text{Preismesszahl}); \quad Q_{0t} = \frac{q_t}{q_0}(\text{Mengenmesszahl})$$

⚠ Preise (p_t) und Mengen (q_t) werden mit Kleinbuchstaben und nur einer Zahl im Index bezeichnet, Indexzahlen für Preise ($P_{0,t}$) und Mengen (Q_{0t}) mit Großbuchstaben und zwei Zahlen im Index (Basis- und Berichtszeitpunkt).

Veränderungen des Basisjahres Bei langen Zeitreihen macht es möglicherweise Sinn, irgendwann ein neues Basisjahr festzulegen (bspw. wegen strukturellen Umbrüchen). Man führt ein neues Basisjahr k ein und berechnet den Index I_{kt} als Quotienten aus Index der Berichtsperiode zur alten Basisperiode (I_{0t}) und dem Index der neuen Basisperiode zur alten Basisperiode (I_{0k}):

$$I_{kt} = \frac{I_{0t}}{I_{0k}}$$

Um die Zeitreihe komplett umzubasieren führt man dies für jeden vorliegenden Zeitpunkt t durch.

8.2 Preisindizes

Zusammengesetzte Indexzahlen verknüpfen einfache Indexzahlen für n verschiedene Güter miteinander. Der naive Ansatz, das Mitteln der Preismesszahlen verschiedener Güter für die verschiedenen Zeitpunkte im Zeitverlauf, wäre problematisch, da alle Güter mit gleicher Gewichtung eingehen würden. Dies wäre jedoch nicht repräsentativ, da nicht für alle Güter die gleichen Mengen angenommen werden. Deshalb wird ein sogenannter repräsentativer **Warenkorb** gebildet, in dem jedes Produkt einzeln, proportional zur jeweiligen Menge, gewichtet wird:

$$P_{0t} = I_{0t}^P(1)\tilde{w}(1) + \dots + I_{0t}^P(n)\tilde{w}(n)$$

Für die Gewichtung existieren zwei Ansätze, der Preisindex nach Laspeyres und der nach Paasche

8.2.1 nach Laspeyres

Der Preisindex nach Laspeyres verwendet für die Gewichtung die Menge aus der *Basisperiode*: Im Zähler steht die Summe der Preise in der Berichtsperiode, jeweils multipliziert mit den Menge aus der Basisperiode, im Nenner dasselbe für die Preise der Basisperiode:

$$P_{0t}^L = \frac{\sum_{i=1}^n p_t(i)q_0(i)}{\sum_{i=1}^n p_0(i)q_0(i)}$$

⚠ Der Preisindex nach Laspeyres kann auch als Summe über die Produkte aus den Preismesszahlen $P_{0t}(i)$ und den Umsatzanteilen aus der *Basisperiode* $\frac{p_0(i)q_0(i)}{\sum_{i=1}^n p_0(i)q_0(i)}$ berechnet werden.

Der Preisindex nach Laspeyres gibt an, wie sich der Wert des Warenkorbs aus Basisperiode in der Berichtsperiode (verglichen mit der Basisperiode) verändert hat.

Der Vorteil dieses Preisindex ist, dass man nach Erheben der Daten für eine neue Berichtsperiode sofort den Preisindex mit früheren Indexwerten vergleichen kann. Der Nachteil ist jedoch, dass der Warenkorb mit der Zeit veraltet, neue Produkte dazukommen und alte außen vorgelassen werden. Daher muss der Warenkorb in regelmäßigen Abständen aktualisiert werden.

8.2.2 nach Paasche

Im Vergleich zum Preisindex nach Laspeyres werden im Preisindex von Paasche die Mengen der Berichtsperiode zur Gewichtung verwendet:

$$P_{0t}^P = \frac{\sum_{i=1}^n p_t(i)q_t(i)}{\sum_{i=1}^n p_0(i)q_t(i)}$$

Der Preisindex nach Paasche gibt an, wie sich der Wert des Warenkorbs aus Berichtsperiode in der Berichtsperiode (verglichen mit der Basisperiode) verändert hat.

Der Vorteil ist, dass der Warenkorb immer aktuell ist, da die Güter und deren Mengen durch jährliche Anpassung nie veralten. Es ist allerdings nicht so einfach diesen Preisindex mit alten Preisindizes zu vergleichen, da man jede Berichtsperiode unterschiedliche Warenkörbe benutzt.

8.3 Mengenindizes

Man erhält den Mengenindizes durch Vertauschen von Preis und Menge, dabei vergleicht man die Mengen von Berichtsperiode & Basisperiode bei gleichbleibendem Preis. Es kann entweder der Preis aus der Basis- (Laspeyres) oder Berichtsperiode (Paasche) herangezogen werden.

8.3.1 nach Laspeyres

Der Mengenindex von Laspeyres wird mit dem konstanten Preis der Basisperiode berechnet.

$$Q_{0t}^L = \frac{\sum_{i=1}^n p_0(i) q_t(i)}{\sum_{i=1}^n p_0(i) q_0(i)}$$

Der Mengenindex nach Laspeyres gibt an, wie sich der Wert des Warenkorbs durch Mengenänderungen verändert hat, bewertet mit gleichbleibenden Preisen aus der *Basisperiode*.

8.3.2 nach Paasche

Der Mengenindex nach Paasche wird mit konstanten Preisen aus der Berichtsperiode berechnet.

$$Q_{0t}^P = \frac{\sum_{i=1}^n p_t(i) q_t(i)}{\sum_{i=1}^n p_t(i) q_0(i)}$$

Der Mengenindex nach Paasche gibt an, wie sich der Wert des Warenkorbs durch Mengenänderungen verändert hat, bewertet mit gleichbleibenden Preisen aus der *Berichtsperiode*.

8.4 Umsatzindex

Beim Umsatzindex berechnet man im Zähler das Produkt aus Preis und Menge der Berichtsperiode und im Nenner das Produkt aus Preis und Menge zur Basisperiode.

$$W_{0t} = \frac{\sum_{i=1}^n p_t(i) q_t(i)}{\sum_{i=1}^n p_0(i) q_0(i)}$$

Der Umsatzindex gibt die Veränderung des Wertes des Warenkorbs in der Berichtsperiode im Verhältnis zum Wert des Warenkorbs aus der Basisperiode an, ohne dabei nach Preis- oder Mengendänderung als Ursache zu differenzieren.

8.5 Spezielle Probleme

8.5.1 Erweiterung des Warenkorbs

In Kapitel 8.2.1 wurde der Nachteil des Preisindex nach Laspeyres betrachtet, dass man den Warenkorb immer wieder aktualisieren muss. Im Folgenden wird erläutert, wie mit diesem Problem neu hinzukommender Waren umgegangen werden kann:

Für die neue Ware liegt üblicherweise kein Preis zur ursprünglichen Basisperiode vor. Man integriert diese dadurch, dass man *nur* für diese Ware die Periode t' , in der sie zum ersten Mal betrachtet wurde, als Basisjahr betrachtet.

Für alle "alten" Güter berechnet man $P_{0t'}^L$, wie gehabt, allerdings nur bis zur Periode t' in der das neue Gut hinzukommt. Ab der Periode t' berechnet man den Index für den erweiterten Warenkorb,

allerdings mit der Besonderheit, dass für alle "alten" Güter weiterhin Periode 0 als Basisperiode dient. Dies führt zu einer Zweiteilung von Zähler & Nenner:

$$P_{t',t'+1}^L = \frac{\sum_{i=1}^n p_{t'+1}(i)q_0(i) + p_{t'+1}(n+1)q_{t'}(n+1)}{\sum_{i=1}^n p_{t'}(i)q_0(i) + p_{t'}(n+1)q_{t'}(n+1)}$$

Um nun $(P_{0,t'+1}^L)$ zu erhalten verkettet man den Index des kleineren Warenkorbs $(P_{0,t'}^L)$ mit dem Preisindex mit dem erweiterten Warenkorb $(P_{t',t'+1}^L)$ durch Multiplikation.

8.5.2 Substitution einer Ware

Man muss eine Substitution durchführen, wenn ein Produkt veraltet und durch eine Innovation ersetzt wird. Ein Beispiel hierfür ist die Substitution des Schwarz-Weiß-Fernsehgerätes durch das Farbfernsehgerät. Bei der Substitution nimmt man an, dass die Mengen bei der Substitution konstant bleiben, sich jedoch die Preise ändern können. Voraussetzung für die Substitution ist, dass für beide Güter in *mindestens einer Periode* t' gleichzeitig die Preise gemessen werden. Dann ist man nämlich in der Lage den Preis des alten Gutes mit der Preissteigerung des neuen Gutes fortzuschreiben:

$$p_t(\text{altes Gut}) = p_{t'}(\text{altes Gut}) \cdot \frac{p_t(\text{neues Gut})}{p_{t'}(\text{neues Gut})} = p_{t'}(\text{altes Gut}) \cdot P_{t't}(\text{neues Gut})$$

Diese Fortschreibung wird dann so lange verwendet, bis der Warenkorb wieder aktualisiert wird und das neue Gut mit aufgenommen werden kann.

8.5.3 Subindizes

Große Warenkörbe bestehen oft aus kleineren Sub-Warenkörben. Dies hilft dabei den Überblick über große Warenkörbe zu behalten. Da die Subkörbe unterschiedlich groß sein können, gehen diese gewichtet in den gesamten Preisindex ein. Dafür berechnet man zuerst den Gesamtumsatz aller Warenkörbe $U = \sum_{i=1}^n p_0(i)q_0(i)$ und berechnet dann jeweils den Anteil eines Sub-Warenkorbes am Gesamtumsatz. Dieser Anteil wird dann bei der Berechnung des Preisindex nach Laspeyres jeweils zu den Werten der Subkörbe multipliziert und die einzelnen Subkörbe addiert. So erhält man dann schlussendlich den Preisindex eines Warenkorbes mit mehreren Subkörben.

8.6 Aufgaben

1. Welche der folgenden Aussagen über Indexzahlen sind wahr?

- a) Zur Umbasierung (Veränderung Basisjahr) werden lediglich die Indexzahlen und nicht die Rohdaten benötigt. ☐
- b) Zur Verkettung von Indexzahlen werden die Rohdaten benötigt ☐
- c) Zur Umbasierung müssen sich alle bereits vorliegenden Indexzahlen auf das gleiche Basisjahr beziehen. ☐
- c) Indexzahlen werden für Mengen und Preise getrennt berechnet. ☐

2. Welche Aussagen bzgl. der verschiedenen Indizes sind wahr?

- a) Der Preisindex nach Laspeyres ist stets größer als der nach Paasche. ☐
- b) Sind die Mengen in Berichts- und Basisperiode gleich, so sind die Preisindizes nach Laspeyres und Paasche ebenfalls identisch. ☐
- c) Bei konstanten Preisen sind sowohl der Mengenindex nach Laspeyres als auch der nach Paasche gleich 1. ☐
- d) Der Mengenindex nach Paasche gewichtet die Mengen mit den Umsatzanteilen aus der Berichtsperiode. ☐

3. Welche Aussagen bzgl. "Spezieller Probleme" sind wahr?

- a) Bei der Substitution muss für mindestens eine Periode der Preis für beide Güter beobachtet werden. ☐
- b) Bei der Erweiterung muss für das neue Produkt auch eine Menge in der Basisperiode bekannt sein. ☐
- c) Bei der Substitution werden die Preissteigerungen des neuen Produkts einfach auf das alte übertragen. ☐
- d) Subindizes können durch Gewichtung mit Mengenanteilen zu einem Gesamtindex kombiniert werden. ☐

9 Zeitreihen

Bei Zeitreihen misst man ein Merkmal wiederholt über die Zeit hinweg und betrachtet wie es sich im Zeitverlauf entwickelt. Diese Entwicklung kann man in *Kurvendiagrammen* darstellen. Dabei befindet sich auf der x-Achse die Zeit (bspw. Tageswerte, Monatswerte, Quartals -/oder Jahreswerte) und auf der y-Achse die Merkmalsausprägungen. Im Folgenden werden einige simple Methoden vorgestellt, mit denen man Zeitreihendaten analysieren kann. Dabei geht man stets von *äquidistanten* Zeitreihen aus, was bedeutet dass die Zeit die zwischen zwei Messungen immer dieselbe ist.

9.1 Zerlegung von Zeitreihen, Komponentenmodell

Zeitreihen eines Merkmals y_t kann man in der Theorie in drei verschiedene Komponenten zerlegen: In die glatte Komponente g_t , die saisonale Komponente s_t und die irreguläre oder Restkomponente r_t . Die glatte Komponente g_t spiegelt den Trend, also die langfristige Entwicklung der Reihe, wieder. Die saisonale Komponente s_t beschreibt die saisonalen Schwankungen (bspw. pro Quartal oder Monat) und erklärt somit wiederkehrende Muster in der Reihe. Die Restkomponente gibt den Anteil an y_t an, der nicht durch die saisonale und die glatte Komponente erklärt beschrieben wird. Diese irregulären Abweichungen sollten jedoch im Mittel 0 sein.


Somit kann die Zeitreihe durch dieses Komponentenmodell (unter der Bedingung, dass die Summe von r_t gleich 0 ist) wie folgt dargestellt werden:

$$y_t = g_t + s_t + r_t .$$

Dies Modell wird als additives Modell bezeichnet, da die Komponenten additiv zusammenhängen. Da es jedoch je nach Datenlage auch zu erforderlich sein kann, einen multiplikativen Ansatz

$$y_t = \tilde{g}_t \cdot \tilde{s}_t \cdot \tilde{r}_t$$

zu verfolgen, wird nachfolgend erläutert wie man unterscheiden kann, welcher Ansatz benötigt wird.

 Es ist möglich, dass multiplikative Modell durch Logarithmieren in additive Schreibweise zu überführen, was mathematisch angenehmer zu handhaben ist.

Um zu unterscheiden welcher Ansatz (additiv vs. multiplikativ) erforderlich ist, sollten die Daten grafisch dargestellt werden um festzustellen ob es im vorliegenden Fall Trend- & Saisonkomponente additiv oder multiplikativ zusammenhängen. In Abbildung 3 sind zwei beispielhafte Szenarien für einen additiven (links) bzw. multiplikativen (rechts) Zusammenhang dargestellt. Von einem additiven Zusammenhang von Trend und Saison kann ausgegangen werden, wenn die saisonalen Schwankungen konstant zu bleiben scheinen, völlig egal ob ein steigender oder fallender Trend vorliegt. Ein multiplikativer Zusammenhang liegt vermutlich vor, falls bei steigendem (fallenden) Trend, die saisonalen Schwankungen im Zeitverlauf entsprechend größer (kleiner) werden. In

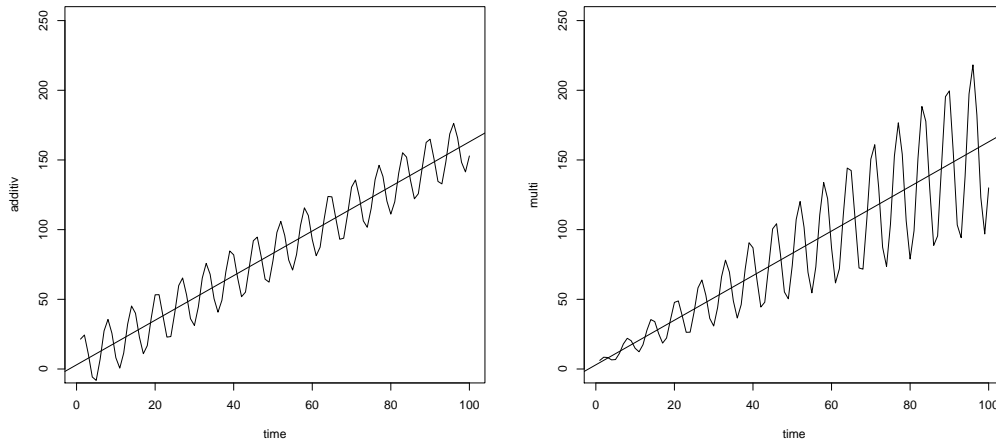


Figure 3: Schaubild für einen additiven (links) bzw. multiplikativen (rechts) Zusammenhang von Trend- und Saisonkomponente

diesem Fall sollte man die Daten logarithmieren und die nachfolgend vorgestellten Methoden auf die logarithmierten Daten anwenden.

9.2 Gleitende Durchschnitte

Gleitende Durchschnitte sind eine Methode um den Trend, also die glatte Komponente g_t , zu schätzen. Durch die Glättung filtert man dabei die saisonalen Schwankungen heraus.

Gleitende Durchschnitte ungerader Ordnung: Um den Wert des gleitenden Durchschnitts für einen bestimmten Zeitpunkt t zu berechnen, berechnet man das arithmetische Mittel (vgl. Kap. 3.5.1) über k Werte vor und k Werte nach diesem Zeitpunkt t , sowie dem Wert y_t selbst. Der Parameter k gibt dabei an wie viele Werte vor und nach dem Wert y_t jeweils in den gleitenden Durchschnitt mit einfließen sollen. Da insgesamt $2 \cdot k + 1$ Werte mit einfließen, spricht man vom gleitenden Durchschnitt $(2k + 1)$ -ter Ordnung.

Beispiel: Würde man jeweils $k = 3$ vor und nach dem Zeitpunkt t mit einfließen lassen, so spräche man von einem gleitenden Durchschnitt 7. Ordnung.

Gleitende Durchschnitte gerader Ordnung: Soll eine gerade Anzahl k an Werten in den gleitenden Durchschnitt mit einfließen, gestaltet sich die Berechnung ein klein wenig komplizierter, da man nun nicht mehr auf natürliche Art und Weise einen mittleren Wert hat wie oben. Daher bedient man sich hier des "Tricks", die Werte am Rand des gewählten Fensters nur mit halbem Gewicht in den Durchschnitt mit einfließen zu lassen. Somit kommt man auf eine gerade Zahl an "ganzen" Werten. Man berechnet also das arithmetische Mittel über je k Werte vor und nach diesem Zeitpunkt t , sowie dem Wert y_t selbst und gewichtet die äußersten Werte mit $\frac{1}{2}$.

Beispiel: Für $k = 2$ fließen fünf Werte in den Durchschnitt mit ein: Die äußersten Werte nur zur Hälfte, die jeweils direkten "Nachbarn" des Wertes y_t ganz und der Wert y_t selbst ebenfalls ganz.

Somit kommt man schließlich auf 4 "ganze" Werte, daher gleitender Durchschnitt 4. Ordnung.

⚠ Obwohl bei obigem Beispiel tatsächlich fünf Werte mit einfließen, wird bei der Berechnung des aithm. Mittels nur durch vier geteilt, da zwei Werte nur jeweils mit Gewicht $\frac{1}{2}$ einfließen.

Diese Berechnung führt man für jeden Zeitpunkt t der Zeitreihe durch und erhält damit eine geglättete Zeitreihe als Schätzung für die Trendkomponente. Wie man sich für eine Ordnung entscheidet, wird in den nachfolgenden Kapiteln erläutert.

⚠ Je höher der gewählte Ordnung, für desto mehr Werte am Rand der Zeitreihe lässt sich kein gleitender Durchschnitt berechnen, da man immer k Werte links & rechts von dem Zeitpunkt t , für den man den gleitenden Durchschnitt berechnen will, braucht.

R-Befehl für den gl. Durchschnitt: <code>> runmean(data, k, ..)</code>	Dokumentation
---	-------------------------------

⚠ Die R-Funktion ist **nicht** Teil von **base-R** sondern Teil des Paketes **caTools**. Dieses muss zunächst installiert (`install.packages("caTools")`) und anschließend importiert werden (`library(caTools)`).

9.3 Saisonale Komponente, konstante Saisonfigur

Gilt $s_t = s_{t+p}$, dann spricht man von einer konstanten Saisonfigur mit Periode p , da nach p Zeitpunkten die Saisonkomponente wieder in etwa den gleichen Wert annimmt. Summiert man alle Werte dazwischenliegenden Werte dieser Periode auf, so sollte 0 herauskommen.

⚠ Auch die Periodendauer p ist etwas, das man durch grafische Darstellung der Zeitreihe herausfinden kann. Man überprüft visuell, ob es saisonale Muster gibt, die sich in regelmäßigen Zeitabständen wiederholen.

9.4 Zerlegung in Trend und Saison

Die saisonale Komponente ist "*eine regelmäßige Wiederholende Schwankung um die glatte Komponente*". Wenn man daher die Ordnung des gleitenden Durchschnitts so wählt, dass sie einem Vielfachen l der Periodendauer entspricht ($2k = l \cdot p$, kann man die Zeitreihe "glätten", also die Saisonkomponente herausrechnen. Übrig bleibt dadurch eine Schätzung für die Trendkomponente g_t . Die kleinste mögliche Ordnung um die Saisonkomponente zu eliminieren entspricht somit der Periodendauer p .

⚠ Höhere Ordnungen $l \cdot p$ sind theoretisch auch möglich (s.o.), würden jedoch dazu führen, dass man für mehr Zeitpunkte am Rand der Reihe keinen gleitenden Durchschnitt berechnen könnte.

Die Differenz d_t zwischen dem urspr. Wert y_t und geschätzten Trendkomponente g_t entspricht bereits grob den saisonale Abweichungen, allerdings noch mit zusätzlichem Fehler.

Durch diese Berechnung der Differenzen erhalten wir nun für jeden Teil der Saisonfigur (bspw. für jedes Quartal) *mehrere* Abweichungen (da in der Zeitreihe z.B. 8 mal ein erstes Quartal auftaucht) für die gilt, dass $d_t \approx d_{t+p}$. Daher mitteln wir für jeden Teil des Saisonfigur alle berechneten Abweichungen, die diesem Teil zuzuordnen sind (bspw. alle d_t die für ein erster Quartal berechnet wurden). Dies führt zu Mittelwerten für jeden der p Teile der Saisonkomponente: $(\bar{d}_1, \dots, \bar{d}_p)$

Da die Summe über alle Teile der Saisonkomponente gleich Null sein muss, müssen diese Mittelwerte noch um die Null zentriert werden: Der Mittelwert der Teile $\bar{d}_1, \dots, \bar{d}_p$, wird von allen Teilen $(\bar{d}_1, \dots, \bar{d}_p)$ subtrahiert. Übrig bleibt dadurch die geschätzte Saisonkomponente \hat{s} .

9.4.1 Trend und Saisonkomponente mit Regression

Eine andere Möglichkeit der Trendschätzung ist die lineare Regression (vgl. Kap. 7). Nimmt man die Zeit t als Einfluss- und die Zeitreihe y_t als Zielgröße können hierdurch z.B. lineare quadratische Trends geschätzt werden. Ein Modell für die Trendkomponente könnte z.B. folgende Form haben:

$$y_t = \hat{a} + \hat{b} \cdot t + e_t$$

Mittels eines multiplen Regressionsmodells (vgl. Kap. 18.2, Statistik II) könnte simultan auch die Saisonkomponente geschätzt werden. Man nimmt dabei zusätzlich zu der Zeit t Dummyvariablen (vgl. Kap. 7.6) für jeden der p Teile der Saisonfigur mit in das Modell auf. Dies führt schlussendlich zu folgendem Trend-Saison-Modell:

$$y_t = a + b \cdot t + \gamma_1 s_1(t) + \dots + \gamma_p s_p(t) + e_t$$

9.5 Alternative Ansätze

Es gibt noch viele weitere (fortgeschrittenere) Ansätze wie Zeitreihen analysiert & modelliert werden können, die jedoch den Umfang & Schwierigkeitsgrad dieser Veranstaltung sprengen würden. Daher nur ein bisschen Name-Dropping:

- Das lokale lineare Trendmodell
- Einbeziehung von trigonometrische Funktionen
- Census X-11 ARIMA
- Census X-12 ARIMA
- BV 4.1

Zu Zeitreihen gibt es eigene ganze Vorlesungen, siehe z.B.:

- [Time series analysis](#)
- [Multivariate time series analysis](#)

9.6 Aufgaben

1. Welche Aussagen bzgl. gleitender Durchschnitte sind wahr?

- a) Durch k wird festgelegt, ob der die Ordnung der gl. Durchschnitt gerade oder ungerade ist. ☐
- b) Bei einem gl. Durchschnitt ungerader Ordnung fallen am Rand der Zeitreihe mehr Werte weg als bei gerader Ordnung. ☐
- c) Je höher die Ordnung, desto mehr Werte fallen am Rand der Zeitreihe weg. ☐
- d) Bei einem gl. Durchschnitt gerader Ordnung gehen (bei gleichem k) mehr Werte in die Berechnung mit ein als bei ungerader Ordnung. ☐

2. Welche Aussagen bzgl. des Zeitreihenmodells sind wahr?

- a) Mit gleitenden Durchschnitten kann die Trendkomponente geschätzt werden. ☐
- b) Mit gleitenden Durchschnitten kann die Saisonkomponente geschätzt werden. ☐
- c) Jede Zeitreihe besitzt eine Trend- und eine Saisonkomponente. ☐
- d) Welche Ordnung für die gleitenden Durchschnitte gewählt wird, hängt von der Saisonkomponente ab. ☐

10 R-Einführung Teil I

Teil der Veranstaltung ist eine Einführung in R sowie eine Computervorlesung. Dieses Kapitel dient als Platzhalter, falls in diesem Skript in Zukunft der Inhalt aus diesem Vorlesungskapitel vertieft werden sollte. Bis dahin wird empfohlen mit den vorhandenen Vorlesungsmaterialien zu arbeiten, da diese bereits sehr ausführlich und weitestgehend selbsterklärend sind.

Statistik II

11 Kombinatorik

Grundlegend ist hier zunächst mal die Unterscheidung von **Permutation & Kombination**. Unter der **Permutation** versteht man eine Möglichkeit, die n Elemente einer Menge M anzuordnen. Von einer **Kombination** spricht man, wenn man nicht (zwingend) alle n Elemente auswählt. Formal ist eine Kombination eine "Möglichkeiten, m Elemente aus n Elementen einer Menge M auszuwählen". In beiden Fällen interessiert man sich für die *Anzahl aller möglichen Permutationen/Kombinationen*.

11.1 Permutation

Da bei der Permutation *alle* Elemente ausgewählt & angeordnet werden, spielt die Reihenfolge per Definition eine Rolle. Man unterscheidet jedoch die Fälle *mit Wiederholung* und *ohne Wiederholung*. Der Begriff *Wiederholung* meint hier, ob ein Element nur einmal oder mehrfach angeordnet werden kann.

Beispiel: Wir möchten die Anzahl der verschiedenen Sitzordnungen im Hörsaal bestimmen.

- a) Zunächst handelt es sich hier (natürlicherweise) um eine Permutation *ohne Wiederholung*, da *alle* Studierenden angeordnet werden und *niemand zwei Plätze einnehmen kann*.
- b) Unterscheiden wir jedoch nicht jeden einzelnen Studierenden, sondern differenzieren nur nach Studienfach (z.B: *BWL*, *VWL*, *Sonstiges*) erzeugen wir den Fall mit Wiederholung. Nun können mehrere Plätze durch z.B. Studierende der VWL eingenommen werden.

Stellt man sich diese Szenarien als Urnen-Experiment vor, so sind im Fall a) z.B. 100 Kugeln in der Urne (also eine pro Studierenden), welche alle angeordnet werden. Im Fall b) wären nur drei Kugeln in der Urne (eine pro Studienfach), welche nach dem Ziehen stets zurückgelegt werden müssten. *Anmerkung:* Zurücklegen nur so oft, wie wir Studierende je Studienfach haben.

11.1.1 Anzahl Permutationen ohne Wiederholungen

Stellt man sich dies als Baumdiagramm vor, so entstehen für die erste Stufe n verschiedene Pfade, für die zweite Stufe jeweils nur noch $n - 1$ (da bereits ein Element angeordnet wurde), für die dritte Stufe jeweils $n - 2$, usw. So geht es weiter bis auf der letzten Stufe jeweils nur noch ein Pfad möglich ist.

Um die Gesamtzahl der möglichen Pfade zu erhalten, muss man die Anzahlen je Stufe miteinander multiplizieren:

$$n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot 1 = n! \quad (\text{ausgesprochen: } n \text{ Fakultät})$$

R-Befehl für Fakultät: <code>> factorial(x)</code>

Dokumentation

11.1.2 Permutation mit Wiederholungen

Bei der Permutation mit Wiederholungen sind nicht mehr alle Elemente unterschiedlich. Daher gibt es (bei gleichem n) hier nun eine geringere Zahl möglicher Permutationen, da es bei einem gleichen

Element egal ist, ob die Elemente z.B. auf Platz 3 und 4 oder auf Platz 4 und 3 sind. Somit dürfen solche beispielhaften Anordnungen nicht mehr doppelt gezählt werden, sondern müssen "herausgerechnet" werden.

Deshalb teilt man $n!$ durch das Produkt der $n_i!$ verschiedenen Gruppengrößen:

$$\frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_s!}$$

Mit n_i berechnet man die Anzahl der möglichen Permutationen in der i -ten Gruppe. Da diese hier nicht mehr von Relevanz sind, werden sie herausgerechnet.

11.2 Kombination

Bei der Kombination werden **nicht mehr** alle n Elemente aus M , sondern nur noch m ausgewählt. Neben der Frage, ob eine Wiederholung möglich ist oder nicht, muss man sich daher hier auch die Frage stellen, ob die Reihenfolge bei der Auswahl eine Rolle spielt oder nicht.

Beispiel: Erweiterung des Beispiels aus Kapitel 11.1.

- a) *Ohne Wiederholung, ohne Reihenfolge:* Wie viele Möglichkeiten gibt es, 10 aus den 100 Studierenden auszuwählen.
- a) *Ohne Wiederholung, mit Reihenfolge:* Wie viele Möglichkeiten gibt es, 10 der 100 Studierenden der Reihe nach auf den Plätzen in der ersten Reihe anzuordnen.
- c) *Mit Wiederholung, ohne Reihenfolge:* Wie viele Möglichkeiten gibt es, 10 aus den 150 Studierenden (jedoch nur unterschieden nach Studienfach, s.o.) auszuwählen.
- d) *Mit Wiederholung, Mit Reihenfolge:* Wie viele Möglichkeiten gibt es, 10 der 150 Studierenden (jedoch nur unterschieden nach Studienfach, s.o.) der Reihe nach auf den Plätzen in der ersten Reihe anzuordnen.

Stellt man sich diese Szenarien als Urnen-Experiment vor, so sind im Fall a) & b) je 100 Kugeln in der Urne (also eine pro Studierenden), von denen in beiden Fällen 10 Stück gezogen werden. Im Fall a) interessiert man sich nicht für die Reihenfolge, in der man die Kugeln zieht, in Fall b) schon. Im Fall c) & d) wären wieder nur drei Kugeln in der Urne (eine pro Studienfach), welche nach dem Ziehen stets zurückgelegt werden müssten. Im Fall c) interessiert man sich nicht für die Reihenfolge, in der man die Kugeln zieht, in Fall d) schon.

⚠ Bei Permutationen wird die Anzahl der möglichen Permutationen durch Wiederholungen geringer, bei Kombinationen wird die Anzahl der möglichen Kombinationen, sowohl für den Fall mit, als auch für den Fall ohne, Berücksichtigung der Reihenfolge, durch Wiederholungen größer.

11.2.1 Kombination ohne Wiederholung und ohne Reihenfolge

Die Anzahl an möglichen Kombinationen berechnet man mit $\binom{n}{m}$. Da es keine Wiederholungen gibt ist in diesem Fall m immer kleiner als n . Im Taschenrechner kann man den Binomialkoeffizient (so heißt die Funktion) mit nCr berechnen z.B. ${}^{24}nCr3 = \binom{24}{3}$.

Man kann den Binomialkoeffizient auch als Permutation mit Wiederholung umschreiben, da man zwei Gruppen hat, zum einen die Gruppe mit den gezogenen Elementen m und die Gruppe $n - m$ mit den nicht gezogenen Elementen. Somit ergibt sich:

$$\binom{n}{m} = \frac{n!}{m! \cdot (n - m)!}$$

R-Befehl für Fakultät: <code>> choose(n, k)</code>

Dokumentation

11.2.2 Kombination ohne Wiederholung und mit Reihenfolge

Dieser Fall ist sehr nahe an *Permutation ohne Wiederholung* (vgl. Kap. 11.1.1), der einzige Unterschied liegt darin, dass das Baumdiagramm nicht "fertig" gezeichnet wird (sondern nach m Stufen abbricht). Multipliziert man den Binomialkoeffizient mit $m!$, so erhält man die passende Formel dafür:

$$m! \cdot \binom{n}{m} = m! \cdot \frac{n!}{m!(n - m)!} = \frac{n!}{(n - m)!}$$

⚠ Die Formel sieht komplizierter aus, als sie letztendlich ist. Im Grund steht im Nenner die Anzahl der nicht ausgewählten Elemente und kürzt dadurch den hinteren Teil von $n!$ weg, sodass diese nach m Stufen abbricht.

11.2.3 Kombination mit Wiederholung und ohne Reihenfolge

Hier kann die Anzahl der ausgewählten Elemente m größer sein als n , da mit Wiederholung gezogen wird. Zusätzlich zum Fall der Kombination ohne Wiederholung vergrößert man künstlich die Menge um $m - 1$ Elemente. Dadurch wird nicht aus n , sondern aus $n + m - 1$ Elementen gezogen:

$$\binom{n + m - 1}{m} = \frac{(n + m - 1)!}{m!(n - 1)!}$$

⚠ Diese Formel mag etwas komisch wirken, eine gute Intuition vermittelt jedoch das Beispiel mit den Eiskugeln aus dem Skript.

11.2.4 Kombination mit Wiederholung und mit Reihenfolge

Betrachtet man diesen Fall wieder als Baumdiagramm, so erhält man einen Baum mit m Stufen und jeweils n Pfaden pro Stufe. Man berechnet die Anzahl der möglichen Kombinationen durch

$$n^m.$$

11.3 Aufgaben

1. Welche Aussagen bzgl. Permutationen sind wahr?

- a) Die Reihenfolge der Elemente kann eine Rolle spielen, muss es aber nicht. ☐
- b) Permutation wird eine mögliche Anordnung von Elementen in einer bestimmten Reihenfolge genannt. ☐
- c) Die Anzahl der Permutationen ohne Reihenfolge und ohne Wiederholung berechnet man mit $n!$ ☐
- d) Die Anzahl der möglichen Permutation mit Wiederholung ist größer als ohne Wiederholungen. ☐

2. Welche Aussagen bzgl. Kombinationen sind richtig?

- a) Die Betrachtung *ohne Wiederholung und ohne Reihenfolge* hat eine größere Anzahl an Kombinationsmöglichkeiten als die Betrachtung *mit Wiederholung und ohne Reihenfolge*. ☐
- b) Wenn die Reihenfolge bei der Kombination mit einbezogen werden soll, dann wird die Anzahl an möglichen Kombinationen größer. ☐
- c) Bei Betrachtung von Kombinationen ohne Wiederholung und mit Reihenfolge erhält man eine höhere Anzahl an Möglichkeiten als bei der Permutation ohne Wiederholung ☐
- d) Bei der Kombination mit Wiederholung und mit Reihenfolge gibt es auf dem "ersten Platz" n verschiedene Möglichkeiten, auf dem "zweiten Platz" $n - 1$, usw. ☐

3. Ein Zahlenschloss besteht aus 4 Rädern mit den Zahlen von 0 bis 9. Welche Aussage ist richtig?

- a) Um die Anzahl der möglichen Kombinationen zu berechnen benutzt man die Permutation mit Wiederholung. ☐
- b) Um die Anzahl der möglichen Kombinationen zu berechnen benutzt man die Kombination mit Wiederholung und mit Reihenfolge. ☐
- c) Um die Anzahl der möglichen Kombinationen zu berechnen benutzt man die Kombination mit Wiederholung und ohne Reihenfolge. ☐
- d) Um die Anzahl der möglichen Kombinationen zu berechnen benutzt man die Kombination ohne Wiederholung und ohne Reihenfolge. ☐
- e) Es gibt 10.000 verschiedene Kombinationen. ☐

4. Bei einem Basketballspiel laufen nacheinander 5 Spieler auf das Spielfeld. Man berechnet die Anzahl der Möglichkeiten für das Einlaufen mit ...

- a) der Kombination ohne Wiederholung und mit Reihenfolge. ☐
- b) der Permutation ohne Wiederholung. ☐
- c) der Permutation mit Wiederholung. ☐
- d) der Kombination ohne Wiederholung und ohne Reihenfolge. ☐
- e) der Kombination mit Wiederholung und mit Reihenfolge. ☐

5. Bei einem Rosenzüchter gibt es 14 verschiedene Rosenarten und man möchte ein Strauß mit 20 Rosen. Wie viele unterschiedliche mögliche Sträuße gibt es?

- a) $2,432902008 \times 10^{18}$ ☐
- b) $5,73166440 \times 10^8$ ☐
- c) $8,366825543 \times 10^{22}$ ☐
- d) $8,71782912 \times 10^{10}$ ☐

12 Wahrscheinlichkeitsrechnung

12.1 Grundlagen & -begriffe

Zunächst ein paar Begriffserklärungen, für das bessere Verständnis:

Ein Elementarereignis ist ein Ereignis, dessen Menge nur aus einem Element besteht (z.B. bei einem Münzwurf: *Kopf* bzw. *Zahl*). Ein **zufälliges Ereignis** ist die Kombination mehrerer Elementarereignisse (z.B. beim Würfel: *Gerade Zahl würfeln* = $\{2, 4, 6\}$).

Der Ereignisraum Ω ist die Menge aller Elementarereignisse (z.B. bei einem Münzwurf die Menge $\{\text{Kopf}, \text{Zahl}\}$).

Ein unmögliches Ereignis ist ein Ereignis, dass kein Elementarereignis enthält (z.B. beim Münzwurf $\{\text{Kreuz}\}$ oder $\{\text{Pik}\}$), ein **sichere Ereignis** hingegen enthält alle möglichen Elementarereignisse (also Ω).

Zufällige Ereignisse können nicht nur alleine auftreten, da diese Mengen von Elementarereignissen sind. Zur Verknüpfung benötigt man sog. Mengenoperation:

A geschnitten B $[A \cap B]$ Dies ist die *Schnittmenge* der beiden Mengen A und B , d.h. diejenigen Elementarereignisse, die sowohl in A als auch in B enthalten sind. Ist $A \cup B = \emptyset$ (d.h. die Schnittmenge ist leer), so spricht man von **disjunkten Ereignissen**.

A vereinigt B $[A \cup B]$ Dies ist die Vereinigungsmenge, d.h. alle Elementarereignisse, die entweder in A oder in B (oder in beiden) enthalten ist. Hierfür werden beide Mengen addiert, wobei Elementarereignisse, die in beiden Mengen vorkommen, nur einmal gezählt werden dürfen.

Nicht A $[\bar{A}]$ Das zufällige Ereignis \bar{A} (sprich: A quer) enthält ausschließlich diejenigen Elementarereignisse, welche in A enthalten sind.

A ohne B $[A \setminus B]$ Die Ereignis bezeichnet alle Elementarereignisse, welche in A , jedoch nicht in B enthalten sind. $B \setminus A$ wären alle, die in B , jedoch nicht in A enthalten sind.

12.2 Relative Häufigkeit

Um (ohne theoretisches Vorwissen) eine Quantifizierung vorzunehmen, wie häufig ein Ereignis zu erwarten ist bzw. wie wahrscheinlich der Ausgang eines Versuchsergebnisses ist, betrachtet man relative Häufigkeiten (vgl. Kap. 2.1.3). Zieht man dabei ausreichend viele Wiederholungen des Zufallsexperiments heran, so nähert sich die relative Häufigkeit der Wahrscheinlichkeit von A an. Diese wird als $P(A)$ geschrieben.

12.3 Laplacesche Wahrscheinlichkeit

Als Laplace-Experiment bezeichnet man ein Zufallsexperiment, bei dem alle Elementarereignisse die gleiche Wahrscheinlichkeit besitzen (& der Ereignisraum endlich ist). Dadurch kann man z.B. die Wahrscheinlichkeit für Ereignis A berechnen, indem man einfach den Quotient aus der Anzahl der für A günstigen Fälle und der Anzahl aller möglichen Elementarereignisse bilden: $\frac{|A|}{|\Omega|} = P(A)$.


Beispiel: Der Würfelwurf ist ein Laplace-Experiment, da jede Zahl die gleiche Wahrscheinlichkeit von $\frac{1}{6}$ besitzt und somit z.B. $P(\text{gerade Zahl}) = \frac{3}{6}$ berechnet werden kann.

12.4 Axiome

Das Axiomensystem von Kolmogorov bietet die formale Grundlage für die Wahrscheinlichkeitsrechnung und besteht aus drei Axiomen.

1. **Axiom:** Jede Wahrscheinlichkeit liegt zwischen 0 und 1.
2. **Axiom:** Die Wahrscheinlichkeit für das sichere Ereignis (vgl. Kap. 12.1) beträgt 1.
3. **Axiom:** Die Wahrscheinlichkeiten zweier disjunkter Ereignisse können einfach addiert werden um die Wahrscheinlichkeit der Vereinigungsmenge zu bestimmen.

Folgerungen: Aus diesen Axiomen lassen sich 5 Folgerungen ableiten:

1. W'keit des Komplementärereignisses: $P(\bar{A}) = 1 - P(A)$, da $A \cup \bar{A} = \Omega$
2. W'keit des unmöglichen Ereignisses: $P(\emptyset) = 0$, da $\emptyset = \bar{\Omega}$ & $P(\Omega) = 1$
3. W'keit der Vereinigungsmenge: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, da die Schnittmenge sowohl in $P(A)$, als auch in $P(B)$ reinspielt & sonst quasi "doppelt" zählen würde.
 Sind A & B disjunkt, reduziert sich dies auf das dritte Axiom, da $P(A \cap B) = 0$
4. W'keit der Teilmenge: Wenn A Teilmenge von B ist, so gilt: $P(A) \leq P(B)$
5. Kann man Ω vollständig in disjunkte Teile A_1, \dots, A_n zerlegen, so gilt für Ereignis B :
 $P(B) = \sum_{i=1}^n P(B \cap A_i)$, d.h. $P(B)$ kann vollständig in die W'keiten der Schnittmengen mit den A_i zerlegt werden.

12.5 Bedingte Wahrscheinlichkeit

Bei der bedingten Wahrscheinlichkeit handelt es sich um eine Wahrscheinlichkeit eines Ereignisses B , bei der schon als Vorinformation bekannt ist, dass ein Ereignis A bereits eingetreten ist. Nun ist die Frage, ob das eingetretene Ereignis A eine Auswirkung auf $P(B)$ hat und wie man diese Auswirkung bei der Berechnung von $P(B)$ berücksichtigt. Formal schreibt man diese sog. **bedingte Wahrscheinlichkeit** als $P(B|A)$.

Man berechnet sie aus dem Quotient der Wahrscheinlichkeit der Schnittmenge (in diesem Fall von A und B) und der Wahrscheinlichkeit für die Bedingung, also hier von $P(A)$.

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

⚠ Andersherum funktioniert es natürlich genauso, d.h. $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Stellt man diese Formel nach der gemeinsamen Wahrscheinlichkeit um, so ergibt sich der sog. **Multiplikationssatz**:

$$P(A \cap B) = P(B|A) \cdot P(A) \quad \text{bzw.} \quad P(A \cap B) = P(A|B) \cdot P(B)$$

Die bedingt Wahrscheinlichkeit, sowie deren Umformung, sind von zentraler Bedeutung für die Herleitung der nächsten beiden Sätze.

12.5.1 Satz von der totalen Wahrscheinlichkeit

Möchte man eine Wahrscheinlichkeit eines Ereignisses berechnen, kann man den Satz der totalen Wahrscheinlichkeit verwenden. Grundlage ist Folgerung 5 aus Kapitel 12.4, in der man die Wahrscheinlichkeit des Ereignisses B als Summe der Wahrscheinlichkeiten der Schnittmengen von B mit jedem disjunkt zerlegten Teil A_i des Ereignisraumes Ω berechnet.

Da die Wahrscheinlichkeit der Schnittmenge auch durch bedingten Wahrscheinlichkeiten berechnet werden kann, setzt man das Produkt aus jeder bedingten Wahrscheinlichkeit von B unter der Bedingung A_i ein & multipliziert dies jeweils mit der Wahrscheinlichkeit der Vorinformation A_i . Anschließend werden diese Produkte (wie in Folgerung 5) aufsummiert, um so die Wahrscheinlichkeit für das Ereignis B zu erhalten:

$$P(B) = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i)$$

12.5.2 Der Satz von Bayes

Den Satz von Bayes wird verwendet, um eine bedingte Wahrscheinlichkeit zu berechnen. Verwendet man beide Versionen des Multiplikationssatzes und setzt diese über die gemeinsame Wahrscheinlichkeit gleich, so erhält man:

$$P(B|A) \cdot P(A) = P(A|B) \cdot P(B)$$

Umgestellt nach einer der beiden bedingten Wahrscheinlichkeiten ergibt sich:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad \text{bzw.} \quad P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

⚠ Ist eine **unbedingte Wahrscheinlichkeit** gesucht, so ist meistens der Satz der totalen Wahrscheinlichkeit hilfreich, sucht man eine **bedingte Wahrscheinlichkeit** so benötigt man meist den Satz von Bayes

12.6 Stochastische Unabhängigkeit

Stochastische Unabhängigkeit liegt vor, wenn die bedingte Wahrscheinlichkeit eines Ereignisses gleich der unbedingten Wahrscheinlichkeit ist, d.h. $P(B) = P(B|A)$. Dies bedeutet, dass das Eintreten eines Ereignisses A kein Einfluss auf die Wahrscheinlichkeit für das Eintreten eines Ereignisses B hat. Grob gesagt heißt das: Egal ob das Ereignis A eintritt oder nicht, weiß ich immer noch nicht mehr über die Wahrscheinlichkeit des Eintretens des Ereignisses B als davor. Die Intuition ist dabei ähnlich zu der in Kapitel 6.2, jedoch bezieht sich dieser Begriff auf *Wahrscheinlichkeiten* und nicht mehr auf *relative Häufigkeiten* (daher auch *stochastische Unabhängigkeit*). Überprüfbar ist stochastische Unabhängigkeit zwischen zwei Ereignissen A & B mittels

$$P(A \cap B) \stackrel{?}{=} P(A) \cdot P(B) .$$

Gilt diese Gleichung, so können A & B als stochastisch unabhängig bezeichnet werden.

12.7 Aufgaben

1. Eine Mutter kauft für den Kindergeburtstag 20 Luftballons, 10 blaue und 10 rote. Welche Aussagen sind richtig?

- a) Ein sicheres Ereignis wäre entweder die Farbe blau oder rot als erstes aufzublasen. ☐
- b) Ein unmögliches Ereignis wäre die Farbe gelb als erstes auszublasen. ☐
- c) Das Komplementärereignis \bar{A} von Ereignis A "*Einen blaue Luftballon als erstes aufblasen*" ist "*Keinen Luftballon aufblasen*". ☐
- d) Bei der Frage, welchen Luftballon die Mutter als erstes aufbläst, gibt es hier zwei Elementarereignisse. ☐
- e) Das Ereignis "*Blauen Luftballon als erstes aufblasen*" ist ein Elementarereignis. ☐

2. Welche Aussagen zur Laplaceschen Wahrscheinlichkeit sind richtig?

- a) Bei der Laplaceschen Wahrscheinlichkeit kann der Ereignisraum unendlich sein, solange die Ereignisse gleich wahrscheinlich sind. ☐
- b) Die einzige Voraussetzung für ein Laplacesche Wahrscheinlichkeit ist, dass die Ereignisse gleich wahrscheinlich sind. ☐
- c) Mit der Anzahl der für A günstigen Fälle und der Anzahl aller möglichen Ereignisse berechnet man die Laplacesche Wahrscheinlichkeit. ☐
- d) Bei der Laplaceschen Wahrscheinlichkeit muss ein Zufallsexperiment zugrunde liegen. ☐

3. Welche Aussagen zur Wahrscheinlichkeitsrechnung sind wahr?

- a) Die Wahrscheinlichkeit eines unmöglichen Ereignisses ist die leere Menge. ☐
- b) Die Wahrscheinlichkeit der Schnittmenge zweier disjunkter Ereignisse ist die Summe der Einzelwahrscheinlichkeiten abzüglich der Schnittmenge der beiden Ereignisse. ☐
- c) Wenn man die Wahrscheinlichkeit des Ereignisses A kennt, kann man auch die Wahrscheinlichkeit des Komplementärereignisses \bar{A} berechnen. ☐
- d) Die Wahrscheinlichkeit eines Ereignisses A kann Zahlen zwischen -1 und 1 annehmen. ☐
- e) Ist B eine Teilmenge von A , so ist die Wahrscheinlichkeit von B kleiner oder gleich der Wahrscheinlichkeit von A . ☐

4. Welche Aussagen zur bedingten Wahrscheinlichkeit sind wahr?

- a) Den Satz der totalen Wahrscheinlichkeit verwendet man, wenn man die unbedingte Wahrscheinlichkeit eines Ereignisses berechnen möchte. ☐
- b) Man kann den Satz von Bayes verwenden, um eine bedingte Wahrscheinlichkeit zu berechnen. ☐
- c) Die bedingte Wahrscheinlichkeit $P(E|G)$ bedeutet die Wahrscheinlichkeit des Ereignisses E , unter der Bedingung, dass das Ereignis G nicht eingetreten ist. ☐
- d) Wenn die bedingte Wahrscheinlichkeit $P(A|B)$ ungleich der bedingten Wahrscheinlichkeit $P(B|A)$, dann sind A & B disjunkt. ☐

5. Welche Aussagen zur stochastischen Unabhängigkeit sind wahr?

- a) Wenn die bedingte Wahrscheinlichkeit $P(A|B)$ ungleich der bedingten Wahrscheinlichkeit $P(B|A)$, dann sind A & B unabhängig. ☐
- b) Ist die bedingte Wahrscheinlichkeit $P(B|G)$ ungleich der Wahrscheinlichkeit $P(B|\bar{G})$, dann liegt keine stochastische Unabhängigkeit zwischen B und G vor. ☐
- c) Ist die bedingte Wahrscheinlichkeit $P(B|G)$ ungleich der Wahrscheinlichkeit $P(B)$, dann liegt keine stochastische Unabhängigkeit zwischen B und G vor. ☐
- d) Liegt stochastische Unabhängigkeit vor, so berechnet man die gemeinsame Wahrscheinlichkeit als die Summe der beiden Einzelwahrscheinlichkeiten. ☐

13 Zufallsvariablen

In der deskriptiven Statistik (Statistik I) haben wir fest vorgegebenes Datenmaterial beschrieben. Im Gegensatz dazu gehen wir in der induktiven Statistik (Statistik II) von Zufallsexperimenten mit sogenannten Zufallsvariablen aus. Mit Zufallsvariablen können Ergebnisse eines noch nicht durchgeführten Zufallsexperiments beschrieben werden.

Beispiel: Roulette Spiel mit der Zufallsvariable Z ("groß Z "): Bevor wir das Roulette drehen, ist der Wert von Z nicht bekannt, sondern nur Wahrscheinlichkeiten über die möglichen Ausprägungen. Nach einer Runde nimmt die Variable Z aber einen Wert zwischen $0, 1, 2, \dots, 36$ an. Dieser Wert heißt auch Realisierung der Zufallsvariable und wird mit z ("klein z ") beschrieben. Grundsätzlich gibt es zwei verschiedene Arten von Zufallsvariablen – **diskrete** und **stetige**.

13.1 Diskrete Zufallsvariablen

Wenn Zufallsvariablen nur eine endliche oder abzählbar unendliche Menge an Werten annehmen, spricht man von diskreten Zufallsvariablen. Das bedeutet grob, dass es z.B. eine fixe Anzahl an Werten gibt (wie z.B. beim Roulette-Spiel oder Würfelwurf), oder dass es sich um sogenannte Zählraten handelt, wie etwa die Anzahl an Versicherungsschäden an einem bestimmten Tag. Jedoch ist zu beachten, dass theoretisch beliebig hohe Werte möglich sind, diese jedoch abzählbar sein müssen.

Abzählbarkeit: Mächtigkeit von Mengen Für Interessierte gibt es hier einen kleinen [Ausflug in die Mathematik](#)

Wahrscheinlichkeitsfunktion Die Funktion $P(X = x_i = p_i)$ ordnet jeder möglichen Ausprägung x_i einer diskreten Zufallsvariable X eine Wahrscheinlichkeit p_i zu. (*Anmerkung:* $\sum_{i=1}^n p_i = 1$)

13.2 Stetige Zufallsvariablen

Stetige Zufallsvariablen sind innerhalb eines beliebigen Intervalls definiert und können unendlich viele verschiedene Werte annehmen. Man nennt diesen Wertebereich überabzählbar unendlich.

Beispiel: Geschwindigkeit von Autos bei einer Radarkontrolle. Hier sind theoretisch unendlich viele Werte zwischen z.B. 50km/h und 60km/h möglich. Obwohl natürlich die Messgenauigkeit durch das verwendete Messgerät beschränkt ist (z.B. auf 1km/h genau), werden solche Variablen in der Statistik also meist als stetige Variablen behandelt.

Dichtefunktion Das stetige Analogon zur Wahrscheinlichkeitsfunktion beschreibt stetige Zufallsvariablen. (*Anmerkung:* $\text{Muss stets } \geq 0 \text{ sein \& } \int_{-\infty}^{\infty} f(x) dx = 1$)

⚠ Bei stetigen Zufallsvariablen sind die Punktwahrscheinlichkeiten stets Null, d.h. wir können lediglich Aussagen über Wahrscheinlichkeiten für Intervalle treffen.

13.3 Träger einer Zufallsvariablen

Der Träger einer Zufallsvariablen bezeichnet alle möglichen Ergebnisse einer Zufallsvariablen. Beim Roulette-Spiel wäre der Träger z.B. $T = 0, 1, 2, \dots, 36$. Für die Geschwindigkeit bei der Radarkontrolle kommen theoretisch alle positiven reellen Zahlen in Frage, hier wäre der Träger also die Menge der reellen Zahlen \mathbb{R}^+ .

13.4 Verteilungsfunktion

Die Interpretation ist relativ ähnlich zur empirischen Verteilungsfunktion aus Kapitel 2.4, mit dem Unterschied, dass wir hier von *Wahrscheinlichkeiten* anstatt von relativen Häufigkeiten sprechen. D.h. es handelt sich um ein rein theoretisches Konstrukt (im Gegensatz zum *empirischen* Verteilungsfunktion).

⚠ Bei diskreten Zufallsvariablen wird über die Punktwahrscheinlichkeiten aufsummiert (\rightarrow Treppenfunktion), bei stetigen Zufallsvariablen über die Dichtefunktion integriert (\rightarrow glatter Verlauf).

13.5 Erwartungswert & Varianz

Ähnlich der zentralen Lagemaße für *Daten* in der deskriptiven Statistik (vgl. Kap. 3.2 oder 3.5.1) dient der Erwartungswert zur Beschreibung des Schwerpunktes der Verteilung von *Zufallsvariablen*. Die Varianz dient dazu, die Streuung von Verteilungen um deren Erwartungswert zu quantifizieren, im Gegensatz zur *empirischen* Varianz (vgl. Kap. 4.4) welche die Streuung von *Daten* um deren Mittelwert quantifiziert.

⚠ Auch hier ist bei der Berechnung wieder die Fallunterscheidung zwischen diskreten Zufallsvariablen (*Summation*; eher simpel) und stetigen Zufallsvariablen (*Integration*; eher aufwändig) nötig.

13.6 Zweidimensionale Zufallsvariablen

Zweidimensionale diskrete Verteilungen können über eine Kontingenztafel (vgl. Kap. 6.1) dargestellt werden, für den stetigen Fall Bedarf es einer gemeinsamen Dichtefunktion $f_{XY}(x, y)$.

Das Prinzip der Unabhängigkeit ist ähnlich definiert wie in Kapitel 6.2, in diesem Fall jedoch wieder umgemünzt auf Wahrscheinlichkeiten anstatt relative Häufigkeiten. Ebenso einfach kann die Idee für die Kovarianz & die Korrelation übertragen werden.

13.7 Aufgaben

Welche Aussagen zu den Zufallsvariablen sind richtig?

- a) Der fünffache Münzwurf ist eine diskrete Zufallsvariable. ☐
- b) Mit Zufallsvariablen können Ergebnisse von Zufallsexperimenten beschrieben werden, die noch nicht durchgeführt wurden. ☐
- c) Der Träger der Zufallsvariable Würfelwurf eines 24-seitigen Würfels ist $T = \{1, 2, 3, \dots, 24\}$ ☐
- d) Eine stetige Zufallsvariable hat abzählbar endliche viele mögliche Ergebnisse. ☐

14 Spezielle Verteilungen

14.1 Diskrete Verteilungen

Diskret heißt grob gesagt, dass ein Experiment eine endliche Zahl an möglichen Ergebnissen hat. Beispiele für diskrete Verteilungen in der Vorlesung sind die diskrete Gleichverteilung, Bernoulliverteilung, Binomialverteilung, (hyper-)geometrische Verteilung, Poissonverteilung und die Multinomialverteilung. Diese diskreten Verteilungen beschreiben Wahrscheinlichkeiten, mit denen die einzelnen Werte von diskreten Zufallsvariablen (=ZV) auftreten können. Diese ZV besitzt abzählbar viele Werte (Beispiel Würfelwurf mit den Zahlen 1–6).

14.1.1 Diskrete Gleichverteilung

Idee/Anwendung: Modellierung einer diskreten ZV, deren mögliche Ausprägungen alle mit derselben Wahrscheinlichkeit auftreten.

Ein Beispiel dafür ist das Würfeln mit einem fairen Würfel. *Fair* bedeutet in diesem Fall, dass jede Augenzahl (das sind die möglichen Ausprägungen) mit derselben Wahrscheinlichkeit vorkommt.

Verteilung:

Wahrscheinlichkeitsfunktion: $P(X = x_i) = p_i = \frac{1}{k}, \quad i = 1, 2, \dots, k$

→ Hier ist die Wahrscheinlichkeitsfunktion an der Stelle x_i einfach die Wahrscheinlichkeit, dass die ZV den Wert x_i annimmt. Und diese Wahrscheinlichkeit ist für alle $x_i, i = 1, 2, \dots, k$, gleich und hat dementsprechend jeweils den Wert $\frac{1}{k}$.

Erwartungswert und Varianz: $E[X] = \frac{k+1}{2}$ und $Var[X] = \frac{1}{12}(k^2 - 1)$

Bemerkung: k ist die Anzahl der möglichen Ausprägungen, d.h. die Mächtigkeit des Ereignisraums

14.1.2 Bernoulliverteilung

Idee/Anwendung: Ein einziges Experiment mit nur zwei möglichen Ergebnissen, wobei wir 0 für "Misserfolg" und 1 für "Erfolg" kodieren. Erfolg tritt mit Wahrscheinlichkeit p und Misserfolg mit der entsprechenden Gegenwahrscheinlichkeit $(1 - p)$ auf. Formell: $P(X = 1) = p$ und $P(X = 0) = 1 - p$.

Ein Beispiel ist der einfache Münzwurf einer fairen Münze. Hierbei spielt es keine Rolle, ob man Kopf oder Zahl als Erfolg (Misserfolg) wählt, es sei denn, man hat eine weitere, persönliche Interpretation der Ereignisse. Wenn man z.B. in einer Wette auf Kopf setzt, ist es sinnvoll, dem Ereignis *Kopf* den Wert 1, also "Erfolg", zuzuweisen.

Verteilung:

Notation: $X \sim B(1, p)$

$$\text{Wahrscheinlichkeitsfunktion: } P(X = x) = \begin{cases} p & \text{für } x = 1, \\ 1 - p & \text{für } x = 0 \end{cases}$$

$$\text{bzw. } P(X = x) = p^x(1 - p)^{1-x} \text{ für } x \in \{0, 1\}$$

→ Wahrscheinlichkeitsfunktion nimmt für beide möglichen Werte (es handelt sich ja um eine binäre ZV) die entsprechende Eintrittswahrscheinlichkeit an. Ansonsten ist sie nicht definiert (was für uns so viel bedeutet, dass sie dort einfach "0" ist).

Erwartungswert und Varianz: $\mathbb{E}[X] = p$ und $\text{Var}[X] = p(1 - p)$

Bemerkung: Binomialverteilung ($B(n, p)$) mit $n = 1$.

14.1.3 Binomialverteilung

Idee/Anwendung: Die Binomialverteilung entsteht durch mehrmaliges (n -maliges) Wiederholen desselben Bernoulli-Experiments. Außerdem gilt, dass diese Wiederholungen unabhängig voneinander sind. Von Interesse ist hierbei die gesamte Anzahl der Erfolge, nicht jedoch die konkrete Abfolge von Erfolgen und Misserfolgen.

Ein Beispiel ist der wiederholte Münzwurf. Angenommen, man ist daran interessiert, wie oft das Ereignis *Kopf* bei 100 Münzwürfen vorkommt. Dann wählt man *Kopf* als "Erfolg" im einfachen Bernoulli-Experiment, d.h. man weist *Kopf* den Wert 1 zu. Die entsprechende Wahrscheinlichkeit bei einer fairen Münze ist $P(X = 1) = 0,5$. Da man an der Anzahl von *Kopf* bei 100 Münzwürfen interessiert ist, setzt man nun noch $n = 100$ in der Binomialverteilung.

Verteilung:

Notation: $X \sim B(n, p)$

Wahrscheinlichkeitsfunktion: $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, $k = 0, 1, \dots, n$

→ Die Wahrscheinlichkeitsfunktion lässt sich intuitiv in ihre einzelnen Faktoren zerlegen:

$\binom{n}{k}$: Dies ist die Anzahl der möglichen Kombinationen, bei denen genau k Erfolge bei n Versuchen eintreten. Der Binomialkoeffizient (Ziehen ohne Zurücklegen und ohne Reihenfolge) wird benutzt, weil uns die Stellen, an denen der Erfolg eintritt, egal sind. So kann z.B. in den ersten k Versuchen Erfolg eintreten und in den letzten $n - k$ nicht. Genauso könnte auch in den ersten $n - k$ Versuchen kein Erfolg eintreten, dafür aber in den letzten k .

p^k : Wenn wir uns unabhängige Ereignisse anschauen, multiplizieren wir einfach die entsprechenden Wahrscheinlichkeiten für das, was eintreten soll, auf. p^k ist das Produkt der k Erfolgswahrscheinlichkeiten für die k Erfolge.

$(1-p)^{n-k}$: Da außerdem $n-k$ Misserfolge eintreten, müssen diese mit $n-k$ Misserfolgswahrscheinlichkeiten, gesammelt in dem Faktor $(1-p)^{n-k}$, berücksichtigt werden.

Erwartungswert und Varianz: $\mathbb{E}[X] = np$ und $\text{Var}[X] = np(1-p)$

Bemerkung: Kann durch $\mathcal{N}(np, np(1-p))$ approximiert werden (vgl. Kapitel 15.2.1).

14.1.4 Geometrische Verteilung

Idee/Anwendung: Die geometrische Verteilung modelliert die *Anzahl der Versuche* bis zum ersten Eintreten des Erfolgs bei unabhängigen (identischen) Bernoulliversuchen.

Ein Beispiel ist hier wieder der wiederholte Münzwurf. Angenommen, man ist daran interessiert, wie oft man eine faire Münze werfen muss, bis das Ereignis *Kopf* vorkommt, dann wählt man erneut *Kopf* als "Erfolg" im einfachen Bernoulli-Experiment. Da die Wahrscheinlichkeit $P(X=1) = 0,5$ ist, erwartet man in diesem Fall, dass zwei Versuche genügen, um den ersten Erfolg zu erzielen.

Verteilung:

Notation: $X \sim G(p)$

Wahrscheinlichkeitsfunktion: $P(X=k) = p(1-p)^{k-1}$, $k \in \mathbb{N}$

mit $P(X \leq k) = 1 - (1-p)^k$

→ Die Wahrscheinlichkeitsfunktion lässt sich intuitiv in ihre einzelnen Faktoren zerlegen:

$(1-p)^{k-1}$: Es geht hier um die Wahrscheinlichkeit, genau im k -ten Versuch den ersten Erfolg zu erzielen. Dementsprechend müssen bis dahin $k-1$ Misserfolge aufgetreten sein. Diese $k-1$ Misserfolge (in *unabhängigen* Bernoulli-Versuchen) treten genau mit der Wahrscheinlichkeit $\underbrace{(1-p) \cdot (1-p) \cdot \dots \cdot (1-p)}_{k-1 \text{ mal}} = (1-p)^{k-1}$ auf.

p : Dann fehlt nur noch die Wahrscheinlichkeit, dass im k -ten Versuch der Erfolg auch eintritt. Dafür wird dann einmal mit der Erfolgswahrscheinlichkeit p multipliziert.

Erwartungswert und Varianz: $\mathbb{E}[X] = \frac{1}{p}$ und $\text{Var}[X] = \frac{1}{p} \left(\frac{1}{p} - 1 \right)$

Bemerkung:

14.1.5 Hypergeometrische Verteilung

Idee/Anwendung: n -maliges Ziehen (ohne Zurücklegen!) aus einem Topf mit N Kugeln, von denen M ($\leq N$) Kugeln das gewünschte Merkmal tragen. Es ist hierbei egal, ob die anderen $N-M$ Kugeln alle dasselbe Merkmal oder verschiedene Merkmale tragen. Die Anzahl der gezogenen

Kugeln mit gewünschtem Merkmal nach n -maligem Ziehen folgt dann einer hypergeometrischen Verteilung.

Ein Beispiel ist hier das Lottospielen. Es gibt insgesamt $N = 49$ Kugeln, von denen $M = 6$ das Merkmal *Auf Lottoschein angekreuzt* tragen und die anderen $N - M = 43$ nicht. Die Anzahl der richtigen Zahlen im Lotto bei $n = 6$ Zügen ist entsprechend hypergeometrisch verteilt ($H(6, 6, 49)$).

Verteilung:

Notation: $X \sim H(n, M, N)$

Wahrscheinlichkeitsfunktion: $P(X = x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$

→ Die Wahrscheinlichkeitsfunktion lässt sich intuitiv in ihre einzelnen Faktoren zerlegen, wobei im Zähler die Anzahl der günstigen Fälle und im Nenner die Anzahl der insgesamt möglichen Fälle stehen:

$\binom{M}{x}$: Der erste Binomialkoeffizient steht für alle möglichen Kombinationen, von den M Einheiten mit dem gewünschten Merkmal genau x zu ziehen.

$\binom{N-M}{n-x}$: Bei N Einheiten und M mit gewünschtem Merkmal bleiben $N - M$ weitere Einheiten übrig. Aus diesen sollen noch $n - x$ Einheiten stammen, also die Anzahl der Züge minus der gewünschten Anzahl der Einheiten mit dem Merkmal (es sind ja bereits x Züge "verbraucht").

$\binom{N}{n}$: Das Produkt der oberen beiden Binomialkoeffizienten ergibt die Anzahl der günstigen Fälle. Da wir eine Wahrscheinlichkeit berechnen, muss diese noch durch die Anzahl der möglichen Fälle geteilt werden. Und diese Gesamtzahl sind einfach alle Möglichkeiten, n -mal aus N zu ziehen.

Erwartungswert und Varianz: $\mathbb{E}[X] = n \frac{M}{N}$ und $\text{Var}[X] = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \left(\frac{N-n}{N-1}\right)$

Bemerkung: Kann durch $B\left(n, \frac{M}{N}\right)$ approximiert werden (vgl Kapitel 15.2.4).

14.1.6 Poissonverteilung

Idee/Anwendung: Mit der Poissonverteilung kann man die Anzahl von Ereignissen in einem gegebenen Zeitintervall modellieren. Man zählt also die Anzahl der Ereignisse, die in einem fest vorgegebenen Zeitintervall eintreten, und möchte die Wahrscheinlichkeiten modellieren, mit der genau x Ereignisse in diesem Zeitraum auftreten. Die Länge des Zeitintervalls kann dabei je nach Anwendung unterschiedlich gewählt werden (z.B. 1 Sekunde, 1 Minute, 1 Stunde, 1 Tag oder auch 1 Jahr).

Beispiele sind die **Anzahl** von Fischen die **täglich** unter einem Sensor in der Isar vorbeischwimmen oder **Anzahl** von Haftpflichtfällen **pro Jahr** eines Versicherungsnehmers.

Verteilung:

Notation: $X \sim Po(\lambda)$, $\lambda > 0$

Wahrscheinlichkeitsfunktion: $P(X = x) = \frac{\lambda^x}{x!} \cdot \exp(-\lambda)$, $x \in \mathbb{N}_0$

Erwartungswert und Varianz: $\mathbb{E}[X] = \lambda$ und $Var[X] = \lambda$

Bemerkung: Kann durch $\mathcal{N}(\lambda, \lambda)$ approximiert werden (vgl. Kapitel 15.2.3). Außerdem gilt, dass die Wartezeit bis zum Eintreten des ersten Ereignisses exponentialverteilt mit Parameter λ ist, wenn die Anzahl der Ereignisse in einem fixem Kontinuum (i.d.R. Zeitraum) der Poissonverteilung mit Parameter λ folgt (vgl. Exponentialverteilung, Kapitel 14.2.2).

14.1.7 Multinomialverteilung

Idee/Anwendung: Allgemein stellt die Multinomialverteilung eine Erweiterung der Binomialverteilung auf mehr als zwei mögliche Ereignisse dar. Man kann sie sich als Verteilung für das Ziehen mit Zurücklegen aus einer Urne mit k Sorten Kugeln vorstellen. Jede dieser verschiedenen Sorten hat eine individuelle Wahrscheinlichkeit p_i , $i = 1, 2, \dots, k$, mit der sie in der Urne vorkommt. Diese Wahrscheinlichkeiten sind einfach nur die relativen Häufigkeiten. Mit der Wahrscheinlichkeitsfunktion berechnet man die Wahrscheinlichkeit für das Eintreten gewisser Anzahlen x_i , $i = 1, 2, \dots, k$. Uns ist hierbei die Gesamtzahl der Kugeln in der Urne egal, da wir mit Zurücklegen ziehen².

Seien zum Beispiel $k = 10$ und $p_i = 0,1$, $i = 1, 2, \dots, 10$, dann ist es egal, ob es insgesamt 100 Kugeln gibt und jede Sorte zehnmal vorkommt oder 1000 Kugeln und jede Sorte 100-mal. Da wir mit Zurücklegen ziehen, ändern sich die Wahrscheinlichkeiten bei den verschiedenen Zügen nicht.

Verteilung:

Notation: $X \sim M(n; p_1, p_2, \dots, p_k)$

Wahrscheinlichkeitsfunktion: $P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1! \cdot x_2! \cdot \dots \cdot x_k!} \cdot p_1^{x_1} \cdot p_2^{x_2} \cdot \dots \cdot p_k^{x_k}$

mit $\sum_{i=1}^k p_i = 1$ und $\sum_{i=1}^k x_i = n$ (vgl. ³ und ⁴)

→ Die Wahrscheinlichkeitsfunktion lässt sich intuitiv in ihre einzelnen Faktoren zerlegen:

$\frac{n!}{x_1! \cdot x_2! \cdot \dots \cdot x_k!}$: "Permutationen mit Wiederholung", da es k unterscheidbare Ausprägungen, die mit einer entsprechenden Häufigkeit x_i , $i = 1, 2, \dots, k$, vorkommen sollen, gibt.

²vgl. <https://de.wikipedia.org/wiki/Multinomialverteilung>

³Die erste Summe ist die übliche Bedingung, die wir an Wahrscheinlichkeiten stellen

⁴Wenn wir n -mal ziehen, muss die Gesamtzahl der Ereignisse natürlich gleich n sein; das stellt die zweite Summe dar

$p_i^{x_i}$: Die Wahrscheinlichkeit, dass Ausprägung i drankommt ist p_i . Nun soll Ausprägung i genau die Anzahl x_i haben. Deshalb nehmen wir wieder das Produkt $\underbrace{p_i \cdot p_i \cdot \dots \cdot p_i}_{x_i \text{ mal}} = p_i^{x_i}$ und zwar für alle $i = 1, 2, \dots, k$.

Erwartungsvektor und Kovarianzmatrix⁵: $\mathbb{E}[X] = (np_1, np_2, \dots, np_k)^\top$ und

$$Cov[X_i, X_j] = \begin{cases} np_i(1 - p_i) & \text{für } i = j, \\ -np_i p_j & \text{für } i \neq j \end{cases}$$

Bemerkung:

⁵Für die vollständige Kovarianzmatrix siehe Vorlesung Slide 13.36

14.1.8 Aufgaben

1. Welche Aussagen bzgl. der Gleichverteilung sind richtig?

- a) Bei der diskreten Gleichverteilung muss es immer zwei mögliche Ausgänge geben, die die gleiche Wahrscheinlichkeit haben. ☐
- b) Ein Würfelwurf mit 88 Seiten ist gleichverteilt. ☐
- c) Ein unfairer Münzwurf mit Kopf und Zahl ist gleichverteilt. ☐
- d) In einem Bällebad sind 500 rote, 300 blaue, 450 gelbe und 350 grüne Bälle.
Zur Berechnung des Erwartungswertes eines gezogenen Balles der ZV: Anzahl der blauen und roten Bälle kann man die diskrete Gleichverteilung verwenden. ☐

2. Welche Aussagen bzgl. der Bernoulli-Verteilung und Binomialverteilung sind wahr?

- a) Die Binomialverteilung besteht aus n abhängige und identische Bernoulli-Experimente. ☐
- b) Der Erwartungswert der Bernoulli-Verteilung ist die Wahrscheinlichkeit, dass das Ereignis eintritt, selbst. ☐
- c) Bei der Bernoulli-Verteilung und bei der Binomialverteilung gibt es jeweils nur zwei mögliche Ergebnisse. ☐
- d) Die Verteilungsfunktion der Binomialverteilung geht von 0 bis 1. ☐

3. Gegeben ist die Gleichung $P(X = 4) = 0,30, 7^{(3)}$ Welche Aussagen dazu sind wahr?

- a) Hierbei handelt es sich um die geometrische Verteilung. ☐
- b) Der Erwartungswert beträgt hier ca. 3,33 ☐
- c) Mit dieser Gleichung berechnet man die Wahrscheinlichkeit, dass das gewünschte Ereignis bei der dritten Wiederholung eintritt. ☐
- d) Die Wahrscheinlichkeit für das gewünschte Ereignis liegt bei 0,7. ☐
- e) Die Varianz beträgt ca. 0,6122. ☐

4. Welche Aussagen bezüglich der geometrischen und hypergeometrischen Verteilung sind wahr?

- a) Je größer die Wahrscheinlichkeit eines Ereignisses ist, desto länger braucht es, bis es zum ersten Mal auftritt. ☐
- b) Die hypergeometrische Verteilung verwendet man, wenn man wissen will, wie oft man ein Experiment wiederholen muss, bis man das gewünschte Ereignis erhält. ☐

- c) Um eine Wahrscheinlichkeit mit der geometrischen Verteilung zu berechnen benötigt man 4 Variable, N Einheiten unter denen M Elemente mit dem einem bestimmten Merkmal sind, man wählt n Elemente aus den N aus und man berechnet die Wahrscheinlichkeit, dass man x Elemente mit dem bestimmten Merkmal gezogen hat. ☐
- d) Die hypergeometrische Verteilung berechnet die Wahrscheinlichkeit unter der Voraussetzung, dass die gezogenen bzw. ausgewählten Elemente wieder zurückgelegt werden. ☐

5. Welche Aussagen zur Poissonverteilung sind richtig?

- a) Die Poissonverteilung gibt die Wahrscheinlichkeit für das Eintreffen eines Ereignisses wieder. ☐
- b) Das λ muss größer gleich 0 sein. ☐
- c) Die Varianz und der Erwartungswert sind immer gleich groß. ☐
- d) Das λ gibt die Intensitätsrate an, das heißt je größer λ ist, desto häufiger wird das Ereignis vorkommen. ☐
- e) Sind zwei Zufallsvariable unabhängig, dann kann man bei der Poissonverteilung die λ addieren und damit die Wahrscheinlichkeit eines Ereignisses berechnen. ☐

6. Welche Aussagen zur Multinomialverteilung sind richtig?

- a) Bei der Multinomialverteilung ist es wichtig zu wissen, wie viele Einheiten insgesamt vorhanden sind. Also wie groß n ist, da wir nicht mit zurücklegen ziehen. ☐
- b) Die Multinomialverteilung ist die Verallgemeinerung der Bernoulliverteilung mit mind. 2 Ereignissen. ☐
- c) k gibt die verschiedenen Sorten die vorhanden sind an. ☐
- d) Man benötigt die Gesamtheit aller Einheiten nicht, da mit Zurücklegen gezogen wird und hierbei die bekannten relativen Häufigkeiten gleich bleiben. ☐
- e) Die Form der Wahrscheinlichkeitsfunktion gleicht der Permutation ohne Wiederholung. ☐

14.2 Stetige Verteilungen

Stetig heißt grob gesagt, dass ein Experiment unendliche viele Zahlen an möglichen Ergebnissen hat. Beispiele aus der Vorlesung sind die stetige Gleichverteilung, Exponentialverteilung, Normalverteilung, χ^2 -Verteilung, t-Verteilung und die F-Verteilung. Diese basieren auf stetigen Zufallsvariablen. Die Menge der Werte, die diese stetigen ZV annehmen können ist unendlich und nicht zählbar (Beispiel: das Intervall $[1,6]$).

14.2.1 Stetige Gleichverteilung

Idee: Die möglichen Ausprägungen der ZV liegen auf einem fixen Intervall $[a;b]$ (z.B. $[0;1]$). Wichtig ist, dass die ZV in diesem Intervall jeden möglichen Wert, also überabzählbar unendlich viele Werte, annehmen kann. In $[0;1]$ kann die Zahl 0,5 genauso vorkommen wie die Zahl 0,4358984. Es sind also beliebig viele Nachkommastellen und Kombinationen dieser möglich, also auch beliebig viele Werte.

Wie bei der diskreten Gleichverteilung hat hier jede mögliche Ausprägung dieselbe Wahrscheinlichkeit. Da es nun unendlich viele Werte k gibt, hat jeder einzelne Wert die Wahrscheinlichkeit: $\lim_{k \rightarrow +\infty} \frac{1}{k} = 0$. Deshalb trifft man in der Regel Aussagen darüber, mit welcher Wahrscheinlichkeit die ZV in einem bestimmten Intervall landet. Zum Beispiel landet eine ZV, die der stetigen Gleichverteilung auf dem Intervall $[0;1]$ folgt, zu 50 % im Intervall $[0;0,5]$, was genau die Hälfte des Ereignisraums ist.

Verteilung:

Notation: $X \sim U(a, b)$

$$\text{Dichtefunktion}^6 : f(x) = \begin{cases} \frac{1}{b-a} & \text{für } x \in [a; b] \\ 0 & \text{sonst} \end{cases}$$

→ Die Dichtefunktion näher erklärt:

Die Dichtefunktion ist in gewissermaßen das stetige Pendant zu Wahrscheinlichkeitsfunktion. Deshalb mag es erstmal komisch anmuten, dass sie positive Werte an allen möglichen Ausprägungen hat, obwohl diese jeweils mit Wahrscheinlichkeit 0 vorkommen. Wenn wir uns aber daran erinnern, dass wir für die Berechnung von der Wahrscheinlichkeit, dass X in einem bestimmten Intervall landet, Integrale benutzen, ergibt das ganze mehr Sinn.

Da X mit Wahrscheinlichkeit 1 im Intervall $[a; b]$ landet, muss gelten: $\int_a^b f(x)dx = 1$. Da die Dichtefunktion hier nicht unterschiedlichen Werten im Ereignisraum $[a; b]$ verschiedene Werte zuweisen darf, erhalten alle denselben Wert $\frac{1}{b-a}$. D.h. auf dem Intervall $[a; b]$ ist die Dichte eine horizontale Linie auf Höhe des y-Wertes $\frac{1}{b-a}$ (und sonst 0). Mit dem Wert $\frac{1}{b-a}$ erhalten wir aber auch, dass die Fläche unter der

⁶Zur Erinnerung: Bei stetigen ZVs gibt es eine Dichte-, bei diskreten ZVs eine Wahrscheinlichkeitsfunktion

Dichte (also das Integral; hier ein Rechteck) genau 1 ergibt, da die Breite (auf der x-Achse) $b - a$ beträgt und die Höhe des Rechtecks eben genau $\frac{1}{b-a}$ ist.

Erwartungswert und Varianz: $\mathbb{E}[X] = \frac{a+b}{2}$ und $\text{Var}[X] = \frac{(b-a)^2}{12}$

Bemerkung:

14.2.2 Exponentialverteilung

Idee/Anwendung: Wird für Warte- und Ausfallzeiten verwendet und kann als stetige Version der geometrischen Verteilung angesehen werden. Zur Erinnerung: Auch bei der stetigen Gleichverteilung hatten wir angenommen, dass man damit eine Wartezeit messen kann (vgl. Vorlesung Slide 13.41 - Wartezeit auf S-Bahn $\sim U(0, 10)$). Der bedeutendste Unterschied bei der Exponentialverteilung ist, dass die weitere Wartezeit unabhängig von der bereits verstrichenen Wartezeit ist (das nennt sich "Gedächtnislosigkeit der Exponentialverteilung").

Ein Beispiel ist die Lebenszeit einer Glühbirne (also die Wartezeit bis zum Ausfall), bzw. allgemein Wartezeiten bis zum Eintreffen eines Ereignisses. Die Annahme der Gedächtnislosigkeit ist natürlich bei solchen Beispielen etwas kritisch zu betrachten.

Verteilung:

Notation: $X \sim \text{Expo}(\lambda)$

Dichtefunktion: $f(x) = \begin{cases} \lambda \cdot \exp(-\lambda x) & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$

Verteilungsfunktion: $F(x) = \begin{cases} 1 - \exp(-\lambda x) & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$

Erwartungswert und Varianz: $\mathbb{E}[X] = \frac{1}{\lambda}$ und $\text{Var}[X] = \frac{1}{\lambda^2}$

Bemerkung: Wenn eine ZV poissonverteilt mit Parameter λ ist ($X \sim \text{Po}(\lambda)$), dann ist die Wartezeit bis zum ersten Eintreffen exponentialverteilt mit demselben Parameter λ . Hierbei ist zu beachten, dass man dann die erwartete Wartezeit (also den Erwartungswert der exponentialverteilten ZV), wenn möglich, in der richtigen (Zeit-)Einheit angibt. Ist der fixe Zeitraum der Poissonverteilung beispielsweise eine Stunde (1 h), dann ist die erwartete Wartezeit in Stunden auch $\frac{1}{\lambda} \cdot 1$ h.

14.2.3 Normalverteilung (aka Gauß'sche Glockenkurve)

Gründe für die Verwendung und Wichtigkeit in der Statistik:

- Oftmals sind normalverteilte Modelle sehr einfach zu rechnen.

- Der Durchschnitt einer Stichprobe mit beliebiger Verteilung folgt einer Normalverteilung, d.h. man kann n Zufallszahlen aus egal welcher Verteilung ziehen, der Mittelwert wird jedoch immer der Normalverteilung folgen (**zentraler Grenzwertsatz**: Idee klassischen t-Test – Bildung des Stichprobenmittelwerts der normalverteilt ist).
- viele Naturphänomene folgen einer Normalverteilung (Bsp: Körpergröße – "Durchschnitt" vieler genetischer Faktoren).

Idee/Anwendung: Eine Intuition wie bei den anderen Funktion gibt es hier nicht wirklich. Aber wie oben beschrieben, gibt es viele gute Gründe, die Normalverteilung zu benutzen. Sie ist deshalb auch eine der häufigsten Verteilungen in der Statistik.

Verteilung:

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$

Dichtefunktion: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Erwartungswert und Varianz: $\mathbb{E}[X] = \mu$ und $\text{Var}[X] = \sigma^2$

Standardisierung: Sei $X \sim \mathcal{N}(\mu, \sigma^2)$, dann gilt: $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$

Bemerkung: Die Dichtefunktion der Standardnormalverteilung ($\mathcal{N}(0, 1)$) wird in der Regel als $\phi(x)$ und die Verteilungsfunktion als $\Phi(x)$ bezeichnet. Die Quantile der Standardnormalverteilung werden in der Regel mit z beschrieben.

14.2.4 Chi-Quadrat-Verteilung

Idee/Anwendung: Ähnlich wie die Normalverteilung haben wir hier keine direkte Intuition, sondern benutzen die Verteilung, weil viele theoretische Konzepte auf ihr beruhen. Die Summe der Quadrate von n standardnormalverteilten ZVs ist χ^2 -verteilt mit n Freiheitsgraden (degrees of freedom, df) (vgl. Vorlesung Slide 13.57).

Bemerkung: Die Quantile der χ^2 -Verteilung werden in der Regel mit c beschrieben.

14.2.5 t-Verteilung

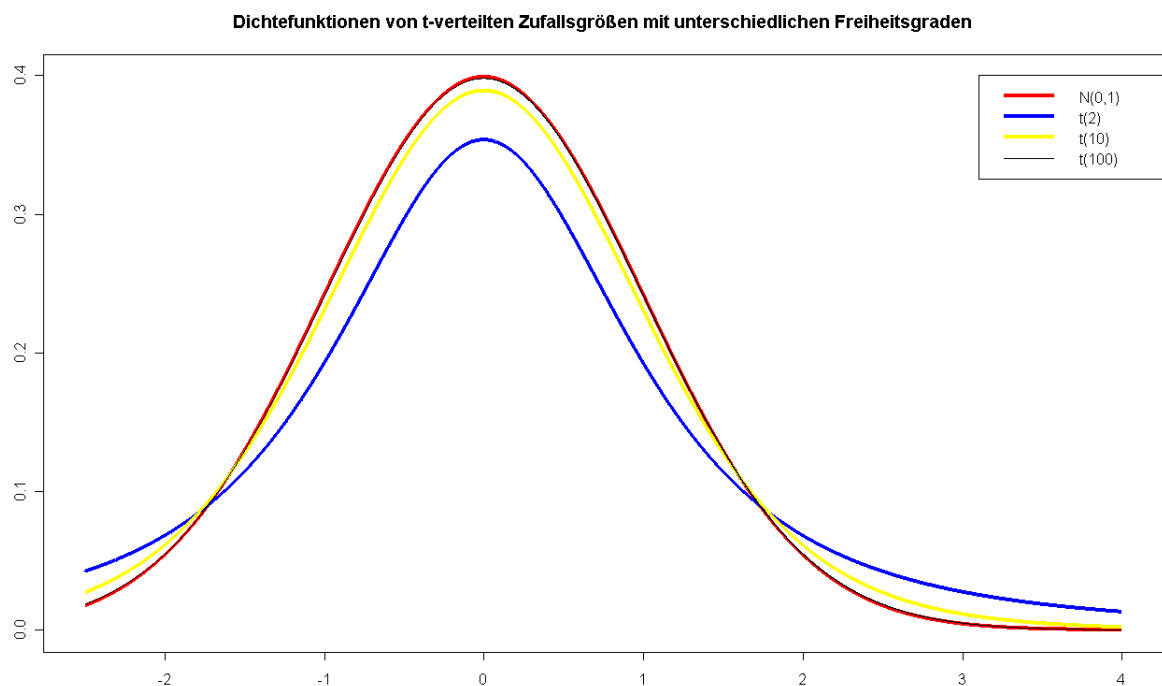
Idee/Anwendung: Ähnlich wie die Standardnormalverteilung und bei uns hauptsächlich für statistische Tests benutzt. In Bezug auf statistische Tests ist hier anzumerken, dass die t-Verteilung der Standardnormalverteilung sehr ähnlich ist. Sie hat allerdings breitere Enden, was bedeutet, dass kleinere Quantile (kleiner als der Median, der 0 ist) einen kleineren Wert haben als bei der Standardnormalverteilung (d.h. dass der Betrag ist größer, da die Quantile ja negativ sind). Größere Quantile (größer als 0) haben analog dazu einen größeren Wert als die der Standardnormalverteilung. Das ist wichtig für den Ablehnbereich in der Testtheorie (vgl. dazu Fig. 4). Dazu später mehr (vgl. Kapitel 17.3).

Verteilung:

Notation: $X \sim t_{df}$

Bemerkung: Die Quantile der t-Verteilung werden in der Regel mit t beschrieben. Außerdem kann für hohe Freiheitsgrade die t-Verteilung durch die Standardnormalverteilung approximiert werden. Das gilt dann natürlich auch für die entsprechenden Quantile.

⚠ Bitte verwendet immer die korrekten Quantile, wenn diese angegeben sind.



Quelle: https://de.wikipedia.org/wiki/Studentsche_t-Verteilung

Figure 4: Annäherung der t-Verteilung an die Standardnormalverteilung

Die Grafik verdeutlicht die Annäherung der t-Verteilung mit zunehmenden Freiheitsgraden an die Standardnormalverteilung.

14.2.6 F-Verteilung

Idee/Anwendung: Auch für die F-Verteilung müssen wir ohne weitere Intuition auskommen. Sie wird wiederum für diverse theoretische Konzepte benutzt.

Verteilung:

Notation: $X \sim F_{df_1, df_2}$

Bemerkung:

⚠ Die F-Verteilung hat zwei Freiheitsgrade. Bitte denkt immer an beide (z.B. bei einem Overall-F-test für die Betakoeffizienten eines multiplen Regressionsmodells)!

14.2.7 Aufgaben

1. Welche Aussagen zur stetigen Gleichverteilung sind richtig?

- a) Jede einzelne Ausprägung kommt mit einer Wahrscheinlichkeit von 0 vor, also eine Punktwahrscheinlichkeit von 0. ☐
- b) Der Unterschied zwischen der stetigen und diskreten Gleichverteilung ist, dass die stetige Gleichverteilung unendlich verschiedene Werte annehmen kann. ☐
- c) Die Dichtefunktion ist das Pendant zur Verteilungsfunktion. ☐
- d) Die Dichtefunktion bei der stetigen Gleichverteilung ist eine monoton steigende Gerade. ☐
- e) Eine stetig gleichverteilte Dichtefunktion hat einen Ereignisraum von $[2,6]$. Die Linie in diesem Intervall ist auf der Höhe 0.25. ☐

2. Welche Aussagen zur Exponentialverteilung sind richtig?

- a) Die Wartezeit ist abhängig von der davor schon verstrichenen Zeit. ☐
- b) Bei der exponential wie auch der geometrischen Verteilung geht es darum, wann das gewünschte Ereignis zum ersten Mal auftritt, bei der Exponentialverteilung ist es eine stetige Spanne und bei der Geometrischen Verteilung sind es diskrete Schritte. ☐
- c) Die Verteilungsfunktion hat ein exponentielles Wachstum. ☐
- d) Ein zentraler Begriff bei der Exponentialverteilung ist die Gedächtnislosigkeit. ☐

3. Welche Aussagen zur Normalverteilung / Chi-Quadrat-Verteilung sind wahr?

- a) Die Dichtefunktion ist symmetrisch um den Erwartungswert verteilt. ☐
- b) Nur der Durchschnitt von normalverteilten Zufallsvariablen berechnet man mit der Normalverteilung. ☐
- c) Um die Dichtefunktion der Normalverteilung aufzustellen benötigt man nur σ und μ . ☐
- d) Die Normalverteilung hat bei μ ihr Maximum. ☐
- e) Bei der Chi-Quadrat-Verteilung werden die QUadrate von mehreren standardnormalverteilten Zufallsvariablen gebildet und aufsummiert. ☐

4. Welche Aussagen über die t-Verteilung sind richtig?

- a) Die t-Verteilung kann immer durch die Standardnormalverteilung approximiert werden. ☐
- b) Je größer die Stichprobenzahl, desto näher kommt sie der Normalverteilung. ☐
- c) Die Wahrscheinlichkeitsmasse die nach außen verteilt ist ist bei der t-Verteilung geringer als bei der Normalverteilung. ☐
- d) Je weniger Freiheitsgrade, desto ähnlicher ist die t-Verteilung der Normalverteilung. ☐

14.3 Wichtige Schlüsselbegriffe und "Konzepte" anhand der diskreten Gleichverteilung

14.3.1 Parameter von Verteilungen

Mögliche Ergebnisse werden oft mit den Variablen x_1, x_2, \dots, x_n bezeichnet. Für das Beispiel des Roulette-Spiels wären $x_1 = 0, x_2 = 1, \dots, x_{37} = 36$. Die **Parameter** im Fall der diskreten Gleichverteilung werden oft als a und b definiert. Beim Roulettespiel gilt: $a = 0$ und $b = 36$, d.h. a und b sind sozusagen die Grenzen des Ergebnisraums.

14.3.2 Träger einer Verteilung

Der **Träger** der diskreten Gleichverteilung sind alle Ausprägungen x_1, \dots, x_n , also alle natürlichen Zahlen zwischen den (und einschließlich der) Grenzen a und b . Beim Roulette-Beispiel also alle Zahlen von $0, \dots, 36$.

14.3.3 Verteilungsfunktion, Wahrscheinlichkeitsfunktion und Dichtefunktion

In der Statistik I Vorlesung haben wir schon die empirische Verteilungsfunktion kennengelernt (z.B. als Treppenfunktion oder Polygonzug). Eine **Verteilungsfunktion** ist grob gesagt einfach nur eine Art Hilfestellung zur Beschreibung von diskreten/stetigen Wahrscheinlichkeitsverteilungen. Sie wird oftmals als Funktion F definiert, die jedem Ergebnis x_i einer Zufallsvariablen X eine Wahrscheinlichkeit $P(X \leq x_i)$ zuordnet. Mathematisch ausgedrückt einfach nur:

$$F : x \rightarrow P(X \leq x_i)$$

Sowohl die **Wahrscheinlichkeitsfunktion**⁷ für diskrete ZVs als auch die **Dichtefunktion** für stetige ZVS dienen als Beschreibung von Wahrscheinlichkeitsverteilungen. Sie beschreiben Wahrscheinlichkeiten für jedes mögliche Ergebnis x_i eines Zufallsexperiments (mathematisch einfach $P(X = x_i)$). Oftmals wird sie aber einfach als $f(x)$ definiert. Liegt eine **stetige ZV** zugrunde, besitzt die Dichtefunktion folgende Eigenschaften:

⁷Die Wahrscheinlichkeitsfunktion wird häufig auch Dichte genannt und mit $f(x)$ notiert

1. Die Funktion besitzt keinen negativen Wert: also $f(x) \geq 0$ für alle $x_i \in \mathbb{R}$
2. Die Fläche unter der Funktion (berechnet als ihr Integral) ergibt 1 (analog diskrete ZV: Summe aller Einzelwahrscheinlichkeiten ergibt ebenfalls 1) – mathematisch: $\int_{-\infty}^{\infty} f(x)dx = 1$
 \rightarrow Die Integralgrenzen können durch die Grenzen des Definitionsbereichs ersetzt werden.

Wichtige mathematische Zusammenhänge zwischen der Verteilungsfunktion, der Dichtefunktion und der Quantilsfunktion sind folgende:

1. $f(x) = \frac{d}{dx}F(x)$, d.h. die Dichte ist also einfach die Ableitung der Verteilungsfunktion.
2. $F(x) = \int_{-\infty}^x f(t)dt$, d.h. die Verteilungsfunktion ist die Fläche unter der Dichte (also das Integral der Dichte).
3. $Q(x) = F^{-1}(x)$, d.h. die Quantilsfunktion ist die Umkehrfunktion der Verteilungsfunktion. (Rückblick Statistik I: Quantile sind einfach nur bestimmte Schwellenwerte, d.h. ein bestimmter Anteil der Werte ist kleiner oder gleich dem Quantil, der Rest ist größer – Beispiel: Median).
4. $F(x) = Q^{-1}(x)$, d.h. die Verteilungsfunktion ist die Umkehrfunktion der Quantilsfunktion.

Aber zurück zum Beispiel der diskreten Gleichverteilung unseres Roulette-Spiels:

Verteilungsfunktion Allgemein ergibt sich für eine diskrete Gleichverteilung auf den ganzen Zahlen $\{a, a+1, a+2, \dots, b-1, b\}$ für die Verteilungsfunktion:

$$F(x) = \begin{cases} 0, & x < a \\ \frac{\lfloor x \rfloor - a + 1}{b - a + 1}, & x \in [a, b] \\ 1, & x > b \end{cases}$$

$\lfloor x \rfloor$ bedeutet, dass wir x abrunden, d.h. $\lfloor 3.7 \rfloor = 3$. Speziell für das Beispiel Roulette: Will man z.B. wissen wie wahrscheinlich eine Zahl kleiner als 3.5 ist berechnet man:

$$F(3.5) = \frac{\lfloor 3.5 \rfloor - 0 + 1}{36 - 0 + 1} = \frac{4}{37}$$

Wahrscheinlichkeitsfunktion

Die Wahrscheinlichkeitsfunktion sieht folgendermaßen aus:

$$f(x) = \begin{cases} \frac{1}{37}, & x \in \{0, 1, 2, \dots, 36\} \\ 0, & \text{sonst} \end{cases}$$

14.3.4 Erwartungswert und Varianz

Der **Erwartungswert** der diskreten Gleichverteilung ist definiert als $E(X) = \frac{a+b}{2}$ und die Varianz ist definiert als $Var(X) = \frac{(b-a+1)^2-1}{12}$. Vorsicht: Dies sind etwas allgemeinere Formeln als der Spezialfall in der Formelsammlung, im Prinzip stellt es aber das Gleiche dar.

14.3.5 Aufgaben

1. Welche Aussagen sind richtig?

- a) Die Träger der diskreten Gleichverteilung eines Münzwurfs sind Kopf und Zahl. ☐
- b) Eine Verteilungsfunktion geht immer von -1 bis 1. ☐
- c) Der Wert einer Verteilungsfunktion an der Stelle $x = 5$ ist die Wahrscheinlichkeit für $X \leq 5$. ☐
- d) Die Werte der Verteilungsfunktion sind die kumulierten Wahrscheinlichkeiten. ☐
- e) Die Wahrscheinlichkeitsfunktion der diskreten Zufallsvariablen ist das gleiche wie die Verteilungsfunktion der stetigen Zufallsvariablen. ☐

2. Welche Aussagen sind richtig?

- a) Die Verteilungsfunktion ist monoton steigend. ☐
- b) Die Verteilungsfunktionen von stetigen Zufallsvariablen sind Treppenfunktionen im Gegensatz zu den diskreten Zufallsvariablen. ☐
- c) Eine Eigenschaft der diskreten Verteilungsfunktion ist, dass die Fläche unterhalb der Funktion immer 1 ergibt. ☐
- d) Die Summe aller Einzelwahrscheinlichkeiten bei einer stetigen Zufallsvariable geben 1. ☐
- e) Eine Eigenschaft der Dichtefunktion einer stetigen ZV ist, dass die Werte immer größer gleich 0 sind. ☐

3. Welche Aussagen sind richtig?

- a) Um die Verteilungsfunktion zu erhalten, leite ich die Dichtefunktion ab. ☐
- b) Die Fläche unter der Dichtefunktion ist die Quantilsfunktion. ☐
- c) Die Quantilsfunktion ist die Umkehrfunktion der Verteilungsfunktion und umgekehrt. ☐
- d) Bei der Verteilungsfunktion gibt es immer mind. 3 Stellen die separat definiert werden müssen. ☐

15 Grenzwertsätze und Approximationen von Verteilungen

Dieses Kapitel der Vorlesung beschreibt Grundbegriffe über das Verhalten von Folgen von Zufallsvariablen, wenn n gegen unendlich strebt. Die "normale" mathematische Folge sollte schon aus der Vorlesung "Mathematik für Wirtschaftswissenschaftler" bekannt sein. Diese Idee wird nun durch das Hinzunehmen von Zufallsvariablen erweitert. Für eine Auffrischung der Theorie von Folgen und Reihen siehe:

https://www.statistik.uni-muenchen.de/formulare/skripte_u_aehnliches/mathehandrechnung_schneider.pdf


15.1 Grenzwertsätze

15.1.1 Gesetz der großen Zahlen

Das Gesetz der großen Zahlen besagt, dass sich beobachtete relative Häufigkeiten mit zunehmendem n immer näher an die theoretischen Wahrscheinlichkeiten annähern. Genau dann konvergiert die durchschnittliche mittlere Abweichung zwischen den Zufallsvariablen X_i und ihrem Erwartungswert $E(X_i)$ gegen Null. Die Gesetzmäßigkeit lässt sich dadurch erklären, dass der Einfluss von Ausreißern mit zunehmendem Stichprobenumfang des Experiments abnimmt. Man kann sich dies z.B. anhand des einfachen Münzwurfs klarmachen: Wirft man eine Münze 10 mal, wird sich nur in seltenen Fällen genau eine relative Häufigkeit von $1/2$ für Kopf oder Zahl ergeben. Wirft man sie dagegen 1000 mal, wird der Wert sehr nahe an $1/2$ liegen.

15.1.2 Zentraler Grenzwertsatz (ZGS)

Der zentrale Grenzwertsatz gehört zu den wichtigsten Aussagen der Wahrscheinlichkeitstheorie. Er gibt eine Charakterisierung der Normalverteilung als Grenzverteilung von Überlagerungen einer Vielzahl unabhängiger zufälliger Einzeleffekte. Der ZGS besagt, dass sich der Mittelwert einer jeden beliebigen Verteilung von iid Zufallsvariablen mit zunehmenden Stichprobenumfang der Normalverteilung annähern wird. Wegen des zentralen Grenzwertsatzes können wir Hypothesentests durchführen, auch wenn die Grundgesamtheit keiner Normalverteilung unterliegt, vorausgesetzt, dass die Stichprobe ausreichend groß ist.

 Vorsicht: Oftmals wird angenommen, dass der zentrale Grenzwertsatz sagt, dass eine Stichprobe ab einer gewissen Größe automatisch normalverteilt sein wird. Dies stimmt aber nicht!

15.2 Approximationen

Wir können unter bestimmten Voraussetzungen Verteilungen durch andere Verteilungen approximieren. Doch warum und wann sind Approximationen sinnvoll?

- Bei Rechenintensiven Tasks.
- Bei diskreten Verteilungen: Addition vieler Einzelwahrscheinlichkeiten wird vermieden.

- Bei Approximation durch Normalverteilung: Standardisierung möglich.

Insgesamt haben wir vier Approximationen kennengelernt, die nachfolgend nochmal – überwiegend bezüglich ihrer Intuition – besprochen werden.

⚠ Sehr wichtig: Immer die Annahmen für die Approximation prüfen.

15.2.1 Approximation der Binomial- durch die Normalverteilung

Voraussetzung: $np(1-p) \geq 9$

Approximation: $B(n, p) \rightarrow \mathcal{N}(np, np(1-p))$

Intuition: Bei erfüllter Voraussetzung ist die Wahrscheinlichkeitsfunktion der Binomialverteilung (annähernd) symmetrisch und sieht der Dichte der Normalverteilung sehr ähnlich. In diesem Fall ist die Approximation zulässig. Um euch davon zu überzeugen, wie die Wahrscheinlichkeitsfunktion der Binomialverteilung für verschiedene n und p aussieht könnt ihr auf der Seite <https://matheguru.com/stochastik/binomialverteilung.html> das Tool "Interaktive Binomialverteilung" nutzen und verschiedene Kombinationen für n und p ausprobieren. Ihr könnt z.B. Werte, die die Voraussetzung nicht erfüllen, mit Werten, die die Voraussetzung erfüllen, vergleichen. Beachtet dabei, dass ihr "PDF" anklickt (und nicht "CDF").

Bemerkung 1: Auch bei z.B. $n = 10$ und $p = 0.5$ ist die Wahrscheinlichkeitsfunktion symmetrisch und sieht der Normalverteilung ähnlich. Wir dürfen trotzdem nicht approximieren (vgl. $10 \cdot 0.5 \cdot 0.5 = 2.5 < 9$). Der Grund dafür, dass wir hier noch nicht approximieren dürfen ist, dass wir eine diskrete Verteilung (Binomialverteilung) durch eine stetige Verteilung (Normalverteilung) approximieren. Die Schritte in der Wahrscheinlichkeitsfunktion sind einfach noch zu groß, um durch eine stetige Verteilung approximiert werden zu dürfen.

Bemerkung 2: Außerdem gilt nach dieser Approximation, dass die ZV $\hat{p} = \frac{1}{n} \sum X_i$ auch normalverteilt ist: $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$

15.2.2 Approximation der Binomial- durch die Poissonverteilung

Voraussetzung: Großes n und kleines p .

Approximation: $B(n, p) \rightarrow Po(np)$

Intuition: Da wir sowohl die Binomial- als auch die Poissonverteilung durch die Normalverteilung approximieren können (wenn die entsprechenden Voraussetzungen erfüllt sind), ist diese Approximation dann einfach nur logisch (vgl. dazu $B(0,05; 200)$ und $Po(10)$). *Aber was ist, wenn die jeweiligen Voraussetzungen nicht erfüllt sind?* In diesem Fall schauen wir uns einfach wieder die Binomialverteilung für ein großes n und ein kleines p in dem Tool "Interaktive Binomialverteilung" an: Die Binomialverteilung ist rechtsschief/linkssteil (für kleine p). Dann ist np auch so klein, dass die Poissonverteilung rechtsschief/linkssteil ist (vgl. Poissonverteilung mit $\lambda = 1$, z.B. als Approximation für die Binomialverteilung mit $p = 0,01$ und $n = 100$).

15.2.3 Approximation der Poisson- durch die Normalverteilung

Voraussetzung: $\lambda \geq 10$

Approximation: $Po(\lambda) \rightarrow \mathcal{N}(\lambda, \lambda)$ ⁸

Intuition: Auch hier schauen wir uns einfach die Wahrscheinlichkeitsfunktion der Poisson-verteilung an und sehen, dass diese mit steigendem λ *immer symmetrischer* wird. Und wenn letztere annähernd symmetrisch ist (ab $\lambda = 10$), können wir durch die symmetrische Normalverteilung approximieren. Außerdem ist hierbei wichtig, dass die Normalverteilung "weit genug" auf dem positiven Teil der x-Achse liegt und negative Werte nur mit einer sehr, sehr geringen Wahrscheinlichkeit vorkommen, da die Poissonverteilung ja Anzahlen modelliert, die immer nicht-negativ sind. Das ist wiederum erfüllt, wenn die Normalverteilung um $\lambda = \mu \geq 10$ zentriert ist.

Zur Veranschaulichung könnt ihr ein ähnliches Tool wie für die Binomialverteilung nutzen. Ihr findet es unter <https://matheguru.com/stochastik/poisson-verteilung.html>.

15.2.4 Approximation der hypergeometrischen durch die Binomialverteilung

Voraussetzungen:

1. $n \leq 0,1 \cdot M$
2. $n \leq 0,1 \cdot (N - M)$

Approximation: $H(n, M, N) \rightarrow B(n, \frac{M}{N})$

Intuition: Hinter dieser Approximation steckt, dass jedes Ziehen (insgesamt n -mal) mit der hypergeometrischen Verteilung ein Bernoulli-Experiment ist, ob das gewünschte Merkmal gezogen wird oder nicht. Beim ersten Ziehen ist noch sehr offensichtlich, dass die Wahrscheinlichkeit, das gewünschte Merkmal zu ziehen, $p = \frac{M}{N}$ ist. Nach jedem Zug reduzieren sich aber N bzw. N und M , d.h. p im einfachen Bernoulli-Experiment verändert sich. Bei erfüllten Faustregeln hingegen ist die Anzahl der Züge im Vergleich zur Anzahl der Objekte mit dem gewünschten Merkmal (M) bzw. ohne das gewünschte Merkmal ($N - M$) so gering, dass angenommen werden kann, dass sich p nicht ändert. D.h. wir nehmen $p = \frac{M}{N}$ als konstant an. Dann sind die n Züge unabhängige Bernoulli-Experimente und daher ist die Approximation durch die Binomialverteilung zulässig.

⁸Zur Erinnerung: Für $Po(\lambda)$ gilt $\mathbb{E}[X] = \text{Var}[X] = \lambda$, d.h. man übernimmt Erwartungswert und Varianz der Poissonverteilung einfach als Parameter für die Normalverteilung.

15.3 Aufgaben

1. Welche Aussagen bzgl. den Grenzwertsätzen ist wahr?

- a) Je größer der Stichprobenumfang, desto größer wird der Einfluss von Ausreißern. ☐
- b) Mit zunehmendem Stichprobenumfang nähert sich die beobachtete relative Häufigkeit der theoretischen Wahrscheinlichkeit an. ☐
- c) $N = 25$ liegt näher an der theoretischen Wahrscheinlichkeit als $N = 26$ ☐
- d) Egal welche Verteilung vorliegt, nähert sich der Mittelwert von iid Zufallsvariablen bei zunehmenden Stichprobenumfang der Normalverteilung an. ☐

2. Welche Aussagen bzgl. Approximationen sind richtig?

- a) Die hypergeometrische Verteilung kann durch die Normalverteilung approximiert werden. ☐
- b) Bei der Approximation der Poissonverteilung durch die Normalverteilung wird das λ für das μ und das σ einfach übernommen. ☐
- c) Die Poissonverteilung darf zu jeder Zeit durch die Normalverteilung approximiert werden. ☐
- d) Die hypergeometrische Verteilung $H(9,120,40)$ darf durch die Binomialverteilung approximiert werden. ☐

16 Schätzen

16.1 Die Maximum Likelihood Schätzung

Die ML Schätzung ist ein Prinzip für die Konstruktion von Parameterschätzern bei gegebener Verteilung einer Zufallsvariable.

Idee: Wähle die Schätzwerte für die wahren Parameter der Grundgesamtheit so, dass unter diesen die beobachtete Stichprobe am wahrscheinlichsten sind (dies erklärt auch den Namen der Methode).

Die ML-Methode beinhaltet 5 Schritte:

1. Liegt eine iid-Stichprobe vor?
2. Bestimme die Dichte für die einzelnen Beobachtungen.
3. Stelle die Likelihoodfunktion auf.
4. Stelle die log-Likelihoodfunktion auf.
5. Maximiere die log-Likelihood (Ableitung und 0 setzen) und prüfe die Bedingung 2. Ordnung (Maximum oder Minimum?).

Beispiele der ML-Schätzung für Normalverteilung, Binomialverteilung, Poissonverteilung und Exponentialverteilung:

<https://mars.wiwi.hu-berlin.de/mediawiki/mmstat3/index.php/Maximum-Likelihood-Methode>

16.2 Konfidenzintervalle (KI)

Wozu braucht man eigentlich Konfidenzintervalle? Im Bereich der induktiven Statistik wird mit Hilfe einer Stichprobe versucht, allgemeine Aussagen über die Grundgesamtheit zu machen. Mit Hilfe der ML-Schätzung haben wir dafür schon eine Methode kennengelernt um Punktschätzer zu "generieren". Die erhobenen Daten einer Stichprobe werden dann in einem Punktschätzer zusammengefasst (z.B. Mittelwert oder Varianz), um damit auf die wahren Werte in der Grundgesamtheit zu schließen.

⚠ Probleme:

- Wie präzise ist diese Punktschätzung eigentlich?
- In welchem Bereich liegt der wahre Mittelwert der Grundgesamtheit höchstwahrscheinlich?
- Ist es möglich, dass in Wirklichkeit im Mittel doch ein anderer Wert herauskommt, aber wir in dieser Stichprobe einfach nur Pech hatten?

⇒ Diese Fragen kann ein Punktschätzer nicht beantworten – aber ein Intervallschätzer kann das!

Doch was ist ein Konfidenzintervall genau? Die folgende Unterscheidung ist sehr wichtig für das Verständnis von Konfidenzintervallen:

- Mit einer Stichprobe schätzen wir einen Parameter, z.B. $\hat{\mu}$.
- Der wahre Parameter μ in der Grundgesamtheit ist dann zwar in der Nähe von $\hat{\mu}$, aber quasi nie genau gleich.

Den wahren Parameter μ werden wir also nie exakt bestimmen können, dennoch können wir versuchen einen Bereich zu bestimmen in dem er ziemlich sicher liegt – und genau das ist die Idee von Konfidenzintervallen.

Definition KI (ohne Formel):

Ein 95%-KI ist ein Intervall $[a,b]$, das, wenn es sehr häufig mit neuen Stichproben berechnet wird, den wahren Parameter, z.B. (μ) , mit einer Wahrscheinlichkeit von 95% auch überdeckt.

⇒ ein einzelnes 95%-KI ist mit 95%-iger Wahrscheinlichkeit eines von denen, das den wahren Parameter (μ) beinhaltet.

Und wie schätzt man ein Konfidenzintervall? Das zentrale Prinzip für alle Konfidenzintervalle:

1. Berechne einen Punktschätzer für einen Parameter, z.B. für den Anteilswert p einer Bernoulli oder Binomialverteilung.

2. Um diese Punktschätzer bildet man dann ein (meistens) symmetrisches Intervall, das abhängig von der Varianz in der Stichprobe und des gewünschten Konfidenzniveaus $1 - \alpha$ enger oder breiter wird.

16.3 Aufgabe

Welche Aussagen ist bzgl. der Konfidenzintervalle richtig?

- a) Die Punktschätzer sind immer richtig und wahrheitsgetreu. ☐
- b) Ein Konfidenzintervall gibt den Bereich vor, in dem der wahre Parameter immer zu finden ist. ☐
- c) Man kann die Länge eines Intervalls variieren, je nach dem mit welcher Wahrscheinlichkeit der wahre Parameter innerhalb des Intervalls liegen soll. ☐
- d) Die Länge eines Intervalls hängt von α , σ^2 , der Verteilung und dem Stichprobenumfang ab. ☐
- e) Ein 99%-KI besagt, dass zu 99% mein Schätzer richtig ist. ☐

17 Testtheorie

17.1 Der p-Wert

Der p-Wert gibt die Wahrscheinlichkeit an, dass die Teststatistik, unter der Annahme, dass H_0 wahr ist, den beobachteten/realisierten Wert t oder einen noch extremeren Wert annimmt. Noch extremer heißt in diesem Fall, dass man sich davon wegbewegt, H_0 anzunehmen. Um das zu verdeutlichen, müssen wir uns erstmal vor Augen halten, wo genau diese Wahrscheinlichkeit herkommt:

Wir wissen, dass die Teststatistik $T(\mathbf{X})$ unter H_0 eine bestimmte Verteilung hat (Merke: $T(\mathbf{X})$ ist eine ZV). Von dieser Verteilung kommen ja dann auch die Quantile, die wir als kritische Werte heranziehen, um die Testentscheidung mittels der Teststatistik zu treffen. Was genau das mit dem p-Wert zu tun hat, gehen wir jetzt am Beispiel eines einseitigen und eines zweiseitigen t-Tests durch.

17.1.1 Der p-Wert beim einseitigen t-Test

Der einseitige t-Test wird genutzt, um zu testen, ob der Erwartungswert einer normalverteilten ZV größer (kleiner) als ein unter H_0 angenommener Wert μ_0 ist, wenn die Varianz σ^2 unbekannt ist und geschätzt werden muss. Man kann die Hypothese (je nachdem, was man zeigen möchte) für beide Richtungen aufstellen. Wir zeigen es an dem Beispiel, dass wir testen wollen, ob der Erwartungswert signifikant kleiner als ein μ_0 ist. D.h. die Hypothesen lauten:

$$H_0: \mu \geq \mu_0 \quad \text{gegen} \quad H_1: \mu < \mu_0$$

Die Teststatistik und ihre Verteilung sind dann unter der H_0 -Hypothese:

$$T(\mathbf{X}) = \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{S_x} \stackrel{H_0}{\sim} t_{n-1}$$

Mit den vorliegenden Daten berechnet man nun eine Realisation t der Zufallsvariable $T(\mathbf{X})$. Dann wird H_0 abgelehnt, wenn $t \in (-\infty, -t_{n-1, 1-\alpha})$. Je kleiner die Realisation t von $T(\mathbf{X})$ also ist, desto eher lehnen wir H_0 ab bzw. desto weiter entfernen wir uns davon, H_0 anzunehmen. Die Wahrscheinlichkeit, dass man dann eine genauso kleine oder noch kleinere Teststatistik (im Vergleich zur Realisation t) erhält, wenn man die unter H_0 angenommene Verteilung (t_{n-1}) zugrunde legt, ist dann der p-Wert, also:

$$p\text{-Wert} = P(T \leq t),$$

wobei $T (\sim t_{n-1})$ wieder die ZV ist und t die Realisation. Ist diese Wahrscheinlichkeit kleiner als das Signifikanzniveau α , dann ist es sehr unwahrscheinlich, dass die Realisation t von der ZV $T(\mathbf{X})$ "unterboten" wird, wenn denn die Verteilungsannahme aus H_0 stimmt. Aber wenn es eben so unwahrscheinlich ist, dass t von der ZV $T(\mathbf{X})$ unterboten wird, muss t selbst (unter der Verteilung

in H_0) ein Ausreißer, also ein seltenes Ereignis sein. D.h. wir hätten mit unseren Daten zufällig ein seltenes Ereignis getroffen. Wir wollen aber nicht glauben, dass wir ausgerechnet ein seltenes Ereignis gefunden haben und verwerfen deshalb die unter H_0 angenommene Verteilung und somit die H_0 -Hypothese.

Merke: $p\text{-Wert} = P(T \leq t) < \alpha \Rightarrow H_0 \text{ verwerfen!}$

⚠ Für die Hypothesen:

$$H_0: \mu \leq \mu_0 \quad \text{gegen} \quad H_1: \mu > \mu_0$$

ist der p-Wert entsprechend:

$$p\text{-Wert} = P(T \geq t) = 1 - P(T < t),$$

da wir ja dann den Ablehnbereich $(t_{n-1, 1-\alpha}, +\infty)$ haben.

17.1.2 Der p-Wert beim zweiseitigen t-Test

Hier werden folgende Hypothesen gegeneinander getestet:

$$H_0: \mu = \mu_0 \quad \text{gegen} \quad H_1: \mu \neq \mu_0$$

Die Teststatistik ist wiederum:

$$T(\mathbf{X}) = \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{S_x} \stackrel{H_0}{\sim} t_{n-1}$$

und der Ablehnbereich ist: $(-\infty, -t_{n-1, 1-\frac{\alpha}{2}}) \cup (t_{n-1, 1-\frac{\alpha}{2}}, +\infty)$.

D.h. dann, dass wir uns sowohl für Teststatistiken, die größer als $t_{n-1, 1-\frac{\alpha}{2}}$ sind, immer mehr vom Annahmebereich entfernen als auch für Teststatistiken, die kleiner als $-t_{n-1, 1-\frac{\alpha}{2}}$ sind. Für den p-Wert bedeutet das, dass wir beide Richtungen berücksichtigen müssen (was wir mit Hilfe des Betrags der Teststatistik machen, wobei wir die Symmetrie der t-Verteilung ausnutzen):

$$\begin{aligned} p\text{-Wert} &= P(|T| \geq |t|) \\ &= P(T \leq -|t|) + P(T \geq |t|) \\ &= 2 \cdot P(T \geq |t|) \\ &= 2 \cdot (1 - P(T < |t|)) \\ &= 2 \cdot (1 - F(|t|)), \end{aligned}$$

wobei $F(\cdot)$ die Verteilungsfunktion der entsprechenden Verteilung ist – hier also der t-Verteilung mit $n - 1$ Freiheitsgraden.

17.2 Hypothesentests

In der Vorlesung haben wir eine Reihe von Hypothesentests kennengelernt. Auf die wichtigsten wird hier noch einmal intuitiv eingegangen. Bevor wir auf die verschiedenen Tests eingehen eine kleine Bemerkung zu einfachen bzw. doppelten und einseitigen bzw. zweiseitigen Tests: Ein einfacher Test (im Vergleich zu einem doppelten Test) betrachtet eine einzige Stichprobe. Der einfache Test kann aber sowohl einseitig ($H_0 : \mu \geq \mu_0$ gg. $H_1 : \mu < \mu_0$ bzw. umgekehrt) als auch zweiseitig ($H_0 : \mu = \mu_0$ gg. $H_1 : \mu \neq \mu_0$) formuliert werden. Ein doppelter Test hingegen betrachtet zwei Stichproben – auch dieser kann sowohl ein- als auch zweiseitig formuliert werden.

17.2.1 Einfacher Gauss-Test

Test des Erwartungswerts μ_x einer (per Annahme) normalverteilten ZV $X (\sim \mathcal{N}(\mu_x, \sigma_x^2))$. Die **Varianz** σ_x^2 ist dabei **bekannt**.

17.2.2 Einfacher t-Test

Test des Erwartungswerts μ_x einer (per Annahme) normalverteilten ZV $X (\sim \mathcal{N}(\mu_x, \sigma_x^2))$. Die **Varianz** σ_x^2 ist dabei **unbekannt** und muss aus der Stichprobe mittels der Stichprobenvarianz S^2 geschätzt werden.

17.2.3 Approximativer Einfacher Binomialtest

Test der Erfolgswahrscheinlichkeit p einer (per Annahme) bernoulliverteilten ZV $X (\sim B(1, p))$. Die Teststatistik ist approximativ standardnormalverteilt.

17.2.4 Chi-Quadrat-Anpassungstest

Test der unter H_0 erwarteten und tatsächlich Beobachteten absoluten Häufigkeiten von verschiedenen Ausprägungen einer ZV. Dadurch kann man testen, ob die Verteilung, von der die Stichprobe stammt, ungleich einer unter H_0 angenommenen Verteilung ist.

17.2.5 F-Test

Testet das Verhältnis (Größer/kleiner bzw. Gleichheit gegen Ungleichheit) der Varianzen zweier unabhängiger normalverteilter ZV. Die Teststatistik ist der Quotient der Stichprobenvarianzen und F-verteilt.

⚠ Die F-Verteilung hat **zwei Freiheitsgrade**. Denkt immer an beide (z.B. auch bei der F-Statistik eines multiplen Regressionsmodells).

17.2.6 Doppelter Gauss-Test

Testet das Verhältnis (Größer/Kleiner bzw. Gleichheit gegen Ungleichheit) der Erwartungswerte zweier unabhängiger normalverteilter ZVs (X und Y). Die **Varianzen** σ_x^2 und σ_y^2 sind dabei **bekannt**.

17.2.7 Doppelter t-Test

Testet das Verhältnis (Größer/Kleiner bzw. Gleichheit gegen Ungleichheit) der Erwartungswerte zweier unabhängiger normalverteilter ZVs (X und Y). Die **Varianzen** σ_x^2 und σ_y^2 sind dabei **unbekannt** und müssen aus der Stichprobe mittels der Stichprobenvarianz S^2 geschätzt werden. Hier wird zusätzlich angenommen, dass die zugrundeliegenden Verteilungen **beider ZVs dieselbe Varianz** haben ($\sigma_x^2 = \sigma_y^2$).

17.2.8 Welch-Test

Testet das Verhältnis (Größer/Kleiner bzw. Gleichheit gegen Ungleichheit) der Erwartungswerte zweier unabhängiger normalverteilter ZV (X und Y). Die **Varianzen** σ_x^2 und σ_y^2 sind dabei **unbekannt** und müssen aus der Stichprobe mittels der Stichprobenvarianzen S_x^2 und S_y^2 geschätzt werden. Hier wird zusätzlich angenommen, dass die zugrundeliegenden Verteilungen **beider ZVs ungleiche Varianzen** haben ($\sigma_x^2 \neq \sigma_y^2$).

17.2.9 Paired t-Test

Testet das Verhältnis (Größer/Kleiner bzw. Gleichheit gegen Ungleichheit) der Erwartungswerte zweier **abhängiger** normalverteilter ZVs (X und Y). Diese Abhängigkeit kann z.B. dadurch entstehen, dass man an denselben Untersuchungseinheiten dasselbe Merkmal zu zwei verschiedenen Zeitpunkten misst.

17.2.10 Approximativer Doppelter Binomialtest

Hier werden statt Mittelwerte Wahrscheinlichkeiten miteinander verglichen. Es handelt sich dabei aber immer noch um Erwartungswerte, da p ja genau der Erwartungswert einer ZV mit der Bernoulli-Verteilung $B(1, p)$ ist.

17.2.11 Mann-Whitney-U-Test für zwei unabhängige Stichproben

Vergleich der Lageparameter (das ist bei der Normalverteilung der Erwartungswert μ) zweier stetig verteilter ZVs, die ansonsten der gleichen Verteilung folgen. Hier wird im Vergleich zu den meisten bisherigen Tests die Annahme, dass die ZVs normalverteilt sein müssen, gelockert.

17.2.12 Kolmogorov-Smirnov-Anpassungstest

Test von Gleichheit gegen Ungleichheit der Verteilung zweier ZVs (X und Y). Wir berechnen dafür die empirischen Verteilungsfunktionen von X und Y (\hat{F} and \hat{G}). Dann schauen wir uns jede einzelne Beobachtung von X und Y an (bzw. wir fassen sie einfach zu einem Pool von Beobachtungen zusammen). Aus diesem Pool nehmen wir jeden Wert (t) einmal und betrachten $|\hat{F}(t) - \hat{G}(t)|$. Die größte dieser Differenzen ist die Teststatistik. Dann wird wie üblich gegen einen kritischen Wert eine Entscheidung getroffen. Das ist aber an dieser Stelle zu komplex \rightarrow wir benutzen Computer dafür.

17.2.13 Chi-Quadrat-Unabhängigkeitstest

Wie beim χ^2 -Anpassungstest werden hier beobachtete absolute Häufigkeiten mit unter H_0 erwarteten absoluten Häufigkeiten verglichen. Es gibt jedoch zwei Unterschiede. Erstens werden die beobachteten Häufigkeiten aus zwei Stichproben (für zwei ZVs X und Y) mit den unter H_0 erwarteten Häufigkeiten verglichen. Zweitens sind die unter H_0 erwarteten Häufigkeiten nicht die einer beliebigen Verteilung, sondern genau die, die man unter Unabhängigkeit beider ZVs erwartet.

17.2.14 Odds-Ratio-Test

Testet ebenfalls die Unabhängigkeit zweier Variablen mittels der Odds-Ratio, kann aber dementsprechend nur für Daten, die sich in einer Vier-Felder-Tafel darstellen lassen, angewendet werden.

17.3 Unterschied zwischen Gauss-Tests und t-Tests

Der Unterschied zwischen Gauss-Tests und t-Tests besteht darin, dass bei Gauss-Tests die Varianz der Verteilung, die der Stichprobe zugrunde liegt, bekannt ist. Bei t-Tests hingegen nicht. Dass führt dazu, dass die Teststatistik bei Gauss-Tests standardnormalverteilt ist. Bei t-Tests hingegen ist sie t-verteilt.

Was genau hat das jetzt mit der Testentscheidung zu tun? Dazu schauen wir uns erst einmal Fig. 4 an. Wir sehen, dass (vor allem für wenig Freiheitsgrade) die t-Verteilung breitere Enden hat als die Standardnormalverteilung, was wiederum bedeutet, dass alle α -Quantile für $\alpha < 0,5$ bei der t-Verteilung kleiner (also weiter im negativen Bereich) sind als die der Standardnormalverteilung. Alle α -Quantile für $\alpha > 0,5$ sind bei der t-Verteilung hingegen größer (also weiter im positiven Bereich) als bei der Standardnormalverteilung⁹. Das gilt dann insbesondere auch für die Quantile, die wir als kritische Werte für die Testentscheidung heranziehen. Somit verkleinert sich der Ablehnbereich bei t-Tests im Vergleich zu Gauss-Tests.

Dafür gibt es auch eine intuitive Begründung: Bei t-Tests müssen wir nicht nur den interessierenden statistischen Parameter (meistens μ) schätzen, sondern auch σ^2 mittels der Stichprobenvarianz S^2 . Da wir mehr Parameter schätzen, wollen wir H_0 nicht so bereitwillig ablehnen wie bei Gauss-Tests. Dementsprechend ist der Ablehnbereich bei t-Tests kleiner.

Aber auch dafür gibt es Abhilfe: Je größer der Stichprobenumfang n , desto größer die Anzahl der Freiheitsgrade und desto ähnlicher sind sich Standardnormal- und t-Verteilung. D.h. dann, dass wir für einen großen Stichprobenumfang (annähernd) dieselben kritischen Werte für Gauss- und t-Tests benutzen. Die zusätzliche Ungewissheit durch die Schätzung der Varianz (im Vergleich zu bekannter Varianz) wird also durch mehr Beobachtungen gemildert. Das ist natürlich ein intuitives Ergebnis: Je mehr Beobachtungen wir haben, desto "sicherer" sind wir uns mit unserer

⁹Der Median, also $\tilde{x}_{0,5}$, ist bei beiden Verteilungen der Wert 0, da sie um 0 zentriert sind

Schlussfolgerung.

17.3.1 Unterschied zwischen doppeltem t-Test und Welch-Test

Eine ähnliche Intuition wie für den Unterschied zwischen Gauss-Tests und t-Tests gibt es auch für den doppelten t-Test und den Welch-Test. Die Anzahl der Freiheitsgrade der Teststatistik vom doppelten t-Test ist größer als die beim Welch-Test. D.h. wir haben beim doppelten t-Test einen größeren Ablehnbereich als beim Welch-Test und wir sind uns somit "etwas eher mit der Entscheidung sicher", H_0 zu verwerfen, falls es denn dazu kommt. Das liegt daran, dass man beim doppelten t-Test zwei Stichproben hat, mit denen man eine unbekannte Varianz schätzt. Beim Welch-Test hingegen hat man zwei Stichproben, mit Hilfe derer man zwei unbekannte Varianzen schätzen muss. Deshalb ist beim letzteren etwas mehr Ungewissheit im Spiel.

17.4 Aufgaben

1. Welche Aussagen zum p-Wert sind richtig?

- a) Der p-Wert gibt die Wahrscheinlichkeit an, mit der mein beobachteter Wert richtig ist. ☐
- b) Der p-Wert gibt $P(X)$ an. ☐
- c) Die Hypothese, die man Beweisen will befindet sich in der Alternativhypothese H_1 . ☐
- d) Wenn der p-Wert kleiner als mein α ist, dann lehne ich die Nullhypothese ab. ☐
- e) Bei einem beidseitigen Test gibt es immer zwei Ablehnbereiche. ☐
- f) Der p-Wert gibt uns darüber Auskunft, mit welcher Wahrscheinlichkeit unser Wert oder ein Wert, der noch weiter von der Nullhypothese entfernt ist, auftritt, wenn ich daran festhalte, dass die Nullhypothese stimmt. ☐

2. Welche Aussagen über die Hypothesentests sind wahr?

- a) Mit Hypothesentest kann man entscheiden, ob eine Nullhypothese richtig ist. ☐
- b) Bei einem einfachen t-Test gibt es immer nur ein Ablehnbereich. ☐
- c) Der Unterschied zwischen dem einfachen Gauss-Test und einfachen t-Test ist die unbekannte Varianz beim t-Test. ☐
- d) Mit dem approximativen einfachen Binomialtest überprüft man, ob ein μ signifikant von μ_0 abweicht, also im Ablehnbereich liegt oder nicht. ☐
- e) Der Gauss-Test ist im Gegensatz zum t-Test normalverteilt. ☐

3. Welche Aussagen über die Hypothesentests sind wahr?

- a) Mit dem F-Test kann man das Verhältnis von Varianzen zweier unabhängigen Zufallsvariablen testen. ☐
- b) Beim doppelten Gauss-Test handelt es sich im Gegensatz zum einfachen Gauss-Test um das Verhältnis der Varianzen zweier unabhängiger, normalverteilter ZV. ☐
- c) Beim doppelten Gauss-Test sind ebenfalls die Varianzen unbekannt. ☐
- d) Der Unterschied zwischen des doppelten t-Test und des Welch-Test ist, dass beim doppelten t-Test die Varianzen bekannt sind und beim Welch-Test diese unbekannt sind. ☐
- e) Die Besonderheit beim Paired t-Test ist, dass es sich hierbei um einen Test mit zwei abhängige normalverteilte ZV handelt. ☐

4. Welche Aussagen über die Hypothesentests sind wahr?

- a) Mit dem Chi-Quadrat-Anpassungstest kann ist testen, ob die Verteilung zweier ZV ungleich ist. ☐
- b) Beim Approximativen Doppelten Binomialtest handelt es sich um den Vergleich von Wahrscheinlichkeiten und somit nicht mehr um den Vergleich von Erwartungswerten. ☐
- c) Der Vorteil beim Mann-Whitney-U-Test ist, dass die ZV nicht mehr unbedingt normalverteilt sein müssen. ☐
- d) Der Kolmogorov-Smirnov-Anpassungstest prüft, ob sich zwei ZV in ihrer Verteilung unterscheiden. ☐
- e) Beim Chi-Quadrat-Unabhängigkeitstest wird die Unabhängigkeit zweier Stichproben überprüft, indem die beobachteten Häufigkeiten mit den zu erwarteten Häufigkeiten ohne Abhängigkeit unter H_0 vergleicht. ☐
- f) Der Odds-Ration-Test testet die Unabhängigkeit von zwei beliebigen Variablen ohne Ausnahme. ☐

5. Welche Aussagen über die Hypothesentests sind wahr?

- a) Der Ablehnbereich ist beim Gauss-Test größer als beim t-Test. ☐
- b) Der Gauss-Test ist etwas unsicherer als der t-Test. ☐
- c) Bei größeren Stichproben nähern sich die Ergebnisse des t-Tests und des Gauss-Tests an. ☐
- d) Beim Welch-Test sind wir mit unserer Entscheidung ein wenig sicherer als beim doppelten t-Test, da wir nur eine unbekannte Varianz schätzen müssen. ☐

18 Lineare Regression

Die Lineare Regression ist bereits aus Kapitel 7 bekannt. Im Folgenden werden wir das aus Statistik I schon gelernte wiederholen und weiter vertiefen.

Bei der linearen Regression benötigt man zwei Variable die abhängige und die unabhängige. Die Annahme ist, dass zwischen diesen beiden eine lineare Beziehung herrscht. Die lineare Regression wird mit der Formel: $Y = \beta_0 + \beta_1 X + \epsilon$, wobei ϵ der unbeobachtete Fehlerterm ist. In Statistik II sind nun die Daten Realisationen von Zufallsvariablen, durch die Schlüsse auf die Grundgesamtheit geschlossen werden, im Gegensatz zu Statistik I. Durch die Regression kann man nun nicht mehr nur Aussagen über die erklärte Streuung machen, sondern auch über Verteilungsannahmen und die statistische Signifikanz.

Der Fehlerterm ϵ ist nun eine Zufallsvariable und für jede Beobachtung gibt es ein Fehlerterm, welche jedoch unabhängig voneinander sind und deren Erwartungswert 0 ist. Die Varianz der ϵ ist σ^2 , da durch den Verschiebungssatz schlussendlich nur noch der Erwartungswert der quadrierten Epsilons übrigbleibt. Zudem wird die Annahme getroffen, dass die Fehlerterme unabhängig voneinander sind, daher ist die Kovarianz bei gleicher Beobachtung σ^2 und bei ungleicher Beobachtung 0 und somit unkorreliert.

Wenn die Varianz σ^2 ist und somit konstant ist, dann ist der Fehlerterm unabhängig von dem Wert der Einflussgröße. Ist dies der Fall, dann spricht man von **Homoskedastizität**. Oftmals werden diese Annahmen jedoch verletzt und die Varianz der Fehlerterme nimmt mit steigendem x-Wert zu. Dies nennt man dann **Heteroskedastizität** was man gut in einem Residualplot oder oftmals auch schon in einem Scatterplot sehen kann.

Nun überträgt man die Verteilung der Epsilons auf die Zufallsvariable Y. Das heißt, wenn die Epsilons unabhängig sind, dann sind das auch die Zufallsvariablen. Eine weitere Annahme ist der Erwartungswert, dabei setzt man den Ansatz des Regressionsmodells ein und da der Erwartungswert von Epsilon gleich 0 ist, ist der Erwartungswert schlussendlich $\beta_0 + \beta_1 x_1$. Da dies als eine Konstante gesehen wird ist die Varianz der Zufallsvariable auch σ^2 . Wie auch bei den Annahmen der Epsilons ist die Kovarianz bei gleichen Beobachtungen σ^2 und bei ungleichen Beobachtungen 0. Somit sind die Zufallsvariablen auch normalverteilt, die Varianz bleibt die gleiche nur der Erwartungswert erfährt eine Verschiebung und liegt bei $\beta_0 + \beta_1 x_i$.

Matrixschreibweise Man kann das Regressionsmodell auch in der Matrixschreibweise schreiben, was bei vielen Beobachtungen einer Zufallsvariablen praktisch und übersichtlich sein kann. Dabei ist in jeder Zeile eine Regressionsgleichung einer Zufallsvariablen abgetragen. I_n ist die nxn Einheitsmatrix. Ein Element in der Einheitsmatrix ist 1, wenn die i-te Zeile gleich der j-ten Spalte ist. Ist dies nicht der Fall, dann ist das Element in der Einheitsmatrix 0

Wie bei Statistik I in Kapitel 7.6, kann auch hier die Dummy und Effektkodierung angewendet werden. Als kleine Wiederholung, man wählt eine Referenzkategorie aus, welche bei der Dummyvariablen durch Nullsetzen der anderen Variablen entsteht und bei der Effektkodierung durch

das -1 setzen aller anderen Variablen. Die Regressionsgleichungen der einzelnen Merkmalsausprägungen kann dann schlussendlich durch einsetzen von 0, 1 und bei Effektkodierung von -1 in die allgemeine Regressionsgleichung gebildet werden.

18.1 Kleinste-Quadrate-Schätzer

Auch das Thema der Kleinsten-Quadraten-Schätzer haben wir schon in Statistik I 7.3 besprochen. Hier ist nun zu beachten, dass wir nicht mehr a und b als Schreibweise benutzen sondern nun $\beta_0, \beta_1, \beta_2, \dots$

Somit ergibt sich als Schätzer für $\hat{\beta}_1$: $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{y} \bar{x}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$ und für den Intercept bzw. y-Achsenabschnitt: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Da σ^2 unbekannt ist und auch geschätzt werden muss, ergibt sich als Verteilungsannahme für die standardisierten Parameterschätzer eine t-Verteilung mit 2 Freiheitsgraden.

18.2 Multiple lineare Regression

Bei der multiplen linearen Regression haben wir nun nicht mehr eine Einflussgröße X, sondern p verschiedene Einflussgrößen. Die verschiedenen Beobachtungen einer kann man dann wieder in eine Matrix übertragen.

Die Annahmen für die lineare Regression sind soweit wieder die gleichen wie bei der einfachen linearen Regression und die Schätzung kann man wieder mit der Kleinste-Quadrate-Methode bestimmen. Daraus ergeben sich dann die Eigenschaften des KQ-Schätzers für β . $\hat{\beta}$ ist erwartungstreu, da dessen Erwartungswert β ist. Der β -Schätzer lässt sich berechnen als X transponiert mal X, woraus sich eine $p+1 \times p+1$ -Matrix ergibt, daraus die Inverse und multipliziert dies mit X transponiert y. Zwei weitere Eigenschaften sind, dass $\hat{\beta}$ unter den linearen Schätzern der bestmögliche unverzerrte Schätzer und normalverteilt ist. Die Kovarianz ist $\sigma^2(X'X)^{-1}$. Da σ^2 ein Schätz ist, ist die Kovarianz auch nur ein Schätzer. Bei der Berechnung des geschätzten σ^2 ist nun die Besonderheit, dass nicht mehr durch n-2 Freiheitsgrade geteilt wird, wie bei der einfachen linearen Regression, sondern durch n-(p+1). P ist dabei die Anzahl der Kovariablen im Modell und +1 bezieht sich auf den Intercept, der zusätzlich noch abgezogen wird, da er nicht in p enthalten ist.

Dies spiegelt sich auch in der Verteilungsannahme für die standardisierte Parameterschätzung wider, da die Verteilung eine t-Verteilung mit n-p-1 Freiheitsgraden ist. Die standardisierte Parameterschätzung berechnet man mit der Differenz aus $\hat{\beta}_p$ und dessen Erwartungswert und teilt durch die geschätzten Standardfehler von $\hat{\beta}_p$.

Bemerkung: Lineare Modelle sind möglicherweise allgemeiner als man denkt, da nur die Parameter β linear sein müssen.

18.3 Signifikanztests und Konfidenzintervalle

Mit dem Signifikanztest soll überprüft werden, ob eine j-te Einflussgröße einen signifikanten Erklärungswert auf die Zielgröße besitzt. Dabei kommt in die Alternativhypothese die Eigenschaft die wir nachweisen wollen, dass β_j einen statistisch signifikanten Einfluss hat. Somit ist die Alterna-

tivhypothese $H_1 : \beta_j \neq 0$ und damit die Nullhypothese $H_0 : \beta_j = 0$. Die Teststatistik mit ihrem Ablehnungsbereich findet man in der Formelsammlung unter *Test für Regressionskoeffizienten* auf Seite 9. Man kann die komplette t-Statistik auch quadrieren, was dann der F-Verteilung mit $1, (n-p-1)$ Freiheitsgraden entspricht. Zudem findet man auch die Konstruktion des Konfidenzintervalls für β_j .

18.4 Bestimmtheitsmaß und Overall F-Test

18.4.1 Bestimmtheitsmaß

Bereits aus Statistik I ist das Bestimmtheitsmaß R^2 zur Beurteilung der Güte der Regression bekannt. Die Berechnung von R^2 ist auch beim multiplen Regressionsmodell möglich. R^2 berechnet sich wieder aus dem Quotienten der erklärten Streuung SS_{Reg} und der Gesamtvarianz von Y SS_Y . Die erklärte Streuung entspricht dabei dem Produkt aus Zeilen und Spaltenvektor $(\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y})$ und SS_Y berechnet sich aus $(Y - \bar{Y})'(Y - \bar{Y})$. R^2 kann man auch mit der Reststreuung RSS berechnen, wobei RSS die Residuen Streuung $\hat{\epsilon}'\hat{\epsilon}$ ist.

Die perfekte Anpassung ist wenn $\hat{\epsilon}$ gleich 0 ist und $R^2 = 1$ ist. Erklärt das Regressionsmodell gar keine Streuung, dann ist die Gesamtstreuung gleich der Residuenstreuung und $R^2 = 0$. Man spricht von Nullanpassung.

18.4.2 Overall F-Test

Mit dem Overall F-Test kann man testen, ob die Regressoren überhaupt einen statistisch signifikanten Erklärungswert für die abhängige Variable Y liefert. Dabei steht in der Nullhypothese die Annahme, dass kein Regressor einen signifikanten Erklärungswert über die abhängige Variable Y liefert: $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$. In der Alternativhypothese steht dann: $H_1 : \beta_j \neq 0$ für mind. ein j. Das heißt, die Nullhypothese wird abgelehnt, wenn mindestens einer der Regressoren eine signifikante Erklärung für die unabhängige Variable liefert. Somit testet man das Modell mit allen Regressoren gegen das Modell nur mit Intercept, da alle Betas null sind.

Die Teststatistik berechnet man mit $F = \frac{R^2}{1-R^2} \cdot \frac{n-(p+1)}{p}$. Somit verwendet man bei der Teststatistik das Bestimmtheitsmaß R^2 . Wie der Name schon sagt, handelt es sich hier um eine F-Verteilung. Um die Nullhypothese ablehnen zu können, muss mein berechnetes F größer sein, als $F_{1-\alpha}$ mit p und $n-(p+1)$ Freiheitsgraden. Hier liegt anders wie beim t-Test nur ein Ablehnungsbereich vor. Die Nullhypothese kann somit nur bei positiven Werten angelehnt werden.

Bemerkung: Hat man einen R-Output gegeben, dann befindet sich meistens am Ende schon das berechnete Bestimmtheitsmaß.

Mit einem Scatterplot kann man den Zusammenhang eines Regressors mit der abhängigen variablen Y graphisch darstellen. Dabei kann man eine Regressionsgleichung mit hineinlegen, um zu sehen, in wie fern diese den Daten entspricht. Somit kann man mehr oder weniger graphisch sehen, wie gut ein Regressor die unabhängige Variable erklärt. Zudem gibt es auch Multivariate Analysen wie die Modellzusammenfassung oder die ANOVA Tabelle, um schnell einen Überblick über den

Erklärungswert der Regressoren über die unabhängige Variable zu bekommen.

Eine gute Übersicht zu den Annahmen und den diversen Plots (Residualplot, QQ-Plot, etc.) findet ihr unter folgendem Link:

<https://wikis.fu-berlin.de/display/fustat/Residuenplots>

18.5 Aufgaben

1. Welche Aussagen über die einfache lineare Regression sind wahr?

- a) Der Erwartungswert der Residuen sollte immer 1 sein. ☐
- b) Bei der Homoskedastizität geht man von einer konstanten Varianz der Fehlertermen aus. ☐
- c) In Statistik II trifft man die Annahme, dass die Fehlerterme eine Zufallsvariable sind. ☐
- d) In Statistik II kann man keine Schlüsse über die Grundgesamtheit ziehen, da schlussendlich alles nur Annahmen sind und nichts sicher ist. ☐

2. Welche Aussagen über die einfache lineare Regression sind wahr?

- a) Mit steigenden x-Werten, sollte auch die Varianz der Residuen steigen. ☐
- b) Die Normalverteilung der Zufallsvariablen lautet $N(0, \sigma^2)$ ☐
- c) Bei der Effektkodierung wählt man im Gegensatz zur Dummykodierung keine Referenzkategorie. ☐
- d) Bei der Effektkodierung wird die Referenzkategorie mit -1 setzen aller Effektvariablen gebildet. ☐

3. Welche Aussagen bezüglich der multiplen linearen Regression sind richtig?

- a) Eine Eigenschaft des KQ-Schätzers ist, dass er erwartungstreu ist. ☐
- b) $\hat{\beta}$ ist der allgemein beste Schätzer für β . ☐
- c) $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$ hat 5 Freiheitsgrade. ☐
- d) $y = \beta_0 + \beta_1 \sqrt{x^2} + \beta_2 + \beta_3 x^{3,75}$ ist eine lineares Modell. ☐
- e) Die Freiheitsgrade der t-Statistik werden mit n-2 berechnet. ☐

4. Welche Aussagen bezüglich des Signifikanztest und der KI-Intervalle sind richtig?

- a) Mit dem Signifikanztest für einen Parameter kann man überprüfen, ob die Nullhypothese signifikant richtig ist. ☐
- b) Da in der Nullhypothese angenommen wird, dass β_j gleich 0 ist, muss man diesen nicht in der Teststatistik nicht weiter beachten. ☐
- c) Das Konfidenzintervall liegt symmetrisch um den Schätzer. ☐
- d) Wenn $t = -4,8$ ist und $t_{n-p-1, 1-\alpha} = 4,5$, dann kann ich die Nullhypothese nicht ablehnen. ☐
- e) Eine Teststatistik mit t_{194} Freiheitsgraden und $n=200$ hat ein $p=6$. ☐

5. Welche Aussagen über das Bestimmtheitsmaß und den Overall F-Test sind richtig?

- a) Wenn die Nullhypothese abgelehnt wird, dann haben alle Regressoren einen statistisch signifikanten Erklärungswert für die Zielvariable Y. ☐
- b) Bei $R^2 = 1$ spricht man von einer Nullanpassung. ☐
- c) Bei $R^2 = 0$ wird die Streuung vom Regressionsmodell komplett erklärt. ☐
- d) $R^2 = \frac{RSS}{SSY}$ ☐
- e) Ist $R^2 = 0$, dann wird H_0 nicht abgelehnt. ☐
- f Ist mein berechneter Wert F negativ, dann wird die H_0 -Hypothese abgelehnt. ☐

19 R-Einführung Teil II

Teil der Veranstaltung ist eine erweiterte Einführung in R sowie eine Computervorlesung. Dieses Kapitel dient als Platzhalter, falls in diesem Skript in Zukunft der Inhalt aus diesem Vorlesungskapitel vertieft werden sollte. Bis dahin wird empfohlen mit den vorhandenen Vorlesungsmaterialien zu arbeiten, da diese bereits sehr ausführlich und weitestgehend selbsterklärend sind.

References

- * Vorlesungsunterlagen "Statistik 1 für Wirtschaftswissenschaftler" (WiSe 2020/21, Prof. Heumann) an der LMU München.
- * Vorlesungsunterlagen "Statistik 2 für Wirtschaftswissenschaftler" (SoSe 2020, Prof. Heumann) an der LMU München.
- * Toutenburg, H. and Heumann, C., 2008. Induktive Statistik: eine Einführung mit R und SPSS. Springer-Verlag.
- * Dr. Alexander Engelhardt: <https://www.crashkurs-statistik.de>
- * https://www.statistik.uni-muenchen.de/formulare/skripte_u_aehnliches/mathehandrechnung_schneider.pdf
- * <https://www.mathebibel.de>
- * https://de.wikibooks.org/wiki/Mathe_f%C3%BCr_Nicht-Freaks
- * <https://www.statistik-nachhilfe.de>
- * <https://matheguru.com>