Statistik

Tristan Hörmann

Contents

Chapter 1

Einführung

1.1 Geschichtliches

Example 1.1.1

"Wenn ein Mensch stirbt, ist es ein Malheur, bei 100 Toten ist es eine Katastrophe, ab 1000 Toten eine Statistik." - Eichmann
Diese Form von Statistiken als Erbsenzählerei gibt es schon sehr lange. Die ersten geschrieben Texte verweisen auf Zahlen, Statistiken wurden des Pyramidenbaus angefertigt, die Bibel beschreibt eine Volkszählung.
Man kann sich nur schwer eine Zeit ohen Zahlen vorstellen, es muss sie aber gegeben haben. Dann haben die Leute Sachen eben nicht durchnummeriert, sondern wahrscheinlich Namen gegeben. Statt "es fehlen 2 Tiere" haben sie vermutlich "es fehlen Peter und Paul" gesagt.

Es gibt aber auch eine Definition von Statistik im instrumentalen Sinne: Verfahren, nach denen empirische Zahlen gewonnen, dargestellt, verarbeitet, analysiert und für Schlussfolgerungen, Prognosen und Entscheidungen verwendet werden.

Diese beiden Definitionen sind nacheinander entstanden. Erst hat man Jahrhunderte lang Daten gesammelt, bis irgendwann bis Notwendigkeit kam, Verfahren zu entwickeln, um die Übersicht zu erhöhen, ohne allzu viel Information zu verlieren.

Claim 1.1.1 Relevanz von Daten

Ohne Daten machen statistische Analyse und deren Methoden keinen Sinn.

Es gibt drei verschiedene Formen des Wissens oder der Erkenntnisgewinnung: *Wissen durch Wahrnehmen, Wissen durch Logik oder Schlussfolgern und Wissen duch Glauben. Diese Formen des Wissens wirken zusammen und beeinflussen sich wechselseitig beim Erzeugen und Verwenden von Wissen.* Leider ist *Wissen durch Glauben vielleicht die häufigste Form des Wissens.*
Die statistische Analyse selber hat mit den beiden ersten Formen des Wissens zu tun. Sie gibt die Instrumente, Informationen zu überprüfen, um sie nicht einfach glauben zu müssen.

1.2 Erste Anwendungen

Definition 1.2.1: Deskriptive und empirische Statistik

Die Analyse der (Daten-)Vorgänge oder Verteilungen innerhalb der Grundgesamtheit bzw innerhalb der Stichprobe nennt man **deskriptive** oder **beschreibende** Statistik.
Wird die Verteilung nicht analysiert, sondern **erhoben**, so wird sie **empirisch** genannt.

Für die Grundgesamtheit ist eine ständige Erhebung oft nicht machbar. Deshalb wird versucht, die Verteilung zu modellieren. Also eine Funktion zu definieren, welche die Verteilung möglichst gut beschreibt. Man spricht dann von einer **theoretischen** Verteilung. Ein berühmtes Beispiel wäre die Normalverteilung.

Ist die Grundgesamtheit bekannt und will man im voraus Aussagen über die zu ziehende Stichprobe treffen, spricht man von **deduktiver Statistik**. Dies ist im bekanntesten Fall die Wahrscheinlichkeitsrechnung.
Ist umgekehrt nur die Stichprobe bekannt und will man eine Aussage über die unbekannte Grundgesamtheit treffen, spricht man von **Inferenzstatistik oder deduktiver Statistik**. Weitere Synonyme für diese sind *analytische Statistik* oder *beurteilende Statistik*.

Example 1.2.1 (Beispielhafte Zuordnung von Inferenzstatistik und deskriptiver Statistik)

- *Mitteilung über die Zahl der Schüler, die in einer Klassenarbeit gut oder schlecht abgeschnitten haben*: Deskriptiv.
- *Berechnung eines Stichprobenmittelwertes zur Bezeichnung der zentralen Tendenz in den Daten*: Deskriptiv.
- *Durchführung einer Untersuchung, um den Zusammenhang zwischen Bildungsniveau und Einkommen in der BRD zu bestimmen*: Inferenzstatistik.
- *Angabe zum Durchschnittsgehalt von Beamten in den neuen Bundesländern auf der Basis der gesamten Gehaltsstatistik*: Deskriptiv.
- *Schätzung der durchschnittlichen Regenmenge in 1999 in Hannover*: Inferenzstatistik

Chapter 2

Empirische Verteilungen

Es handelt sich hier überwiegend um die Bearbeitung von größeren gesammelten Datenmengen, wie man sie ordnet, wie man sie darstellt und wie man sie in Tabellen oder gar Kennzahlen zusammen fassen kann. Ganz allgemein entstehen diese Daten durch Stichprobenerhebung oder Befragung. Wie die Erhebung oder Befragung organisier wird, ist Sache von Marktforschungsabteilungen.

Claim 2.0.1 Zusammenhang Datenerhebung und Datenqualität

Man muss sich darüber im Klaren sein, dass die Daten niemals besser sein können, als die Erhebungsmethode, aus welcher sie hervorgingen.

Daraus ergibt sich, dass Marktforscher im Idealfall gute Statistikkenntnisse aufweisen. Diese Vorlesung arbeitet jedoch nach dem Leitsatz: *Statistik beginnt erst, wenn die Daten vorliegen!*

2.1 Entstehung der Daten

Man muss sich in diesem Rahmen

1. zuerst darüber im Klaren sein, welche Frage eigentlich beantwortet werden muss. (**Forschungshypothese**)
2. dan Gedanken darüber machen, wie diese Frage beantwortet werden kann. Die Entstehung der **Untersuchungsvariablen**

Um über die Qualität der Operationalisierung bzw der Messung urteilen zu können, müssen Gütekriterien her:

1. **Objektivität:** Merkmalsauswahl bzw deren Auswerten und Interpretation sollte unabhängig vom jeweiligen Forscher erfolgen
2. **Zuverlässigkeit:** Wird die Messung wiederholt, sollten ähnliche Ergebnisse rauskommen
3. **Gültigkeit:** Messfehler sollten so klein wie möglich sein

2.2 Merkmale und ihre Eigenschaften

Die empirische Statistik untersucht die Verteilung von Merkmalen oder Variablen, die nicht vorher durch Theorien bekannt sein können. Die Ergebnisse sind nicht vorhersehbar, da das gemessene unter den Merkmalsträgern variiert.
Merkmale können als **Untersuchungsgegenstand** definiert werden. Es ist eine Eigenschaft, die an einem Untersuchungssubjekt (**Merkmalsträger**) beobachtet wird. Die Merkmale und ihre **Merkmalsausprägungen** werden in der Planung der Untersuchung festgelegt.

Example 2.2.1 (Merkmalsträger und -ausprägung)

Bei einer Umfrage sind die befragten Personen die Merkmalsträger. Die abgefragten Gegenstände (Haarfarbe, Kinderzahl, Alter, Geschlecht) sind die Merkmale. Die Ausprägungen sind entsprechend

(blond, schwarz, rot,..) und (0,1,2,..) usw.

Wie man sieht, können sehr verschiedene Eigenschaften gemessen werden. Eine Untersuchung kann nur eine begrenzte Anzahl von Merkmalen aufnehmen und kann demnach nur als vereinfachtes Abbild der Realität fungieren.
Bei dem obigen Beispiel sind die Merkmale bewusst so gewählt worden. Es zeigt, dass die Ausprägungen sehr verschiedener Art sein können. Diese Unterscheidung wird unter dem Begriff **Messniveau** zusammengefasst:

Definition 2.2.1: Messniveaus / Skalierungen

Es wird in unterschiedliche Skalierungen unterteilt:

- **qualitativ** (nicht-metrische Skalierung)
- **quantitativ** (metrische Skalierung)

Die nicht-metrische Skalierung wird wie folgt unterschieden:

- **nominal**: Die Ausprägungen werden lediglich kategorisiert, ohne den Kategorien eine Rangfolge oder einen numerischen Wert zuzuweisen. (Geschlecht, Haarfarbe, etc)
- **ordinal**: Die Ausprägungen haben eine Rangordnung, allerdings ist der Abstand zwischen den Ausprägungen nicht gleichmäßig oder gar unbekannt. (Schulnoten (1, 2, 3, 4, 5, 6), Zufriedenheitsstufen (sehr zufrieden, zufrieden, unzufrieden, sehr unzufrieden) oder sozioökonomischer Status (niedrig, mittel, hoch))

Auf nicht-metrischen Skalen sind arithmetische Operationen nicht sinnvoll und teilweise auch nicht möglich.
Die metrische Skalierung wird wie folgt unterschieden:

- **Intervallskala**: Die Ausprägungen haben eine Rangordnung und einen bekannten, gleichmäßigen Abstand. Es gibt jedoch keinen absoluten Nullpunkt. (IQ-Wert, Temperatur in Celsius)
- **Verhältnisskala**: Die Ausprägungen haben eine Rangordnung und einen bekannten, gleichmäßigen Abstand **und** einen absoluten Nullpunkt. Zusätzlich wird hier in **diskret** und **stetig** unterschieden.

Auf metrischen Skalen sind arithmetische Operationen sinnvoll. Ist kein Nullpunkt vorhanden, so sind lediglich Addition und Subtraktion sinnig. Ist ein Nullpunkt vorhanden, so können die Operationen um Multiplikation und Division ergänzt werden.

Note:-

Bei Untersuchungen wird man oft feststellen, dass ein diskretes Mekrmal, welches sehr viele Ausprägungen (Geldbeträge in Cent) annehmen kann, häufig als stetiges Merkmal behandelt wird. Man spricht bei ihnen von **quasistetigen** Merkmalen. Andererseits werden auch einige stetige Merkmale als diskret erhoben (Lebensdauer).

2.3 Definition und Notationen

Definition 2.3.1: Grundbegriffe

Die Menge aller für die Untersuchung relevanten Merkmalsträger ist die **Grundgesamheit**.
Die Menge der in der Untersuchung betrachteten Merkmalsträger nennt man die **Strichprobe**.
Die Gesamtheit aller Daten über Merkmale, Merkmalsträger und Merkmalsausprägungen wird als Beobachtungsdaten oder **Urliste** bezeichnet.

Man spricht immer von n vorhandenen Merkmalsträgern. Es wird ein **Laufindex** j für den j-ten Merkmalsträger definiert:

$$j = 1, 2, \ldots, n-1, n$$

X und Y werden immer die Merkmale allgemein bezeichnen. x_j bezeichnet die Ausprägung von Merkmal X für den j-ten Merkmalsträger.
Eine Datenreihe für ein Merkmal X lautet dann entsprechend:

$$x_1, x_2, \ldots, x_{n-1}, x_n$$

2.4 Univariate Darstellung für verschiedene Merkmalsniveaus

2.4.1 Nominalskalierte Merkmale

Nominalskalierte Merkmale lassen sich in einer Häufigkeitstabelle (2.4.1) darstellen und zusammenfassen.

x_i	0	1	2	3	$\sum$	aufgetretene Ausprägungen
n_i						absolute Häufigkeiten
$p_i = \frac{n_i}{n}$						relative Häufigkeiten
$100p_i$						prozentuale Häufigkeiten
p_i^2						quadrierte rel. Häufigkeiten

Table 2.1: Häufigkeitstabelle nominaler Merkmale

Die Häufigkeitstabelle besteht aus k Klassen bzw Kategorien.
Aus der Tabelle kann man mehrere Sachen ablesen, die man ohnehin schon wusste:

Claim 2.4.1 Summe der absoluten Häufigkeiten

Die absoluten Häufigkeiten, die der Tabelle ihren Namen geben, addierern sich zu n (dem Stichprobenumfang)

$$\sum_{i=1}^{k} n_i = n$$

Beispiel: 3 Leute, davon 2 blond, einer dunkelhaarig. $2 + 1 = 3$.

Claim 2.4.2 Relative Häufigkeiten aus absoluten Häufigkeiten

Dividiert man die absoluten Häufigkeiten durch n, erhält man die relativen Häufigkeiten. Damit kann man Stichproben mit verschiedenen Umfängen vergleichen. Die relativen Häufigkeiten addieren sich zu 1:

$$\frac{1}{n} \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} p_i = 1$$

Um eine Idee von der Streuung der Daten zu bekommen, wird ein Maß eingeführt, welches man den **Index of Diversity** nennt:

Definition 2.4.1: Index of Diversity

$$D = 1 - \sum_{i=1}^{k} p_i^2$$

Der Index of Diversity liegt in einem Wertebereich $0 \leqslant D < 1$. Ein Wert von 0 bedeutet, dass alle Merkmalsausprägungen identisch sind (keine Vielfalt), während ein Wert nahe 1 auf eine hohe Vielfalt in der Verteilung der Merkmalsausprägungen hindeutet. Es ist ratsam, den Wert D nicht an 1, sondern an $1 - k^{-1}$ zu vergleichen, weil k in den seltensten Fällen unendlich groß sein wird.

2.4.2 Ordinalskalierte Merkmale

Ordinalskalierte Werte sind Werte, die eine klare Hierarchie aufweisen, wobei die Abstände zwischen den Werten nicht gleich interpretiert werden können.

x_i	0	1	2	3	4	$\sum$	aufgetretene Ausprägungen
n_i							absolute Häufigkeiten
$p_i = \frac{n_i}{n}$							relative Häufigkeiten (Prozentwert)
F_i							kumulierte relative Häufigkeiten

Table 2.2: Häufigkeitstabelle ordinaler Merkmale

Das einzig Neue in dieser Tabelle (2.2) sind die **kumulierten relativen Häufigkeiten** (auch Summenhäufigkeiten). Diese sind vor allem interessant, weil sie Vergleiche zwischen Stichproben unterschiedlicher Umfänge erlauben. Dies ist aufgrund der eindeutigen Ordnung der Werte möglich.

Definition 2.4.2: Kumulierte relative Häufigkeit

$F(x_i)$ stellt den Anteil der Werte dar, die höchstens die Merkmalsausprägung x_i aufweisen.

$$F(x_i) = F_i = A(X \leqslant x_i) = \sum_{j=1}^{i} p_j$$

Üblicherweise wird die Summenhäufigkeit als Treppenfunktion dargestellt.

Übliche Kennwerte sind die **Quantile** (oder Prozentpunkte). Die repräsentieren bestimmte Punkte einer Verteilung und teilen eben diese, basierend auf den kumulierten relativen Häufigkeiten, in gleich große Abschnitte ein.

Definition 2.4.3: Quantile

Quantile X_w werden als die erste Merkmalsausprägung definiert, für welche die kumulierte relative Häfigkeit größer gleich w ist.

$$F_i \geqslant w$$

Dies bedeutet, dass sie den Punkt in der Verteilung angeben, an dem mindestens ein bestimmter Prozentsatz w der Daten erreicht oder überschritten wird.

Claim 2.4.3 Quartile

Zu den wichtigsten Quantilen gehören die Quartile. Sie werden als Q_i mit $i \in \{1, 2, 3\}$ angegeben. Man spricht dann von dem ersten, zweiten oder dritten Quartil. Die Quartile teilen die Rangwertreihe in 4 gleichmäßige Abschnitte auf und trennen bei 25%, 50% und 75%.

Der **Median** ist ein bekanntes Beispiel für ein Quantil. Er ist der Wert, bei dem die kumulierte relative Häufigkeit erstmals 50% erreicht oder überschreitet. Das bedeutet, dass 50% der Daten kleiner oder gleich dem Median sind, während die anderen 50% der Daten größer oder gleich dem Median sind. Der Median ist somit Q_2 und berechnet sich als mittelster Wert der sortierten Liste, wenn die Anzahl der ungerade ist und als Mittelwert der beiden mittleren Werte, wenn die Anzahl gerade ist.

2.4.3 Streuungsmaße

- Die **Spann- oder Streuweite** R (wie Range) ist die Differenz zwischen den Extremwerten:

$$R = x^{max} - x^{min}$$

Die Streuweite wird üblicherweise zusammen mit dem Mittelwert angegeben, weil sie andernfalls nicht mehr aussagekräftigt wäre.

- Der **Interquartilsabstand** IQA gibt den Bereich an, der von den mittleren 50% der Werte bedeckt wird.

$$IQA = Q_3 - Q_1$$

- Der **relative Quartilsabstand** berechnet sich aus dem Verhältnis von IQA und R. Je näher diese Zahl an 1 ist, desto gestreuter ist die Verteilung. Je näher diese Zahl an 0 ist, desto konzentrierter ist die Verteilung.

$$IQA_r = \frac{IQA}{R}$$

- Der **mittlere Quartilsabstand** MQA wird als die Hälfte des IQA definiert.

$$MQA = \frac{IQA}{2} = \frac{Q_3 - Q_1}{2}$$

  Bei symmetrischen Verteilungen gilt entsprechend mit Md als Median:

$$MQA = \frac{Q_3 - Q_1}{2} = Q_3 - Md$$

2.4.4 Intervallskalierte und diskret proportionalskalierte Merkmale

Auch hier kann man den Median, Index of Diversity und bisher bekanntes berechnen. Neu ist nachfolgend das arithmetische Mittel und die Standardabweichung.

Definition 2.4.4: Arithmetisches Mittel (Durchschnitt)

Für den Durchschnitt, genauer: das arithmetische Mittel, wird die Summe aller Werte der Urliste berechnet und durch die Anzahl der summierten Werte addiert. Mathematisch ausgedrückt:

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

Falls die Daten nur in Häufigkeitstabellen bzw relativen Häufigkeiten (p_z) mit k Klassen gegeben sind, ändert sich die Formel wie folgt:

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j = \frac{1}{n} \sum_{z=1}^{k} x_z n_z = \sum_{z=1}^{k} x_z \frac{n_z}{n}$$
$$= \sum_{z=1}^{k} x_z p_z$$

Claim 2.4.4 Symmstrie

Ist eine Verteilung symmetrisch, so ist der häufigste Wert in der Mitte, der Median auch und der Durchschnitt auch. Deshalb gilt:

$$\overline{x} = \tilde{x} = x_D$$

Wobei $\overline{x}$ der Durchschnitt ist, $\tilde{x}$ der Median und x_D der häufigste Wert ist.

Zu beachten ist außerdem, dass das arithmetische Mittel sehr empfindlich gegenüber Ausreißern in den Daten ist. Eine eigentlich sehr konzentrierte Verteilung mit einzelnen Werten weit ab vom Rest, kann den Durchschnitt erheblich beeinflussen. So wäre der Durchschnitt von $X_1 = \{1, 2, 3, 4, 4, 3, 2, 1\}$ gleich 2.5 ist, während der Durchschnitt von $X_2 = \{1, 2, 3, 4, 4, 3, 2, 25\}$ gleich 5.5 ist. Hieran sieht man deutlich, dass ein solcher Ausreißer den Durchschnitt in seiner Aussagekraft deutlich mindert.

Claim 2.4.5 Summe der Abweichungen vom arithmetischen Mittel ist immer 0

Gegeben ist eine Urliste von Werten X mit Mächtigkeit n und dem zugehörigen arithmetischen Mittel $\overline{x}$. Es gilt immer:

$$\sum_{j=1}^{n}(x_j - \overline{x}) = 0$$

Im gleichen Zuge die **Abweichung vom Durchschnitt** für die j-te Ausprägung wie folgt definiert:

$$x^* = x_j - \overline{x}$$

Verdeutlicht an einem Beispiel mit $X = \{1, 2, 3\}$ und entsprechend $\overline{x} = 2$:

$$\begin{aligned} Abweichung &= \sum_{j=1}^{n}(x_j - \overline{x}) = \sum_{j=1}^{3}(x_j - \overline{2}) \\ &= (1-2) + (2-2) + (3-2) \\ &= -1 + 0 + 1 \\ &= 0 \end{aligned}$$

Weil mit dieser Darstellung der Abweichung vom Durchschnitt nicht gut zu arbeiten ist (und sie maximal Rechenfehler aufzeigen kann). Diese Tatsache zeigt lediglich, dass das arithmetische Mittel einen Gleichgewichtspunkt in der Verteilung der Urliste darstellt. Es gibt jedoch Möglichkeiten, die Abweichung vom Durchschnitt so zu modifizieren, dass sie ein aussagekräftiger Indikator im Messen von Abweichungen werden kann.

Definition 2.4.5: Mittlere quadratische Abweichung

Es wird d^2 als mittlere quadratische Abweichung definiert, in dem die Abweichung vom Durchschnitt quadriert wird und man den Durchschnitt der quadrierten Abweichungen berechnet. x_j^* stellt in diesem Fall die Abweichung vom Durchschnitt für die j-te Ausprägung dar:

$$d^2 = \frac{1}{n}\sum_{j=1}^{n} x_j^{*2} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2$$

Aus der mittleren quadratischen Abweichung lässt sich anschließend die Stichprobenvarianz berechnen.

Definition 2.4.6: Stichprobenvarianz

Die Stichprobenvarianz s^2 wird wie folgt berechnet:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n} x_j^{*2} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$$

Der offensichtlich einzige Unterschied zu d^2 liegt darin, dass die Summe keinen Faktor mehr von n^{-1} aufweist, sondern einen Faktor von $(n-1)^{-1}$. Entsprechend lässt sich die Stichprobenvarianz auch wie folgt beschreiben:

$$s^2 = \frac{n}{n-1} \cdot d^2$$

Mit der definierten Stichprobenvarianz lässt sich das bekannte Abweichungsmaß der **Stichprobenstandardabweichung** bestimmen.

Definition 2.4.7: Stichprobenstandardabweichung (Standardabweichung)

Die Stichprobenstandardabweichung s ist wie folgt definiert:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2} = \sqrt{\frac{n}{n-1} \cdot d^2}$$

Man kann bei s annehmen, dass es in den sleben Einheiten ausgedrückt wird, wie die Daten selber. Die Standardabweichung stellt die durchschnittliche Abweichung aller Werte von dem arithmetischen Mittel dar.

Claim 2.4.6 Mittleres Schwankungsintervall

Am aussagekräftigsten ist die Angabe des mittleren Schwankungintervalls:

$$\overline{x} \pm s = [\overline{x} - s, \overline{x} + s]$$

Je mehr sich die Grenzen dieses Intervalls von den Grenzen des Definitionsbereichs der Urliste absetzt, desto weniger streut die Verteilung.

Die praktische Vorgehensweise zur Berechnung von s benötigt zusätzlich den **Durchschnitt der Quadrate** der Merkmalsausprägungen:

$$\overline{x^2} = \frac{1}{n}\sum_{j=1}^{n} x_j^2$$

um d^2 wie folgt zu definieren:

$$d^2 = \overline{x^2} - \overline{x}^2$$

Definition 2.4.8: Variationskoeffizient

Der Variationskoeffzient v ist ein Maß, welches ein Vergleich zwischen zwei Verteilungen schaffen kann. Er gibt an, wie viele % des Arithmetischen Mittels von der Standardabweichung ausgemacht werden und wird wie folgt berechnet:

$$v = \frac{s}{\overline{x}}$$

v kann < 1 oder > 1 sein, allerdings niemals negativ. Ist v kleiner als 1, so ist die Standardabweichung kleiner als der Mittelwert ist und die Daten relativ eng um den Mittelwert herum verteilt sind. Umso weiter v gegen 0 geht, desto enger liegen die Werte um $\overline{x}$ herum. Gegenteilig sind die Werte weit um $\overline{x}$ herum gestreut, wenn v größer als 1 sein sollte. Der Variationskoeffizient sollte also genutzt werden, um zwei Verteilungen auf ihre Streuung hin zu vergleichen.

Example 2.4.1 (Beispiel zum Variationskoeffizienten)

Gegeben sind die beiden Datensätze $X_1 = \{1, 2, 2, 2, 3, 4\}$ und $X_2 = \{-20, 1, 1, 1, 40, 400\}$. Für diese wird nun jeweils der Variationskoeffizient berechnet:
Datensatz 1:

- $\overline{x} = 2.33$
- $s = 1.03$
- $v = 1.03 \cdot 2.33^{-1} = 0.44$

Datensatz 2:

- $\overline{x} = 70.5$

- $s = 146.26$
- $v = 146.26 \cdot 70.5^{-1} = 2.07$

Hier ist deutlich zu sehen, was das Betrachten der Datensätze schon vermuten lässt: Datensatz 2 ist deutlich weiter gestreut, als Datensatz 1. Die Streuung von Datensatz 1 ist sogar als recht eng zu beurteilen, weil v eher gegen 0, als gegen 1 geht.

Chapter 3

Empirische bivariate Analysen

3.1 Einführung in die Korrelationsanalyse

Die Korrelationsanalyse ist eine zentrale Methode in der empirischen bivariaten Analyse und ermöglicht es, den Zusammenhang zwischen zwei Merkmalen zu untersuchen. Dabei wird zunächst betrachtet, inwieweit die Merkmale voneinander abhängig sind und ob sie in einer Beziehung zueinander stehen. Diese Beziehung kann sowohl linear als auch nicht-linear sein. Ziel der Korrelationsanalyse ist es, den Grad und die Art der Beziehung zwischen den Merkmalen zu bestimmen und zu quantifizieren.

Um die Stärke des Zusammenhangs zwischen zwei Merkmalen zu messen, wird der Korrelationskoeffizient verwendet. Dieser Wert liegt typischerweise im Bereich von -1 bis 1, wobei ein Wert von 0 auf keinen Zusammenhang, negative Werte auf eine negative Korrelation (wenn ein Merkmal steigt, fällt das andere) und positive Werte auf eine positive Korrelation (wenn ein Merkmal steigt, steigt auch das andere) hindeuten.

In der bivariaten Analyse werden außerdem verschiedene grafische Darstellungen verwendet, um den Zusammenhang zwischen den Merkmalen visuell zu erfassen. Hierzu gehören beispielsweise Streudiagramme, die die Wertepaare der Merkmale X und Y in einem kartesischen Koordinatensystem darstellen, oder Kontingenztafeln, die Häufigkeiten von Wertepaaren in einer tabellarischen Form zeigen.

Es ist wichtig zu beachten, dass Korrelation nicht gleich Kausalität bedeutet. Ein signifikanter Zusammenhang zwischen zwei Merkmalen impliziert nicht zwangsläufig, dass das eine Merkmal das andere verursacht. Daher sind bei der Interpretation der Ergebnisse einer Korrelationsanalyse immer auch alternative Erklärungen und mögliche Störvariablen in Betracht zu ziehen.

3.1.1 Kovarianz

Definition 3.1.1: Empirische Kovarianz

Die empirische Kovarianz ist wie folgt definiert:

$$\operatorname{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Eine umgeformte kürzere Fassung:

$$\operatorname{cov}(x, y) = \overline{xy} - \overline{x} \cdot \overline{y}$$

mit

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i y_i)$$

In Worten ausgedrückt: Die Differenz vom Durschschnitt der Produkte und dem Produkt der Durschschnitte

Parallel dazu wird, analog zum vorherigen Kapitel, die Stichprobenkovarianz definiert:

Definition 3.1.2: Stichprobenkovarianz

Die Stichprobenkovarianz s_{xy} ist wie folgt definiert:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Dieser Wert gibt an, inwieweit zwei Merkmale X und Y gemeinsam variieren. Ein positiver Wert zeigt an, dass die Merkmale sich tendentiell proportional verhalten, während ein negativer Wert auf ein antiproportionales Verhalten hindeutet.

Example 3.1.1 (Beispiel zur Stichprobenkovarianz)

Angenommen, wir untersuchen den Zusammenhang zwischen der Anzahl der Stunden, die jemand lernt, und der erreichten Punktzahl bei einer Prüfung. Eine positive Kovarianz würde darauf hindeuten, dass im Allgemeinen mehr Lernstunden zu einer höheren Punktzahl führen, während eine negative Kovarianz bedeuten würde, dass mehr Lernstunden mit einer niedrigeren Punktzahl verbunden sind (was in der Praxis unwahrscheinlich wäre).

3.1.2 Stichprobenkorrelationskoeffizient

Die Stichprobenkovarianz hat zwei Nachteile: Sie ist unbegrenzt und maßstabsabhängig. Der Korrelationskoeffizient wurde entwickelt, um diese Probleme zu beheben. Um die Einheiten von X und Y zu eliminieren, wird die Stichprobenkovarianz durch Größen dividiert, die in Einheiten von X und Y ausgedrückt sind, wie z.B. die Standardabweichungen, die ebenfalls Abweichungen in ihren Berechnungen verwenden.

Definition 3.1.3: Stichprobenkorrelationskoeffizient

Der Stichproben-Korrelationskoeffizient r ist ein Maß für den Grad und die Richtung des linearen Zusammenhangs zwischen zwei Variablen und wie folgt definiert:

$$r = \frac{\mathrm{cov}(x,y)}{d_x \cdot d_y} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2} \cdot \sqrt{\overline{y^2} - \overline{y}^2}} = \frac{s_{xy}}{s_x \cdot s_y}$$

Er variiert zwischen -1 und 1, wobei r = 1 eine perfekte positive Korrelation (alle Punkte liegen auf einer Linie mit positiver Steigung), r = -1 eine perfekte negative Korrelation (alle Punkte liegen auf einer Linie mit negativer Steigung) und r = 0 keinen linearen Zusammenhang zwischen den Variablen bedeutet.

Es ist zu beachten, dass r maßstabsunabhängig ist und dadurch stabil gegenüber Datentransformationen (wie der Währungsumrechnung)ist. Der Stichprobenkorrelationskoeffizient ist ebenfalls sehr empfindlich gegenüber ausreißern.
Für die Werte von $|\,r\,|$ spricht man von folgenden Zusammenhängen:

- 1: vollkommene Korrelation (alle Werte auf einer Geraden)
- 0.66-0.99: sehr starke Korrelation
- 0.36-0.65: starke Korrelation
- 0.16-0.35: mäßige Korrelation
- 0.00-0.15: vernachlässigbare Korrelation

Ist r negativ, so spricht man von negativen Korrelationen, ist r positiv, so spricht man von positiven Korrelationen.

3.2 Einführung in die Regressionsanalyse

Die Regressionsanalyse ist eine statistische Methode, die den Zusammenhang zwischen einer abhängigen Variable und einer oder mehreren unabhängigen Variablen untersucht. Der grundlegende Zweck der Regressionsanalyse besteht darin, Vorhersagen zu treffen und die Beziehung zwischen den Variablen zu modellieren.

Die Regression unterscheidet sich von der Korrelation insofern, als sie nicht nur den Grad und die Richtung des Zusammenhangs zwischen zwei Variablen misst, sondern auch ein mathematisches Modell zur Vorhersage der Werte einer Variablen (die abhängige Variable) auf der Grundlage der Werte einer oder mehrerer anderer Variablen (die unabhängigen Variablen) bietet. Während die Korrelation nur den Grad des linearen Zusammenhangs zwischen zwei Variablen misst, ohne eine Kausalität zu implizieren, ermöglicht die Regression die Untersuchung kausaler Beziehungen.

Beispiele für kausale Zusammenhänge, die durch eine Regressionsanalyse untersucht werden können, sind:

- Die Beziehung zwischen Bildungsstand und Einkommen: Hier könnte der Bildungsstand die unabhängige Variable und das Einkommen die abhängige Variable sein. Die Regressionsanalyse könnte verwendet werden, um zu untersuchen, wie Veränderungen im Bildungsstand das Einkommen beeinflussen.
- Die Wirkung von Werbeausgaben auf die Verkaufszahlen: In diesem Fall könnten die Werbeausgaben die unabhängige Variable und die Verkaufszahlen die abhängige Variable sein. Die Regressionsanalyse könnte genutzt werden, um zu analysieren, wie eine Erhöhung oder Reduzierung der Werbeausgaben die Verkaufszahlen beeinflusst.
- Der Einfluss der Außentemperatur auf den Energieverbrauch eines Gebäudes: Hier könnte die Außentemperatur die unabhängige Variable und der Energieverbrauch die abhängige Variable sein. Eine Regressionsanalyse könnte zeigen, wie Änderungen der Außentemperatur den Energieverbrauch beeinflussen.

3.2.1 Berechnung des Regressionskoeffizienten

Definition 3.2.1: Regressionskoeffizient

Der Regressionskoeffizient, auch bekannt als Beta-Koeffizient (β oder b), ist ein Parameter in der Regressionsgleichung, der die Steigung der Regressionslinie angibt. Er gibt den durchschnittlichen Zusammenhang zwischen der unabhängigen und der abhängigen Variable an, wenn alle anderen Variablen konstant gehalten werden.
Der Regressionskoeffizient ist ein Maß für die Änderung der abhängigen Variable, die durch eine Einheit Änderung der unabhängigen Variable verursacht wird.

$$b = \frac{\mathrm{cov(x, y)}}{d_x^2}$$

Diese Formel sagt aus, dass die Änderung von y in Bezug auf x proportional zur Kovarianz von x und y ist, skaliert durch die Varianz von x.

Definition 3.2.2: Bestimmtheitsmaß eines Regressionsmodells

Das Bestimmtheitsmaß (auch als R-Quadrat oder R^2 bekannt) ist ein statistisches Maß in der Regressionsanalyse, das die Güte der Anpassung eines Regressionsmodells an die tatsächlichen Daten angibt. Es misst den Anteil der Varianz in der abhängigen Variable, der durch das Regressionsmodell erklärt wird.

$$r^2 = \left(\frac{\text{cov}(x,y)}{s_x \cdot s_y}\right)^2 = \frac{\text{cov}(x,y)^2}{s_x^2 \cdot s_y^2}$$

Hier ist cov(x, y) die Kovarianz zwischen x und y, s_x ist die Standardabweichung von x und s_y ist die Standardabweichung von y.
Dieses Maß kann zwischen 0 und 1 liegen. Ein Wert von 0 bedeutet, dass das Modell überhaupt keine Varianz erklärt, während ein Wert von 1 bedeutet, dass das Modell die gesamte Varianz erklärt.
Ein höheres R^2 zeigt also eine bessere Anpassung des Modells an die Daten. Es ist jedoch wichtig zu beachten, dass ein hohes R^2 nicht unbedingt bedeutet, dass das Modell korrekt oder nützlich ist, da es auch bei nicht sinnvollen oder überangepassten Modellen hoch sein kann.

3.2.2 Rangkorrelationskoeffizient nach Spearman

Der Rangkorrelationskoeffizient nach Spearman, oft als Spearman's Rho bezeichnet, ist ein nicht-parametrisches Maß der statistischen Abhängigkeit zwischen den Rängen zweier Variablen. Es misst die Stärke und Richtung der Assoziation zwischen zwei ranggeordneten Variablen und ist somit besonders nützlich für die Analyse ordinaler Daten.
In der Praxis ersetzt Spearman's Rho die Daten im gewöhnlichen Korrelationskoeffizienten durch ihre Ränge. Dies bedeutet, dass anstatt die tatsächlichen Werte der Variablen zu vergleichen, ihre Positionen oder Ränge innerhalb der Datensätze verglichen werden. Diese Methode ist besonders robust gegenüber Ausreißern, da diese in der Rangreihenfolge weniger Gewicht erhalten.
Die Berechnung von Spearman's Rho ähnelt der Berechnung des Pearson-Korrelationskoeffizienten, wobei jedoch die Ränge der Beobachtungen anstelle der tatsächlichen Werte verwendet werden. Wichtig ist jedoch, dass weiterhin die Rangpärchen gegenübersgestellt werden, wie vorher die Wertpärchen kombiniert wurde.

Die Hauptmerkmale von Spearman's Rho sind:

- Er kann Werte zwischen -1 und 1 annehmen, wobei -1 eine perfekte negative Korrelation, 0 keine Korrelation und 1 eine perfekte positive Korrelation anzeigt.
- Er ist ein nicht-parametrisches Maß und setzt daher keine bestimmte Verteilungsform voraus.
- Er ist robust gegenüber Ausreißern.
- Er kann zur Analyse ordinaler Daten verwendet werden.
- Er misst monotone Zusammenhänge, also ob die Beziehung zwischen den Variablen stetig ist, unabhängig davon, wie schnell oder langsam sich die Veränderung vollzieht.

3.2.3 Kontingenz für nominale Merkmale

Das bekannte Konzept von Vierfeldertafeln für binäre Merkmale erweitern wir nun um Merkmale mit mehr als zwei Ausprägungen. Dazu verwenden wir Kreuztabellen, die auch als Kontingenztabellen bekannt sind.
Kreuztabellen sind eine Erweiterung der Vierfeldertafeln und ermöglichen es uns, die Beziehungen zwischen zwei oder mehr Merkmalen zu analysieren, die jeweils mehrere Ausprägungen haben können. Sie bieten eine übersichtliche Darstellung der Häufigkeiten der verschiedenen Kombinationen von Merkmalsausprägungen.
Als Beispiel betrachten wir die Merkmale *Geschlecht* (mit den Ausprägungen "männlich" und "weiblich") und *Berufsposition* (mit den Ausprägungen "leitender Angestellter", "Tarifangestellter" und "Arbeiter").

Eine entsprechende Kreuztabelle könnte so aussehen:

	Leitender Angestellter	Tarifangestellter	Arbeiter	Summe
Männlich	$n_{11} = 20$	$n_{12} = 30$	$n_{13} = 50$	$n_{1+} = 100$
Weiblich	$n_{21} = 15$	$n_{22} = 35$	$n_{23} = 45$	$n_{2+} = 95$
Summe	$n_{+1} = 35$	$n_{+2} = 65$	$n_{+3} = 95$	$n_{++} = 195$

Definition 3.2.3: Kontingenzkoeffizient C

Um den Zusammenhang zwischen den Merkmalen zu messen, führen wir den Kontingenzkoeffizienten C ein. Er ist ein Maß für die Stärke des Zusammenhangs zwischen den beiden Merkmalen und kann Werte zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang) annehmen.

$$C = \sqrt{\frac{\sum_{i,j}(n_{ij} - \frac{n_{i+} \cdot n_{+j}}{n_{++}})^2}{(n_{++}^2 - \sum_i n_{i+}^2) \cdot \sum_j n_{+j}^2}}$$

In dieser Formel wird über alle Zellen i, j der Tabelle summiert. Der Ausdruck in der Klammer im Zähler ist die Differenz zwischen der beobachteten Häufigkeit n_{ij} und der erwarteten Häufigkeit unter der Annahme der Unabhängigkeit der Merkmale (gegeben durch das Produkt der Summenzeile n_{i+} und der Summenspalte n_{+j} geteilt durch die Gesamtzahl der Beobachtungen n_{++}). Im Nenner wird die quadratische Summe der Summenzeilen und Summenspalten vom Quadrat der Gesamtzahl der Beobachtungen subtrahiert. Der gesamte Ausdruck wird quadriert, um negative Werte zu vermeiden. Die Wurzel sorgt dafür, dass der Wertebereich des Kontingenzkoeffizienten wieder zwischen 0 und 1 liegt.

Durch die Verwendung von Kreuztabellen und dem Kontingenzkoeffizienten können wir also den Zusammenhang zwischen nominalen Merkmalen quantitativ erfassen und interpretieren.
Die Kontingenz ist allgemeiner als die Korrelation. Man spricht hier meist von stochastischer (in diesem Fall linearer) Abhängigkeit. Wichtig zu beachten ist hierbei, dass der Zusammenhang zwischen den Merkmalen immer theoretisch begründbar sein muss. Andernfalls wäre es Nachweisbar, dass es einen Zusammenhang zwischen beispielsweise der Anzahl an Krebs-Neuerkrankungen in Deutschland und den Wahlergebnissen in den USA gibt. Grundsätzlich muss man immer in Betracht ziehen, dass in der Realität meist mehr als 2 Merkmale gegenseitigen Einfluss ausüben und eine Analyse von nur zwei Merkmalen somit nur begrenzte Aussagekraft hat. In der Regel ist man bereits froh, wenn man für $|r|$ einen Wert von ungefähr 0.7 oder sogar höher findet.