MASTERARBEIT | MASTER'S THESIS

Titel | Title

Ausgerechnet Fußball!
Stochastische Modellierung von Spielausgängen

verfasst von | submitted by

Fabian Strobl BEd

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Education (MEd)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 199 500 520 02
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Lehramt Sek (AB) Unterrichtsfach Bewegung und Sport Unterrichtsfach Mathematik
Betreut von | Supervisor:	ao. Univ.-Prof. Mag. Dr. Peter Raith

Zusammenfassung

Die vorliegende Masterarbeit untersucht den Ausgang von Fußballspielen aus einer wahrscheinlichkeitstheoretischen Perspektive am Beispiel der deutschen Bundesliga. Ausgangspunkt ist die Beobachtung, dass Fußball trotz vergleichsweise geringer Toranzahl und stark zufallsbeeinflusster Spielsituationen statistische Regelmäßigkeiten aufweist, die eine mathematische Modellierung ermöglichen. Ziel der Arbeit ist es, geeignete stochastische Modelle zur Beschreibung von Torverteilungen und Spielausgängen theoretisch zu fundieren, empirisch zu überprüfen und hinsichtlich ihrer Prognosefähigkeit zu bewerten.

Hierzu wird eine literaturgestützte Aufarbeitung zentraler Konzepte der Wahrscheinlichkeitstheorie mit einer datenbasierten Analyse realer Spieldaten kombiniert. Auf theoretischer Ebene werden grundlegende Begriffe wie Wahrscheinlichkeitsräume, die Kolmogorov-Axiome, Zufallsvariablen und deren Verteilungen sowie die stochastische Unabhängigkeit behandelt. Im Fokus stehen dabei insbesondere die Binomial- und die Poisson-Verteilung, die als Modelle für die Torentstehung herangezogen werden.

Zur empirischen Überprüfung der Modellannahmen wird der Chi-Quadrat-Anpassungstest eingesetzt, um die Übereinstimmung zwischen beobachteten und modellierten Torhäufigkeiten zu beurteilen. Die Analyse von Bundesligadaten zeigt, dass die Poisson-Verteilung die charakteristische Form der Torverteilung insgesamt gut approximiert, wenngleich vereinzelt Abweichungen auftreten. Gleichzeitig wird auf Basis langfristig aggregierter Daten deutlich, dass auch statistisch signifikante Abweichungen zwischen empirischen und theoretischen Verteilungen auftreten können.

Aufbauend darauf wird ein probabilistisches Prognosemodell entwickelt, das die Torentstehung zunächst vereinfacht als Bernoulli-Prozess modelliert und anschließend in ein Poisson-Modell überführt. Zur differenzierteren Abbildung der Teamstärken werden neben klassischen Kennzahlen insbesondere Expected-Goals-basierte Größen sowie der Heimvorteil berücksichtigt. Das resultierende Modell ermöglicht die Berechnung von Wahrscheinlichkeiten für konkrete Spielausgänge und liefert realistische Verteilungen möglicher Ergebnisse.

Insgesamt zeigt die Arbeit, dass sich Fußball trotz seiner inhärenten Zufälligkeit erfolgreich mit stochastischen Modellen beschreiben lässt. Auch wenn einzelne Spielergebnisse nicht deterministisch vorhersagbar sind, erlaubt der probabilistische Ansatz eine strukturierte und fundierte Bewertung von Spielausgängen.

Abstract

This master's thesis examines the outcomes of football matches from a probabilistic perspective, using the German Bundesliga as a case study. The starting point is the observation that, despite a relatively low number of goals and game situations heavily influenced by chance, football exhibits statistical regularities that allow for mathematical modeling. The aim of this thesis is to theoretically ground, empirically test, and evaluate suitable stochastic models for describing goal distributions and match outcomes in terms of their predictive power.

To this end, a literature-based review of central concepts in probability theory is combined with a data-driven analysis of real match data. At the theoretical level, fundamental concepts such as probability spaces, the Kolmogorov axioms, random variables and their distributions, as well as stochastic independence are addressed. The focus is particularly on the binomial and Poisson distributions, which are used as models for goal generation.

To empirically test the model assumptions, the chi-square goodness-of-fit test is used to assess the agreement between observed and modeled goal frequencies. Analysis of Bundesliga data shows that the Poisson distribution generally approximates the characteristic shape of the goal distribution well, although isolated deviations occur. At the same time, based on long-term aggregated data, it becomes clear that statistically significant deviations between empirical and theoretical distributions can also occur.

Building on this, a probabilistic forecasting model is developed that initially models goal generation in a simplified manner as a Bernoulli process and subsequently transforms it into a Poisson model. To provide a more nuanced representation of team strengths, expected-goals-based metrics and home-field advantage are taken into account in addition to traditional metrics. The resulting model enables the calculation of probabilities for specific match outcomes and provides realistic distributions of possible results.

Overall, the study demonstrates that football, despite its inherent randomness, can be successfully described using stochastic models. Even though individual match results cannot be predicted deterministically, the probabilistic approach allows for a structured and well-founded assessment of match outcomes.

Inhaltsverzeichnis

1 Einleitung

Fußball ist nicht nur die weltweit beliebteste Sportart, sondern auch ein Phänomen, das Menschen über Generationen und Kulturen hinweg verbindet. Auch für mich persönlich spielt dieser Sport seit meiner Kindheit eine zentrale Rolle: sei es als aktiver Spieler, als Zuschauer im Stadion oder vor dem Fernseher – und zunehmend auch als analytisch Denkender. Parallel zu meiner Begeisterung für den Fußball entwickelte sich nämlich eine besondere Affinität für Zahlen und mathematische Strukturen. In dieser Arbeit vereine ich zwei auf den ersten Blick gegensätzliche Interessenbereiche, die sich bei genauerem Hinsehen jedoch auf faszinierende Art und Weise ergänzen.

Trotz klarer Regeln und professioneller Strukturen ist Fußball ein Spiel voller Spannung und Emotionen, das in seinem Ausgang oft unvorhersehbar bleibt. Ein Ausdruck dessen sind die immer wieder auftretenden überraschenden Spielverläufe, die allen Erwartungen und Prognosen widersprechen. So etwa im Jahr 2016, als der zuvor kaum beachtete Leicester City Football Club direkt nach dem Aufstieg völlig unerwartet die englische Premier League gewann. Ein Titelgewinn, dem die Buchmacher anfänglich lediglich eine Wahrscheinlichkeit von gerade mal 0,02 % bzw. eine Quote von 5 000 : 1 zugeschrieben hatten [1, 6]. Als weiteres Paradebeispiel lässt sich das legendäre Champions-League-Rückspiel zwischen dem FC Barcelona und Paris Saint-Germain am 8. März 2017 nennen, in dem Barcelona nach einem 0:4 im Hinspiel mit einem 6:1-Heimsieg noch die nächste Runde erreichte – ein aus statistischer Sicht nahezu unmögliches Szenario [19, 51].

Abgesehen davon zeigen statistische Analysen, dass in einem Bundesligaspiel durchschnittlich etwa drei Tore erzielt werden, während die mittlere Anzahl an Abschlussversuchen pro Team bei zehn bis zwanzig liegt [18]. Diese Diskrepanz zwischen der Häufigkeit von Torversuchen und der tatsächlich erzielten Trefferzahl erklärt die Rolle von Zufallseinflüssen sowie von Glück und Pech im Spielverlauf. Im Gegensatz dazu fallen in Sportarten wie Handball etwa 50 Tore pro Spiel, wodurch die Wahrscheinlichkeit, dass ein nominell unterlegenes Team gewinnt, deutlich geringer ist. Dementsprechend setzen sich in Spielen mit hoher Trefferzahl aus statistischer Sicht häufiger die stärkeren Mannschaften durch. Die vergleichsweise geringe Torfrequenz im Fußball hingegen begünstigt unvorhersehbare Spielverläufe – ein zentrales Merkmal, das wesentlich zur Faszination und Attraktivität dieser Sportart beiträgt. In Verbindung mit der hohen Komplexität des Spiels erschwert dies die Berechnung von Wahrscheinlichkeiten und macht verlässliche Vorhersagen über Spielausgänge zu einer anspruchsvollen Aufgabe.

Gleichzeitig wächst das Interesse an quantitativer Spielanalyse, sodass sich der Fußball zunehmend zu einem datengetriebenen Sport entwickelt. Selbst vormals einfache Platt-

formen, auf denen früher lediglich Endstände abrufbar waren, stellen heute umfangreiche statistische Informationen zu Torchancen, Passmustern oder Positionsdaten bereit und schaffen damit die Grundlage für mathematisch fundierte Analysen. An dieser Stelle setzt die vorliegende Arbeit an, indem sie zwei scheinbar gegensätzliche Disziplinen miteinander verbindet: die emotionale Unvorhersehbarkeit des Spiels und die rationale Struktur der Mathematik. Ziel ist es, den Spielausgang im Fußball mithilfe stochastischer Modelle zu beschreiben, insbesondere unter Anwendung der Poisson-Verteilung. In diesem Zusammenhang werden nicht nur historische Tordaten berücksichtigt, sondern auch moderne sportanalytische Kennzahlen wie der sogenannte Expected-Goals-Wert (xG) als Maß für die Spielstärke in die Modellierung einbezogen.

2 Fußball als Ausgangspunkt mathematischer Modellierung

Heutzutage steht der Fußball mit einer Fangemeinschaft von schätzungsweise 3,5 Milliarden Menschen an der Spitze der beliebtesten Sportarten weltweit [45]. Laut dem „Big Count“-Projekt der FIFA gab es im Jahr 2006 bereits über 265 Millionen aktive Fußballspieler:innen, die wiederum dafür verantwortlich sind, Millionen und Abermillionen Zuschauer:innen regelmäßig vor den Fernseher oder in die Stadien zu locken [52].

Die Zahlen der österreichischen Fußballgemeinschaft sind zwar überschaubarer, aber dennoch beeindruckend. Ende 2024 waren rund 270 500 organisierte Fußballspieler:innen in 1 985 registrierten Fußballvereinen zu verzeichnen [66, 77]. Damit rangiert der Österreichische Fußball-Bund (ÖFB) als größter Einzelsportfachverband Österreichs. Angesichts der großen Vereinsstruktur stellt Fußball die beliebteste Vereinssportart der Österreicher:innen dar [77].

Die Faszination des Fußballs beruht nicht zuletzt auf seiner Einfachheit: Es bedarf nur wenig Ausrüstung, wodurch das Spiel weltweit – auch in wirtschaftlich schwächeren Regionen – großen Anklang fand und sich verbreitete. Seine leicht verständlichen Regeln erleichtern den Zugang für Spieler:innen wie Zuschauer:innen gleichermaßen [94]. Diese breite Anziehungskraft ist jedoch keineswegs ein Phänomen der Neuzeit – schon seit Jahrtausenden übten fußballähnliche Aktivitäten in verschiedenen Kulturen eine besondere Begeisterung auf Menschen aus und legten damit den Grundstein für die spätere Entwicklung des modernen Fußballs.

2.1 Geschichte

Nach Kovar und Zart [49] lassen sich die ersten belegten Ursprünge bis ins alte China zurückverfolgen. Bereits 3000 v. Chr. wurde dort ein Spiel namens „Ts'uh-küh“ (Ts'uh: mit dem Fuß stoßen; küh: Ball) praktiziert, das charakteristische Parallelen zum Fußball aufwies und zu militärischen Trainingszwecken diente [9, 82].

Obwohl sich weitere fußballähnliche Spiele über die Zeit auf der ganzen Welt verbreiteten, bildete Großbritannien die Keimzelle des modernen Fußballs, so wie er heute bekannt ist. Was zuvor ein wildes, unstrukturiertes und von Gewalt geprägtes Spiel war, entwickelte sich im 19. Jahrhundert zu einer gesitteten Variante, die sich durch entscheidende Regeländerungen allmählich vom Rugby abgrenzte [49]. Im Jahr 1848 wurden an der

Universität Cambridge die gleichnamigen „Cambridge Rules“ formuliert, die das Spiel mit dem Fuß bevorzugten [83]. Dieses einheitliche Regelwerk wurde in den darauffolgenden Jahren immer differenzierter, bis im Jahr 1863 schließlich die Geburtsstunde des modernen Fußballs in London schlug. Dort wurde nämlich nicht nur der erste nationale Fußballverband der Welt, die „Football Association“ (FA), gegründet, sondern mit ihm auch 14 allgemeingültige Regeln festgelegt. Durch das Verbot, den Ball zu halten oder zu tragen, wurde der Fußball reformiert und emanzipierte sich endgültig von seiner einst verwandten Sportart, dem Rugby [68, 83].

Von England aus verbreitete sich Fußball rasch in andere Länder, zunächst über Handels- und Bildungsnetzwerke, später auch durch gezielte Vereinsgründungen. In Kontinentaleuropa war insbesondere die Schweiz ein frühes Zentrum der Entwicklung, während sich der Sport in Deutschland und Österreich allmählich gegen etablierte Turntraditionen behaupten musste [83, 97].

Einen weiteren Meilenstein markierte die Gründung des Fußball-Weltverbandes, der „Fédération Internationale de Football Association“ (FIFA), im Jahr 1904 in Paris [25]. Die FIFA trieb die internationale Standardisierung des Regelwerks voran, organisierte ab 1930 Weltmeisterschaften und etablierte sich in der Folge als zentraler Akteur der weltweiten Fußballkoordination und -entwicklung. Damit wurde der Grundstein für die spätere Globalisierung des Sports gelegt, die sich heute in der Zusammenarbeit der FIFA mit ihren 211 Mitgliedsverbänden auf der ganzen Welt widerspiegelt [25, 49].

2.2 Spielregeln

Bis heute legt die FIFA in Zusammenarbeit mit dem „International Football Association Board“ (IFAB) das offizielle Regelwerk fest. Dieses setzt sich aus insgesamt 17 konstitutiven Spielregeln zusammen, die 1938 vom IFAB neu verfasst und seitdem nur geringfügig modifiziert wurden. In jährlich stattfindenden Konferenzen werden die Regeln hinsichtlich ihrer Aktualität und Spielangemessenheit überprüft und bei Bedarf adaptiert, wie etwa durch den Einsatz von zusätzlichen Video-Schiedsrichterassistent:innen (VAR) im Jahr 2018 [26, 49, 96].

Grundsätzlich ist Fußball eine Ballsportart, bei der zwei Mannschaften mit jeweils elf Spieler:innen gegeneinander antreten, um mehr Tore als der Gegner zu erzielen und damit das Spiel zu gewinnen. Mit Ausnahme von sogenannten K.-o.-Runden, in denen zwingend ein Sieger ermittelt werden muss, endet ein Spiel unentschieden, wenn beide Teams am Ende der regulären Spielzeit die gleiche Anzahl an Toren erzielt haben. Der Ball wird zwar vorwiegend mit dem Fuß getreten, darf aber abgesehen von den Armen und Händen mit dem ganzen Körper gespielt werden. Lediglich der Torwart darf den Ball innerhalb des eigenen Strafraums zusätzlich mit den Händen berühren. Auf einem rechteckigen freien Feld, das durch seine Seiten- und Torlinien begrenzt ist, werden zwei Halbzeiten von jeweils 45 Minuten gespielt, unterbrochen von einer 15-minütigen Pause. Der genaue Aufbau sowie etwaige Abmessungen des Spielfelds sind in der folgenden

Abbildung 2.1 dargestellt. Der Größenrahmen des abgebildeten Spielfelds bezieht sich jedoch auf nationale Spiele, bei denen vergleichsweise große Toleranzbereiche gelten. Für internationale Begegnungen hingegen sind die zulässigen Abmessungen etwas enger gefasst, sodass die Länge zwischen 100 und 110 Metern und die Breite zwischen 64 und 75 Metern liegen muss. Die Spielleitung obliegt einer Schiedsrichterin bzw. einem Schiedsrichter, die bzw. der gemeinsam mit den Assistent:innen für die Einhaltung der Fußballregeln sorgt. Etwaige Verstöße gegen das Regelwerk werden mit persönlichen Strafen sanktioniert. Während weniger gravierende Regelverstöße mit einer formellen Verwarnung in Form einer „Gelben Karte“ geahndet werden, führen schwerwiegende Vergehen zu einem Platzverweis, der durch eine „Rote Karte“ signalisiert wird. Erhält ein:e Spieler:in zweimal im selben Spiel die gelbe Karte, zieht dies eine rote Karte nach sich, sodass diese:r durch die sogenannte „Gelb-Rote Karte“ ebenfalls des Platzes verwiesen wird [84, 96].

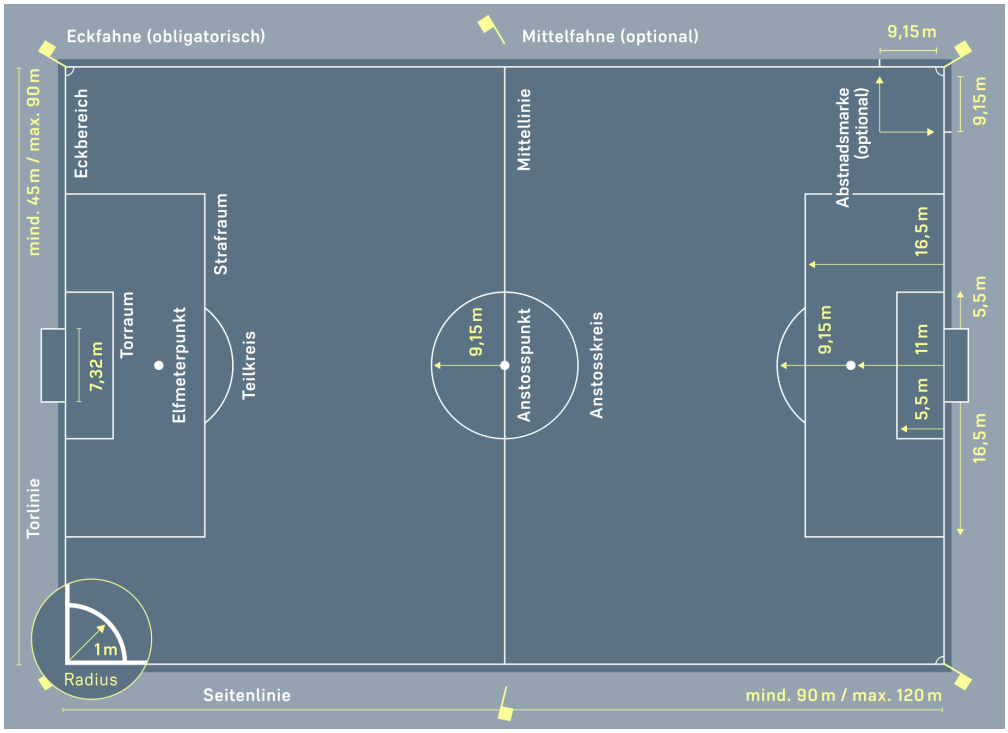

Abbildung 2.1: Grundriss eines Fußballfeldes nach dem IFAB [85]

Da eine detaillierte Darstellung des gesamten Regelwerks den Rahmen dieser Arbeit überschreiten würde, sei an dieser Stelle auf die offizielle Fassung des IFAB verwiesen, die unter folgender Internetadresse abrufbar ist: www.theifab.com.

2.3 Deutsche Fußball-Bundesliga

Im Rahmen dieser Arbeit dient die deutsche Fußball-Bundesliga als Grundlage für mathematische Analysen. Im Gegensatz etwa zur österreichischen Bundesliga, bei der nach 22 Spieltagen die Punkte halbiert und die Liga in zwei Gruppen (Meister- und Qualifikationsgruppe) mit jeweils sechs Vereinen aufgeteilt wird, zeichnet sich das deutsche Pendant durch eine konstante Ligastruktur und stabilere Wettbewerbsbedingungen aus [65]. Der lineare Saisonverlauf sowie die daraus resultierenden konsistenten statistischen Daten unterliegen keiner strukturellen Verzerrung, wodurch sie eine geeignete Grundlage für mathematische Modellierungen bilden [20]. Darüber hinaus weist die deutsche Bundesliga eine höhere internationale Relevanz auf, was sich unter anderem in ihrer Platzierung auf Rang vier der UEFA-Fünfjahreswertung widerspiegelt und sie zu einem interessanten Untersuchungsobjekt macht [92].

Die heutige Bundesliga ist das Resultat eines tiefgreifenden Reformprozesses, der bereits lange vor ihrer Gründung einsetzte. In den 1920er-Jahren existierten im Deutschen Reich über 70 sogenannte „Erste Ligen“ [3]. Die Ermittlung des deutschen Meisters erfolgte damals über ein K.-o.-System mit einem Endspiel, das vielfach kritisiert wurde. Insbesondere unterlegene Teams fühlten sich benachteiligt und forderten aus sportlichen Erwägungen eine einheitliche Liga auf nationaler Ebene [3, 95]. Trotz mehrfacher Reformvorschläge wurde erst im Jahr 1962 ein entscheidender Schritt zur Professionalisierung des deutschen Vereinsfußballs unternommen. Unter dem Eindruck des enttäuschenden Ausscheidens der deutschen Nationalmannschaft bei der Weltmeisterschaft 1962 beschloss der Deutsche Fußball-Bund (DFB) am 28. Juli 1962 in Dortmund die Gründung einer bundesweiten Fußball-Profiliga. Ziel war es, sowohl die strukturellen Schwächen des bisherigen Systems zu beheben als auch die internationale Wettbewerbsfähigkeit zu stärken. Mit Beginn der Saison 1963/64 nahm die Bundesliga ihren Spielbetrieb auf und löste den bis dahin bestehenden Pokalmodus ab [3, 5, 98]. Seither bildet die Bundesliga die höchste Spielklasse im deutschen Fußball. Der grundlegende Austragungsmodus blieb dabei über die Jahrzehnte hinweg weitgehend konstant; lediglich die Anzahl der teilnehmenden Teams (16, 18 oder 20) und der Absteiger (zwischen zwei und vier) wurden im Laufe der Zeit mehrfach angepasst [95, 98]. Seit ihrer Gründung hat sich die deutsche Bundesliga – neben der englischen Premier League, der italienischen Serie A und der spanischen La Liga – zu einer der konkurrenzfähigsten und leistungsstärksten Ligen Europas entwickelt [92]. Heute umfasst sie insgesamt 18 Vereine, die im Rahmen eines Meisterschaftsjahres jeweils zweimal gegeneinander antreten – einmal im eigenen Stadion und einmal auswärts. Dadurch ergeben sich insgesamt 34 Spieltage pro Team [20]. Die Saison läuft typischerweise von August bis Mai, wobei sie in Jahren mit internationalen Turnieren (Europa- oder Weltmeisterschaft) bereits im April enden kann [95].

Die Gesamttabelle basiert primär auf Punkten. Wie im Fußball üblich, beschert ein Sieg drei Punkte, während ein Unentschieden jeweils einen Punkt für beide Kontrahenten und eine Niederlage keine Zähler bringt. Auf dieser Grundlage werden der deutsche Fußballmeister, die Absteiger sowie die Teilnehmer an den Europapokalwettbewerben ausge-

spielt. Jenes Team, das nach dem letzten Spieltag auf Tabellenplatz eins steht, darf sich bis zum Abschluss der darauffolgenden Saison „Deutscher Fußballmeister“ nennen. Mit bislang 34 Meistertiteln ist der FC Bayern München nicht nur amtierender Titelträger der Saison 2024/25, sondern auch Rekordmeister der deutschen Bundesligageschichte. Die beiden Letztplatzierten müssen in die 2. Bundesliga absteigen und werden im Gegenzug von deren zwei erstplatzierten Teams ersetzt. Zudem bestreitet der Drittletzte der Bundesliga (Platz 16) zwei Relegationsspiele gegen den Tabellendritten der 2. Bundesliga, um sich einen Startplatz in der kommenden Bundesligasaison zu sichern [20, 95]. Im Hinblick auf die Startplätze für internationale Vereinswettbewerbe ergibt sich folgendes Bild: Die vier bestplatzierten Teams, einschließlich des Meisters, erhalten jeweils ein Ticket für die Ligaphase der UEFA Champions League. Der Meisterschaftsfünfte sowie der Sieger des DFB-Pokals dürfen sich auf die Ligaphase der UEFA Europa League freuen, während das sechstplatzierte Team an den Play-offs der UEFA Conference League teilnimmt. Sollte der DFB-Pokalsieger gleichzeitig einen der ersten sechs Tabellenplätze belegen, rückt der Siebtplatzierte in das deutsche Starterfeld nach, sodass auf jeden Fall sieben deutsche Vereine in europäischen Pokalwettbewerben vertreten sind [20]. Diese vergleichsweise hohe Teilnehmerquote ist nicht zuletzt dem stabilen vierten Platz der UEFA-Fünfjahreswertung zu verdanken, der das Leistungsniveau und die internationale Wettbewerbsfähigkeit der Liga widerspiegelt [92].

Bei einer quantitativen Betrachtung der deutschen Bundesliga zeigt sich, dass Anfang August 2025 der gesamte Marktwert aller Spieler bei rund 4,48 Milliarden Euro lag. Mit einem individuellen Marktwert von 140 Millionen Euro gilt der deutsche Nationalspieler Jamal Musiala als derzeit wertvollster Akteur der Liga. Der bis dato kostspieligste Bundesliga-Neuzugang ist der englische Stürmer Harry Kane, der im Sommer 2023 für 95 Millionen Euro vom FC Bayern verpflichtet wurde. Demgegenüber steht mit dem Franzosen Ousmane Dembélé der bislang teuerste Abgang der Bundesligageschichte, der im Jahr 2017 für 148 Millionen Euro zum spanischen Topclub FC Barcelona wechselte. Im Laufe der Bundesligageschichte avancierte der deutsche Stürmer Gerd Müller mit 365 Treffern in 427 Spielen zum erfolgreichsten Torschützen aller Zeiten. Aktuell umfasst die Liga 542 Spieler, von denen 301 aus dem Ausland stammen (ca. 56 %). In der vergangenen Saison 2024/25 verfolgten insgesamt 11,8 Millionen Zuschauer:innen die 306 Bundesliga-Partien live in den Stadien, was einem durchschnittlichen Zuschaueraufkommen von etwa 39 000 Personen pro Spiel entspricht. Vor diesem Hintergrund wird deutlich, welchen hohen gesellschaftlichen und wirtschaftlichen Stellenwert der Fußball nicht nur in Deutschland, sondern weltweit einnimmt [34, 90].

2.4 Statistische Analyse der Toranzahlen im Fußball

Ein bekanntes Zitat der deutschen Trainerlegende Sepp Herberger – „Das Runde muss in das Eckige“ – bringt auf einfache und prägnante Weise auf den Punkt, worum es im Fußball im Kern geht [20]. Was schlicht klingt, ist in der Realität jedoch das Ergebnis zahlreicher, vielschichtig miteinander verflochtener Faktoren, die darüber entscheiden, ob

ein Spielzug mit einem Torerfolg endet oder nicht. Selbst kleinste Einflüsse können eine aussichtslose Spielsituation in eine scheinbar sichere Torchance verwandeln – und vice versa. Unumstritten spielen dabei Zufallsphänomene eine wesentliche Rolle im Verlauf eines Spiels. Wenige Millimeter oder ein leichter Windstoß können darüber entscheiden, ob ein Ball vom Pfosten hinter die Linie springt oder nicht. Ebenso kann ein Schuss unglücklich abgefälscht werden, sodass der Torwart keine Abwehrchance hat, oder der Schiedsrichter eine strittige Szene falsch einschätzt und einen Elfmeter beziehungsweise eine Rote Karte verhängt. All dies sind vertraute Szenarien im Fußball, denen die Akteure auf dem Feld wahllos ausgesetzt sind und die den Ausgang eines Spiels entscheidend prägen können.

Um den Beitrag solcher Faktoren quantifizierbar zu machen, untersuchte der Sportwissenschaftler Martin Lames in der Saison 2011/12 insgesamt 1 931 Treffer in der deutschen Bundesliga und der englischen Premier League. Zu diesem Zweck entwickelte er ein Beobachtungssystem, das den Zufallseinfluss bei Torerfolgen operationalisiert [54]. Dabei definierte er sechs Zufallskriterien, die ein nicht kontrollierbares oder nicht geplantes Zustandekommen eines Tores beschreiben:

- Abgefälschter Schuss
- Abpraller vor dem Schuss
- Ballkontakt mit Pfosten oder Latte
- Starker Ballkontakt durch den Torwart
- Tor aus sehr großer Distanz
- Fehler des Gegners [54, 73]

Die Auswertung ergab, dass 47 % der untersuchten Treffer von mindestens einem dieser Zufallsfaktoren begleitet waren. Auf Basis dieser Erkenntnisse betont Lames [54], dass der Erfolg im Fußball nicht ausschließlich auf schematische Spielzüge und taktische Raffinesse zurückgeführt werden sollte, sondern der Faktor Zufall ebenfalls als reguläre Einflussgröße zu berücksichtigen ist. Paradoxerweise ist es gerade diese Zufallskomponente, die eine mathematische Modellierung des Spiels besonders lohnend macht und zugleich wertvolle Anknüpfungspunkte für die Wahrscheinlichkeitstheorie bietet.

Im Folgenden wird die deutsche Bundesliga als Untersuchungsobjekt herangezogen, wobei zunächst die Torhäufigkeiten analysiert werden. Schließlich gelten das Erzielen und Verhindern von Toren als oberste Prämisse im Fußball. Das Ergebnis eines Spiels und damit die Anzahl der vergebenen Punkte wird allein durch die Torerfolge definiert: Wer mehr Treffer erzielt als der Gegner, gewinnt und erhält drei Punkte. Tore sind fundamentale Elemente des Fußballspiels und auf mehreren Ebenen von wesentlicher Bedeutung. Beispielsweise können sie psychologische Auswirkungen auf die Spieler haben und deren Motivation steigern oder senken, taktische Anpassungen hervorrufen, Spielstrategien beeinflussen oder einfach den Unterhaltungswert für Zuschauer:innen und Medien prägen.

Zudem bergen sie aus analytischer Perspektive Hinweise auf die Offensiv- oder Defensivstärke eines Teams und dienen als Basisgröße für zahlreiche statistische Kennzahlen, wie etwa die Torquote, die Tordifferenz oder den Expected-Goals-Wert (xG).

Darauf aufbauend werden im nächsten Schritt zentrale Begriffe der deskriptiven Statistik am Beispiel der Bundesliga eingeführt. Grundlage bildet eine Datenerhebung mit dem Umfang $n = 306$, bei der alle Spiele der Saison 2024/25 im Hinblick auf das Merkmal „X = Torsumme pro Spiel“, also die Gesamtanzahl der in einem Spiel erzielten Tore beider Teams, untersucht wurden. Die beobachteten Merkmalswerte (Realisationen) $x_1, x_2, \ldots, x_n$ werden auch als *Urliste* oder *Rohdaten* bezeichnet [24]. Mit zunehmender Anzahl n der Untersuchungseinheiten – hier der Spiele – verliert eine bloße Auflistung dieser Daten jedoch schnell an Übersichtlichkeit, sodass alternative Darstellungsformen zweckmäßig sind. Um die erhobenen Werte strukturiert zusammenzufassen, werden sie nach den verschiedenen auftretenden *Merkmalsausprägungen (Modalitäten)*

$$a_1, a_2, \ldots, a_k \quad \text{mit } k \leq n$$

geordnet [24]. Eine Merkmalsausprägung ist dabei ein möglicher Wert, den das Merkmal X annehmen kann [53]. Für die Torsumme pro Spiel kommen in Abhängigkeit von der torreichsten Begegnung der Saison die natürlichen Zahlen in Betracht. Da in der Spielzeit 2024/25 höchstens neun Treffer in einem Spiel erzielt wurden, ergibt sich die Menge der beobachteten Merkmalsausprägungen zu

$$a_1 = 0,\ a_2 = 1,\ \ldots,\ a_{10} = 9.$$

Anschließend wird die Anzahl der Merkmalswerte der Urliste ermittelt, die mit einer Ausprägung a_j für $j \in \{1, \ldots, 10\}$ übereinstimmen. Allgemein wird jeder Untersuchungseinheit i (mit $i = 1, \ldots, n$) die entsprechende Ausprägung a_j als Merkmalswert x_i zugeordnet. Bei einem nicht allzu großen Umfang n eignet sich zur Darstellung der Häufigkeiten eine Strichliste, bei der für jede Untersuchungseinheit ein Strich bei der beobachteten Merkmalsausprägung notiert wird. Jeder fünfte Strich wird quer durch die vorangegangenen vier gezogen, sodass durch die Bündelung eine Struktur entsteht, die das Ablesen erleichtert. Dabei vermittelt sie zugleich einen anschaulichen optischen Eindruck der Häufigkeitsverteilung [24, 32, 53]. Wendet man dieses Verfahren auf alle Spielbegegnungen an, ergibt sich die in Tabelle 2.1 (S. 10) dargestellte Übersicht.

Da die Lesbarkeit der Strichliste bei steigendem n schnell abnimmt, werden die eingetragenen Anzahlen üblicherweise in komprimierter Form zusammengefasst. Hierzu werden die *absoluten* und *relativen Häufigkeiten* der jeweiligen Merkmalsausprägungen gebildet und in einer Häufigkeitstabelle anschaulich dargestellt. Während die absolute Häufigkeit durch Abzählen ermittelt werden kann, lässt sich die relative Häufigkeit als Verhältnis der absoluten Häufigkeit zur Gesamtzahl der Untersuchungseinheiten n berechnen.

Definition 2.4.1 (**Absolute und relative Häufigkeit**)**.** Es sei n die Anzahl der statistischen Einheiten, und es seien a_j mit $j = 1, 2, \ldots, k$ und $k \leq n$ mögliche Merkmalsausprägungen. Dann heißt die Anzahl der statistischen Einheiten mit der Merkmalsausprägung a_j die *absolute Häufigkeit* $h(a_j) = h_j$ der Merkmalsausprägung a_j. Es

entsteht eine *absolute Häufigkeitsverteilung* $h_1, \ldots, h_k$.
Weiters gilt: $\sum_{j=1}^{k} h_j = n$.

Der Anteil der statistischen Einheiten mit der Merkmalsausprägung a_j an der Gesamtanzahl n der statistischen Einheiten

$$f_j := \frac{h_j}{n}$$

heißt *relative Häufigkeit* (syn.: Anteil) $f(a_j) = f_j$ der Merkmalsausprägung a_j. Es entsteht eine *relative Häufigkeitsverteilung* $f_1, \ldots, f_k$.
Weiters gilt: $\sum_{j=1}^{k} f_j = 1$ (vgl. [24, S. 32–33], [47, S. 39–40], [53, S. 12]).

Bemerkung 2.4.2. Gegebenenfalls kann die Summe der relativen Häufigkeiten aufgrund von Rundungsfehlern geringfügig von 1 abweichen [53].

Tabelle 2.1: Strichliste der Torsumme pro Spiel (Saison 2024/25)

Torsumme pro Spiel a_j	**Strichliste** Anzahl der Spiele mit a_j Toren
0	𝍸 𝍸 𝍸 𝍸 ||
1	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||
2	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||||
3	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||
4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |
5	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||
6	𝍸 𝍸 𝍸 𝍸 |
7	𝍸 ||||
8	||
9	|

In Tabelle 2.2 (S. 11) sind die absoluten und relativen Häufigkeitsverteilungen des Merkmals „X = Torsumme pro Spiel“ dargestellt. Die relativen Häufigkeiten f_j können aufgrund ihrer Normierung als prozentuale Anteile interpretiert werden, indem sie mit dem Faktor 100 multipliziert werden [47]. Aus der Tabelle lassen sich zudem zentrale Eigenschaften der Verteilungen ablesen: Die Summe der absoluten Häufigkeiten $h_1 + h_2 + \ldots + h_{10} = 306$ entspricht der Gesamtzahl der statistischen Einheiten n, also der Anzahl der Spiele der Saison 2024/25. Gleichzeitig ergibt sich für die relativen Häufigkeiten $f_1 + f_2 + \ldots + f_{10} = 1$, was der Gesamtheit aller Anteile von 100 % entspricht. Da die relativen Häufigkeiten auf vier Nachkommastellen gerundet wurden, können Rundungsfehler auftreten, die die exakte Normierung beeinflussen. Im vorliegenden Fall summieren sich die Werte jedoch trotz möglicher Rundungsabweichungen exakt zu eins. Sollte dies nicht der Fall sein, wird gemäß Kosfeld et al. [47] dennoch der Wert eins angesetzt.

Tabelle 2.2: Häufigkeitsverteilung der Torsumme pro Spiel (Saison 2024/25)

Torsumme pro Spiel a_j	**abs. Häufigkeit** h_j	**rel. Häufigkeit** f_j
0	22	0,0719
1	37	0,1209
2	64	0,2092
3	52	0,1699
4	66	0,2157
5	32	0,1046
6	21	0,0686
7	9	0,0294
8	2	0,0065
9	1	0,0033
Summe	**306**	**1**

Während der Spielzeit 2024/25 der deutschen Bundesliga wurden in 306 Spielen insgesamt 959 Tore erzielt. Ganze 66-mal fielen in einer Begegnung genau vier Tore (Endstände 4:0, 3:1, 2:2, 1:3 oder 0:4), dicht gefolgt von 64 Partien mit insgesamt zwei Treffern. Man kann daher festhalten, dass der *Modus* (Modalwert) bei vier liegt, da diese Merkmalsausprägung am häufigsten auftritt [47]. Obwohl diese Information in der Tabelle nicht ausgewiesen ist, sei angemerkt, dass das häufigste konkrete Ergebnis – unabhängig von der Torsumme – ein 1:1 war, das insgesamt 26-mal vorkam.

Weist ein Merkmal besonders viele unterschiedliche Ausprägungen auf, kann die zugehörige Häufigkeitsverteilung schnell unübersichtlich werden. Vor allem bei metrischen, (quasi-)stetigen Merkmalen ist es häufig nicht möglich, die Urliste auf eine überschaubare Menge einzelner Modalitäten zu reduzieren. Um eine Balance zwischen Anschaulichkeit und Informationsgehalt zu wahren, bietet es sich an, mehrere Merkmalsausprägungen in geeigneten Klassen zusammenzufassen. Auf diese Weise entsteht eine gruppierte Häufigkeitsverteilung, die eine kompaktere und anschaulichere Darstellung der Daten ermöglicht. Im Kontext der Gesamttreffer pro Spiel wäre dieses Vorgehen auch bei Ausreißern nach oben sinnvoll. Begegnungen mit außergewöhnlich hohen Ergebnissen mögen zwar Anlass zu unterhaltsamen Diskussionen unter Stammtisch-Expert:innen geben, tragen jedoch nur wenig zur Beschreibung der grundlegenden statistischen Strukturen der Torhäufigkeiten bei [24].

Abgesehen davon lassen sich die ermittelten Häufigkeiten zur besseren Veranschaulichung nicht nur tabellarisch, sondern auch grafisch aufbereiten. Für metrisch skalierte Merkmale mit einer begrenzten Anzahl an unterschiedlichen Ausprägungen, wie die Torsumme pro Spiel, stehen verschiedene Möglichkeiten offen, etwa Stab-, Säulen-, Kreis-, Punkt- oder Liniendiagramme [53]. Mit Ausnahme des Kreisdiagramms werden diese Diagramm-

typen in der Regel in ein kartesisches Koordinatensystem eingebettet. Auf der Abszisse (x-Achse) sind dabei die Merkmalsausprägungen a_j verzeichnet, während die Ordinate (y-Achse) die Häufigkeiten h_j beziehungsweise f_j für $j = 1, \ldots, k$ angibt. Im vorliegenden Zusammenhang wird ein Säulendiagramm verwendet. Es besteht aus Rechtecken mit konstanter horizontaler Breite, deren vertikale Länge proportional zur absoluten oder relativen Häufigkeit der jeweiligen Merkmalsausprägung ist. Gemäß Definition 2.4.1 ergibt die Summe aller Rechteckslängen die Gesamtzahl der statistischen Einheiten n im Fall absoluter Häufigkeiten beziehungsweise 1 bei relativen Häufigkeiten [47, 53]. Das in Abbildung 2.2 dargestellte Säulendiagramm zeigt die relative Häufigkeitsverteilung der Gesamttreffer pro Spiel in der Saison 2024/25 der deutschen Bundesliga.

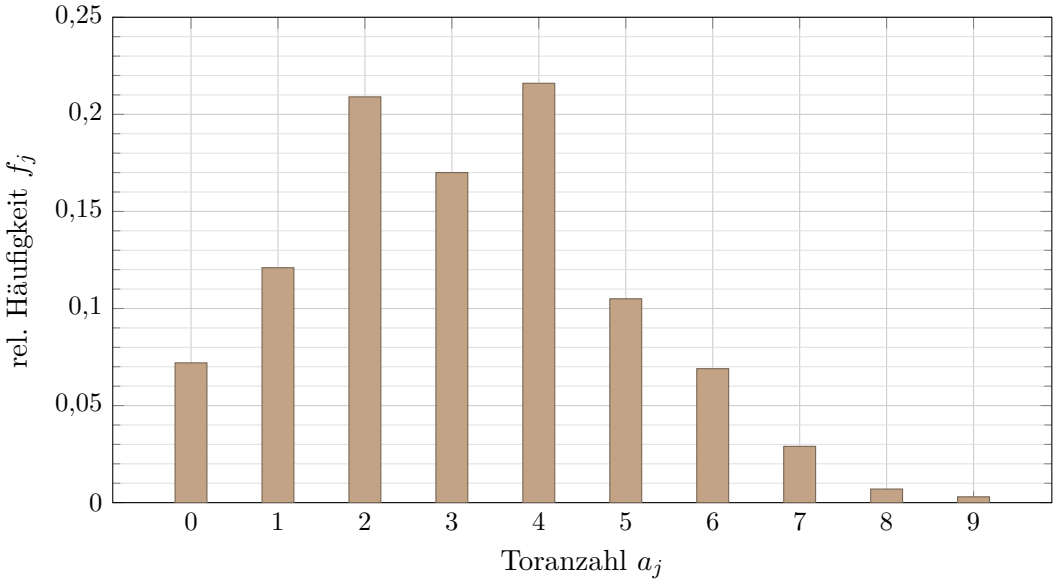

Abbildung 2.2: Relative Häufigkeitsverteilung der Torsumme pro Spiel (Saison 2024/25)

Neben dem übersichtlichen Eindruck der relativen Häufigkeiten wird bei genauerem Hinsehen deutlich, dass in etwa 60 % aller Spiele insgesamt zwei, drei oder vier Tore erzielt wurden. Die größten Anteile liegen demnach leicht links vom Zentrum der Merkmalsverteilung. Daraus lässt sich die Tendenz ableiten, dass im Fußball eher wenige Tore fallen. Diese Annahme wird durch die Beobachtung gestützt, dass in 57 % der Spiele höchstens drei Treffer erzielt wurden.

Auf Basis der erhobenen Daten der deutschen Bundesliga können darüber hinaus weitere Kennzahlen berechnet werden, die eine kompakte Beschreibung der Verteilung ermöglichen. Bei der Analyse mehrerer Datensätze ist es häufig unübersichtlich, direkt mit umfangreichen Listen zu arbeiten. Zwar kann es zweckmäßig sein, die Daten vollständig aufzulisten, doch erweist es sich oft als praktisch, ein geeignetes Lagemaß heranzuziehen, das die Verteilung der Daten möglichst repräsentativ zusammenfasst. Neben dem bereits erwähnten Modus eignet sich für die Toranzahlen insbesondere das *arithmeti-*

sche Mittel, das in zahlreichen Alltagssituationen am gebräuchlichsten ist. Da es für metrische Merkmale sinnvoll definiert ist, können die Einzelwerte oder Merkmalsausprägungen direkt in die Berechnung einbezogen werden. Das arithmetische Mittel gibt einen Durchschnittswert an, der durch die Summe aller Realisationen geteilt durch die Anzahl der betrachteten Merkmalsträger bestimmt wird und somit eine Orientierung über die typische Ausprägung innerhalb der Datenmenge bietet [24, 47].

Definition 2.4.3 (Arithmetisches Mittel). Seien $x_1, x_2, \ldots, x_n$ Daten eines quantitativen Merkmals. Dann heißt

$$\bar{x} := \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

arithmetisches Mittel dieser Daten [53, S. 30].

Da in der Saison 2024/25 für $n = 306$ Beobachtungseinheiten die Summe aller Merkmalswerte x_i mit $i \in \{1, \ldots, n\}$ insgesamt 959 Treffer beträgt, ergibt sich das arithmetische Mittel wie folgt:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{306} \cdot 959 \approx 3{,}13.$$

Demnach fielen in der deutschen Bundesliga durchschnittlich etwa 3,13 Tore pro Spiel.

Werden identische Merkmalswerte mehrfach beobachtet, ist es zur Vereinfachung der Berechnung nicht erforderlich, jeden Wert einzeln zu summieren. Stattdessen können gleiche Realisationen mit der Anzahl ihres Auftretens in der Urliste – vereinfacht ausgedrückt die Merkmalsausprägungen a_j mit ihren absoluten Häufigkeiten h_j für $j = 1, \ldots, k$ – multipliziert werden, und die resultierenden Produkte werden anschließend summiert [47, 53].

Satz 2.4.4. *Seien $x_1, x_2, \ldots, x_n$ Daten eines quantitativen Merkmals, die in den Ausprägungen $a_1, a_2, \ldots, a_k$ mit den absoluten Häufigkeiten $h_1, h_2, \ldots, h_k$ für $k \leq n$ vorkommen. Dann gilt für das arithmetische Mittel $\bar{x}$ [50]:*

$$(1) \quad \bar{x} = \frac{1}{n} \sum_{j=1}^{k} a_j \cdot h_j, \qquad (2) \quad \bar{x} = \sum_{j=1}^{k} a_j \cdot f_j.$$

Beweis. O. B. d. A. sei die Datenliste $x_1, x_2, \ldots, x_n$ geordnet. Dann gilt nach Kronfellner et al. [50]

$$\begin{aligned} \bar{x} &= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (a_1 + \ldots + a_1 + a_2 + \ldots + a_2 + a_k + \ldots + a_k) \\ &= \frac{1}{n} \cdot (h_1 \cdot a_1 + h_2 \cdot a_2 + \ldots + h_k \cdot a_k) = \frac{1}{n} \cdot \sum_{j=1}^{k} a_j \cdot h_j \end{aligned}$$

und (1) ist gezeigt. Wegen $\frac{h_j}{n} = f_j$ folgt dann für (2):

$$\begin{aligned}\bar{x} &= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{n} \cdot (n \cdot f_1 \cdot a_1 + n \cdot f_2 \cdot a_2 + \ldots + n \cdot f_k \cdot a_k) \\ &= f_1 \cdot a_1 + f_2 \cdot a_2 + \ldots + f_k \cdot a_k = \sum_{j=1}^{k} a_j \cdot f_j.\end{aligned}$$

□

Welche Formel für das arithmetische Mittel heranzuziehen ist, hängt von der Form der vorliegenden Daten ab. Liegen Einzelwerte vor, lässt sich das arithmetische Mittel direkt aus diesen bestimmen. Bei unklassierten Häufigkeiten erfolgt die Berechnung hingegen über die Kombination der Merkmalsausprägungen mit ihren zugehörigen Häufigkeiten [47].

Auf der Grundlage der Daten von Tabelle 2.2 (S. 11) und mithilfe der Aussagen aus Satz 2.4.4 kann das arithmetische Mittel der Torsumme pro Spiel nun auch auf folgende Weise bestimmt werden:

$$\begin{aligned}(1)\quad \bar{x} &= \frac{1}{n} \cdot \sum_{j=1}^{k} a_j \cdot h_j = \frac{1}{306} \cdot \sum_{j=1}^{10} a_j \cdot h_j \\ &= \frac{1}{306} \cdot (0 \cdot 22 + 1 \cdot 37 + 2 \cdot 64 + 3 \cdot 52 + 4 \cdot 66 + 5 \cdot 32 + 6 \cdot 21 + 7 \cdot 9 \\ &\qquad + 8 \cdot 2 + 9 \cdot 1) \\ &= \frac{1}{306} \cdot (37 + 128 + 156 + 264 + 160 + 126 + 63 + 16 + 9) \\ &= \frac{1}{306} \cdot 959 \approx 3{,}1340\end{aligned}$$

$$\begin{aligned}(2)\quad \bar{x} &= \sum_{j=1}^{k} a_j \cdot f_j = \sum_{j=1}^{10} a_j \cdot f_j \\ &\approx (0 \cdot 0{,}0719 + 1 \cdot 0{,}1209 + 2 \cdot 0{,}2092 + 3 \cdot 0{,}1699 + 4 \cdot 0{,}2157 + 5 \cdot 0{,}1046 \\ &\qquad + 6 \cdot 0{,}0686 + 7 \cdot 0{,}0294 + 8 \cdot 0{,}0065 + 9 \cdot 0{,}0033) \\ &= (0{,}1209 + 0{,}4184 + 0{,}5097 + 0{,}8628 + 0{,}5230 + 0{,}4116 + 0{,}2058 \\ &\qquad + 0{,}0520 + 0{,}0297) \\ &= 3{,}1339\end{aligned}$$

Wie in der zweiten Berechnung ersichtlich, ergibt sich aufgrund der bereits gerundeten relativen Häufigkeiten f_j aus Tabelle 2.2 eine geringfügige Abweichung im Resultat.

Da im Rechenvorgang des arithmetischen Mittels alle Einzelwerte $x_1, x_2, \ldots, x_n$ gleich gewichtet werden, lässt sich eine zentrale Eigenschaft direkt aus der Definition ableiten.

Es zeigt sich nämlich, dass die Summe der Abweichungen aller Merkmalswerte x_i mit $i = 1, \ldots, n$ vom arithmetischen Mittel $\bar{x}$ den Wert Null annimmt [24, 53, 56].

Satz 2.4.5 (**Schwerpunkteigenschaft des arithmetischen Mittels**). *Seien $x_1, x_2, \ldots, x_n \in \mathbb{R}$ die Daten eines quantitativen Merkmals und $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ihr arithmetisches Mittel. Dann gilt nach Leiner [56]:*

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Beweis. Aus der Definition des arithmetischen Mittels folgt unmittelbar

$$n \cdot \bar{x} = \sum_{i=1}^{n} x_i.$$

Daher gilt nach Leiner [56]

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n \cdot \bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0.$$

□

Geometrisch betrachtet entspricht dies der Situation, dass alle Werte $x_1, x_2, \ldots, x_n$ als Massenpunkte mit gleicher Masse auf der Zahlengeraden liegen und ihr gemeinsamer Schwerpunkt genau im arithmetischen Mittel liegt. Das arithmetische Mittel ist somit der Punkt, an dem die Abweichungen nach links und rechts im Gleichgewicht sind, weshalb man auch von der Schwerpunkteigenschaft spricht. Diese Eigenschaft könnte zudem als Ausgangspunkt für die Definition des arithmetischen Mittels dienen; ein Nachteil bestünde jedoch darin, dass daraus keine unmittelbare Berechnungsvorschrift ableitbar ist [53].

Aus der Schwerpunkteigenschaft folgt, dass sich für einen beliebigen Punkt $M \in \mathbb{R}$ die Abweichungen von x_i zu M genau dann im besten Sinne aufheben, wenn M mit dem Schwerpunkt $\bar{x}$ übereinstimmt. Formal lässt sich diese Ausgleichseigenschaft durch die Minimierung der quadrierten Abweichungen charakterisieren [56].

Satz 2.4.6 (**Minimaleigenschaft des arithmetischen Mittels**). *Seien $x_1, \ldots, x_n \in \mathbb{R}$ die Daten eines quantitativen Merkmals und $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ihr arithmetisches Mittel. Dann gilt für jeden Wert $M \in \mathbb{R}$ mit $M \neq \bar{x}$:*

$$\sum_{i=1}^{n}(x_i - M)^2 \; > \; \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

wobei Gleichheit genau dann eintritt, wenn $M = \bar{x}$ [56].

Beweis. Schreibe für $M \in \mathbb{R}$ mit $M \neq \bar{x}$ die Abweichung als

$$x_i - M \;=\; (x_i - \bar{x}) + (\bar{x} - M).$$

Quadrieren und summieren liefert

$$\begin{aligned}\sum_{i=1}^{n}(x_i - M)^2 &= \sum_{i=1}^{n}\big[(x_i - \bar{x}) + (\bar{x} - M)\big]^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - M)\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(\bar{x} - M)^2.\end{aligned}$$

Da per Schwerpunkteigenschaft des arithmetischen Mittels

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0,$$

verschwindet der Kreuzterm. Außerdem ist $(\bar{x} - M)^2$ unabhängig von i und kann aus der Summe gezogen werden:

$$\sum_{i=1}^{n}(x_i - M)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n\,(\bar{x} - M)^2.$$

Der zweite Summand ist für $M \neq \bar{x}$ stets positiv und wird genau dann null, wenn $M = \bar{x}$. Damit ist die Behauptung gezeigt [56]. □

Da für eine belastbare Datengrundlage der Stichprobenumfang n möglichst groß sein sollte, wird nun unter Verwendung der zuvor erläuterten statistischen Kennzahlen eine umfassende Analyse von 1 530 Spielen aus fünf aufeinanderfolgenden Spielzeiten (2020/21 bis 2024/25) vorgenommen. Die Ergebnisse sind in Tabelle 2.3 (S. 17) übersichtlich zusammengefasst. Dort werden die absoluten Häufigkeiten $h_j^{(s)}$ mit $s = 1, \ldots, 5$ für die jeweiligen Spielzeiten angegeben, wobei sich diese auf Spiele mit einer Torsumme von a_j Treffern ($j = 1, \ldots, 10$) beziehen.

Die dargestellten Torhäufigkeiten weisen über die letzten Jahre eine weitgehend konsistente Verteilung auf. Betrachtet man alle Spielzeiten gemeinsam, lässt sich ein erkennbares Muster feststellen: Mit Ausnahme der Saison 2024/25 liegt der Modus – also die am häufigsten auftretende Torsumme – stets bei zwei Treffern pro Spiel. An zweithäufigster Stelle stehen ebenfalls in diesen Spielzeiten drei Tore pro Partie. Die durchschnittliche Toranzahl variiert dabei nur geringfügig zwischen 3,03 und 3,22 Treffern pro Spiel. Das arithmetische Mittel über alle fünf Spielzeiten beträgt 3,14 Tore pro Spiel. Zum Vergleich: Die bislang torreichste Saison der deutschen Bundesligageschichte wurde in den Jahren 1983/84 verzeichnet, mit insgesamt 1 097 Treffern und einem Durchschnitt von etwa 3,59 Toren pro Spiel [34]. Die hier präsentierten Daten dienen unter anderem als Grundlage für die wahrscheinlichkeitstheoretischen Modelle im Hauptteil dieser Arbeit.

Im folgenden Abschnitt wird die Entwicklung der durchschnittlichen Toranzahl im Fußball näher untersucht, wobei auch mögliche Ursachen für die beobachteten Veränderungen diskutiert werden.

Tabelle 2.3: Häufigkeitsverteilungen $h_j^{(s)}$ der Toranzahlen a_j über fünf Spielzeiten

Torsumme pro Spiel a_j	**20/21** $h_j^{(1)}$	**21/22** $h_j^{(2)}$	**22/23** $h_j^{(3)}$	**23/24** $h_j^{(4)}$	**24/25** $h_j^{(5)}$	**Summe** $\sum_{s=1}^{5} h_j^{(s)}$
0	18	16	15	13	22	84
1	37	35	32	32	37	173
2	71	76	74	70	64	355
3	68	59	69	65	52	313
4	54	52	51	60	66	283
5	32	42	30	32	32	168
6	15	13	18	24	21	91
7	9	11	14	7	9	50
8	2	1	3	2	2	10
9	0	1	0	1	1	3
Summe $\sum_{j=1}^{10} h_j^{(s)}$	306	306	306	306	306	1 530
durchschn. Torsumme $\bar{x}^{(s)}$	3,03	3,12	3,17	3,22	3,13	3,14

2.5 Fußball(-tore) im Wandel der Zeit

Seit dem Start der deutschen Bundesliga im Jahr 1963 sind bis 2025 insgesamt 62 Spielzeiten ohne Unterbrechung verstrichen. Obwohl in diesem Zeitraum die Organisationsstruktur und der Austragungsmodus weitgehend konstant blieben, zeigen sich Veränderungen in den durchschnittlichen Torzahlen pro Spiel. Betrachtet man den Zeitverlauf, liegt die Torquote bis 1987 auf einem vergleichsweise hohen Niveau. In den folgenden Jahrzehnten sinkt sie deutlich und bleibt bis 2017 meist unter der Drei-Tore-Marke. Seit 2018 ist hingegen ein erneuter Anstieg zu verzeichnen, sodass die „magische“ Drei-Tore-Grenze durchgängig überschritten wird. In der Saison 2024/25 wurde mit 504 Treffern sogar die torreichste Hinrunde seit 33 Jahren registriert. Auch im Vergleich mit den europäischen Top-5-Ligen setzt sich die Bundesliga an die Spitze [90]. In fünf Spielzeiten von 2020/21 bis 2024/25 wurde die Drei-Tore-Marke außerhalb Deutschlands lediglich zweimal überschritten: in der italienischen Serie A 2020/21 mit durchschnittlich 3,06 Toren pro Spiel und in der englischen Premier League 2023/24 mit 3,28 Toren pro Spiel. Im Mittel derselben fünf Spielzeiten liegt die Premier League bei 2,91 Treffern pro Partie, gefolgt von der französischen Ligue 1 (2,81). Dahinter rangieren die Serie A (2,73) und die spanische La Liga mit lediglich 2,56 Toren pro Spiel. Im selben Zeitraum verzeichnet die Bundesliga im Durchschnitt 3,14 Treffer pro Partie [29].

2.5.1 Lineare Regression

Zur übersichtlichen Darstellung der 62 Bundesliga-Saisons werden die Datenpunkte zunächst in ein Koordinatensystem eingetragen: Auf der y-Achse ist die durchschnittliche Toranzahl pro Spiel abgetragen, auf der x-Achse die Spielzeiten, beginnend mit 1963/64 und jeweils durch das Startjahr gekennzeichnet. Um die Beziehung zwischen den beiden Variablen „X = Saisonjahr“ und „Y = Torquote“ zu erfassen, wird eine Ausgleichsgerade durch die Punktwolke gelegt, die möglichst nahe an den tatsächlichen Beobachtungswerten verläuft. Dieses Vorgehen basiert auf dem *Regressionsmodell*, mit dessen Hilfe langfristige Entwicklungen einer Variablen in Abhängigkeit von einer anderen quantifiziert werden können. Auf Grundlage der erhobenen Daten wird somit eine Gerade bestimmt, die den Trend der Torquoten im Zeitverlauf bestmöglich beschreibt. Diese Annäherung in Form einer linearen Funktion zweier Variablen wird als *lineare Regressionsgerade* bezeichnet [24, 47].

Die Berechnung dieser Regressionsgeraden erfolgt nach dem Prinzip der *Methode der kleinsten Quadrate*, die Ende des 18. Jahrhunderts von Carl Friedrich Gauss entwickelt und 1809 in seiner Abhandlung zur Bahnbestimmung von Himmelskörpern erstmals publiziert wurde. Ziel dieser Methode ist es, jene Gerade zu bestimmen, bei der die Summe der quadrierten Abweichungen zwischen den beobachteten Werten und den durch die Gerade prognostizierten Werten minimiert wird [53].

Hervorzuheben ist, dass die Herleitung der Regressionsgeraden eng an die zuvor dargestellten Eigenschaften des arithmetischen Mittels anknüpft. So garantiert die Schwerpunkteigenschaft, dass die durch die Methode der kleinsten Quadrate bestimmte Gerade stets den Punkt $(\bar{x}, \bar{y})$ durchläuft. In diesem Sinne bildet das Paar der arithmetischen Mittelwerte den „Schwerpunkt“ der Daten, durch den die bestangepasste Gerade verläuft [53]. Darüber hinaus stellt die Minimaleigenschaft des Mittels eine direkte Vorstufe des Regressionsansatzes dar: Während in einer Dimension das arithmetische Mittel derjenige Wert ist, der die Summe der quadrierten Abweichungen minimiert,

$$\bar{y} = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n} (y_i - c)^2,$$

verallgemeinert die lineare Regression dieses Prinzip auf zwei Dimensionen. Hier werden die Parameter α und β der Geraden

$$y = \alpha + \beta x$$

so gewählt, dass die Gesamtabweichung in Form der quadrierten vertikalen Abstände zu den Beobachtungspunkten minimiert wird:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha,\beta) \in \mathbb{R}^2} \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2.$$

Die Regressionsgerade kann somit als eine natürliche Erweiterung der Minimaleigenschaft des arithmetischen Mittels verstanden werden, wobei die Schwerpunkteigenschaft die geometrische Verankerung dieses Zusammenhangs liefert [47, 53].

Definition 2.5.1 (**Lineares Regressionsmodell**)**.** Seien $x_1, \ldots, x_n \in \mathbb{R}$ die Ausprägungen einer unabhängigen Variablen und $y_1, \ldots, y_n \in \mathbb{R}$ die zugehörigen Ausprägungen einer abhängigen Variablen. Das *lineare Regressionsmodell* ist gegeben durch

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{für } i = 1, \ldots, n,$$

wobei $\alpha \in \mathbb{R}$ das Absolutglied (Ordinatenabschnitt) und $\beta \in \mathbb{R}$ den Steigungsparameter (Regressionskoeffizienten) der linearen Beziehung darstellen und ε_i zufällige Fehler sind [32, S. 574].

Für die Begründung der Eindeutigkeit und Globalität der Lösung der Methode der kleinsten Quadrate werden im Folgenden einige grundlegende Begriffe und Resultate aus der Analysis und linearen Algebra zusammengefasst.

Definition 2.5.2 (**r-malige stetige Differenzierbarkeit**)**.** Eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ heißt *r-mal stetig differenzierbar*, wenn alle partiellen Ableitungen bis zur Ordnung r existieren und stetig sind. In diesem Fall schreibt man $f \in C^r(\mathbb{R}^n)$ [2, S. 315].

Bemerkung 2.5.3. Sei $f \in C^1(\mathbb{R}^n)$. Die Ableitung von f an der Stelle $x \in \mathbb{R}^n$ ist gegeben durch

$$\mathrm{d}f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right),$$

die nach Arens et al. [2, S. 878] auch als *Jacobi-Matrix* bezeichnet wird. Besitzt f an der Stelle $x \in \mathbb{R}^n$ ein lokales Extremum, so gilt notwendig

$$\mathrm{d}f(x) = 0.$$

Ein Punkt $x \in \mathbb{R}^n$ mit $\mathrm{d}f(x) = 0$ wird nach Arens et al. [2, S. 341] als kritischer Punkt von f bezeichnet.

Definition 2.5.4 (**Konvexe Funktion**)**.** Nach Arens et al. [2, S. 335] heißt eine Funktion $f : \mathbb{R}^n \to \mathbb{R}$ *konvex*, wenn für alle $x, y \in \mathbb{R}^n$ und alle $t \in [0, 1]$ gilt

$$f\big(x + t(y - x)\big) = f\big((1 - t)x + ty\big) \leq (1 - t)f(x) + tf(y).$$

Bemerkung 2.5.5. Gilt die Konvexitätsungleichung für alle $t \in (0, 1)$ sogar in strenger Form, so spricht man von einer *strikt* bzw. *streng konvexen* Funktion. Analog heißt eine Funktion auf einem Intervall *(streng) konkav*, wenn die zugehörige Funktion $-f$ *(streng) konvex* ist [2].

Definition 2.5.6 (**Positive Definitheit**)**.** Nach Arens et al. [2, S. 735] heißt eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ *positiv definit*, wenn

$$x^\top A x > 0 \quad \text{für alle } x \in \mathbb{R}^n \setminus \{0\}.$$

Bemerkung 2.5.7 (**Kriterien für positive Definitheit**)**.** Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Dann sind die folgenden Aussagen nach Arens et al. [2] äquivalent:

1. A ist positiv definit, d. h. $x^\top Ax > 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$.
2. Alle Eigenwerte von A sind positiv.
3. Alle führenden Hauptminoren $\det(A_k)$, $k = 1, \ldots, n$, sind positiv.
4. Beim strikten Gauß-Verfahren (ohne Zeilenvertauschungen) sind alle auftretenden Pivotelemente, d. h. die Diagonalelemente der resultierenden Dreiecksmatrix, positiv.

Definition 2.5.8 (Hesse-Matrix). Sei $f \in C^2(\mathbb{R}^n)$. Die *Hesse-Matrix* von f an der Stelle $x \in \mathbb{R}^n$ ist die Matrix der zweiten partiellen Ableitungen und nach Arens et al. [2, S. 892] gegeben durch

$$H_f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

Satz 2.5.9. *Sei $f \in C^2(\mathbb{R}^n)$ und die Hesse-Matrix $H_f(x)$ für alle $x \in \mathbb{R}^n$ positiv definit. Dann ist f streng konvex auf $\mathbb{R}^n$ [2, S. 892].*

Beweis. Seien $x, y \in \mathbb{R}^n$ mit $x \neq y$, und sei $t \in (0, 1)$. Betrachte die Funktion

$$g(t) := f((1-t)x + ty) - ((1-t)f(x) + tf(y)).$$

Da $f \in C^2(\mathbb{R}^n)$ gilt, ist auch $g \in C^2([0, 1])$. Weiters gelten nach der mehrdimensionalen Kettenregel:

$$\begin{aligned} g'(t) &= \mathrm{d}f((1-t)x + ty)(y - x) + f(x) - f(y), \\ g''(t) &= (y - x)^\top H_f((1-t)x + ty)(y - x). \end{aligned}$$

Aus $x \neq y$ folgt $y - x \neq 0$. Wegen der positiven Definitheit der Hesse-Matrix für alle $x \in \mathbb{R}^n$ gilt daher

$$g''(t) > 0 \quad \text{für alle } t \in (0, 1).$$

Seien nun $0 < t < s < 1$. Nach dem Mittelwertsatz existiert ein $\xi \in (t, s)$ mit

$$g'(s) - g'(t) = g''(\xi)(s - t).$$

Da $g''(\xi) > 0$ folgt $g'(s) > g'(t)$, also ist g' streng monoton steigend.

Da $g(0) = g(1) = 0$, existiert nach dem Mittelwertsatz ein $t_0 \in (0, 1)$ mit

$$g'(t_0) = g'(t_0)(1 - 0) = g(1) - g(0) = 0.$$

Für $t \in (0, t_0]$ existiert nach dem Mittelwertsatz ein $\xi \in (0, t)$ mit

$$g(t) = g(t) - g(0) = g'(\xi)(t - 0).$$

Wegen $\xi < t_0$ gilt $g'(\xi) < g'(t_0) = 0$, also $g(t) < 0$.

Für $t \in [t_0, 1)$ gibt es nach dem Mittelwertsatz ein $\xi \in (t, 1)$ mit

$$-g(t) = g(1) - g(t) = g'(\xi)(1 - t).$$

Hier ist $t_0 < \xi$ und deshalb $0 = g'(t_0) < g'(\xi)$, woraus sich $-g(t) > 0$, also $g(t) < 0$, ergibt.

Also gilt $g(t) < 0$ für alle $t \in (0, 1)$, und nach der Definition von g erhält man

$$f((1 - t)x + ty) < (1 - t)f(x) + tf(y).$$

Folglich ist f streng konvex. □

Für den folgenden Satz würde es genügen vorauszusetzen, dass $f \in C^1(\mathbb{R}^n)$ und streng konvex ist. Im vorliegenden Zusammenhang ist es jedoch zweckmäßig, ihn unter den Voraussetzungen von Satz 2.5.9 zu formulieren und zu beweisen.

Satz 2.5.10. *Sei $f \in C^2(\mathbb{R}^n)$ und die Hesse-Matrix $H_f(x)$ für alle $x \in \mathbb{R}^n$ positiv definit. Besitzt f einen kritischen Punkt x_0, so ist dieser eindeutig bestimmt und stellt das globale Minimum von f dar.*

Beweis. Es sei x_0 ein kritischer Punkt von f, also $\mathrm{d}f(x_0) = 0$. Sei nun $x \in \mathbb{R}^n$ mit $x \neq x_0$. Definiere $g : [0, 1] \to \mathbb{R}$ durch

$$g(t) := f((1 - t)x_0 + tx).$$

Dann ist $g \in C^2([0, 1])$ und nach der mehrdimensionalen Kettenregel gelten:

$$\begin{aligned} g'(t) &= \mathrm{d}f((1 - t)x_0 + tx)(x - x_0), \\ g''(t) &= (x - x_0)^\top H_f((1 - t)x_0 + tx)(x - x_0). \end{aligned}$$

Da $H_f((1 - t)x_0 + tx)$ positiv definit ist, folgt

$$g''(t) > 0 \quad \text{für alle } t \in [0, 1].$$

Seien jetzt $0 \leq t < s \leq 1$. Nach dem Mittelwertsatz existiert ein $\xi \in (t, s)$ mit

$$g'(s) - g'(t) = g''(\xi)(s - t).$$

Wegen $g''(\xi) > 0$ ergibt sich daraus $g'(s) > g'(t)$, also ist g' streng monoton steigend.

Da $g'(0) = \mathrm{d}f(x_0)(x - x_0) = 0$ und g' streng monoton steigend ist, gilt

$$g'(t) > 0 \quad \text{für alle } t > 0.$$

Insbesondere folgt $g'(1) = \mathrm{d}f(x)(x - x_0) > 0$, und damit $\mathrm{d}f(x) \neq 0$. Somit existiert kein weiterer kritischer Punkt.

Nach dem Mittelwertsatz existiert ein $\xi \in (0, 1)$ mit

$$g(1) - g(0) = g'(\xi)(1 - 0) = g'(\xi).$$

Da $g'(\xi) > 0$, folgt $g(1) > g(0)$. Mit $g(0) = f(x_0)$ und $g(1) = f(x)$ ergibt sich

$$f(x) > f(x_0).$$

Damit ist x_0 das eindeutige globale Minimum von f. □

Die obigen Resultate bilden die theoretische Grundlage für die folgende Herleitung der Kleinste-Quadrate-Schätzer im linearen Regressionsmodell.

Satz 2.5.11 (Methode der kleinsten Quadrate). *Seien $(x_1, y_1), \ldots, (x_n, y_n)$ Beobachtungswerte einer linearen Beziehung, wobei die x_i nicht alle gleich sind, d. h.*

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0.$$

Die Regressionsgerade

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

wird durch die Parameter

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \text{und} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

bestimmt, wobei $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ die arithmetischen Mittelwerte sind. Die beiden Parameter $\hat{\alpha}$ und $\hat{\beta}$ heißen auch Kleinste-Quadrate-Schätzer für α und β [32, S. 575].

Beweis. Die Schätzung der Regressionsgeraden erfolgt durch Minimierung der Summe der quadrierten Abweichungen

$$S(\alpha, \beta) = \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2.$$

Da S eine Summe quadratischer Polynome ist, gilt $S \in C^2(\mathbb{R}^2)$, siehe Definition 2.5.2.

Zur Bestimmung der kritischen Punkte werden die partiellen Ableitungen nach α und β gebildet und gleich null gesetzt:

$$\frac{\partial S}{\partial \alpha} = -2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0, \qquad \frac{\partial S}{\partial \beta} = -2\sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0.$$

Durch Auflösen dieses linearen Gleichungssystems erhält man die bekannten Formeln aus Satz 2.5.11:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Zur Untersuchung der Art des kritischen Punktes $(\hat{\alpha}, \hat{\beta})$ betrachten wir die Hesse-Matrix von S:

$$H(\alpha, \beta) = \begin{pmatrix} \frac{\partial^2 S}{\partial \alpha^2} & \frac{\partial^2 S}{\partial \alpha \partial \beta} \\ \frac{\partial^2 S}{\partial \beta \partial \alpha} & \frac{\partial^2 S}{\partial \beta^2} \end{pmatrix} = \begin{pmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix}.$$

Da H symmetrisch ist, genügt nach Bemerkung 2.5.7 zur Prüfung der positiven Definitheit, dass die führenden Hauptminoren positiv sind. Hier gilt

$$\frac{\partial^2 S}{\partial \alpha^2} = 2n > 0 \quad \text{für alle } \alpha, \beta$$

sowie

$$\det H = \frac{\partial^2 S}{\partial \alpha^2} \cdot \frac{\partial^2 S}{\partial \beta^2} - \left(\frac{\partial^2 S}{\partial \alpha \partial \beta}\right)^2 = 4n\Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\Big) = 4n\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Damit ist $H(\alpha, \beta)$ für alle (α, β) positiv definit, wodurch S nach Satz 2.5.9 streng konvex ist. Zudem folgt aus Satz 2.5.10, dass der gefundene kritische Punkt $(\hat{\alpha}, \hat{\beta})$ eindeutig ist und ein globales Minimum der Fehlerfunktion darstellt. Folglich liefern $(\hat{\alpha}, \hat{\beta})$ die eindeutigen Kleinste-Quadrate-Schätzer der linearen Regression [53]. □

Bemerkung 2.5.12. Die Parameter $\hat{\alpha}$ und $\hat{\beta}$ ergeben sich aus der Minimierung der quadrierten vertikalen Abweichungen zwischen den Beobachtungswerten (x_i, y_i) und den durch die Regressionsgerade prognostizierten Werten [47].

Korollar 2.5.13. *Die durch die Methode der kleinsten Quadrate bestimmte Regressionsgerade verläuft stets durch den Punkt $(\bar{x}, \bar{y})$. Dieser Punkt $(\bar{x}, \bar{y})$ heißt Schwerpunkt [53].*

Beweis. Setzt man $x = \bar{x}$ in $\hat{y} = \hat{\alpha} + \hat{\beta}x$ ein, folgt

$$\hat{y} = \hat{\alpha} + \hat{\beta}\bar{x} = (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}\bar{x} = \bar{y}.$$

□

Zur Analyse des zeitlichen Verlaufs der durchschnittlichen Torsumme pro Spiel werden die folgenden Variablen definiert:

- x_i bezeichnet das Startjahr der i-ten Saison, z. B. 1965 für die Saison 1965/1966.
- y_i bezeichnet die durchschnittliche Toranzahl pro Spiel in der i-ten Saison.

Die Beobachtungswerte (x_i, y_i) bilden somit die Grundlage für die lineare Regression, die durch die Methode der kleinsten Quadrate bestimmt wird. Die vollständigen Bundesliga-Daten der Torquote pro Saison (1963–2024),

$$\{(1963; 3{,}56), (1964; 3{,}32), (1965; 3{,}23), \ldots, (2024; 3{,}13)\},$$

sind als Punkte im Koordinatensystem in der untenstehenden Abbildung 2.3 dargestellt. Im Folgenden wird die Regressionsgerade schrittweise berechnet.

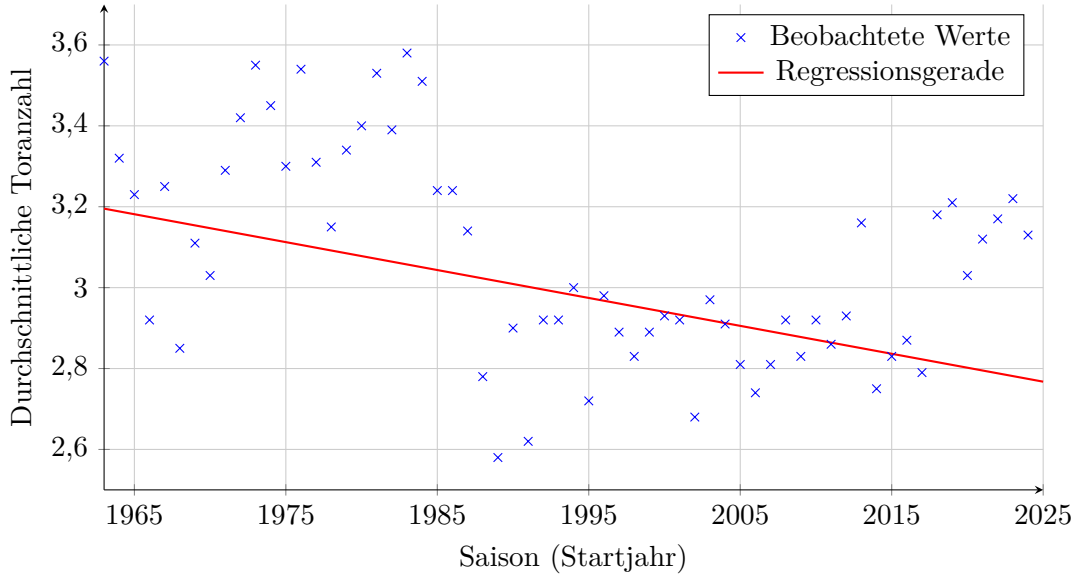

Abbildung 2.3: Lineare Regression der durchschnittlichen Toranzahl pro Spiel in der deutschen Bundesliga (1963–2024)

Schritt 1: Mittelwerte berechnen

$$\bar{x} = \frac{1}{62}\sum_{i=1}^{62} x_i = 1993{,}5 \qquad \text{und} \qquad \bar{y} = \frac{1}{62}\sum_{i=1}^{62} y_i \approx 3{,}07$$

Schritt 2: Steigungsmaß $\hat{\beta}$ berechnen

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{62}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{62}(x_i - \bar{x})^2} \\ &= \frac{(1963 - 1993{,}5)(3{,}56 - 3{,}07) + \ldots + (2024 - 1993{,}5)(3{,}13 - 3{,}07)}{(1963 - 1993{,}5)^2 + \ldots + (2024 - 1993{,}5)^2} \\ &\approx -0{,}00686\end{aligned}$$

Schritt 3: Ordinatenabschnitt $\hat{\alpha}$ berechnen

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 3{,}07 - (-0{,}00686) \cdot 1993{,}5 \approx 16{,}738$$

Schritt 4: Regressionsgerade aufstellen

$$\hat{y} = 16{,}738 - 0{,}00686x$$

Abbildung 2.3 zeigt die Regressionsgerade und veranschaulicht den Trend der durchschnittlichen Toranzahl im Zeitverlauf. Der leicht fallende Funktionsgraph deutet darauf hin, dass die durchschnittliche Toranzahl pro Spiel im Mittel um etwa 0,007 pro Saisonjahr abnimmt. Der Ordinatenabschnitt ist hingegen rein mathematischer Natur und daher nicht inhaltlich interpretierbar.

Ob der seit 2018 beobachtete empirische Anstieg der Torquote anhält oder sich der nur schwach ausgeprägte negative Trend der Regressionsgerade langfristig bemerkbar macht, wird sich in den kommenden Jahren zeigen.

2.6 Ursachen der geringen Toranzahl

Regelmäßig wird im professionellen Fußball über Regeländerungen diskutiert, um bestimmte Effekte im Spiel zu erzielen – insbesondere, um die Zahl der Tore zu erhöhen. Mehrere wissenschaftliche Studien haben sich mit den Auswirkungen solcher Maßnahmen auseinandergesetzt. Besonders hervorzuheben ist die Arbeit von Palacios-Huerta [69], der auf Grundlage von mehr als 100 000 Spielen aus den vier höchsten englischen Ligen im Zeitraum von 1888 bis 1996 einen kontinuierlichen Rückgang der durchschnittlichen Toranzahl nachweisen konnte. Seine Analyse ergab folgende Mittelwerte:

- 1888–1915: durchschnittlich 3,28 Tore pro Spiel
- 1920–1939: durchschnittlich 3,16 Tore pro Spiel
- 1947–1982: durchschnittlich 2,88 Tore pro Spiel
- 1983–1996: durchschnittlich 2,67 Tore pro Spiel

Dieser langfristige Abwärtstrend bildete die Grundlage für eine der bedeutendsten Regeländerungen der FIFA: die Einführung der Drei-Punkte-Regel. Während bis Mitte der 1990er Jahre ein Sieg mit zwei Punkten, ein Unentschieden mit einem Punkt und eine Niederlage mit null Punkten bewertet wurde, beschloss der Weltverband FIFA 1994, die Punktevergabe zu überarbeiten. Zur Saison 1995/96 übernahm auch der Deutsche Fußball-Bund (DFB) dieses System. Von nun an wurden Siege mit drei Punkten belohnt, wodurch Unentschieden deutlich an Wert verloren [22]. Ziel der Reform war es, offensivere Spielweisen zu fördern, die Zahl torloser Partien zu reduzieren und insgesamt attraktivere Begegnungen zu ermöglichen.

Die empirische Untersuchung von Palacios-Huerta und Garicano [70] zeigt jedoch, dass diese Erwartung nicht erfüllt wurde. Im Gegenteil: Anstatt höhere Offensivrisiken einzugehen, konzentrierten sich die Mannschaften verstärkt auf ihre Defensivarbeit. Ein knapper Vorsprung wurde zunehmend durch eine tiefstehende Formation verteidigt, was

zu einem weiteren Rückgang der Torquoten führte. In der französischen Ligue 1 wurden in der Saison 2005/06 lediglich 2,13 Treffer pro Spiel erzielt. Der zusätzliche Punkt für einen Sieg bewirkte somit keine Steigerung der Offensivleistung, sondern führte zu einem strategischen Umdenken zugunsten der Defensive. Biermann [12] bringt dies auf den Punkt, indem er die Situation mit anderen Lebenssituationen vergleicht, in denen eine Belohnung nicht zwangsläufig zu einer Steigerung der eigenen Anstrengungen führt, sondern vielmehr zu einer gezielten Einschränkung der Leistungen des Gegners.

Ein ähnliches Bild zeigt sich in der deutschen Bundesliga. Eine Untersuchung von Dilger und Geyer [22], die den Zeitraum von zehn Jahren vor und nach der Einführung der Drei-Punkte-Regel umfasst, konnte zwar einen leichten Rückgang der Torquote von 2,93 auf 2,87 Tore pro Spiel feststellen, dieser erwies sich jedoch als statistisch insignifikant. Anhand von Abbildung 2.3 (S. 24) ist zudem erkennbar, dass die Toranzahlen in der Bundesliga insbesondere vor Beginn der 1990er Jahre höher lagen, was vor allem auf taktische Veränderungen im Profifußball zurückgeführt wird [46]. Auch aktuelle Daten verdeutlichen die Problematik. Zwischen den Spielzeiten 2020/21 und 2024/25 verzeichneten die 18 Bundesligavereine durchschnittlich etwa 16 Torabschlüsse pro Spiel, von denen jedoch nur ein vergleichsweise geringer Anteil erfolgreich verwertet wurde [90]. Dieser Befund lässt darauf schließen, dass das Erzielen von Toren nicht allein von der Qualität der Offensivaktionen abhängt, sondern vor allem auch durch die gestiegenen Defensivfähigkeiten der Profispieler:innen erschwert wird. Physische Höchstleistungen über die gesamte Spieldauer, detaillierte Gegneranalysen mithilfe moderner Technologien sowie taktische Anpassungen wie kompakte Abwehrformationen und ein akribisches Einüben von Standardsituationen tragen dazu bei, Torchancen effektiver zu verhindern.

Die Folge dieser Entwicklung ist, dass Spiele zunehmend durch minimale Vorsprünge entschieden werden. Ein einzelnes Tor kann den Ausschlag über Sieg, Unentschieden oder Niederlage geben, wodurch sich die Wahrscheinlichkeit unerwarteter Ergebnisse unweigerlich erhöht. Ben-Naim et al. [10] konnten in einer vergleichenden Studie aufzeigen, dass Fußball mit einer Wahrscheinlichkeit von 45 % jene Sportart ist, in der Favoriten am häufigsten gegen Außenseiter verlieren. Diese hohe Unsicherheit trägt maßgeblich zur Attraktivität des Fußballs bei und erklärt seine anhaltende Popularität. Gemeinsam mit der im vorherigen Abschnitt angesprochenen Zufallskomponente verleiht die vergleichsweise geringe Toranzahl dem Spiel seine besondere Spannung und erhöht dessen Unterhaltungswert. Aufbauend auf diesen Erkenntnissen wird im Hauptteil der Arbeit dargelegt, welche Auswirkungen die geringe Toranzahl auf die Modellierung des Fußballspiels hat und welche besonderen Implikationen sich daraus ergeben.

2.7 Heimvorteil

Der Heimvorteil gilt im Fußball seit Langem als zentrale Einflussgröße auf den Spielerfolg. Ob und in welchem Ausmaß dieses Phänomen im modernen Fußball weiterhin besteht, lässt sich anhand empirischer Untersuchungen und aktueller Daten analysieren.

Historische Analysen der deutschen Bundesliga zeigen einen Rückgang des Heimvorteils. So stellte Dilger und Geyer [22] im Zusammenhang mit der Einführung der Drei-Punkte-Regel fest, dass die Quote der Heimsiege von 49,23 % auf 48,50 % leicht zurückging, während die Auswärtssiege signifikant von 21,54 % auf 25,75 % anstiegen. Ähnliche Befunde berichteten Dewenter (2003) für die portugiesische Liga sowie Amann, Dewenter und Namini (2004) ebenfalls für die Bundesliga (zitiert nach [22]). Diese Ergebnisse deuten darauf hin, dass die Drei-Punkte-Regel den Heimvorteil tendenziell verringerte, da sie den Anreiz erhöhte, auch auswärts aktiv auf Sieg zu spielen. Gleichzeitig nahm die durchschnittliche Toranzahl ab, wobei der Rückgang der Heimtore signifikant ausfiel, derjenige der Auswärtstore hingegen nur geringfügig.

Langfristige Entwicklungen belegen einen deutlichen Rückgang des Heimvorteils: In den 1960er- und 1970er-Jahren lag die Heimsiegquote bei durchschnittlich etwa 56 %, während die Spielzeiten 2020/21 bis 2024/25 im Mittel nur noch rund 44 % aufwiesen (vgl. Tabelle 2.4, S. 28). Die Ursachen sind vielschichtig, lassen sich jedoch im Wesentlichen mit der zunehmenden Professionalisierung des Sports in Verbindung bringen. Technische Innovationen, globalisierte Spielerkarrieren, moderne Stadioninfrastrukturen sowie die Einführung des Video Assistant Referee (VAR) haben den Einfluss des Spielorts relativiert. Hinzu kommt die COVID-19-Pandemie: Während der Geisterspielphasen 2019/20 und 2020/21 sank die Heimsiegquote auf historische Tiefstwerte von teils unter 40 % [33]. Untersuchungen von Stickel und Nufer [79] belegen zwar einen Anstieg des Heimvorteils mit der Rückkehr der Zuschauer:innen, langfristig jedoch eine erneute Abnahme – ungeachtet der steigenden Zuschauerzahlen.

Trotz dieses Rückgangs bleibt der Heimvorteil statistisch nachweisbar. Seit Beginn der Bundesliga 1963 liegt die Heimsiegquote knapp über 50 %. Im Dezember 2024 betrug sie 50,17 % (9 394 von 18 725 Spielen). Da Auswärtssiege und Unentschieden die übrigen rund 50 % nahezu gleichmäßig teilen, erzielen Heimteams im Mittel mehr Punkte. Zu den zentralen Faktoren zählen die Unterstützung durch das Publikum, die Vertrautheit mit der Spielumgebung und die Vermeidung längerer Anreisen. Gleichzeitig haben taktische Entwicklungen, die weniger an den Spielort gebunden sind, sowie strukturelle Veränderungen im Zuge der Professionalisierung des Fußballs den Effekt abgeschwächt [4].

Neben den spielpraktischen Faktoren beeinflusst auch die Reisedistanz den Heimvorteil. Oberhofer et al. [63] zeigten anhand von über 6 000 Bundesliga-Auswärtsspielen, dass die Leistung eines Teams in Bezug auf das Erzielen und Verhindern von Toren mit zunehmender Entfernung abnimmt. Beckmann [7] bestätigte diesen Zusammenhang in einer Analyse der deutschen Bundesliga über 57 Jahre, weist jedoch darauf hin, dass die Bedeutung der Reisedistanz in den letzten Jahren durch verbesserte Reisebedingungen, eine wachsende Zahl mitreisender Auswärtsfans und die gesteigerte Professionalisierung des Sports abgenommen hat.

Insgesamt lässt sich festhalten, dass der Heimvorteil im modernen Fußball zwar rückläufig ist, jedoch weiterhin besteht und durch ein Zusammenspiel historischer, taktischer, infrastruktureller und psychologischer Faktoren geprägt wird.

Um den aktuellen Heimvorteil in der deutschen Bundesliga zu quantifizieren und für die Modellierung eines Fußballspiels nutzbar zu machen, wurden erneut die fünf Spielzeiten von 2020/21 bis 2024/25 analysiert. Die in der folgenden Tabelle 2.4 dargestellte Verteilung der Ergebnisse und Punkte zeigt, dass mit Ausnahme der Saison 2024/25 stets mehr als 40 % aller Spiele im eigenen Stadion gewonnen wurden. Im Mittel beträgt das Verhältnis von Heimsiegen (HS) zu Auswärtssiegen (AS) etwa 1,41 : 1, was bedeutet, dass ein Team zu Hause rund 41 % häufiger gewinnt als auswärts. Dies stellt ein deutliches Indiz dafür dar, dass der Heimvorteil nach wie vor eine relevante Rolle spielt. Bestätigt wird diese Annahme durch den Anteil der im eigenen Stadion erzielten Punkte, der im Mittel bei 56,96 % liegt. Über die vergangenen fünf Spielzeiten haben die 18 Bundesliga-Teams demnach durchschnittlich mehr als die Hälfte ihrer Punkte vor heimischem Publikum erzielt [20, 90].

Tabelle 2.4: Heimvorteil der deutschen Bundesliga (2020/21 – 2024/25): Ergebnis- und Punkteverteilung

Saison	Spiele	HS (%)	Remis (%)	AS (%)	Punkte zuhause
2020/21	306	129 (42)	81 (26)	96 (31)	55,91 %
2021/22	306	143 (47)	73 (24)	90 (29)	59,41 %
2022/23	306	145 (47)	75 (25)	86 (28)	60,50 %
2023/24	306	134 (44)	81 (26)	91 (30)	57,71 %
2024/25	306	118 (39)	77 (25)	111 (36)	51,25 %
Gesamt	**1 530**	**669 (44)**	**387 (25)**	**474 (31)**	**56,96 %**

Tabelle 2.5: Heimvorteil der deutschen Bundesliga (2020/21 – 2024/25): Torverteilung

Saison	Gesamttore	Heimtore (%)	Auswärtstore (%)
2020/21	928	513 (55)	415 (45)
2021/22	954	539 (56)	415 (44)
2022/23	971	568 (59)	403 (41)
2023/24	985	553 (56)	432 (44)
2024/25	959	499 (52)	460 (48)
Gesamt	**4 797**	**2 672 (56)**	**2 125 (44)**

Da die Toranzahl eine noch aussagekräftigere Kennzahl für die Spielstärke eines Teams darstellt, ist es ebenso aufschlussreich, die Anzahl der Tore in Abhängigkeit vom Spielort zu analysieren. Im betrachteten Zeitraum wurden insgesamt 4 797 Treffer registriert, davon 2 672 im eigenen Stadion und 2 125 auswärts (vgl. Tabelle 2.5). Damit entfielen mit 56 % mehr als die Hälfte aller Tore auf Heimteams. Im Untersuchungszeitraum erzielte ein Heimteam durchschnittlich etwa 1,75 Tore pro Spiel, während ein Auswärtsteam auf rund 1,39 Tore kam, was einem Verhältnis von Heim- zu Auswärtstoren von etwa 1,26 : 1

entspricht. Heimteams erzielen demnach etwa 26 % mehr Tore als Auswärtsteams [20, 90]. Dies verdeutlicht, dass der Heimvorteil über die reine Punkteausbeute hinausgeht und sowohl offensiv (höhere Torzahl) als auch defensiv (geringere Gegentoranzahl) seine Wirkung offenbart. Somit unterstreicht auch die Torstatistik den Einfluss des Heimvorteils.

Die dargestellten Analysen liefern hinreichende Evidenz dafür, den Heimvorteil als wesentliche Einflussgröße im Grundgerüst der Modellierung eines Fußballspiels zu berücksichtigen.

3 Grundlagen der Wahrscheinlichkeitstheorie

Bevor die Wahrscheinlichkeit verschiedener Spielausgänge im Fußball untersucht werden kann, ist es erforderlich, zentrale Begriffe zu definieren und grundlegende Konzepte der Wahrscheinlichkeitsrechnung zu erläutern.

3.1 Wahrscheinlichkeitsbegriff

> Einer der großen Vorteile der Wahrscheinlichkeitsrechnung ist der, daß man lernt, dem ersten Anschein zu mißtrauen [55, S. 127].

Dieses Zitat von Pierre-Simon Laplace verdeutlicht den Kern der Wahrscheinlichkeitstheorie. Sie fordert dazu auf, intuitiven Einschätzungen kritisch zu begegnen und Zufallsphänomene nicht bloß auf Grundlage subjektiver Wahrnehmungen zu beurteilen. Die Wahrscheinlichkeitstheorie beschäftigt sich mit Vorgängen, deren Ausgänge vom Zufall abhängen und daher nicht eindeutig vorhergesagt werden können. Ihr Ziel besteht darin, Zufälligkeiten und Unsicherheiten systematisch zu erfassen, mathematisch zu modellieren und daraus objektive Prognosen über das Eintreten bestimmter Ereignisse abzuleiten [43].

Historisch betrachtet entstanden die ersten Überlegungen zur Wahrscheinlichkeit im Zusammenhang mit Glücksspielen. Aus diesen frühen Ansätzen entwickelte sich ein hoch formalisiertes Teilgebiet der Mathematik, das heute in nahezu allen Wissenschaften Anwendung findet – überall dort, wo Zufall eine Rolle spielt [56]. Auch im alltäglichen Sprachgebrauch wird der Begriff *wahrscheinlich* verwendet, um Unsicherheiten und Vermutungen auszudrücken, wie etwa: „wahrscheinlich wird es morgen regnen“ oder „wahrscheinlich sehen wir uns morgen“ [16].

Im Zentrum der Wahrscheinlichkeitstheorie stehen Experimente, bei denen die Menge aller möglichen Ergebnisse bereits im Vorfeld bekannt ist, der konkrete Ausgang jedoch vor der Durchführung ungewiss bleibt und erst durch die Ausführung bestimmt wird. Ein Vorgang, der diese Eigenschaften aufweist und zudem real oder gedanklich unter denselben Bedingungen beliebig oft wiederholt werden kann, wird als *Zufallsexperiment* bezeichnet [14].

Beispiele für Zufallsvorgänge sind etwa das Werfen eines Würfels, die Ziehung der Lottozahlen oder – im Fußballkontext – der Münzwurf zu Spielbeginn beziehungsweise der

Spielausgang selbst. Bei einem Fußballspiel ist bekannt, welche Ergebnisse grundsätzlich möglich sind (Sieg der Heimmannschaft, Unentschieden oder Sieg der Auswärtsmannschaft). Vor dem Anpfiff ist jedoch offen, welches dieser Ergebnisse tatsächlich eintreten wird. Ein solches Spiel kann zudem theoretisch mehrfach wiederholt werden [48].

Im nachfolgenden Abschnitt sollen zentrale Begriffe der Wahrscheinlichkeitstheorie exemplarisch am Beispiel eines klassischen Zufallsexperiments, dem Werfen eines Würfels, erläutert werden.

Definition 3.1.1 (**Grundraum, Ergebnis, Ereignis, Elementarereignis**)**.** Die Menge aller möglichen Ergebnisse eines Zufallsexperiments wird als *Grundraum* oder *Ergebnismenge* bezeichnet und üblicherweise mit Ω symbolisiert. Sie ist eine nichtleere Menge, die sämtliche beim Zufallsexperiment denkbaren Ergebnisse ω_i enthält [17, 48].

$$\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$$

Ein *Ergebnis* ω_i beschreibt den konkreten Ausgang eines Zufallsexperiments und entspricht einem Element der Ergebnismenge [17].

Jedes einzelne dieser möglichen Ergebnisse wird als *Elementarereignis* bezeichnet. Formal entsprechen die Elementarereignisse den einelementigen Teilmengen $\{\omega_i\}$, deren Schnittmengen jeweils leer sind. Man sagt auch, dass die Elementarereignisse paarweise disjunkt sind, d. h. für $i \neq j$ gilt $\{\omega_i\} \cap \{\omega_j\} = \varnothing$ [56].

Ein *Ereignis* hingegen ist eine Teilmenge des Grundraums Ω, deren enthaltene Elementarereignisse eine gemeinsame Eigenschaft erfüllen. Die Menge aller zulässigen Ereignisse wird durch die *σ-Algebra* festgelegt, die im weiteren Verlauf der Arbeit noch genauer definiert wird [58].

Das Ereignis, das bei jeder Durchführung des Zufallsexperiments eintritt, wird als *sicheres Ereignis* bezeichnet und umfasst alle Elementarereignisse aus Ω. Ein Ereignis, das niemals eintreten kann, heißt *unmögliches Ereignis* und beinhaltet kein Element aus Ω; wird also durch die leere Menge $\varnothing = \{\}$ dargestellt [14, 48].

Hat man ein Ereignis $A \subseteq \Omega$ gegeben und wird ein Ergebnis $\omega \in \Omega$ beobachtet, so lassen sich zwei grundlegende Fälle unterscheiden:

1. Liegt ω in A, so sagt man, dass das Ereignis A *eingetreten* ist.
2. Liegt ω nicht in A, das heißt $\omega \in A'$, so ist das Ereignis A *nicht eingetreten* [58].

Beispiel 3.1.2 (Einfacher Würfelwurf)**.** Zur Veranschaulichung sei das Zufallsexperiment des Werfens eines idealen Würfels betrachtet. Der Grundraum bzw. die Ergebnismenge umfasst alle möglichen Augenzahlen von eins bis sechs und lautet daher

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Dementsprechend setzen sich die Elementarereignisse aus den einelementigen Mengen

$$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$$

zusammen. Ist A das Ereignis „es wird eine gerade Zahl geworfen“, so tritt A genau dann ein, wenn das Ergebnis in der Menge

$$A = \{2, 4, 6\}$$

enthalten ist. Das Ereignis B „die Augenzahl ist kleiner als drei“ tritt entsprechend genau dann ein, wenn das Ergebnis ein Element der Menge

$$B = \{1, 2\}$$

ist.

Obwohl Wahrscheinlichkeiten intuitiv oft genutzt werden, wurde die mathematische Untersuchung von Zufallsvorgängen erst 1933 durch den russischen Mathematiker Andrej N. Kolmogorov auf ein axiomatisches Fundament gestellt.

Er postulierte, dass sich Zufallsexperimente durch geeignete *Wahrscheinlichkeitsräume* beschreiben lassen, die formal durch das Tripel $(\Omega, \mathcal{A}, P)$ gegeben sind. Dabei bezeichnet Ω den *Grundraum*, $\mathcal{A}$ die *Ereignis-σ-Algebra* und P eine Abbildung von $\mathcal{A}$ in das Intervall $[0, 1]$, das sogenannte *Wahrscheinlichkeitsmaß* [58].

Damit Wahrscheinlichkeiten mathematisch sinnvoll definiert werden können, muss zunächst festgelegt werden, für welche Teilmengen des Grundraums Ω überhaupt Wahrscheinlichkeiten angegeben werden dürfen. Hierfür wird eine spezielle Mengensystematik benötigt – die sogenannte *σ-Algebra*.

Definition 3.1.3 (σ-Algebra). Nach Leiner [56, S. 69] nennt man ein System $\mathcal{A}$ von Teilmengen eines Raumes Ω eine σ-Algebra, wenn folgende Eigenschaften erfüllt sind:

1. $\Omega \in \mathcal{A}$,
2. Ist $A \in \mathcal{A}$, so gilt für das Komplement $A' = \Omega \setminus A$, dass $A' \in \mathcal{A}$,
3. Für jede abzählbare Folge $(A_i)_{i \in \mathbb{N}}$ mit $A_i \in \mathcal{A}$ gilt

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}.$$

Die Elemente der σ-Algebra $\mathcal{A}$ werden als *Ereignisse* bezeichnet. In endlichen Grundräumen, wie beim Würfelwurf, kann die Potenzmenge $\mathcal{P}(\Omega)$ als σ-Algebra gewählt werden. Für unendliche Grundräume, beispielsweise $\Omega = \mathbb{R}$, verwendet man üblicherweise die *Borel-σ-Algebra*, die im Folgenden definiert wird [58].

Definition 3.1.4 (Borel-σ-Algebra). Sei $\mathcal{C} \subseteq \mathcal{P}(\mathbb{R})$ die Menge aller abgeschlossenen endlichen Intervalle in $\mathbb{R}$, also

$$\mathcal{C} = \{[a, b] : a < b\} \quad \text{mit } a, b \in \mathbb{R}.$$

Die von $\mathcal{C}$ auf $\mathbb{R}$ erzeugte σ-Algebra $\sigma(\mathcal{C})$ heißt *Borelsche σ-Algebra* und wird mit $\mathcal{B}(\mathbb{R})$ bezeichnet. Die Elemente von $\mathcal{B}(\mathbb{R})$ heißen *Borel-Mengen* [58, S. 5].

Definition 3.1.5 (**Paarweise disjunkte Mengen**)**.** Eine Familie von Mengen $(A_i)_{i\in I}$ heißt *paarweise disjunkt*, wenn gilt:

$$A_i \cap A_j = \varnothing \quad \text{für alle } i, j \in I \text{ mit } i \neq j.$$

Das bedeutet, dass keine zwei Mengen dieser Familie gemeinsame Elemente besitzen [17, S. 163].

Definition 3.1.6 (**Kolmogorov-Axiome**)**.** Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, wobei $\mathcal{A}$ eine σ-Algebra auf Ω und $P : \mathcal{A} \to [0, 1]$ eine Abbildung ist. Nach Leiner [56, S. 74] und Bosch [14, S. 9] erfüllt das Wahrscheinlichkeitsmaß P folgende Axiome:

1. **Nichtnegativität:** Für jedes Ereignis $A \in \mathcal{A}$ gilt
$$P(A) \geq 0.$$
2. **Normiertheit:** Das sichere Ereignis Ω hat Wahrscheinlichkeit 1, also
$$P(\Omega) = 1.$$
3. **Additivität:** P ist σ-additiv, d. h. für jede abzählbar unendliche Folge paarweise disjunkter Ereignisse $A_1, A_2, A_3, \ldots \in \mathcal{A}$ gilt
$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Aus den Kolmogorov-Axiomen lassen sich nun mehrere grundlegende Eigenschaften ableiten, die für die Berechnung von Wahrscheinlichkeiten von großem Nutzen sind. Die folgenden Sätze und Beweise orientieren sich inhaltlich an den Darstellungen von Linde [58, S. 9–10], Bosch [14, S. 9] und Leiner [56, S. 75–76].

Satz 3.1.7. *Für jedes Ereignis $A \in \mathcal{A}$ gilt:*

$$P(A') = 1 - P(A).$$

Beweis. Da A und A' disjunkt sind und $A \cup A' = \Omega$ gilt, folgt mit der σ-Additivität und der Normiertheit:

$$1 = P(\Omega) = P(A \cup A') = P(A) + P(A').$$

Daraus ergibt sich unmittelbar $P(A') = 1 - P(A)$. □

Satz 3.1.8. *Für das unmögliche Ereignis $\varnothing$ gilt:*

$$P(\varnothing) = 0.$$

Beweis. Aus Satz 3.1.7 und der Normiertheit folgt:

$$P(\varnothing) = 1 - P(\Omega) = 1 - 1 = 0.$$

□

Satz 3.1.9. *Das Wahrscheinlichkeitsmaß ist monoton, d. h. für $A, B \in \mathcal{A}$ mit $A \subseteq B$ gilt:*

$$P(A) \leq P(B).$$

Beweis. Da B disjunkt zerlegt werden kann in $B = A \cup (B \setminus A)$, folgt mit der Additivität:

$$P(B) = P(A) + P(B \setminus A).$$

Wegen der Nichtnegativität ist $P(B \setminus A) \geq 0$, also $P(A) \leq P(B)$. □

Satz 3.1.10. *Für $A, B \in \mathcal{A}$ mit $A \subseteq B$ gilt:*

$$P(B \setminus A) = P(B) - P(A).$$

Beweis. Aus der Zerlegung $B = A \cup (B \setminus A)$ folgt direkt:

$$P(B) = P(A) + P(B \setminus A).$$

Nach Umstellen ergibt sich die Behauptung. □

Satz 3.1.11. *Für alle $A, B \in \mathcal{A}$ gilt:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Beweis. Die Mengen $A \setminus B$, $B \setminus A$ und $A \cap B$ sind paarweise disjunkt und bilden zusammen $A \cup B$. Mit der Additivität folgt:

$$P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B).$$

Außerdem gilt:

$$P(A) = P(A \setminus B) + P(A \cap B), \qquad P(B) = P(B \setminus A) + P(A \cap B).$$

Durch Addition und Umstellen erhält man:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

□

3.2 Bestimmung von Wahrscheinlichkeiten

Nachdem die zentralen Eigenschaften des Wahrscheinlichkeitsraums $(\Omega, \mathcal{A}, P)$ eingeführt wurden, soll im Folgenden der Frage nachgegangen werden, wie Wahrscheinlichkeiten in der Praxis oder auf Basis theoretischer Überlegungen bestimmt werden können. Dabei lassen sich zwei grundlegende Zugänge unterscheiden: die empirisch-statistische (*a posteriori*) und die modelltheoretische (*a priori*) Bestimmung von Wahrscheinlichkeiten.

3.2.1 Empirische (frequentistische) Wahrscheinlichkeit

Der frequentistische Wahrscheinlichkeitsbegriff, auch als *statistische* oder *a-posteriori*-Wahrscheinlichkeit bezeichnet, wurde maßgeblich durch Richard von Mises (1931) geprägt. Er beruht auf der Annahme, dass sich die Wahrscheinlichkeit eines Ereignisses aus der wiederholten Durchführung desselben Zufallsexperiments erschließen lässt [48].

Wird ein Zufallsexperiment n-mal wiederholt, so kann gezählt werden, wie oft ein bestimmtes Ereignis A eintritt. Diese Anzahl lässt sich als Verhältnis der absoluten Häufigkeit $h_n(A)$ zur Gesamtzahl der Durchführungen n in Form der *relativen Häufigkeit* darstellen:

$$f_n(A) = \frac{h_n(A)}{n}.$$

Da sich die relativen Häufigkeiten nach jeder Wiederholung des Zufallsexperiments ändern können, entsteht eine Folge von Werten $f_n(A)$. Zu Beginn schwanken diese meist stark, doch mit zunehmender Anzahl der Wiederholungen stabilisieren sie sich und nähern sich einem festen Wert $P(A)$ an [43, 48, 56]. Dieses Stabilitätsverhalten der relativen Häufigkeit bildet die Grundlage des frequentistischen Wahrscheinlichkeitsverständnisses und ist in Abbildung 3.1 (S. 37) dargestellt.

Definition 3.2.1 (**Empirische Wahrscheinlichkeit**)**.** Sei $A \in \mathcal{A}$ ein Ereignis und $f_n(A)$ die relative Häufigkeit von A nach n Wiederholungen des Zufallsexperiments. Dann ist die *empirische Wahrscheinlichkeit* von A definiert als

$$P(A) = \lim_{n \to \infty} f_n(A),$$

sofern der Grenzwert existiert [48, S. 15].

Da ein Zufallsexperiment in der Praxis nicht unendlich oft wiederholt werden kann, lässt sich der Grenzwert empirisch nicht exakt bestimmen. Selbst wenn man seine Existenz annimmt, ist diese Annahme ohne Informationen über die Konvergenzgeschwindigkeit von begrenztem praktischem Nutzen. Nichtsdestotrotz wird sie im frequentistischen Ansatz vorausgesetzt, um Wahrscheinlichkeiten in anwendungsbezogenen Kontexten zu modellieren, wie sie insbesondere in der Statistik, der Datenanalyse, im Ingenieurwesen oder bei der Analyse sportlicher Ereignisse von Bedeutung sind. Die angenommene Stabilisierung der relativen Häufigkeiten findet ihren Ausdruck im *empirischen Gesetz der großen Zahlen*, das auf Jacob Bernoulli zurückgeht [43].

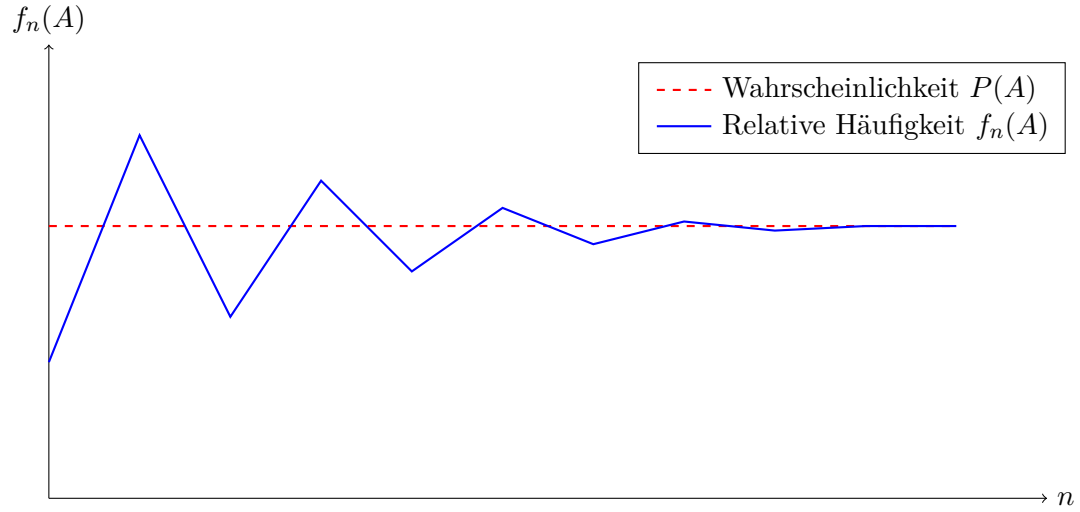

Abbildung 3.1: Stabilitätseigenschaft der relativen Häufigkeit (eigene Darstellung in Anlehnung an Kosfeld et al. [48, S. 15])

Beispiel 3.2.2 (Münzwurf)**.** Wir betrachten einen fairen Münzwurf. Um einen Richtwert für die *empirische Wahrscheinlichkeit* zu erhalten, wird das Zufallsexperiment mehrfach durchgeführt und gezählt, wie oft das Ereignis „Z = Zahl“ auftritt. Es ergibt sich somit eine Folge relativer Häufigkeiten

$$f_n(Z) = \frac{h_n(Z)}{n},$$

wobei $h_n(Z)$ die absolute Häufigkeit nach n Versuchen darstellt. Die entsprechenden Werte sind in der nachstehenden Tabelle 3.1 zusammengefasst.

Tabelle 3.1: Absolute und relative Häufigkeiten des Ereignisses „Zahl“ bei Münzwürfen

Anzahl der Versuche n	Absolute Häufigkeit h_n	Relative Häufigkeit f_n
50	29	0,58
100	54	0,54
300	136	0,4533
600	289	0,4817
1000	507	0,507

Man erkennt, dass die relative Häufigkeit $f_n(Z)$ mit zunehmendem n dem theoretischen Wert $P(Z) = 0{,}5$ immer näher kommt. Dies veranschaulicht die *Stabilitätseigenschaft der relativen Häufigkeit*.

Beispiel 3.2.3 (Fußball: Torschüsse und Tore)**.** Betrachtet man alle Torschüsse eines Teams über mehrere Spiele, so kann das Ereignis A als „Torschuss führt zu einem Tor“

definiert werden. Die absolute Häufigkeit $h_n(A)$ entspricht dabei der Anzahl der Torschüsse, die zu einem Tor führen. Die relative Häufigkeit $f_n(A)$ gibt den Anteil der erzielten Tore an allen abgegebenen Torschüssen an. Mit zunehmender Anzahl von Versuchen nähert sich diese relative Häufigkeit einem stabilen Wert, der die *empirische Wahrscheinlichkeit* des Ereignisses „Torerfolg“ repräsentiert.

In einem Spiel der deutschen Bundesliga fallen durchschnittlich 3,14 Tore pro Spiel, also 1,57 Tore pro Team (vgl. Tabelle 2.3 auf S. 17). Laut Dambeck [18] liegt die mittlere Anzahl an Schussversuchen – einschließlich verfehlter und geblockter Abschlüsse – bei etwa 14 pro Team und Spiel. Geht man davon aus, dass ein beliebiges Team in einem Spiel 14 Schüsse abgibt und dabei zwei Tore erzielt, so ergibt sich die relative Häufigkeit

$$f_{14}(A) = \frac{2}{14} \approx 0{,}1429.$$

Da die zugrunde liegenden Durchschnittswerte für Tore und Torschüsse auf Daten über mehrere Jahre beruhen, kann dieser Wert bereits als gute Annäherung an die tatsächliche Trefferwahrscheinlichkeit betrachtet werden.

3.2.2 Modelltheoretische (Laplacesche) Wahrscheinlichkeit

Im Gegensatz zur empirischen Betrachtung wird bei der modelltheoretischen oder *a-priori*-Auffassung der Wahrscheinlichkeit nicht auf Beobachtungen zurückgegriffen, sondern ein theoretisches Modell zugrunde gelegt. Diese Herangehensweise geht auf den französischen Mathematiker Pierre-Simon Laplace (1749–1827) zurück und wird auch als *Gleichwahrscheinlichkeitsmodell* bezeichnet [48, 56].

Dabei wird angenommen, dass alle möglichen Ergebnisse bzw. Elementarereignisse eines Zufallsexperiments gleich wahrscheinlich sind. Diese Annahme ist insbesondere bei idealisierten Experimenten mit endlichem Ergebnisraum plausibel, wie etwa beim Würfeln eines fairen Würfels oder bei der Ziehung der Lottozahlen. In diesem Fall ergibt sich die Wahrscheinlichkeit als Verhältnis der Anzahl der für ein Ereignis günstigen Fälle zur Gesamtzahl aller möglichen Fälle [48, 56].

Definition 3.2.4 (Laplacesche Wahrscheinlichkeit). Sei $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ die endliche Menge gleichwahrscheinlicher Elementarereignisse und $A \subseteq \Omega$ ein Ereignis. Nach Cramer und Kamps [17, S. 164] ist die *Laplacesche Wahrscheinlichkeit* von A gegeben durch:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der für A günstigen Fälle}}{\text{Anzahl der möglichen Fälle}}.$$

Die Formel der Laplaceschen Wahrscheinlichkeit lässt sich, in Anlehnung an Hofbauer und Greschonig [43, S. 14] und Bosch [14, S. 12–13], unmittelbar aus den Grundannahmen dieses Ansatzes herleiten.

Sei der Ergebnisraum

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

eine endliche Menge aus n gleichwahrscheinlichen Elementarereignissen. Da sich der Ergebnisraum als disjunkte Vereinigung dieser Elementarereignisse darstellen lässt,

$$\Omega = \{\omega_1\} \cup \{\omega_2\} \cup \ldots \cup \{\omega_n\},$$

folgt nach den Kolmogorov-Axiomen (Definition 3.1.6, S. 34):

$$1 = P(\Omega) = P(\{\omega_1\}) + P(\{\omega_2\}) + \ldots + P(\{\omega_n\}).$$

Da alle Elementarereignisse gleich wahrscheinlich sind, gilt:

$$P(\{\omega_i\}) = p \quad \text{für alle } i = 1, \ldots, n,$$

und somit:

$$1 = P(\Omega) = p + p + \ldots + p = n \cdot p \quad \Rightarrow \quad p = \frac{1}{n}.$$

Für ein Ereignis $A \subseteq \Omega$, das aus k dieser Elementarereignisse besteht, ergibt sich daraus:

$$P(A) = k \cdot \frac{1}{n} = \frac{k}{n}.$$

Mit $n = |\Omega|$ und $k = |A|$ erhält man schließlich die bekannte Formel

$$P(A) = \frac{|A|}{|\Omega|}.$$

Beispiel 3.2.5 (Würfelwurf)**.** Beim Wurf eines fairen sechsseitigen Würfels ist der Ergebnisraum $\Omega = \{1, 2, 3, 4, 5, 6\}$. Für das Ereignis $A = \{\text{gerade Zahl}\} = \{2, 4, 6\}$ gibt es drei günstige von insgesamt sechs möglichen Ausgängen. Damit ergibt sich die Laplacesche Wahrscheinlichkeit zu

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}.$$

Aus diesem Beispiel wird deutlich, dass sich die Berechnung der Wahrscheinlichkeit bei einem Laplace-Experiment auf das Abzählen der Elementarereignisse reduzieren lässt.

Beispiel 3.2.6 (Zweimaliger Münzwurf)**.** Beim zweimaligen Wurf einer fairen Münze ist der Ergebnisraum

$$\Omega = \{(K, K), (K, Z), (Z, K), (Z, Z)\},$$

wobei K für Kopf und Z für Zahl steht. Sei das Ereignis A definiert als „mindestens einmal Kopf“. Anstatt $P(A)$ direkt zu berechnen, kann man hier auch mit der Gegenwahrscheinlichkeit des Komplementärereignisses arbeiten:

$$A' = \{\text{kein Kopf}\} = \{(Z, Z)\}.$$

Dann gilt

$$P(A) = 1 - P(A') = 1 - \frac{|A'|}{|\Omega|} = 1 - \frac{1}{4} = \frac{3}{4}.$$

Vor allem bei Beispielen mit einem größeren Ergebnisraum kann der Umweg über die Gegenwahrscheinlichkeit die Berechnung eleganter machen.

3.2.3 Bedingte Wahrscheinlichkeit

In vielen Zufallsexperimenten liegt neben dem interessierenden Ereignis A zusätzliche Information über ein weiteres Ereignis B vor. In einer solchen Situation stellt sich die Frage, wie sich diese Vorabinformation auf die Wahrscheinlichkeit von A auswirkt, da B das Eintreten von A sowohl begünstigen als auch benachteiligen kann. Intuitiv betrachtet richtet sich das Augenmerk daher auf die Wahrscheinlichkeit, dass A eintritt, *unter der Voraussetzung (Bedingung)*, dass das Ereignis B bereits eingetreten ist [48]. Vor dem Hintergrund des axiomatischen Wahrscheinlichkeitsbegriffs nach Kolmogorov wird dieser Zugang durch folgende Definition formalisiert.

Definition 3.2.7. Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und seien $A, B \in \mathcal{A}$ mit $P(B) > 0$. Nach Hofbauer und Greschonig [43, S. 17] ist die *bedingte Wahrscheinlichkeit* von A unter der Bedingung B definiert durch

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

Die bedingte Wahrscheinlichkeit misst damit den Anteil der Wahrscheinlichkeit von B, der zugleich auf das Ereignis A entfällt. Zudem folgt aus der Definition unmittelbar eine grundlegende Rechenregel in der Wahrscheinlichkeitstheorie, die es erlaubt, Wahrscheinlichkeiten für das gemeinsame Eintreten von Ereignissen systematisch zu berechnen.

Satz 3.2.8 (Multiplikationssatz). *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und seien $A, B \in \mathcal{A}$ mit $P(B) > 0$. Dann gilt nach Bosch [14, S. 31]*

$$P(A \cap B) = P(A \mid B) \cdot P(B).$$

Beispiel 3.2.9. Es werde ein fairer sechsseitiger Würfel einmal geworfen. Der Ergebnisraum ist

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Zunächst wird das Ereignis A „es wird eine 4 gewürfelt“ betrachtet. Für $A = \{4\}$ beträgt die Wahrscheinlichkeit für das Eintreten des Ereignisses A klarerweise

$$P(A) = \frac{1}{6}.$$

Nun stellt sich die Frage, wie sich diese Wahrscheinlichkeit verändert, wenn man erfährt, dass die gewürfelte Zahl gerade ist. Sei daher das Ereignis B „die gewürfelte Zahl ist gerade“, also $B = \{2, 4, 6\}$. Damit gilt

$$P(B) = \frac{3}{6} = \frac{1}{2} \qquad \text{und} \qquad A \cap B = \{4\} \text{ mit } P(A \cap B) = \frac{1}{6}.$$

Die bedingte Wahrscheinlichkeit von A unter der Bedingung B ergibt sich somit zu

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

Kennt man also die Information, dass eine gerade Zahl gewürfelt wurde, so beträgt die Wahrscheinlichkeit, dass es sich dabei um die Augenzahl 4 handelt, $\frac{1}{3}$.

3.3 Kombinatorik

Die Kombinatorik ist ein Teilgebiet der Mathematik, das sich mit dem Zählen von Möglichkeiten beschäftigt. Sie spielt insbesondere in der Wahrscheinlichkeitstheorie eine zentrale Rolle, da sich viele Wahrscheinlichkeiten auf das systematische Abzählen von Stichproben oder Anordnungen zurückführen lassen. Im Zusammenhang mit der Laplaceschen Wahrscheinlichkeit besteht der grundlegende Ansatz ja gerade darin, günstige und mögliche Ereignisse abzuzählen. Kombinatorische Methoden bilden die Grundlage vieler stochastischer Überlegungen, da sie es ermöglichen, die Anzahl der Möglichkeiten zu bestimmen, Elemente einer Menge zu *ordnen* oder aus ihr *auszuwählen* [35, 48].

Zur Veranschaulichung typischer Fragestellungen wird häufig das *Urnenmodell* herangezogen. Dabei enthält eine Urne n Kugeln, die sich nach bestimmten Merkmalen (z. B. Farbe oder Nummerierung) unterscheiden, aus denen k-mal gezogen wird. Jeder Ziehvorgang kann in einem Baumdiagramm als Pfad dargestellt werden [17].

Um Anordnungs- und Auswahlprobleme korrekt analysieren zu können, ist es sinnvoll, zunächst die *Pfadregeln* zu erläutern, da sie die Grundlage für die Wahrscheinlichkeitsberechnung bei mehrstufigen Zufallsexperimenten bilden. Diese Regeln lassen sich in Anlehnung an Henze [35] wie folgt zusammenfassen:

1. *Multiplikationsregel:* Die Wahrscheinlichkeit eines Pfades ergibt sich als Produkt der entlang der einzelnen Stufen auftretenden (bedingten) Wahrscheinlichkeiten. Dies folgt unmittelbar aus der Definition der bedingten Wahrscheinlichkeit und dem Multiplikationssatz (vgl. Definition 3.2.7 und Satz 3.2.8).

2. *Additionsregel:* Besteht ein Ereignis aus mehreren Pfaden, die paarweise disjunkt sind, so ergibt sich die Wahrscheinlichkeit als Summe der Wahrscheinlichkeiten aller zugehörigen Pfade. Formal beruht dies auf der σ-Additivität des Wahrscheinlichkeitsmaßes P (vgl. Axiom 3 in Definition 3.1.6): Sei $E = \bigcup_{i=1}^{n} A_i$ ein Ereignis, das aus den paarweise disjunkten Pfaden $A_1, \dots, A_n$ besteht. Dann gilt

$$P(E) = P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i).$$

Beispiel 3.3.1 (Pfadregeln bei einem zweistufigen Urnenexperiment)**.** Eine Urne enthält eine rote und zwei blaue Kugeln. Es wird zweimal mit Zurücklegen gezogen. Die Wahrscheinlichkeit, zweimal hintereinander eine rote Kugel zu ziehen, ergibt sich mit der *Multiplikationsregel* zu

$$P(\text{rot, rot}) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}.$$

Das Baumdiagramm zeigt, dass sich insgesamt neun mögliche Pfade ergeben. Unter diesen Ereignissen existiert genau ein günstiger Fall, der der Farbreihenfolge rot–rot entspricht. Da es sich um ein Laplace-Experiment handelt, entspricht die Wahrscheinlichkeit dieses Ereignisses dem Bruch $\frac{1}{9}$.

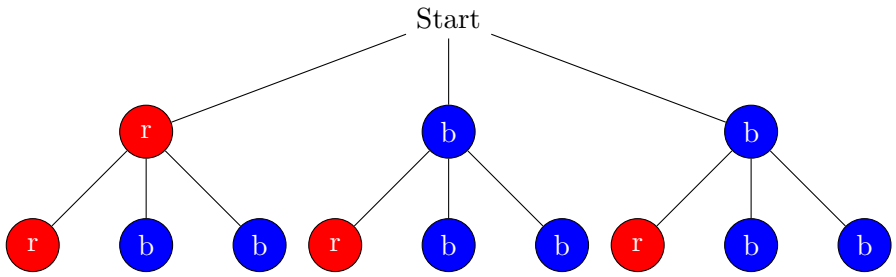

Abbildung 3.2: Baumdiagramm für zweimaliges Ziehen mit Zurücklegen aus einer Urne mit einer roten und zwei blauen Kugeln

Die Wahrscheinlichkeit, mindestens einmal eine rote Kugel zu ziehen, ergibt sich mithilfe der *Additionsregel* als Summe der Wahrscheinlichkeiten der paarweise disjunkten Ereignisse „rot - rot“, „rot - blau“ und „blau - rot“ und beträgt

$$P(\text{mind. einmal rot}) = \frac{1}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{2}{3} + \frac{2}{3}\cdot\frac{1}{3} = \frac{5}{9}.$$

3.3.1 Permutationen

Unter einer Permutation versteht man die Anordnung einer Menge von Objekten in einer bestimmten Reihenfolge. Im Folgenden werden die beiden grundlegenden Typen von Permutationen – ohne und mit Wiederholung – anhand zweier Beispiele veranschaulicht.

Beispiel 3.3.2 (Permutation ohne Wiederholung)**.** Wir betrachten drei verschiedenfarbige Kugeln: eine rote (r), eine blaue (b) und eine grüne (g). Untersucht werden soll, auf wie viele verschiedene Arten diese Kugeln in einer Reihe angeordnet werden können.

Für die *erste Position* stehen alle drei Kugeln zur Verfügung, also $n = 3$ Besetzungsmöglichkeiten. Ist die erste Position belegt, so verbleiben für die *zweite Position* nur noch $n-1 = 2$ Möglichkeiten. Sobald die ersten beiden Positionen festgelegt sind, ist die Farbe der *dritten Kugel* eindeutig bestimmt; es verbleibt somit genau eine Möglichkeit, sie zu platzieren, da $n - 2 = 1$. Damit ergeben sich insgesamt

$$3 \cdot 2 \cdot 1 = 3! = 6$$

mögliche Anordnungen der drei verschiedenfarbigen Kugeln:

$$r\,b\,g, \quad r\,g\,b, \quad b\,r\,g, \quad b\,g\,r, \quad g\,r\,b, \quad g\,b\,r.$$

Da alle sechs Anordnungen gleich wahrscheinlich sind, beträgt die Wahrscheinlichkeit dafür, dass die Kugeln in der Reihenfolge rot, blau, grün gezogen werden,

$$P(r\,b\,g) = \frac{1}{6},$$

da es sich um einen günstigen Fall unter insgesamt sechs möglichen handelt.

Allgemein gilt: Ordnet man n verschiedene Objekte in einer bestimmten Reihenfolge an, so spricht man von einer **Permutation ohne Wiederholung**. Die Anzahl der möglichen Anordnungen beträgt

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1,$$

wobei per Konvention $0! := 1$ gesetzt wird [58]. Permutationen dieser Art bilden die Grundlage vieler kombinatorischer Überlegungen.

Beispiel 3.3.3 (Permutation mit Wiederholung)**.** Im nächsten Schritt betrachten wir den Fall, dass nicht alle Elemente voneinander unterscheidbar sind. Die n-elementige Menge lässt sich dabei in j Teilgruppen einteilen, die sich jeweils aus gleichartigen Objekten zusammensetzen.

Wir nehmen an, es liegen insgesamt 12 Kugeln vor, wovon drei rot, vier blau und fünf grün sind. Gesucht ist erneut die Anzahl der unterschiedlichen Reihenfolgen, in denen diese Kugeln angeordnet werden können.

Wären alle Kugeln verschiedenfarbig und somit unterscheidbar, ergäben sich 12! mögliche Anordnungen. Da nun aber mehrere Kugeln gleichfarbig sind, entstehen durch das Vertauschen gleichartiger Kugeln keine neuen Anordnungen. Diese redundanten Vertauschungen müssen daher aus der Gesamtzahl herausgekürzt werden. Für die drei roten, vier blauen und fünf grünen Kugeln ergibt sich somit

$$3!, \quad 4!, \quad 5!$$

als Anzahl der nicht unterscheidbaren Vertauschungen.

Die Gesamtzahl der verschiedenen Anordnungen beträgt daher

$$\frac{12!}{3! \cdot 4! \cdot 5!} = 27\,720.$$

Da in dieser Formel einzelne Elemente mehrfach vorkommen, spricht man von einer **Permutation mit Wiederholung**.

Nach Bosch [14] gilt allgemein: Sind $n = k_1 + k_2 + \ldots + k_j$ Elemente in j Gruppen mit jeweils $k_1, k_2, \ldots, k_j$ gleichartigen Objekten unterteilt, so ergibt sich die Anzahl der möglichen Anordnungen zu

$$\frac{n!}{k_1!\, k_2!\, \ldots\, k_j!}.$$

3.3.2 Stichprobenmodelle

Die Betrachtung von Permutationen hat gezeigt, wie sich die Anzahl aller möglichen Anordnungen einer Menge von n unterscheidbaren Objekten bestimmen lässt. In vielen praktischen Situationen interessiert man sich jedoch nicht für die vollständige Anordnung

aller Elemente, sondern lediglich für die Auswahl von $k < n$ Elementen. Solche Auswahlvorgänge werden in der Kombinatorik als *Stichproben* bezeichnet. Abhängig davon, ob die Reihenfolge der Auswahl berücksichtigt wird und ob die Elemente nach jedem Zug wieder zurückgelegt werden, unterscheidet man verschiedene Arten von Stichproben [43].

Geordnete Stichprobe ohne Zurücklegen: Betrachten wir zunächst den Fall, dass gezogene Elemente nicht wieder in die Ausgangsmenge zurückgelegt werden. Für die erste Position stehen dann n Möglichkeiten zur Verfügung, für die zweite noch $n-1$, für die dritte $n-2$ und so weiter, bis die k-te Position mit $(n-k+1)$ Möglichkeiten besetzt werden kann. Diese Überlegung entspricht einer verkürzten Permutation, da nur die ersten k Plätze der Gesamtheit besetzt werden [48]. Die Anzahl der möglichen Stichproben beträgt somit:

$$n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

Beispiel 3.3.4 (Urnenmodell: geordnet, ohne Zurücklegen)**.** Eine Urne enthält 5 verschiedenfarbige Kugeln. Es sollen 3 Kugeln nacheinander gezogen werden, ohne dass eine Kugel zurückgelegt wird. Für die erste Ziehung stehen 5, für die zweite 4 und für die dritte 3 Möglichkeiten zur Verfügung.

$$\text{Anzahl der möglichen Ziehungen} = 5 \cdot 4 \cdot 3 = 60.$$

Da die Reihenfolge der gezogenen Kugeln berücksichtigt wird, handelt es sich um eine *geordnete Stichprobe ohne Zurücklegen.*

Geordnete Stichprobe mit Zurücklegen: Werden gezogene Elemente nach jeder Ziehung wieder in die Urne zurückgelegt, so steht bei jedem Zug erneut die gesamte Menge von n Elementen zur Verfügung. Es gibt also für jede der k Positionen n Auswahlmöglichkeiten, was zu

$$\underbrace{n \cdot n \cdot \ldots \cdot n}_{k \text{ Positionen}} = n^k$$

möglichen geordneten Stichproben führt [48].

Beispiel 3.3.5 (Urnenmodell: geordnet, mit Zurücklegen)**.** Eine Urne enthält 3 verschiedenfarbige Kugeln. Es sollen 2 Kugeln nacheinander gezogen werden, wobei jede Kugel nach dem Ziehen wieder zurückgelegt wird. Somit stehen bei jeder Ziehung 3 Möglichkeiten zur Verfügung.

$$\text{Anzahl der möglichen Ziehungen} = 3^2 = 9.$$

Da die Reihenfolge beachtet wird und jedes Element mehrfach auftreten kann, liegt eine *geordnete Stichprobe mit Zurücklegen* vor.

In manchen Fällen spielt nicht die Reihenfolge der Auswahl, sondern lediglich die Zusammensetzung der gezogenen Elemente eine Rolle. Man spricht dann von *ungeordneten Stichproben*, auch *Kombinationen* genannt. Wie zuvor unterscheidet man dabei Ziehungen *mit* und *ohne Zurücklegen*.

Ungeordnete Stichproben ohne Zurücklegen: Werden aus einer Menge von n Elementen k ausgewählt, ohne dass die Reihenfolge berücksichtigt wird, so spricht man von einer *Kombination ohne Zurücklegen*. Die entsprechende Zählformel lässt sich in Anlehnung an Kosfeld et al. [48] unmittelbar aus den Permutationen herleiten: Zunächst betrachtet man alle möglichen Anordnungen der k gezogenen Elemente – also eine geordnete Stichprobe ohne Zurücklegen. Da jedoch jede Kombination auf $k!$ verschiedene Arten angeordnet werden kann, ohne dass sich dabei die Zusammensetzung ändert, muss diese Zahl herausdividiert werden. Damit ergibt sich die Anzahl der verschiedenen Kombinationen zu:

$$\frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

Der Ausdruck $\binom{n}{k}$ wird als *Binomialkoeffizient* bezeichnet und gibt an, auf wie viele Arten k Elemente aus einer n-elementigen Menge ausgewählt werden können, wenn die Reihenfolge unbeachtet bleibt. Da nach der Definition $0! = 1$ gilt, ergibt sich insbesondere $\binom{n}{0} = 1$. Für alle $k < 0$ oder $k > n$ wird der Binomialkoeffizient per Konvention als $\binom{n}{k} = 0$ gesetzt [58].

Anschaulich betrachtet teilt man die Menge der n Objekte in zwei Gruppen auf: eine Gruppe der ausgewählten (k) und eine Gruppe der nicht ausgewählten $(n-k)$ Elemente. Die Formel für den Binomialkoeffizienten lässt sich damit auch als Spezialfall einer *Permutation mit Wiederholung* verstehen, bei der die Menge in genau zwei Teilmengen gleichartiger Objekte unterteilt ist [14].

Beispiel 3.3.6 (Urnenmodell: ungeordnet, ohne Zurücklegen)**.** Aus einer Urne mit 5 verschiedenfarbigen Kugeln sollen 3 Kugeln gezogen werden, wobei die Reihenfolge keine Rolle spielt.

$$\text{Anzahl der möglichen Auswahlen} = \binom{5}{3} = 10.$$

Ungeordnete Stichproben mit Zurücklegen: Werden gezogene Elemente nach jeder Ziehung wieder in die Urne zurückgelegt und spielt die Reihenfolge keine Rolle, ergibt sich eine etwas komplexere Situation. In diesem Fall kann jedes Element mehrfach gewählt werden, sodass die Auswahl einer k-elementigen Stichprobe einer Verteilung von k gleichartigen Objekten auf n verschiedene Kategorien entspricht. Cramer und Kamps [17] und Kosfeld et al. [48] zeigen, dass dies zum folgenden kombinatorischen Ausdruck führt:

$$\binom{n+k-1}{k}.$$

Beispiel 3.3.7 (Urnenmodell: ungeordnet, mit Zurücklegen)**.** Aus einer Urne mit 3 verschiedenfarbigen Kugeln sollen 2 Kugeln gezogen werden, wobei Kugeln mehrfach gezogen werden können und die Reihenfolge keine Rolle spielt.

$$\text{Anzahl der möglichen Auswahlen} = \binom{3+2-1}{2} = \binom{4}{2} = 6.$$

3.4 Konzept der Zufallsvariable

Nach der Einführung in die Grundlagen des Wahrscheinlichkeitsraums, der Laplace'schen Wahrscheinlichkeit und der Kombinatorik wird nun der Begriff der *Zufallsvariablen* eingeführt. Dieser stellt eine zentrale Verbindung zwischen den theoretischen Grundlagen der Wahrscheinlichkeitstheorie und der quantitativen Beschreibung zufälliger Phänomene her.

In vielen Zufallsexperimenten liegen die Ergebnisse bereits in numerischer Form vor, etwa beim Werfen eines Würfels, bei der Bestimmung der Anzahl defekter Produkte in einer Stichprobe oder der erzielten Tore in einem Fußballspiel [87]. In anderen Fällen hingegen treten als Versuchsausgänge zunächst nichtnumerische Ergebnisse auf, beispielsweise beim Münzwurf mit den Ausgängen „Kopf“ oder „Zahl“. Auch beim gleichzeitigen Werfen zweier Würfel ergibt sich zunächst eine Menge möglicher Zahlenpaare

$$\Omega = \{(x, y) \mid x, y \in \{1, 2, 3, 4, 5, 6\}\},$$

die 36 mögliche Ergebnisse umfasst. Häufig interessiert man sich in solchen Fällen nicht für das konkrete Ergebnis selbst, sondern für eine daraus abgeleitete Größe – etwa die Summe oder das Produkt der beiden geworfenen Augenzahlen [53].

Für die Berechnung von Wahrscheinlichkeiten spielt die konkrete Gestalt der Ergebnismenge in der Regel keine wesentliche Rolle. Entscheidend ist vielmehr die Art und Weise, wie den einzelnen Ereignissen Wahrscheinlichkeiten zugeordnet werden [53]. Diese Abstraktion erlaubt es, den Fokus von der Struktur der Ergebnisse auf die zugrunde liegende Wahrscheinlichkeitsverteilung zu lenken – ein Gedanke, der im Konzept der *Zufallsvariablen* formalisiert wird.

Zufallsvariablen dienen der Quantifizierung zufälliger Vorgänge und ermöglichen eine mathematisch präzise Beschreibung der Wahrscheinlichkeit, mit der bestimmte Werte auftreten. Intuitiv lassen sie sich als Variablen auffassen, deren Werte vom Zufall abhängen. Formal werden sie als Funktionen auf der Ergebnismenge eines Zufallsexperiments definiert, die jedem möglichen Ergebnis eine reelle Zahl zuordnen [14, 48].

3.4.1 Definition der Zufallsvariablen

Sei Ω die Menge aller möglichen Ergebnisse eines Zufallsexperiments. Eine Zufallsvariable X ist eine Abbildung, die jedem Ergebnis $\omega \in \Omega$ eine reelle Zahl $X(\omega) = x \in \mathbb{R}$

zuordnet. Damit entspricht jeder mögliche Versuchsausgang einem Zahlenwert, der das Ergebnis des Zufallsexperiments numerisch beschreibt. Da das tatsächliche Ergebnis ω zufallsabhängig ist, gilt dies ebenso für den Wert $X(\omega) = x$ [14, 35].

Definition 3.4.1 (Zufallsvariable). Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine Abbildung

$$X : \Omega \to \mathbb{R}$$

heißt *Zufallsvariable*, wenn für jede Borelmenge $B \subseteq \mathbb{R}$ das Urbild

$$X^{-1}(B) = \{\, \omega \in \Omega \mid X(\omega) \in B \,\}$$

ein Element der σ-Algebra $\mathcal{A}$ ist, d. h. $X^{-1}(B) \in \mathcal{A}$ für alle $B \in \mathcal{B}(\mathbb{R})$ [58, S. 98].

Die Menge aller möglichen Werte einer Zufallsvariablen X wird als *Wertemenge* W bezeichnet. Eine Zahl x gehört genau dann zu W, wenn ein $\omega \in \Omega$ existiert, sodass $X(\omega) = x$ gilt. Zufallsvariablen werden üblicherweise mit Großbuchstaben X, Y, Z bezeichnet, ihre Realisierungen mit den entsprechenden Kleinbuchstaben x, y, z [14, 48].

3.4.2 Rechenregeln für Zufallsvariablen

Wie Henze [35, S. 33] zeigt, sind für Zufallsvariablen X und Y auf dem Grundraum Ω auch deren Summe, Differenz und Produkt wieder Zufallsvariablen, definiert durch:

$$\begin{aligned}(X + Y)(\omega) &:= X(\omega) + Y(\omega),\\ (X - Y)(\omega) &:= X(\omega) - Y(\omega),\\ (XY)(\omega) &:= X(\omega) \cdot Y(\omega), \quad \omega \in \Omega.\end{aligned}$$

Ebenso ist für jedes $a \in \mathbb{R}$ auch das a-Fache einer Zufallsvariablen wieder eine Zufallsvariable:

$$(aX)(\omega) := a \cdot X(\omega), \quad \omega \in \Omega.$$

3.4.3 Verteilungsfunktion

Bei der Analyse von Zufallsvariablen ist es häufig zweckmäßig, kumulierte Wahrscheinlichkeiten zu betrachten, anstatt sich auf einzelne Realisationen zu beschränken. Zu diesem Zweck wird die *Verteilungsfunktion* eingeführt, mit deren Hilfe die Verteilung einer reellwertigen Zufallsvariable beschrieben werden kann. Sie ordnet jedem reellen Wert x die Wahrscheinlichkeit zu, dass die Zufallsvariable einen Wert annimmt, der kleiner oder gleich x ist. Auf diese Weise wird die gesamte Verteilung der Zufallsvariablen in einer einzigen Funktion zusammengefasst [48].

Definition 3.4.2 (Verteilungsfunktion). Sei X eine reellwertige Zufallsvariable auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Die Abbildung $F : \mathbb{R} \to [0, 1]$ mit

$$F(x) := P(X \leq x),$$

heißt *Verteilungsfunktion* von X [35, S. 380].

Satz 3.4.3. *Für jede Verteilungsfunktion F einer Zufallsvariablen X gelten nach Leiner [56] und Henze [35] die folgenden Eigenschaften:*

1. *F ist monoton wachsend, d. h. für alle $a \leq b$ gilt $F(a) \leq F(b)$.*
2. *F ist rechtsseitig stetig, d. h. für jedes $x \in \mathbb{R}$ gilt*
$$F(x) = \lim_{n\to\infty} F(x_n),$$
wobei (x_n) eine beliebige Folge mit $x_1 \geq x_2 \geq \ldots$ und $\lim_{n\to\infty} x_n = x$ ist.
3. $\lim\limits_{x\to-\infty} F(x) = 0$.
4. $\lim\limits_{x\to+\infty} F(x) = 1$.

Beweis. Die folgende Beweisstruktur ist angelehnt an Henze [35].

1. Sei $a \leq b$. Dann gilt aufgrund der Inklusion der Ereignisse
$$\{X \leq a\} \subseteq \{X \leq b\}.$$
Daraus folgt wegen der Monotonie des Wahrscheinlichkeitsmaßes unmittelbar
$$F(a) = P(X \leq a) \leq P(X \leq b) = F(b).$$

2. Sei $(x_n)_{n\in\mathbb{N}}$ eine Folge mit $x_1 \geq x_2 \geq \ldots$ und $\lim_{n\to\infty} x_n = x$. Damit bilden die Ereignisse $\{X \leq x_n\}$ eine absteigende Folge, und die Stetigkeit des Wahrscheinlichkeitsmaßes von oben liefert
$$F(x) = P(X \leq x) = P\left(\bigcap_{n=1}^{\infty} \{X \leq x_n\}\right) = \lim_{n\to\infty} P(X \leq x_n) = \lim_{n\to\infty} F(x_n).$$

3. Für $x \to -\infty$ bildet die Menge $\{X \leq x\}$ eine absteigende Familie von Ereignissen mit
$$\bigcap_{x\to-\infty} \{X \leq x\} = \varnothing.$$
Mit der Stetigkeit von oben folgt
$$\lim_{x\to-\infty} F(x) = \lim_{x\to-\infty} P(X \leq x) = P(\varnothing) = 0.$$

4. Für $x \to +\infty$ bildet $\{X \leq x\}$ eine aufsteigende Ereignisfamilie mit
$$\bigcup_{x\to+\infty} \{X \leq x\} = \Omega.$$
Die Stetigkeit von unten liefert damit
$$\lim_{x\to+\infty} F(x) = \lim_{x\to+\infty} P(X \leq x) = P(\Omega) = 1.$$

□

3.5 Diskrete Zufallsvariablen

Zufallsvariablen werden in der Wahrscheinlichkeitstheorie häufig als diskrete oder stetige Zufallsvariablen betrachtet, wenngleich auch Zufallsvariablen existieren, die keiner dieser beiden Klassen eindeutig zugeordnet werden können. Die Unterscheidung zwischen diskreten und stetigen Zufallsvariablen ist dennoch von zentraler Bedeutung für die Berechnung von Wahrscheinlichkeiten und die Beschreibung von Wahrscheinlichkeitsverteilungen.

Eine Zufallsvariable heißt *diskret*, wenn sie nur endlich oder abzählbar unendlich viele verschiedene Werte annehmen kann. Typischerweise treten diskrete Zufallsvariablen in Experimenten auf, bei denen die möglichen Ausgänge klar voneinander unterscheidbar und zählbar sind, etwa beim Werfen eines Würfels, beim Ziehen von Losen oder beim Zählen von Treffern in einer Stichprobe [16, 48].

Im Gegensatz dazu können *stetige* Zufallsvariablen überabzählbar viele Werte annehmen. In diesem Fall bilden die möglichen Werte ein Intervall der reellen Zahlen [16, 48].

Definition 3.5.1 (Diskrete Zufallsvariable). Eine Zufallsvariable X, deren Werte in einer endlichen oder abzählbar unendlichen Menge $W \subseteq \mathbb{R}$ liegen, wird als *diskret* bezeichnet. Die Gesamtheit der Zahlenpaare

$$(x_i, P(X = x_i)), \quad x_i \in W,$$

heißt *Wahrscheinlichkeitsfunktion* (bzw. Verteilung) der diskreten Zufallsvariablen X [14].

Satz 3.5.2. *Sei X eine diskrete Zufallsvariable mit Wertemenge $W = \{x_i \mid i \in I\}$, wobei I endlich oder abzählbar unendlich ist. Dann gilt nach Bosch [14]:*

$$\sum_{x_i \in W} P(X = x_i) = 1.$$

Beweis. In Anlehnung an Bosch [14] sei für jeden Wert $x_i \in W$

$$A_{x_i} = \{\omega \in \Omega \mid X(\omega) = x_i\}$$

das Ereignis, dass die Zufallsvariable X den Wert x_i annimmt. Da X für jedes $\omega \in \Omega$ genau einen Funktionswert besitzt, sind die Ereignisse A_{x_i} *paarweise disjunkt*, das heißt

$$A_{x_i} \cap A_{x_j} = \varnothing \quad \text{für alle } i \neq j.$$

Da X notwendigerweise einen der Werte aus W annimmt, gilt außerdem

$$\Omega = \bigcup_{x_i \in W} A_{x_i}.$$

Unter Berücksichtigung von Definition 3.1.6 (S. 34) zur σ-Additivität des Wahrscheinlichkeitsmaßes folgt:

$$1 = P(\Omega) = P\left(\bigcup_{x_i \in W} A_{x_i}\right) = \sum_{x_i \in W} P(A_{x_i}) = \sum_{x_i \in W} P(X = x_i).$$

□

Satz 3.5.3. *Sei X eine diskrete Zufallsvariable mit Wertemenge $W = \{x_i \mid i \in I\}$, wobei I endlich oder abzählbar unendlich ist. Nach Bosch [14] gilt für ein Intervall $S = [a, b] \subseteq \mathbb{R}$:*

$$P(a \leq X \leq b) = \sum_{a \leq x_i \leq b} P(X = x_i).$$

Beweis. Nach Bosch [14] gilt für diskrete Zufallsvariablen:

Die Ereignisse

$$A_{x_i} = \{\omega \in \Omega \mid X(\omega) = x_i\}, \quad x_i \in W,$$

sind paarweise disjunkt. Das Ereignis $\{a \leq X \leq b\}$ kann daher als disjunkte Vereinigung derjenigen A_{x_i} geschrieben werden, deren Werte im Intervall $[a, b]$ liegen:

$$\{a \leq X \leq b\} = \bigcup_{a \leq x_i \leq b} A_{x_i}.$$

Unter Verwendung der σ-Additivität des Wahrscheinlichkeitsmaßes (vgl. Definition 3.1.6, S. 34) ergibt sich unmittelbar:

$$P(a \leq X \leq b) = P\left(\bigcup_{a \leq x_i \leq b} A_{x_i}\right) = \sum_{a \leq x_i \leq b} P(A_{x_i}) = \sum_{a \leq x_i \leq b} P(X = x_i).$$

□

Beispiel 3.5.4. In einem Glücksspiel werden zwei faire Würfel gleichzeitig geworfen. Die Zufallsvariable X bezeichnet die Augensumme der beiden Würfel. Damit ergibt sich der Wertebereich von X mit

$$W = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

In Tabelle 3.2 (S. 51) sind die möglichen Kombinationen zweier Würfel für die jeweilige Augensumme dargestellt.

Da es sich um ein *Laplace-Experiment* handelt, sind alle 36 Grundereignisse gleich wahrscheinlich. Die Wahrscheinlichkeit eines Ereignisses ergibt sich somit als Quotient aus der Anzahl der für das Ereignis günstigen und der insgesamt möglichen Ergebnisse. Entsprechend gilt für die Wahrscheinlichkeitsverteilung von X:

$$P(X = x_i) = \frac{\text{Anzahl der Paare mit Summe } x_i}{36}.$$

Tabelle 3.2: Augensumme für zwei Würfel

Augensumme x_i	**mögliche Würfelpaare**
2	$(1,1)$
3	$(1,2),(2,1)$
4	$(1,3),(2,2),(3,1)$
5	$(1,4),(2,3),(3,2),(4,1)$
6	$(1,5),(2,4),(3,3),(4,2),(5,1)$
7	$(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$
8	$(2,6),(3,5),(4,4),(5,3),(6,2)$
9	$(3,6),(4,5),(5,4),(6,3)$
10	$(4,6),(5,5),(6,4)$
11	$(5,6),(6,5)$
12	$(6,6)$

Die Werte der Zufallsvariablen X lassen sich nicht nur in Form der nachfolgenden Tabelle 3.3, sondern auch durch das Säulendiagramm in Abbildung 3.3 (S. 52) veranschaulichen.

Tabelle 3.3: Wahrscheinlichkeitsverteilung der Augensumme zweier Würfel

x_i	2	3	4	5	6	7	8	9	10	11	12
$P(X=x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Das Glücksspiel gilt als gewonnen, wenn die Augensumme höchstens 4 beträgt. Die dazugehörige Wahrscheinlichkeit beträgt:

$$P(2 \leq X \leq 4) = P(X=2) + P(X=3) + P(X=4) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36} = \frac{1}{6}.$$

Schließlich zeigt sich, dass die Summe aller Einzelwahrscheinlichkeiten 1 ergibt:

$$\begin{aligned} P(X=2) + P(X=3) + \ldots + P(X=12) &= \\ &= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = 1. \end{aligned}$$

3.5.1 Wahrscheinlichkeitsverteilung

Für diskrete Zufallsvariablen lässt sich die Verteilungsfunktion besonders einfach darstellen, da die Zufallsvariable nur einzelne, abzählbare Werte annimmt. Der folgende Satz, der unmittelbar aus Satz 3.5.3 folgt, formuliert die Verteilungsfunktion für diesen diskreten Fall.

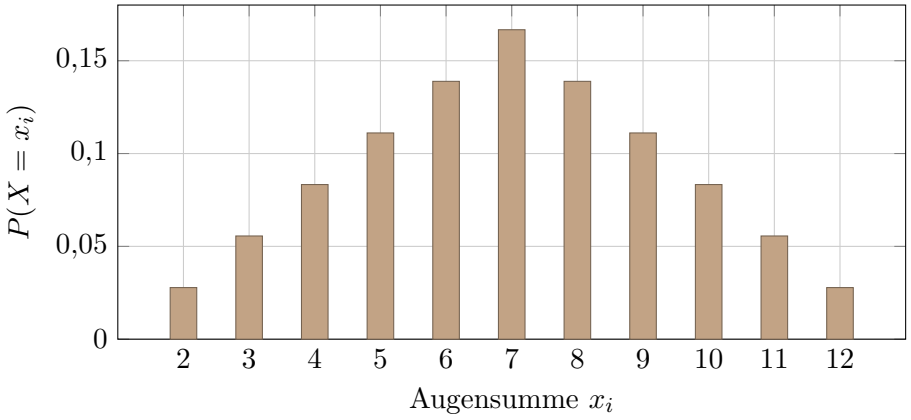

Abbildung 3.3: Wahrscheinlichkeitsverteilung der Augensumme zweier Würfel

Satz 3.5.5. *Sei X eine diskrete Zufallsvariable mit Wertemenge $W = \{x_i \mid i \in I\}$, wobei I endlich oder abzählbar unendlich ist. Dann gilt für alle $x \in \mathbb{R}$ [56]:*

$$F(x) = \sum_{x_i \leq x} P(X = x_i).$$

Für diskrete Zufallsvariablen besitzt die Verteilungsfunktion F die Gestalt einer *Treppenfunktion*: An den möglichen Werten x_i springt sie jeweils um den Betrag $P(X = x_i)$ nach oben und bleibt zwischen diesen Stellen konstant [35, 48].

Satz 3.5.6. *Für eine diskrete Zufallsvariable X und deren Verteilungsfunktion F gilt für $a < b$:*

1. $P(a < X \leq b) = F(b) - F(a)$,
2. $P(a \leq X \leq b) = F(b) - F(a^-)$,
3. $P(X > a) = 1 - F(a)$,

wobei $F(a^-) = \lim_{x \to a^-} F(x)$ den linksseitigen Grenzwert von F an der Stelle a bezeichnet. Dieser Grenzwert entspricht bei diskreten Zufallsvariablen dem Wert der Treppenstufe unmittelbar links von a [14, S. 60].

Beweis. Die folgende Beweisstruktur ist angelehnt an Bosch [14].

1. Das Ereignis $\{a < X \leq b\}$ ist eine disjunkte Zerlegung von $\{X \leq b\}$ abzüglich der Realisationen $X \leq a$:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

2. Das Ereignis $\{a \leq X \leq b\}$ umfasst zusätzlich die Realisation $X = a$. Deshalb gilt

$$P(a \leq X \leq b) = P(X = a) + P(a < X \leq b).$$

Da $P(X = a) = F(a) - F(a^-)$, erhält man

$$P(a \leq X \leq b) = [F(a) - F(a^-)] + [F(b) - F(a)] = F(b) - F(a^-).$$

3. Für das Komplement $\{X > a\}$ gilt

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a).$$

□

Beispiel 3.5.7. Für das oben beschriebene Würfelexperiment, bei dem die Zufallsvariable X die Augensumme zweier fairer Würfel bezeichnet, lässt sich die zugehörige Verteilungsfunktion sowohl in tabellarischer Form als auch grafisch mittels einer Treppenfunktion darstellen, wie im Folgenden gezeigt.

Tabelle 3.4: Werte der Verteilungsfunktion der Zufallsvariablen X

x_i	2	3	4	5	6	7	8	9	10	11	12
$F(x_i)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

Die Sprünge der Verteilungsfunktion treten an den Ausprägungen $x_i = 2, 3, \ldots, 12$ auf und entsprechen jeweils den Wahrscheinlichkeiten $P(X = x_i)$. Zudem lässt sich anhand der Abbildung 3.4 erkennen, dass die Verteilungsfunktion monoton wachsend ist und für $x \to +\infty$ gegen 1 konvergiert. Da bei $x = 12$ sämtliche Ausprägungen der Zufallsvariablen berücksichtigt sind, erreicht sie an dieser Stelle bereits ihren Grenzwert.

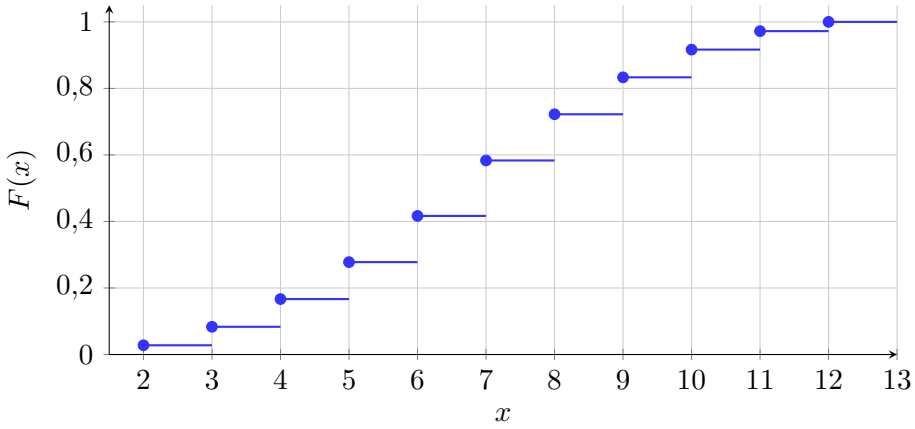

Abbildung 3.4: Graph der Verteilungsfunktion der Zufallsvariablen X

Beispielhaft lassen sich auf Basis der Verteilungsfunktion Intervallwahrscheinlichkeiten für die Augensumme X berechnen:

$$P(3 < X \leq 7) = F(7) - F(3) = \frac{21}{36} - \frac{3}{36} = \frac{18}{36} = \frac{1}{2}$$

$$P(4 \leq X \leq 6) = F(6) - F(3) = \frac{15}{36} - \frac{3}{36} = \frac{12}{36} = \frac{1}{3}$$
$$P(X > 9) = 1 - F(9) = 1 - \frac{30}{36} = \frac{6}{36} = \frac{1}{6}.$$

3.5.2 Maßzahlen diskreter Zufallsvariablen

Zur Beschreibung diskreter Zufallsvariablen werden häufig Kenngrößen herangezogen, die einerseits die zentrale Tendenz und andererseits die Streuung der Verteilung quantifizieren. Die wichtigsten Maßzahlen sind der *Erwartungswert* sowie die *Varianz* und die *Standardabweichung*.

Der Erwartungswert charakterisiert jene typische Ausprägung einer Zufallsvariablen, die sich bei sehr vielen Wiederholungen eines Zufallsexperiments als langfristiger Mittelwert herausbildet. Während er somit das zentrale Lagemaß der Verteilung darstellt, erfassen Varianz und Standardabweichung das Ausmaß der zufälligen Schwankungen um diesen Mittelwert und beschreiben damit die Streuung der möglichen Werte [48].

Definition 3.5.8 (**Erwartungswert**)**.** Sei X eine diskrete Zufallsvariable mit Wertemenge $W = \{x_i \mid i \in I\}$ und Wahrscheinlichkeiten $P(X = x_i)$, wobei $\sum_{i \in I} |x_i| \cdot P(X = x_i)$ konvergiert. Dann heißt

$$\mu = E(X) = \sum_{i \in I} x_i \cdot P(X = x_i)$$

der *Erwartungswert* von X [53, S. 240].

Satz 3.5.9 (**Eigenschaften des Erwartungswertes**)**.** *Seien X und Y diskrete Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und $a, b \in \mathbb{R}$. Dann gelten nach Henze [35, S. 105] und Tijms [87, S. 44]:*

1. *Linearität: $E(a \cdot X + b \cdot Y) = a \cdot E(X) + b \cdot E(Y)$.*
2. *Monotonie: aus $X \leq Y$ folgt $E(X) \leq E(Y)$.*

Beweis. Sei $W_X = \{x_i \mid i \in I\}$ die Wertemenge von X und $W_Y = \{y_j \mid j \in J\}$ die Wertemenge von Y.

1. Nach Definition des Erwartungswertes diskreter Zufallsvariablen gilt

$$E(aX + bY) = \sum_{i \in I, j \in J} (ax_i + by_j)\, P(X = x_i, Y = y_j),$$

wobei $P(X = x_i, Y = y_j)$ die gemeinsame Wahrscheinlichkeit von $X = x_i$ und $Y = y_j$ bezeichnet. Zunächst kann die Summe folgendermaßen aufgeteilt werden:

$$\sum_{i,j} (ax_i + by_j) P(X = x_i, Y = y_j) =$$

$$a\sum_{i,j} x_i\, P(X = x_i, Y = y_j) + b\sum_{i,j} y_j\, P(X = x_i, Y = y_j).$$

Mit den Randwahrscheinlichkeiten

$$\sum_j P(X = x_i, Y = y_j) = P(X = x_i), \qquad \sum_i P(X = x_i, Y = y_j) = P(Y = y_j)$$

folgt schließlich:

$$a\sum_{i,j} x_i P(X = x_i, Y = y_j) = a\sum_i x_i \underbrace{\sum_j P(X = x_i, Y = y_j)}_{=P(X=x_i)} = aE(X),$$

$$b\sum_{i,j} y_j P(X = x_i, Y = y_j) = b\sum_j y_j \underbrace{\sum_i P(X = x_i, Y = y_j)}_{=P(Y=y_j)} = bE(Y).$$

Also insgesamt:

$$E(aX + bY) = aE(X) + bE(Y).$$

2. Angenommen, $X(\omega) \leq Y(\omega)$ für alle $\omega \in \Omega$. Betrachten wir die Erwartungswerte von X und Y über die gemeinsamen Werte (x_i, y_j) und deren Wahrscheinlichkeiten:

$$E(X) = \sum_{i,j} x_i\, P(X = x_i, Y = y_j), \qquad E(Y) = \sum_{i,j} y_j\, P(X = x_i, Y = y_j).$$

Da $X(\omega) \leq Y(\omega)$ für alle ω gilt, folgt für jede Kombination (x_i, y_j) mit $P(X = x_i, Y = y_j) > 0$, dass $x_i \leq y_j$. Somit gilt

$$x_i\, P(X = x_i, Y = y_j) \leq y_j\, P(X = x_i, Y = y_j),$$

und Summation über alle Paare (i, j) liefert

$$E(X) = \sum_{i,j} x_i\, P(X = x_i, Y = y_j) \leq \sum_{i,j} y_j\, P(X = x_i, Y = y_j) = E(Y),$$

womit die Monotonie gezeigt ist [35].

□

Während der Erwartungswert die mittlere Ausprägung einer Zufallsvariablen kennzeichnet, liefert er allein noch keine Auskunft darüber, wie stark die tatsächlichen Werte um diesen Mittelpunkt schwanken können. Genau dieses Streuverhalten wird durch die Varianz beschrieben. Sie misst nämlich die mittlere quadratische Abweichung der Zufallsvariablen von ihrem Erwartungswert und gibt damit an, wie breit oder konzentriert die Verteilung im Durchschnitt liegt. Die Standardabweichung ergibt sich als Quadratwurzel der Varianz und quantifiziert, wie stark die möglichen Werte einer Zufallsvariablen im Durchschnitt vom Erwartungswert abweichen. Sie fasst damit das typische Ausmaß der Streuung in einer einzigen Kennzahl zusammen [48, 53].

Definition 3.5.10 (**Varianz und Standardabweichung**)**.** Sei X eine diskrete Zufallsvariable mit Wertemenge $W = \{x_i \mid i \in I\}$ und Wahrscheinlichkeiten $P(X = x_i)$, wobei $\sum_{i \in I} x_i^2 \cdot P(X = x_i)$ konvergiert. Unter dieser Voraussetzung besitzt X einen Erwartungswert $\mu = E(X)$.

Nach Bosch [14, S. 70] ist die *Varianz* von X definiert durch

$$\sigma^2 = V(X) = \sum_{i \in I} (x_i - \mu)^2 \cdot P(X = x_i),$$

und die *Standardabweichung* durch

$$\sigma(X) = \sqrt{V(X)}.$$

Satz 3.5.11. *Für die Varianz $V(X)$ einer diskreten Zufallsvariablen X gilt die folgende alternative Darstellung:*

$$V(X) = E(X^2) - (E(X))^2,$$

wobei $E(X^2) = \sum_{i \in I} x_i^2 \cdot P(X = x_i)$ ist [14, 80, 87].

Beweis. In Anlehnung an Bosch [14, S. 70] folgt aus der Definition der Varianz:

$$V(X) = \sum_{i \in I} (x_i - \mu)^2 \cdot P(X = x_i) = \sum_{i \in I} (x_i^2 - 2\mu x_i + \mu^2) \cdot P(X = x_i).$$

Weiter ergibt sich unter Anwendung der Linearität des Erwartungswertes:

$$V(X) = \sum_{i \in I} x_i^2 \cdot P(X = x_i) - 2\mu \cdot \sum_{i \in I} x_i \cdot P(X = x_i) + \mu^2 \cdot \sum_{i \in I} P(X = x_i).$$

Da $\sum_{i \in I} x_i \cdot P(X = x_i) = \mu$ und $\sum_{i \in I} P(X = x_i) = 1$ gilt, erhalten wir

$$V(X) = E(X^2) - 2 \cdot \mu^2 + \mu^2 = E(X^2) - \mu^2 = E(X^2) - (E(X))^2.$$

□

Definition 3.5.12 (**Stochastische Unabhängigkeit**)**.** Nach Bosch [14, S. 74] heißen zwei diskrete Zufallsvariablen X und Y *stochastisch unabhängig*, wenn für alle Wertepaare (x_i, y_j) gilt:

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j).$$

Satz 3.5.13. *Seien X und Y diskrete Zufallsvariablen, die stochastisch unabhängig sind und für die der Erwartungswert existiert. Dann gilt nach Bosch [14, S. 77]*

$$E(X \cdot Y) = E(X) \cdot E(Y).$$

Beweis. Nach Bosch [14] wird zunächst die Zufallsvariable $Z = X \cdot Y$ mit Wertemenge $\{z_k\}_{k \in K}$ betrachtet, wobei jeder Wert z_k als Produkt zweier Realisierungen x_i und y_j auftreten kann. Der Erwartungswert von Z lässt sich daher als

$$E(X \cdot Y) = E(Z) = \sum_{k \in K} z_k \cdot P(Z = z_k)$$

schreiben.

Jeder Wert z_k ergibt sich aus bestimmten Paaren (x_i, y_j), für die $x_i \cdot y_j = z_k$ gilt. Damit kann $P(Z = z_k)$ als Summe über alle diese Paare ausgedrückt werden:

$$P(Z = z_k) = \sum_{\substack{i,j \\ x_i \cdot y_j = z_k}} P(X = x_i, Y = y_j).$$

Setzt man diese Darstellung in den Erwartungswert ein und fasst alle Terme über die Paare (x_i, y_j) zusammen, so erhält man unter der Annahme der Unabhängigkeit von X und Y, also $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$, weiter

$$\begin{aligned}
E(X \cdot Y) &= \sum_{k \in K} z_k \cdot \sum_{\substack{i,j \\ x_i \cdot y_j = z_k}} P(X = x_i, Y = y_j) \\
&= \sum_{k \in K} \sum_{\substack{i,j \\ x_i \cdot y_j = z_k}} x_i \cdot y_j \cdot P(X = x_i) \cdot P(Y = y_j) \\
&= \sum_{i} \sum_{j} x_i \cdot y_j \cdot P(X = x_i) \cdot P(Y = y_j) \\
&= \left(\sum_{i} x_i \cdot P(X = x_i) \right) \cdot \left(\sum_{j} y_j \cdot P(Y = y_j) \right) = E(X) \cdot E(Y).
\end{aligned}$$

□

Satz 3.5.14. *Seien X und Y diskrete, stochastisch unabhängige Zufallsvariablen mit existierenden Varianzen. Dann gilt nach Kütting und Sauer [53, S. 255]*

$$V(X + Y) = V(X) + V(Y).$$

Beweis. Wir verwenden die alternative Darstellung der Varianz über den zweiten Moment:

$$V(X + Y) = E\big((X + Y)^2\big) - \big(E(X + Y)\big)^2.$$

Durch Ausmultiplizieren und unter Anwendung der Linearität des Erwartungswertes ergibt sich

$$E\big((X + Y)^2\big) = E(X^2) + 2E(X \cdot Y) + E(Y^2),$$

sowie

$$\big(E(X + Y)\big)^2 = \big(E(X) + E(Y)\big)^2 = E(X)^2 + 2 \cdot E(X) \cdot E(Y) + E(Y)^2.$$

Setzt man beides ein, ergibt sich

$$\begin{aligned} V(X+Y) &= (E(X^2) + 2 \cdot E(X \cdot Y) + E(Y^2)) - (E(X)^2 + 2 \cdot E(X) \cdot E(Y) + E(Y)^2) \\ &= (E(X^2) - E(X)^2) + (E(Y^2) - E(Y)^2) + 2 \cdot (E(X \cdot Y) - E(X) \cdot E(Y)) \\ &= V(X) + V(Y) + 2 \cdot (E(X \cdot Y) - E(X) \cdot E(Y)). \end{aligned}$$

Für stochastisch unabhängige Variablen gilt nach Satz 3.5.13 $E(X \cdot Y) = E(X) \cdot E(Y)$, sodass der letzte Term verschwindet und damit

$$V(X+Y) = V(X) + V(Y)$$

folgt [14]. □

Bemerkung 3.5.15. Der im Beweis auftretende Term $E(X \cdot Y) - E(X) \cdot E(Y)$ entspricht der *Kovarianz* von X und Y, die wie folgt definiert ist:

$$\mathrm{Cov}(X,Y) = E((X - E(X)) \cdot (Y - E(Y))) = E(X \cdot Y) - E(X) \cdot E(Y).$$

Für stochastisch unabhängige Variablen ist die Kovarianz gleich null, weshalb dieser Ausdruck im Verlauf des vorherigen Beweises wegfällt [53].

Beispiel 3.5.16. Betrachtet wird das Experiment, bei dem zwei faire Würfel geworfen werden. Sei X die Augensumme der beiden Würfel. Der Wertebereich lautet

$$W = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\},$$

mit Wahrscheinlichkeiten

$$P(X=2) = \frac{1}{36},\ P(X=3) = \frac{2}{36},\ \ldots,\ P(X=7) = \frac{6}{36},\ \ldots,\ P(X=12) = \frac{1}{36}.$$

Der Erwartungswert der Zufallsvariablen X berechnet sich zu

$$\begin{aligned} E(X) = \sum_{x_i \in W} x_i \cdot P(X = x_i) &= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} \\ &\quad + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7. \end{aligned}$$

Der Erwartungswert $E(X) = 7$ zeigt, dass die Augensumme bei sehr vielen Würfen im Mittel 7 beträgt. Dies entspricht dem typischen Zentrum der Verteilung: Werte um 7 treten häufiger auf als die extremen Ausprägungen 2 oder 12.

Für die Varianz und die Standardabweichung der Zufallsvariablen X ergibt sich

$$\begin{aligned} V(X) = \sum_{x_i \in W} (x_i - 7)^2 \cdot P(X = x_i) &= (2-7)^2 \cdot \frac{1}{36} + (3-7)^2 \cdot \frac{2}{36} \\ &\quad + (4-7)^2 \cdot \frac{3}{36} + \ldots + (11-7)^2 \cdot \frac{2}{36} + (12-7)^2 \cdot \frac{1}{36} = \frac{35}{6} \approx 5{,}8333 \end{aligned}$$

$$\sigma(X) = \sqrt{V(X)} \approx 2{,}4152.$$

Die Varianz $V(X)$ zeigt, wie stark die Augensummen typischerweise um den Mittelwert streuen. Die Standardabweichung $\sigma(X) \approx 2{,}4152$ gibt anschaulich an, dass die Augensummen meist etwa 2 bis 3 Einheiten vom Mittelwert 7 abweichen.

Nachdem im vorangegangenen Abschnitt grundlegende Eigenschaften diskreter Zufallsvariablen untersucht wurden, folgt nun die Betrachtung zweier Verteilungen, die in einer Vielzahl praktischer Anwendungen auftreten: der Binomial- und der Poisson-Verteilung. Beide beruhen auf wiederholten, voneinander unabhängigen Versuchsdurchführungen und bilden damit einen zentralen Baustein vieler stochastischer Modelle.

3.5.3 Binomialverteilung

Zu Beginn wird jenes probabilistische Grundmodell betrachtet, das den Ausgangspunkt für alle folgenden Überlegungen bildet: das *Bernoulli-Experiment*. Im Zentrum steht dabei ein Zufallsversuch mit genau zwei möglichen Ausgängen, etwa dem Eintreten oder Nicht-Eintreten eines bestimmten Ereignisses. Diese dichotome Struktur erlaubt eine präzise mathematische Beschreibung wiederholter, voneinander unabhängiger Zufallsversuche. Insbesondere die Binomial- und die Poisson-Verteilung lassen sich unmittelbar aus dieser elementaren Versuchsanordnung ableiten [16, 35].

Sei p die Wahrscheinlichkeit für das Eintreten eines Ereignisses A und entsprechend $1 - p$ die Wahrscheinlichkeit für das Komplementärereignis A'. Wiederholt man dieses Experiment n-mal unabhängig voneinander, entsteht eine *Bernoulli-Kette* der Länge n. Wir betrachten die Zufallsvariable X, die angibt, wie oft das Ereignis A bei diesen n Wiederholungen auftritt. Im Folgenden wird das Auftreten des Ereignisses A auch als *Erfolg* bezeichnet [48].

Zunächst analysieren wir die Wahrscheinlichkeit einer konkreten Ergebnisfolge, in der das Ereignis A genau k-mal und das Ereignis A' entsprechend $(n - k)$-mal auftritt. Aufgrund der Unabhängigkeit der einzelnen Versuche gilt nach Kosfeld et al. [48]

$$\underbrace{p \cdot p \cdot \ldots \cdot p}_{k \text{ Faktoren}} \cdot \underbrace{(1-p) \cdot (1-p) \cdot \ldots \cdot (1-p)}_{n-k \text{ Faktoren}} = p^k \cdot (1-p)^{n-k}.$$

Da das Ereignis A in beliebiger Reihenfolge innerhalb der n Wiederholungen auftreten kann, ist zusätzlich die Anzahl der möglichen Anordnungen zu berücksichtigen. Diese wird durch den Binomialkoeffizienten $\binom{n}{k}$ beschrieben. Insgesamt ergibt sich damit die Wahrscheinlichkeitsfunktion der Binomialverteilung als natürliche Verallgemeinerung des einmaligen Bernoulli-Experiments auf n unabhängige Wiederholungen [48].

Definition 3.5.17 (Binomialverteilung). Eine diskrete Zufallsvariable X heißt *binomialverteilt* mit Parametern $n \in \mathbb{N}$ und $p \in (0, 1)$, wenn ihre Wahrscheinlichkeitsfunktion durch

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \quad \text{für } k = 0, 1, \ldots, n$$

gegeben ist. In diesem Fall schreibt man kurz $X \sim \mathrm{B}(n,p)$ [53, S. 259].

Bemerkung 3.5.18. Da $p \in (0,1)$ gilt, folgt unmittelbar $P(X=k) \geq 0$ für alle k. Außerdem ergibt sich mit dem binomischen Lehrsatz $\sum_{k=0}^{n} P(X=k) = (p+(1-p))^n = 1$. Damit sind die beiden grundlegenden Anforderungen an ein Wahrscheinlichkeitsmaß erfüllt [43].

Satz 3.5.19. *Sei ein Zufallsexperiment gegeben, das n-mal unter identischen Bedingungen wiederholt wird, und sei A ein Ereignis mit $P(A) = p$, dessen Auftreten in jedem Durchgang unabhängig von den übrigen Durchgängen ist. Dann ist die Zufallsvariable X, die die Anzahl des Auftretens von A in den n Wiederholungen zählt, binomialverteilt mit den Parametern n und p, also $X \sim \mathrm{B}(n,p)$ [43].*

Beweis. Für $k \in \{0,1,\ldots,n\}$ ist $P(X=k)$ die Wahrscheinlichkeit dafür, dass das Ereignis A in genau k der n Wiederholungen eintritt.

Aufgrund der Unabhängigkeit der einzelnen Durchführungen besitzt jede konkrete Folge von n Versuchsausgängen, in der A genau k-mal und das Komplementärereignis A' genau $(n-k)$-mal auftritt, die Wahrscheinlichkeit

$$p^k (1-p)^{n-k}.$$

Die Anzahl derartiger Folgen entspricht der Anzahl der Möglichkeiten, k Erfolge auf n Positionen zu verteilen, und ist gleich dem Binomialkoeffizienten $\binom{n}{k}$. Da diese Ereignisse paarweise disjunkt sind, ergibt sich durch Addition ihrer Wahrscheinlichkeiten

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Damit besitzt X genau die Wahrscheinlichkeitsfunktion der Binomialverteilung mit den Parametern n und p, also $X \sim \mathrm{B}(n,p)$ [43]. □

Satz 3.5.20. *Nach Bosch [14, S. 87] lassen sich der Erwartungswert, die Varianz sowie die Standardabweichung einer binomialverteilten Zufallsvariable X mit den Parametern n und p wie folgt berechnen:*

$$(1)\quad E(X) = n \cdot p, \qquad (2)\quad V(X) = n \cdot p \cdot (1-p), \qquad (3)\quad \sigma(X) = \sqrt{n \cdot p \cdot (1-p)}.$$

Beweis. Die folgende Beweisstruktur ist angelehnt an Kütting und Sauer [53, S. 260–261].

(1) Für eine binomialverteilte Zufallsvariable $X \sim \mathrm{B}(n,p)$ gilt der Erwartungswert

$$E(X) = \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Da für $k = 0$ der Summand verschwindet, kann die Summe ab $k = 1$ geschrieben werden:

$$\begin{aligned} E(X) &= \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} \cdot p^k \cdot (1-p)^{n-k} \\ &= \sum_{k=1}^{n} n \cdot \frac{(n-1)!}{(k-1)!(n-k)!} \cdot p^k \cdot (1-p)^{n-k} \\ &= \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} \cdot p^k \cdot (1-p)^{n-k} \\ &= \sum_{k=1}^{n} n \cdot p \cdot \binom{n-1}{k-1} \cdot p^{k-1} \cdot (1-p)^{n-k} \\ &= n \cdot p \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} \cdot p^{k-1} \cdot (1-p)^{n-k}. \end{aligned}$$

Setze $m = n - 1$ und $j = k - 1$. Dann ergibt sich die Summe

$$\sum_{j=0}^{n-1=m} \binom{m}{j} \cdot p^j \cdot (1-p)^{m-j} = (p + (1-p))^m = 1,$$

nach dem binomischen Lehrsatz. Somit gilt

$$E(X) = n \cdot p.$$

(2) Zur Herleitung der Varianz verwenden wir den Varianzverschiebungssatz 3.5.11:

$$V(X) = E(X^2) - (E(X))^2.$$

Da bereits $E(X) = n \cdot p$ gezeigt wurde, bestimmen wir nun $E(X^2)$. Aus $k^2 = k \cdot (k-1) + k$ folgt

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \\ &= \sum_{k=0}^{n} k \cdot (k-1) \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} + \underbrace{\sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}}_{= E(X) = n \cdot p}. \end{aligned}$$

Für die erste Summe gilt die Identität

$$k \cdot (k-1) \cdot \binom{n}{k} = n \cdot (n-1) \cdot \binom{n-2}{k-2},$$

weshalb wir schreiben können:

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

$$= \sum_{k=2}^{n} k \cdot (k-1) \cdot \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k}$$

$$= n \cdot (n-1) \cdot p^2 \cdot \underbrace{\sum_{k=2}^{n} \binom{n-2}{k-2} \cdot p^{k-2} \cdot (1-p)^{n-k}}_{= (p+(1-p))^{n-2} = 1 \text{ nach binom. Lehrsatz}} = n \cdot (n-1) \cdot p^2.$$

Damit ergibt sich

$$E(X^2) = n \cdot (n-1) \cdot p^2 + n \cdot p = n^2 \cdot p^2 - n \cdot p^2 + n \cdot p.$$

Schließlich folgt für die Varianz:

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= (n^2 \cdot p^2 - n \cdot p^2 + n \cdot p) - (n \cdot p)^2 = n \cdot p - n \cdot p^2 = n \cdot p \cdot (1-p). \end{aligned}$$

(3) Aus der Definition der Standardabweichung als Quadratwurzel der Varianz folgt unmittelbar

$$\sigma(X) = \sqrt{V(X)} = \sqrt{n \cdot p \cdot (1-p)}.$$

□

Beispiel 3.5.21. Ein Multiple-Choice-Test besteht aus $n = 9$ Fragen. Zu jeder Frage gibt es genau eine richtige und zwei falsche Antwortmöglichkeiten. Eine Kandidatin bzw. ein Kandidat beantwortet jede Frage zufällig mit einer der drei Optionen. Damit ergibt sich die Wahrscheinlichkeit für eine richtige Antwort zu $p = \frac{1}{3}$. Der Test gilt als bestanden, wenn mindestens 5 der 9 Fragen richtig beantwortet werden. Die Zufallsvariable X bezeichne die Anzahl der richtig beantworteten Fragen.

a) Bestimme die Wahrscheinlichkeit, dass alle 9 Fragen richtig beantwortet sind.

Zunächst sei festgehalten, dass X binomialverteilt ist mit Parametern $n = 9$ und $p = \frac{1}{3}$, also

$$P(X = k) = \binom{9}{k} \cdot \left(\frac{1}{3}\right)^k \cdot \left(\frac{2}{3}\right)^{9-k}, \qquad k = 0, 1, \ldots, 9.$$

Für $k = 9$ erhält man

$$P(X = 9) = \binom{9}{9} \cdot \left(\frac{1}{3}\right)^9 \cdot \left(\frac{2}{3}\right)^0 \approx 0{,}000016935.$$

b) Bestimme die Wahrscheinlichkeit, dass die Person den Test besteht (d. h. $X \geq 5$) bzw. den Test nicht besteht.

$$P(X \geq 5) = \sum_{k=5}^{9} \binom{9}{k} \cdot \left(\frac{1}{3}\right)^k \cdot \left(\frac{2}{3}\right)^{9-k}$$

$$= \binom{9}{5} \cdot \left(\frac{1}{3}\right)^5 \cdot \left(\frac{2}{3}\right)^4 + \binom{9}{6} \cdot \left(\frac{1}{3}\right)^6 \cdot \left(\frac{2}{3}\right)^3 + \binom{9}{7} \cdot \left(\frac{1}{3}\right)^7 \cdot \left(\frac{2}{3}\right)^2 + \binom{9}{8} \cdot \left(\frac{1}{3}\right)^8 \cdot \left(\frac{2}{3}\right)^1 + \binom{9}{9} \cdot \left(\frac{1}{3}\right)^9 \cdot \left(\frac{2}{3}\right)^0 \approx 0{,}054253$$

Somit beträgt die Wahrscheinlichkeit, den Test zu bestehen, rund 5,43 %. Entsprechend folgt mit der Gegenwahrscheinlichkeit

$$P(X \leq 4) = 1 - P(X \geq 5) \approx 1 - 0{,}054253 = 0{,}945747.$$

c) Bestimme Erwartungswert, Varianz und Standardabweichung der Zufallsvariablen X für das zufällige Ankreuzen.

$$E(X) = n \cdot p = 9 \cdot \frac{1}{3} = 3$$

Im Mittel sind bei rein zufälligem Ankreuzen etwa 3 richtige Antworten zu erwarten – also weniger als die zum Bestehen notwendigen 5.

$$V(X) = n \cdot p \cdot (1-p) = 9 \cdot \frac{1}{3} \cdot \frac{2}{3} = 9 \cdot \frac{2}{9} = 2$$

$$\sigma = \sqrt{V(X)} = \sqrt{2} \approx 1{,}4142$$

Diese Kennzahlen beschreiben die Streuung um den Erwartungswert. Typischerweise liegen zufällig erzielte Punktzahlen etwa 1 bis 2 Punkte vom Mittelwert entfernt. Folglich ist es äußerst unwahrscheinlich, durch diese Vorgehensweise die für das Bestehen erforderlichen 5 oder mehr richtigen Antworten zu erreichen, was sich in der geringen Bestehenswahrscheinlichkeit von rund 5,43 % widerspiegelt.

3.5.4 Poisson-Verteilung

Die Poisson-Verteilung wird insbesondere zur Modellierung der Anzahl von Ereignissen verwendet, die innerhalb eines festgelegten Zeit- oder Raumintervalls auftreten. Sie eignet sich vor allem für Situationen, in denen viele potenzielle Ereignismöglichkeiten vorliegen, während die Wahrscheinlichkeit des Eintretens eines einzelnen Ereignisses pro Teilintervall gering ist. Typische Beispiele sind die Anzahl fehlerhafter Produkte innerhalb eines Tages [16], Verkehrsunfälle an einer Kreuzung innerhalb eines Monats oder eingehende Telefonanrufe während einer Stunde [48]. Die Poisson-Verteilung kann zudem als Grenzfall der Binomialverteilung aufgefasst werden, wenn die Erfolgswahrscheinlichkeit p sehr klein, gleichzeitig aber die Anzahl der Versuche n groß wird. Dieser Grenzübergang beschreibt Situationen mit vielen unabhängigen Versuchen bei gleichzeitig geringer Eintrittswahrscheinlichkeit pro Einzelversuch, in denen die Formeln der Binomialverteilung rechnerisch unhandlich werden. Ausgangspunkt ist dabei eine Folge binomialverteilter Zufallsvariablen mit Parameterpaaren (n, p_n), deren Erwartungswert konstant gehalten wird [16, 48].

Zu diesem Zweck wird eine Folge von Wahrscheinlichkeiten $p_n = \frac{\lambda}{n}$ betrachtet, wobei $\lambda > 0$ konstant ist. Dies gewährleistet, dass die durchschnittliche Anzahl an Erfolgen in den n Versuchen gleich λ beträgt. Die folgende Aussage zeigt, dass die zugehörigen Wahrscheinlichkeiten gegen eine explizit angegebene Grenzverteilung konvergieren [35].

Satz 3.5.22 (**Poisson-Approximation**). *Sei $X_n \sim \mathrm{B}(n, p_n)$ eine Folge binomialverteilter Zufallsvariablen mit $p_n = \frac{\lambda}{n}$. Dann gilt nach Leiner [56, S. 137] und Bosch [14, S. 93] für $\lambda > 0$ und $k \in \mathbb{N}_0$:*

$$\lim_{n\to\infty} \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

Beweis. In Anlehnung an Leiner [56, S. 137] und Bosch [14, S. 93] schreibt man zunächst die Wahrscheinlichkeitsfunktion der Binomialverteilung mit $p_n = \frac{\lambda}{n}$ um:

$$\begin{aligned} P(X_n = k) &= \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{1 \cdot 2 \cdot \ldots \cdot k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{n-k+1}{n} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

Für $n \to \infty$ gilt:

$$\lim_{n\to\infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{n-k+1}{n} = 1, \qquad \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n} = e^{-\lambda}, \qquad \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1.$$

Daraus folgt

$$\lim_{n\to\infty} \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} \cdot e^{-\lambda},$$

womit die Behauptung gezeigt ist. □

Die in diesem Grenzübergang auftretende Verteilung motiviert die folgende Definition.

Definition 3.5.23 (**Poisson-Verteilung**). Eine Zufallsvariable X heißt *Poisson-verteilt* mit Parameter $\lambda > 0$, wenn für alle $k \in \mathbb{N}_0$ gilt

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Man schreibt kurz $X \sim \mathrm{P}(\lambda)$ [87, S. 64].

Bemerkung 3.5.24. Für alle $k \in \mathbb{N}_0$ gilt $P(X = k) \geq 0$, und wegen $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ gilt $\sum_{k=0}^{\infty} P(X = k) = 1$. Damit sind die beiden grundlegenden Eigenschaften eines Wahrscheinlichkeitsmaßes erfüllt [43].

In der Theorie wird beim Einsatz der Poisson-Verteilung vorausgesetzt, dass ein Zeitintervall t in n gleich große Teilintervalle zerlegt werden kann, sodass in jedem dieser Intervalle höchstens ein Ereignis auftreten kann. Die Wahrscheinlichkeit dafür ist proportional zur Länge des Teilintervalls $\frac{t}{n}$ und über alle Teilintervalle hinweg konstant. Darüber hinaus wird angenommen, dass das Auftreten eines Ereignisses in einem Teilintervall unabhängig vom Auftreten in allen übrigen Teilintervallen ist [48].

Für die Approximation der Binomialverteilung durch die Poisson-Verteilung gilt nach Kosfeld et al. [48] folgende Faustregel: Die Approximation ist geeignet, wenn die Einzelereigniswahrscheinlichkeit $p \leq 0{,}1$ gilt und zugleich $n \cdot p \leq 5$ erfüllt ist. Demgegenüber weisen Hofbauer und Greschonig [43] darauf hin, dass eine Approximation der Binomialverteilung durch die Poisson-Verteilung bereits für $p < 0{,}05$ und $n > 10$ zulässig ist.

Die Maßzahlen dieser Verteilung weisen eine für die Poisson-Struktur charakteristische Eigenschaft auf: Erwartungswert und Varianz stimmen überein. Dies steht in engem Zusammenhang mit der Interpretation von λ als mittlere Ereignisrate [16].

Satz 3.5.25. *Für eine Poisson-verteilte Zufallsvariable $X \sim \mathrm{P}(\lambda)$ gilt nach Henze [35, S. 257]:*

$$(1) \quad E(X) = \lambda, \qquad (2) \quad V(X) = \lambda, \qquad (3) \quad \sigma(X) = \sqrt{\lambda}.$$

Beweis. Die folgende Beweisführung orientiert sich an Henze [35, S. 257–258] und Leiner [56, S. 138–139].

(1) Zunächst bestimmen wir den Erwartungswert:

$$E(X) = \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda} \cdot \lambda^k}{k!}.$$

Da der Summand für $k = 0$ verschwindet, folgt

$$E(X) = e^{-\lambda} \cdot \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{k!} = e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda \cdot \lambda^{k-1}}{(k-1)!}.$$

Mit der Substitution $j = k - 1$ ergibt sich

$$E(X) = \lambda \cdot e^{-\lambda} \cdot \underbrace{\sum_{j=0}^{\infty} \frac{\lambda^j}{j!}}_{= e^{\lambda}} = \lambda.$$

(2) Zur Herleitung der Varianz verwenden wir die alternative Darstellung aus Satz 3.5.11:

$$V(X) = E(X^2) - \big(E(X)\big)^2.$$

Da $(E(X))^2 = \lambda^2$ bereits aus (1) folgt, sei noch $E(X^2)$ zu bestimmen:

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} = \sum_{k=1}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \\ &= e^{-\lambda} \cdot \sum_{k=1}^{\infty} k \cdot \frac{\lambda \cdot \lambda^{k-1}}{(k-1)!} = \lambda \cdot e^{-\lambda} \cdot \sum_{k=1}^{\infty} k \cdot \frac{\lambda^{k-1}}{(k-1)!}. \end{aligned}$$

Mit der Substitution $j = k - 1$ folgt

$$\begin{aligned} E(X^2) &= \lambda \cdot e^{-\lambda} \cdot \sum_{j=0}^{\infty} (j+1) \cdot \frac{\lambda^j}{j!} = \lambda \cdot e^{-\lambda} \cdot \sum_{j=0}^{\infty} j \cdot \frac{\lambda^j}{j!} + \lambda \cdot e^{-\lambda} \cdot \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \\ &= \lambda \cdot \underbrace{\sum_{j=0}^{\infty} j \cdot \frac{\lambda^j}{j!} \cdot e^{-\lambda}}_{=\lambda} + \lambda \cdot e^{-\lambda} \cdot \underbrace{\sum_{j=0}^{\infty} \frac{\lambda^j}{j!}}_{=e^{\lambda}} = \lambda \cdot \lambda + \lambda \cdot e^{-\lambda} \cdot e^{\lambda} = \lambda^2 + \lambda. \end{aligned}$$

Damit ergibt sich

$$V(X) = E(X^2) - \big(E(X)\big)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

(3) Aus der Definition der Standardabweichung als Quadratwurzel der Varianz folgt unmittelbar

$$\sigma(X) = \sqrt{V(X)} = \sqrt{\lambda}.$$

□

Die Gleichheit von Erwartungswert und Varianz unterstreicht die zentrale Rolle des Parameters λ: Er bestimmt sowohl die mittlere Ereignisrate als auch die Streuung der betrachteten Zufallsvariablen [48]. Zur Veranschaulichung dieser Bedeutung zeigt Abbildung 3.5 zwei beispielhafte Säulendiagramme der Poisson-Verteilung mit den Parametern $\lambda_1 = 1$ (links) und $\lambda_2 = 2{,}5$ (rechts).

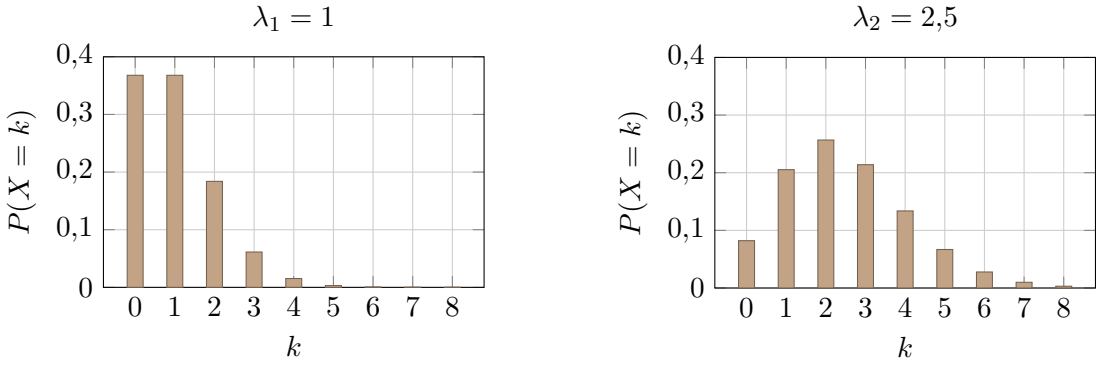

Abbildung 3.5: Säulendiagramme der Poisson-Verteilung für $\lambda_1 = 1$ und $\lambda_2 = 2{,}5$

Insbesondere für kleine Werte des Parameters λ weist die Verteilung eine ausgeprägte Rechtsschiefe auf, während sie mit steigender Ereignisrate zunehmend symmetrischer wird. Das Maximum der Wahrscheinlichkeitsfunktion liegt in unmittelbarer Nähe des Erwartungswertes λ [48].

Beispiel 3.5.26. Es soll untersucht werden, wie groß die Wahrscheinlichkeit ist, dass in einer Gruppe von $n = 730$ Personen mindestens zwei am Silvestertag (31. Dezember) Geburtstag haben. Die Wahrscheinlichkeit, dass eine zufällige Person an diesem Tag Geburtstag feiert, beträgt – unter Vernachlässigung von Schaltjahren –

$$p = \frac{1}{365} \approx 0{,}00274.$$

Die Anzahl der Personen mit Geburtstag am Silvestertag kann als Zufallsvariable X modelliert werden. Da es sich um ein Ereignis mit sehr kleiner Einzelwahrscheinlichkeit bei großer Personenzahl handelt, wird die Poisson-Verteilung als Approximation der Binomialverteilung verwendet. Der Parameter der Poisson-Verteilung ergibt sich zu

$$\lambda = n \cdot p = 730 \cdot \frac{1}{365} = 2.$$

Gesucht ist die Wahrscheinlichkeit, dass mindestens zwei Personen dieses Geburtsdatum teilen. Zweckmäßig wird dazu die komplementäre Wahrscheinlichkeit betrachtet:

$$P(X \geq 2) = 1 - P(X < 2) = 1 - \big(P(X = 0) + P(X = 1)\big).$$

Für eine Poisson-verteilte Zufallsvariable $X \sim \mathrm{P}(\lambda)$ gilt:

$$P(X = 0) = \frac{2^0}{0!} \cdot e^{-2} = e^{-2}, \qquad P(X = 1) = \frac{2^1}{1!} \cdot e^{-2} = 2 \cdot e^{-2}.$$

Damit ergibt sich

$$P(X \geq 2) = 1 - \big(e^{-2} + 2 \cdot e^{-2}\big) = 1 - 3e^{-2} \approx 1 - 0{,}4060 = 0{,}594.$$

Demnach beträgt die Wahrscheinlichkeit, dass in einer Gruppe von 730 Personen mindestens zwei am Silvestertag Geburtstag haben, etwa 59,4 %.

3.6 Stetige Zufallsvariablen

Im Folgenden wenden wir uns den kontinuierlichen Zufallsvariablen zu. Diese stellen ein analoges Modell zu diskreten Zufallsvariablen dar, bei dem abzählbare Wertebereiche durch Intervalle der reellen Zahlen ersetzt werden und Wahrscheinlichkeiten nicht mehr durch Summation, sondern durch Integration von sogenannten Wahrscheinlichkeitsdichtefunktionen bestimmt werden [43].

Stetige Zufallsvariablen werden zur Modellierung von Größen verwendet, die prinzipiell jeden reellen Wert innerhalb eines Intervalls annehmen können. Typische Anwendungsbereiche sind Niederschlagsmengen, Lebens- und Reaktionszeiten oder Zerfallsprozesse

radioaktiver Teilchen, die in der Praxis durch kontinuierliche Größen wie Längen-, Zeit- oder Gewichtseinheiten beschrieben werden, auch wenn reale Messungen stets nur mit endlicher Genauigkeit erfolgen [16, 87].

3.6.1 Wahrscheinlichkeitsverteilung

Definition 3.6.1 (**Stetige Zufallsvariable**)**.** Eine Zufallsvariable X heißt *stetig* bzw. *kontinuierlich*, wenn eine nichtnegative, integrierbare Funktion $f : \mathbb{R} \to [0, \infty)$ existiert mit

$$\int_{-\infty}^{\infty} f(x)\,dx = 1,$$

sodass die Verteilungsfunktion F von X für alle $t \in \mathbb{R}$ durch

$$F(t) = P(X \leq t) = \int_{-\infty}^{t} f(x)\,dx$$

gegeben ist. Die Funktion f wird in diesem Fall als *Dichtefunktion* von X bezeichnet [35, 43].

Die Dichtefunktion liefert dabei keine Wahrscheinlichkeiten einzelner Werte, sondern beschreibt, wie sich die Wahrscheinlichkeit entlang der reellen Achse verteilt. Der Funktionswert $f(x)$ gibt an, wie stark die Umgebung von x zur Gesamtwahrscheinlichkeit beiträgt [16].

Wie im diskreten Fall beschreibt die Verteilungsfunktion $F(t)$ auch bei stetigen Zufallsvariablen die Wahrscheinlichkeit, dass die Zufallsvariable X einen Wert kleiner oder gleich t annimmt. Geometrisch entspricht $F(t)$ dem bis zur Stelle $t \in \mathbb{R}$ akkumulierten Flächeninhalt unter der Dichte $f(x)$ [48].

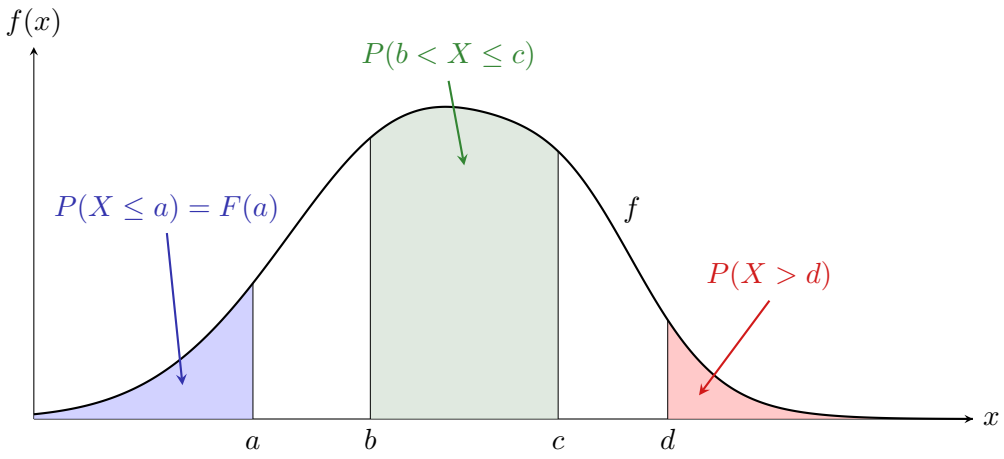

Abbildung 3.6: Wahrscheinlichkeitsdichte einer stetigen Zufallsvariablen

Ein wesentliches Merkmal stetiger Zufallsvariablen besteht darin, dass Punktwahrscheinlichkeiten stets verschwinden, d. h. $P(X = t) = 0$ für alle $t \in \mathbb{R}$ [48]. Folglich spielt es für die Berechnung von Intervallwahrscheinlichkeiten keine Rolle, ob die Grenzen ein- oder ausgeschlossen sind [87]:

$$P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

Damit lassen sich Wahrscheinlichkeiten für beliebige Intervalle direkt durch Integration der Dichtefunktion bestimmen, was im folgenden Satz präzisiert wird.

Satz 3.6.2. *Sei X eine stetige Zufallsvariable mit Dichtefunktion $f : \mathbb{R} \to [0, \infty)$ und zugehöriger Verteilungsfunktion $F : \mathbb{R} \to [0, 1]$. Dann gilt (vgl. [43, S. 29]) für beliebige $a, b \in \mathbb{R}$ mit $a \leq b$:*

1. *Für ein Intervall $I = (a, b]$ gilt*

$$P(X \in I) = P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x.$$

2. *Für jedes $b \in \mathbb{R}$ gilt*

$$P(X \leq b) = P(X < b) = F(b) = \int_{-\infty}^{b} f(x)\,dx.$$

3. *Für jedes $a \in \mathbb{R}$ gilt*

$$P(X \geq a) = P(X > a) = 1 - F(a) = \int_a^{\infty} f(x)\,dx.$$

Beweis. Die folgende Beweisstruktur orientiert sich an Hofbauer und Greschonig [43, S. 29] und Bosch [14, S. 102].

1. Zunächst zeigen wir, dass Punktwahrscheinlichkeiten verschwinden: Für jedes $t \in \mathbb{R}$ gilt $P(X = t) = 0$. Sei dazu ein beliebiges $\varepsilon > 0$. Dann ist $\{X = t\} \subseteq \{X \in (t - \varepsilon, t]\}$, woraus mit Satz 3.1.9 folgt:

$$P(X = t) \leq P(X \in (t - \varepsilon, t]) = \int_{t-\varepsilon}^{t} f(x)\,dx.$$

   Da f integrierbar ist, geht der Integralwert für $\varepsilon \to 0$ gegen null, und somit $P(X = t) = 0$. Für ein Intervall $I = (a, b]$ spielt es daher keine Rolle, ob die Endpunkte ein- oder ausgeschlossen werden.

   Betrachte die disjunkten Ereignisse $\{X \leq a\}$ und $\{a < X \leq b\}$. Dann gilt

$$\{X \leq a\} \cup \{a < X \leq b\} = \{X \leq b\},$$

   und mit der Additivität des Wahrscheinlichkeitsmaßes (vgl. Definition 3.1.6) folgt

$$P(X \leq b) = P(X \leq a) + P(a < X \leq b).$$

Weiter gilt nach Satz 3.4.3

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \geq 0.$$

Damit folgt für das Intervall $I = (a, b]$

$$\begin{aligned} P(X \in I) = P(X \in (a,b]) = P(a < X \leq b) &= F(b) - F(a) \\ &= \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = \int_{a}^{b} f(x)\,dx, \end{aligned}$$

was den ersten Punkt des Satzes zeigt.

2. Aus der Definition der Verteilungsfunktion folgt direkt

$$P(X \leq b) = F(b) = \int_{-\infty}^{b} f(x)\,dx.$$

Da $P(X = b) = 0$ gilt, ergibt sich $P(X < b) = P(X \leq b)$, womit der zweite Punkt gezeigt ist.

3. Hierzu wird das Gegenereignis betrachtet und Punkt 2 verwendet. Mit Satz 3.1.7 folgt dann

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a) = \int_{-\infty}^{\infty} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = \int_{a}^{\infty} f(x)\,dx.$$

Da $P(X = a) = 0$, gilt auch $P(X \geq a) = P(X > a)$, womit der dritte Punkt gezeigt ist.

□

3.6.2 Maßzahlen stetiger Zufallsvariablen

Die Definitionen der Begriffe Erwartungswert und Varianz lassen sich direkt vom diskreten auf den stetigen Fall übertragen, indem endliche oder abzählbare Summen durch Integrale ersetzt werden und an die Stelle der Punktwahrscheinlichkeiten eine Wahrscheinlichkeitsdichtefunktion tritt [43].

Definition 3.6.3 (Erwartungswert). Sei X eine stetige Zufallsvariable mit Wahrscheinlichkeitsdichtefunktion f, für die das uneigentliche Integral $\int_{-\infty}^{\infty} |x| \cdot f(x)\,\mathrm{d}x$ konvergiert. Dann heißt

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x$$

der *Erwartungswert* von X [14, S. 107].

Definition 3.6.4 (Varianz und Standardabweichung). Sei X eine stetige Zufallsvariable mit Wahrscheinlichkeitsdichtefunktion f, für die das uneigentliche Integral

$\int_{-\infty}^{\infty} x^2 \cdot f(x)\,\mathrm{d}x$ konvergiert. In diesem Fall ist der Erwartungswert von X wohldefiniert und wird mit $\mu = E(X)$ bezeichnet.

Nach Bosch [14, S. 107] ist die *Varianz* von X definiert durch

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,\mathrm{d}x,$$

und die *Standardabweichung* durch

$$\sigma(X) = \sqrt{V(X)}.$$

3.6.3 Normalverteilung

In vielen Bereichen der Natur- und Sozialwissenschaften treten Messwerte und beobachtbare Merkmale auf, die durch zahlreiche kleine, unabhängige Einflüsse variieren. Diese Schwankungen führen dazu, dass einzelne Messungen nicht exakt gleich ausfallen, sondern um einen mittleren Wert streuen. Um solche Streuungen systematisch zu beschreiben und zu analysieren, werden die entsprechenden Größen als stetige Zufallsvariablen betrachtet. Die Normalverteilung stellt ein besonders geeignetes Modell dar, da sie die typischen Eigenschaften solcher zufälligen Abweichungen mathematisch erfasst und daher häufig als Approximation für Messwerte oder Merkmalsausprägungen, wie etwa Füllmengen einer Abfüllanlage oder Abmessungen eines Werkstücks, verwendet wird [16].

Die Normalverteilung ist eine kontinuierliche Wahrscheinlichkeitsverteilung, die durch ihre charakteristische symmetrische, glockenförmige Dichtefunktion gekennzeichnet ist. Sie wird vollständig durch zwei Parameter beschrieben: den Erwartungswert μ, der den Lageparameter der Verteilung angibt, und die Standardabweichung σ, die die Breite bzw. Streuung der Verteilung bestimmt. Mathematisch führt dies zur folgenden Definition [16].

Definition 3.6.5 (Normalverteilung). Eine stetige Zufallsvariable X mit den Parametern $\mu \in \mathbb{R}$ und $\sigma > 0$ heißt normalverteilt, falls ihre Wahrscheinlichkeitsdichte die Form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

annimmt. Man schreibt kurz $X \sim \mathrm{N}(\mu, \sigma)$ [56, S. 155].

Die Dichtefunktion der Normalverteilung weist mehrere charakteristische Eigenschaften auf. Sie ist symmetrisch bezüglich der Achse $x = \mu$ und nimmt an dieser Stelle ihr Maximum an. Der Erwartungswert μ stellt somit nicht nur die Symmetrieachse der Verteilung dar, sondern entspricht zugleich dem Modalwert und dem Median [87]. Zu beiden Seiten von μ fällt die Dichtefunktion monoton ab und nähert sich der x-Achse asymptotisch an. Die Wendepunkte der Dichtefunktion liegen jeweils im Abstand einer Standardabweichung vom Erwartungswert, das heißt an den Stellen $x = \mu - \sigma$ und $x = \mu + \sigma$ [56].

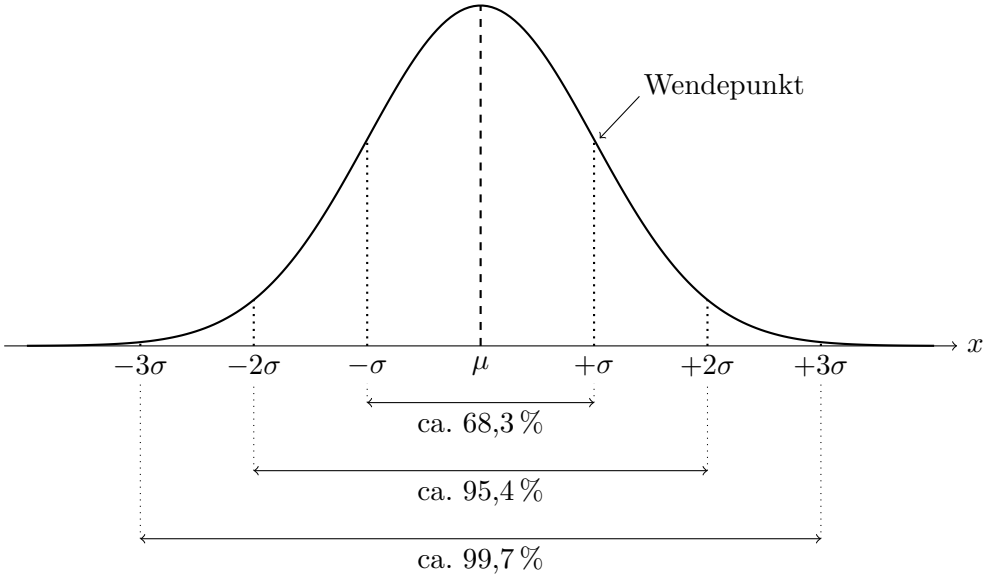

Abbildung 3.7: Schematische Darstellung der Dichtefunktion einer Normalverteilung (eigene Darstellung in Anlehnung an Cleff [16, S. 68])

Ein zentrales Merkmal der Normalverteilung ist die Konzentration der Wahrscheinlichkeitsmasse in der Umgebung des Erwartungswertes. Für eine normalverteilte Zufallsvariable mit Erwartungswert μ und Standardabweichung σ liegt etwa 68,3 % der Wahrscheinlichkeit im Intervall $[\mu - \sigma, \mu + \sigma]$, etwa 95,4 % im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$ und ungefähr 99,7 % im Intervall $[\mu - 3\sigma, \mu + 3\sigma]$ [16].

Eine besondere Stellung innerhalb der Klasse der Normalverteilungen nimmt die sogenannte *Standardnormalverteilung* mit Erwartungswert $\mu = 0$ und Standardabweichung $\sigma = 1$ ein. Sie dient als Referenzverteilung, auf die sich jede Normalverteilung eindeutig zurückführen lässt [16]. Ist $X \sim \mathrm{N}(\mu, \sigma)$ eine normalverteilte Zufallsvariable, so lässt sich durch die Transformation

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

eine neue Zufallsvariable Z definieren, die standardnormalverteilt ist, also $Z \sim \mathrm{N}(0, 1)$. Dieses Vorgehen wird als *Standardisierung* oder *z-Transformation* bezeichnet [48].

Der wesentliche Vorteil dieses Zusammenhangs besteht darin, dass Wahrscheinlichkeiten für beliebige normalverteilte Zufallsvariablen mithilfe der Verteilungsfunktion der Standardnormalverteilung bestimmt werden können. Für diese liegen tabellierte Werte vor, sodass auf die explizite Berechnung von Integralen verzichtet werden kann. Die Standardnormalverteilung stellt somit ein zentrales Hilfsmittel für die praktische Anwendung der Normalverteilung dar [16].

Definition 3.6.6 (Standardnormalverteilung). Eine stetige Zufallsvariable Z heißt *standardnormalverteilt*, wenn sie der Normalverteilung mit $\mu = 0$ und $\sigma = 1$ folgt. Die

zugehörige Wahrscheinlichkeitsdichte ist gegeben durch

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}}, \quad z \in \mathbb{R}.$$

Man schreibt kurz $Z \sim \mathrm{N}(0,1)$ [14, S. 133]. Die Verteilungsfunktion der Standardnormalverteilung ist definiert durch

$$\Phi(t) = P(Z \leq t) = \int_{-\infty}^{t} \varphi(z)\,\mathrm{d}z, \quad t \in \mathbb{R}.$$

Der Graph der Dichtefunktion φ einer standardnormalverteilten Zufallsvariablen wird durch die sogenannte *Gaußsche Glockenkurve* in Abbildung 3.8 dargestellt. Diese ist symmetrisch um $z = 0$ und besitzt Wendepunkte bei $z = -1$ und $z = 1$. Die Fläche unter der Kurve beträgt 1, wodurch die gesamte Wahrscheinlichkeit abgedeckt wird. Zur Berechnung konkreter Wahrscheinlichkeiten können Teilflächen unter der Kurve betrachtet werden, die über die Verteilungsfunktion Φ bestimmt werden [14].

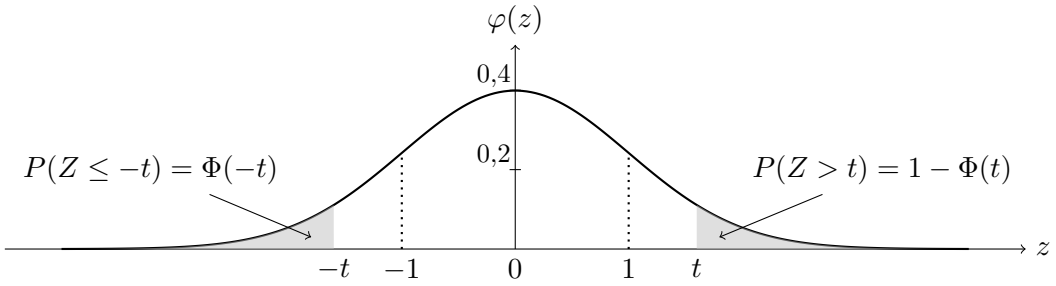

Abbildung 3.8: Dichtefunktion der Standardnormalverteilung

Aufgrund der Symmetrie um $z = 0$ ergeben sich die folgenden Eigenschaften der Standardnormalverteilung:

Satz 3.6.7. *Nach Hofbauer und Greschonig [43, S. 42] gilt:*

1. $\varphi(-z) = \varphi(z), \quad z \in \mathbb{R}.$
2. $\Phi(-t) = 1 - \Phi(t), \quad t \in \mathbb{R}.$

Beweis. Die folgende Beweisstruktur orientiert sich an Hofbauer und Greschonig [43, S. 42].

1. Die Dichtefunktion der Standardnormalverteilung ist definiert durch

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z \in \mathbb{R}.$$

Da $(-z)^2 = z^2$ gilt, folgt unmittelbar

$$\varphi(-z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(-z)^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \varphi(z).$$

2. Die Verteilungsfunktion Φ der Standardnormalverteilung ist definiert als

$$\Phi(t) = \int_{-\infty}^{t} \varphi(z)\, dz.$$

Unter Verwendung von Punkt 1 und der Substitution $u = -z$ ergibt sich

$$\begin{aligned}\Phi(-t) &= \int_{-\infty}^{-t} \varphi(z)\, dz = \int_{-\infty}^{-t} \varphi(-z)\, dz \\ &= \int_{+\infty}^{t} \varphi(u)\,(-du) = -\int_{+\infty}^{t} \varphi(u)\, du = \int_{t}^{+\infty} \varphi(u)\, du = 1 - \Phi(t),\end{aligned}$$

womit die Aussage gezeigt ist.

□

Aus der zweiten Eigenschaft folgt für symmetrische Intervallwahrscheinlichkeiten insbesondere

$$P(-t \leq Z \leq t) = 2 \cdot \Phi(t) - 1,$$

wodurch sich entsprechende Wahrscheinlichkeiten effizient berechnen lassen. Ferner genügt es aufgrund dieser Eigenschaft, Tabellen der Verteilungsfunktion Φ auf nichtnegative Argumente zu beschränken [14].

Beispiel 3.6.8. Der Durchmesser eines offiziellen Fußballs der Größe 5 unterliegt produktionstechnischen Schwankungen. Der zulässige Umfang liegt laut Regularien zwischen 68 cm und 70 cm, was mit $d = \frac{u}{\pi}$ einem Durchmesserintervall von etwa 21,7 cm bis 22,3 cm entspricht [21].

Dementsprechend wird der ideale Durchmesser eines Fußballs als mittlerer Wert dieses Intervalls angenommen, der zugleich dem Erwartungswert μ entspricht:

$$\mu = \frac{21{,}7 + 22{,}3}{2} = 22.$$

Der Durchmesser wird als stetige Zufallsvariable X modelliert und als normalverteilt angenommen:

$$X \sim \mathrm{N}(\mu, \sigma).$$

Zur Festlegung der Standardabweichung σ wird vereinfachend angenommen, dass etwa $99,7\,\%$ der produzierten Bälle einen Durchmesser innerhalb des Intervalls [21,7; 22,3] besitzen. Mit Bezug auf die zuvor erläuterten symmetrischen σ-Intervalle der Normalverteilung ergibt sich näherungsweise:

$$3\sigma = 0{,}3 \quad \Rightarrow \quad \sigma = 0{,}1.$$

Damit gilt für das Modell:

$$X \sim \mathrm{N}(22;\, 0{,}1).$$

a) Berechne die Wahrscheinlichkeit, dass der Durchmesser maximal 21,8 cm beträgt.

Zur Berechnung wird die Zufallsvariable X standardisiert:

$$Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$

Es gilt

$$P(X \leq 21{,}8) = P\left(Z \leq \frac{21{,}8 - 22}{0{,}1}\right) = P(Z \leq -2).$$

Mit der Symmetrie der Standardnormalverteilung folgt

$$P(Z \leq -2) = \Phi(-2) = 1 - \Phi(2) \approx 0{,}0228,$$

wobei $\Phi(2) \approx 0{,}9772$ aus Tabellen entnommen wurde.

b) Berechne die Wahrscheinlichkeit, dass der Durchmesser eines zufällig ausgewählten Balles zwischen 21,9 cm und 22,1 cm liegt.

Es gilt

$$P(21{,}9 \leq X \leq 22{,}1) = P\left(\frac{21{,}9 - 22}{0{,}1} \leq Z \leq \frac{22{,}1 - 22}{0{,}1}\right) = P(-1 \leq Z \leq 1)\,.$$

Mit der Verteilungsfunktion der Standardnormalverteilung ergibt sich

$$P(-1 \leq Z \leq 1) = \Phi(1) - \underbrace{\Phi(-1)}_{=1-\Phi(1)} = 2 \cdot \Phi(1) - 1 \approx 0{,}6826,$$

wobei $\Phi(1) \approx 0{,}8413$ aus Tabellen entnommen wurde.

Dieses Ergebnis bestätigt die bereits zuvor erläuterte Tatsache für Wahrscheinlichkeiten in symmetrischen Standardabweichungsintervallen (vgl. Abbildung 3.7 auf S. 72), wonach bei normalverteilten Zufallsvariablen etwa 68,3 % aller Beobachtungen innerhalb eines Intervalls von einer Standardabweichung um den Erwartungswert liegen.

3.6.4 Chi-Quadrat-Verteilung

Die Chi-Quadrat-Verteilung zählt zu den klassischen kontinuierlichen Verteilungen der mathematischen Statistik und tritt in vielfältigen Zusammenhängen auf, insbesondere in der Varianzanalyse, der Konstruktion von Konfidenzintervallen sowie in einer Reihe von Hypothesentests, darunter dem Chi-Quadrat-Anpassungstest, der im weiteren Verlauf dieser Arbeit genauer betrachtet wird. Von zentraler Bedeutung ist dabei ihre Interpretation als Verteilung einer Summe quadrierter standardnormalverteilter Zufallsvariablen, was den engen Bezug zur Normalverteilung verdeutlicht [16].

Definition 3.6.9 (**Chi-Quadrat-Verteilung**)**.** Seien $Z_1, \ldots, Z_n$ unabhängige standardnormalverteilte Zufallsvariablen mit $Z_i \sim \mathrm{N}(0,1)$ für $i = 1, \ldots, n$. Dann ist die Zufallsvariable

$$X = Z_1^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

Chi-Quadrat-verteilt mit n Freiheitsgraden. Man schreibt kurz $X \sim \chi_n^2$ [24, S. 307].

Die Dichte einer Chi-Quadrat-verteilten Zufallsvariablen ist durch folgenden Satz gegeben:

Satz 3.6.10. *Eine Chi-Quadrat-verteilte Zufallsvariable $X \sim \chi_n^2$ mit n Freiheitsgraden besitzt die Wahrscheinlichkeitsdichte*

$$f_n(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)}\, x^{\frac{n}{2}-1}\, e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

wobei $\Gamma(a) = \int_0^\infty e^{-y} \cdot y^{a-1}\, dy$ die sogenannte Gammafunktion *darstellt [14, S. 154].*

Ein Beweis dieses Resultats findet sich beispielsweise im Skriptum von Hofbauer [42].

Für spezielle Argumente a kann die Gammafunktion $\Gamma(a)$ explizit berechnet werden. Nach Bosch [14, S. 154] gilt:

- $\Gamma(a+1) = a \cdot \Gamma(a)$,
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ und $\Gamma(1) = 1$,
- $\Gamma(n) = (n-1)!$ für $n \in \mathbb{N}$.

Die obige Definition zeigt, dass die Freiheitsgrade n die Anzahl der beteiligten unabhängigen standardnormalverteilten Zufallsvariablen wiedergeben. Aus der Form der Dichte wird ersichtlich, dass die Chi-Quadrat-Verteilung ausschließlich auf dem Intervall $(0, \infty)$ liegt, was mit der Interpretation als Summe quadrierter Zufallsvariablen konsistent ist [56].

Die Charakteristik der Verteilung hängt wesentlich von der Anzahl der Freiheitsgrade ab. Für kleine Werte von n weist sie eine ausgeprägte Rechtsschiefe auf: Ein großer Teil der Wahrscheinlichkeitsmasse liegt in der Nähe des Nullpunkts, während ein langer rechter Verteilungsschwanz entsteht [48]. Mit wachsender Anzahl an Freiheitsgraden n steigen sowohl der Erwartungswert als auch die Varianz und die Standardabweichung:

$$(1) \quad E(X) = n, \qquad (2) \quad V(X) = 2n, \qquad (3) \quad \sigma(X) = \sqrt{2n}.$$

Gleichzeitig nimmt die Schiefe ab, wodurch die Chi-Quadrat-Verteilung für große n immer symmetrischer wird und sich dem Verlauf einer Normalverteilung annähert [16].

Darüber hinaus bildet sie die theoretische Grundlage zahlreicher statistischer Verfahren, die Abweichungen empirischer Werte von theoretischen Modellen auswerten. Für viele

Anwendungen sind dabei neben der Dichte insbesondere die Quantile von Bedeutung. Diese sind wiederum über die Verteilungsfunktion der Chi-Quadrat-Verteilung definiert und können analog zur Normalverteilung aus einer χ^2-Tabelle abgelesen werden [16].

4 Chi-Quadrat-Anpassungstest

Um die im weiteren Verlauf dieser Arbeit untersuchten Torhäufigkeiten im Fußball adäquat modellieren zu können, ist ein Verfahren erforderlich, das die Übereinstimmung zwischen empirischen Häufigkeiten und den durch eine theoretische Verteilung implizierten Erwartungswerten systematisch überprüft. Hierfür hat sich der Chi-Quadrat-Anpassungstest (kurz: χ^2-Anpassungstest), der auf der gleichnamigen Chi-Quadrat-Verteilung basiert, als geeignetes inferenzstatistisches Instrument etabliert. Mithilfe dieses Tests lässt sich beurteilen, ob eine gegebene Stichprobe mit einer vorab spezifizierten theoretischen Verteilung vereinbar ist und durch diese hinreichend genau beschrieben werden kann [24, 48].

Zunächst wird die Nullhypothese H_0 formuliert, wonach die Verteilung der beobachteten Daten mit der angenommenen Wahrscheinlichkeitsverteilung übereinstimmt. Die Alternativhypothese H_1 besagt demgegenüber, dass eine signifikante Abweichung vorliegt [48]. Da Stichprobendaten zufallsbedingten Schwankungen unterliegen, sind geringfügige Differenzen zwischen beobachteten und erwarteten Häufigkeiten unvermeidbar und statistisch zu tolerieren. Erst wenn diese Abweichungen ein bestimmtes Ausmaß überschreiten, wird H_0 verworfen. Die Testentscheidung erfolgt unter Kontrolle einer vorab festgelegten Irrtumswahrscheinlichkeit α, dem sogenannten Signifikanzniveau. Dieses gibt die Wahrscheinlichkeit an, eine tatsächlich zutreffende Nullhypothese fälschlicherweise zurückzuweisen, und stellt damit ein zentrales Maß für die statistische Zuverlässigkeit des Tests dar [24].

Die dem χ^2-Anpassungstest zugrunde liegenden Daten stammen aus einem sogenannten *Multinomial-Experiment*. Ein solches Experiment ist dadurch charakterisiert, dass ein Zufallsexperiment mit $k \geq 2$ möglichen Ausgängen $x_1, \ldots, x_k$ insgesamt n-mal unter identischen Bedingungen und unabhängig voneinander durchgeführt wird. Die möglichen Ausgänge werden auch als Zellen oder Klassen bezeichnet [28].

Jedem Ausgang x_j ist eine Wahrscheinlichkeit p_j zugeordnet, wobei $p_j \geq 0$ für $j = 1, \ldots, k$ und $\sum_{j=1}^{k} p_j = 1$ gilt. Bezeichnet N_j die absolute Häufigkeit des Auftretens der Ausprägung x_j in den n Wiederholungen, so ist der Zufallsvektor

$$(N_1, \ldots, N_k) \sim \text{Mult}(n; p_1, \ldots, p_k)$$

multinomialverteilt. Für die Erwartungswerte gilt insbesondere $E(N_j) = n \cdot p_j$ [28].

Zur Überprüfung der Hypothese wird das zugrunde liegende Zufallsexperiment n-mal unabhängig durchgeführt, wobei die absoluten Häufigkeiten $N_1, N_2, \ldots, N_k$ erfasst werden. Dabei gilt stets $N_1 + N_2 + \ldots + N_k = n$. Die Verwendung von Großbuchstaben

verdeutlicht, dass es sich hierbei um Zufallsvariablen handelt, deren konkrete Werte vom Zufall abhängen. Unter der Nullhypothese sei die Wahrscheinlichkeit für das Auftreten der Ausprägung x_j gleich p_j, wodurch die erwartete absolute Häufigkeit dann $n \cdot p_j$ beträgt. Je geringer die Abweichungen zwischen den beobachteten Häufigkeiten N_j und den erwarteten Häufigkeiten $n \cdot p_j$ ausfallen, desto eher spricht dies für die Gültigkeit der Nullhypothese [43]. Zur quantitativen Erfassung dieser Abweichungen wird die Teststatistik

$$D = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

verwendet, wobei k die Anzahl der Klassen bezeichnet. Große Werte von D deuten auf eine mangelnde Anpassung der postulierten Wahrscheinlichkeitsverteilung hin [87, 89].

Gegebenenfalls ist eine Zusammenfassung der Daten in disjunkte Klassen erforderlich. Dabei sollte die Klasseneinteilung so vorgenommen werden, dass die erwarteten Häufigkeiten keine zu kleinen Werte annehmen [56]. In der Praxis gelten bestimmte Mindestanforderungen an die erwarteten Zellbesetzungen als hinreichende Voraussetzung für eine zuverlässige Approximation der Teststatistik. Diese folgt asymptotisch einer χ^2-Verteilung, sofern $n \cdot p_j \geq 1$ für alle Klassen und $n \cdot p_j \geq 5$ in mindestens 80% der Zellen gilt [48, 87, 89]. Formal ergibt sich die asymptotische χ^2-Verteilung der Prüfgröße aus einem zentralen Grenzwertsatz für den multinomialverteilten Häufigkeitsvektor $(N_1, \ldots, N_k)$: Für hinreichend große Stichproben nähert sich die Verteilung von D der χ^2-Verteilung mit $k-1$ Freiheitsgraden, sodass $D \sim \chi^2_{k-1}$ gilt. Werden Parameter der zugrunde liegenden theoretischen Verteilung aus den Daten geschätzt, so reduziert sich die Anzahl der Freiheitsgrade um die Zahl r der geschätzten Parameter auf $k-1-r$ [87].

Zur Bestimmung des Verwerfungsbereichs sei $f_{k-1}(x)$ die Dichtefunktion der χ^2-Verteilung mit $k-1$ Freiheitsgraden. Für ein vorgegebenes Signifikanzniveau $\alpha \in (0,1)$ wird der kritische Wert $c_{1-\alpha;k-1}$ durch

$$P(\chi^2_{k-1} \leq c_{1-\alpha;k-1}) = \int_0^{c_{1-\alpha;k-1}} f_{k-1}(x)\,\mathrm{d}x = 1-\alpha$$

definiert. Der Wert $c_{1-\alpha;k-1}$ wird als $(1-\alpha)$-Quantil der χ^2-Verteilung mit $k-1$ Freiheitsgraden bezeichnet [28]. Äquivalent gilt

$$P(\chi^2_{k-1} > c_{1-\alpha;k-1}) = \int_{c_{1-\alpha;k-1}}^{\infty} f_{k-1}(x)\,\mathrm{d}x = \alpha,$$

woraus sich der Verwerfungsbereich

$$V = (c_{1-\alpha;k-1}, \infty)$$

ergibt [43].

Große Werte der Prüfgröße D sprechen gegen die Nullhypothese, da sie auf erhebliche Abweichungen zwischen beobachteten und erwarteten Häufigkeiten hinweisen. Zugleich

gilt unter Gültigkeit der Nullhypothese $P(D \in V) = \alpha$. Der Verwerfungsbereich ist somit so definiert, dass er diejenigen Werte enthält, die am stärksten gegen H_0 sprechen, während die Wahrscheinlichkeit einer fälschlichen Zurückweisung auf das vorgegebene Signifikanzniveau begrenzt bleibt [43].

Die Entscheidungsregel des χ^2-Anpassungstests lautet folglich: Die Nullhypothese wird auf dem Signifikanzniveau α verworfen, wenn der beobachtete Wert der Teststatistik D das $(1-\alpha)$-Quantil $c_{1-\alpha;k-1}$ der entsprechenden χ^2-Verteilung überschreitet. Andernfalls besteht auf Basis der vorliegenden Daten kein hinreichender Anlass, die angenommene theoretische Verteilung infrage zu stellen. Aus mathematischer Sicht bedeutet dies, dass im Falle der Verwerfung der Nullhypothese die Alternativhypothese mit einer Sicherheit von mindestens $1-\alpha$ als bewiesen gilt. Wird die Nullhypothese hingegen nicht verworfen, so sprechen die vorliegenden Daten im Rahmen des gewählten Signifikanzniveaus mit einer Sicherheit von mindestens $1-\alpha$ nicht gegen die Nullhypothese [48, 89].

Beispiel 4.0.1. Ein Teilnehmer an einem Glücksspiel vermutet, dass der verwendete Würfel nicht fair ist. In einer Datenbank sind die letzten $n = 1\,500$ Würfe dieses Würfels dokumentiert. Anhand dieser umfangreichen Stichprobe soll überprüft werden, ob die beobachteten Ergebnisse mit dem Modell eines idealen, fairen Würfels vereinbar sind.

Die aufgezeichneten absoluten Häufigkeiten der Augenzahlen seien:

Augenzahl j	1	2	3	4	5	6
N_j	226	263	248	257	239	267

Es gilt $\sum_{j=1}^{6} N_j = 1\,500$.

1. Signifikanzniveau und Nullhypothese: Es werde das Signifikanzniveau $\alpha = 0{,}05$ festgelegt. Die Nullhypothese lautet

$$H_0 : p_1 = p_2 = \ldots = p_6 = \frac{1}{6},$$

d. h. jede Augenzahl tritt mit gleicher Wahrscheinlichkeit auf. Die Alternativhypothese H_1 besagt, dass mindestens eine der Wahrscheinlichkeiten von $\frac{1}{6}$ abweicht.

Unter der Nullhypothese ist der Häufigkeitsvektor $(N_1, \ldots, N_6) \sim \text{Mult}(n; p_1, \ldots, p_6)$ multinomialverteilt. Insbesondere ist jede einzelne Zufallsvariable N_j binomialverteilt mit $N_j \sim \text{B}(n, p_j)$, woraus sich für den Erwartungswert $E(N_j) = n \cdot p_j$ ergibt.

2. Teststatistik und ihre Verteilung: Unter der Nullhypothese ergeben sich die theoretisch erwarteten Häufigkeiten zu

$$n \cdot p_j = 1\,500 \cdot \frac{1}{6} = 250 \quad \text{für alle } j = 1, \ldots, 6.$$

Die Teststatistik des χ^2-Anpassungstests lautet

$$D = \sum_{j=1}^{6} \frac{(N_j - np_j)^2}{np_j}.$$

Einsetzen der Werte liefert

$$\begin{aligned} D = {} & \frac{(226-250)^2}{250} + \frac{(263-250)^2}{250} + \frac{(248-250)^2}{250} \\ & + \frac{(257-250)^2}{250} + \frac{(239-250)^2}{250} + \frac{(267-250)^2}{250}. \end{aligned}$$

Damit ergibt sich

$$\begin{aligned} D &= \frac{576}{250} + \frac{169}{250} + \frac{4}{250} + \frac{49}{250} + \frac{121}{250} + \frac{289}{250} \\ &= 2{,}304 + 0{,}676 + 0{,}016 + 0{,}196 + 0{,}484 + 1{,}156 = 4{,}832. \end{aligned}$$

Da keine Parameter aus den Daten geschätzt werden, ist die Prüfgröße unter der Nullhypothese näherungsweise χ^2-verteilt mit $k - 1 = 6 - 1 = 5$ Freiheitsgraden.

3. Verwerfungsbereich: Für $\alpha = 0{,}05$ und $k = 6$ ergibt sich der kritische Wert als $(1 - \alpha)$-Quantil der χ^2-Verteilung mit $k - 1 = 5$ Freiheitsgraden zu

$$c_{1-\alpha;k-1} = c_{0{,}95;5} \approx 11{,}07.$$

Die entsprechenden Quantile können geeigneten Tabellen der χ^2-Verteilung entnommen werden.

Der Verwerfungsbereich lautet somit

$$V = (c_{1-\alpha;k-1}; \infty) = (11{,}07; \infty).$$

4. Testentscheidung: Da der beobachtete Wert der Teststatistik kleiner ist als der kritische Wert $c_{1-\alpha;k-1}$, gilt

$$D = 4{,}832 \notin V = (11{,}07; \infty).$$

Die Nullhypothese wird daher auf dem Signifikanzniveau von 5 % nicht verworfen. Die in der Datenbank dokumentierten 1 500 Würfe liefern keinen statistisch signifikanten Hinweis darauf, dass der im Glücksspiel verwendete Würfel von der Annahme eines fairen Würfels abweicht. Die beobachteten Abweichungen zwischen empirischen und theoretischen Häufigkeiten sind mit zufälligen Schwankungen vereinbar. Insbesondere sprechen die vorliegenden Daten mit einer Sicherheit von mindestens 0,95 nicht gegen die Nullhypothese eines fairen Würfels.

Der χ^2-Anpassungstest stellt ein zentrales inferenzstatistisches Verfahren zur Beurteilung der Übereinstimmung zwischen empirischen Beobachtungen und den durch ein theoretisches Wahrscheinlichkeitsmodell implizierten Erwartungswerten dar. Im Kontext der Analyse von Torhäufigkeiten im Fußball ermöglicht er insbesondere zu prüfen, inwieweit die tatsächlich beobachteten Spielergebnisse mit stochastischen Modellen vereinbar sind. Damit bildet der χ^2-Anpassungstest eine wesentliche methodische Grundlage für die empirische Untersuchung von Torverteilungen und schafft zugleich die Voraussetzung für die im folgenden Kapitel entwickelte wahrscheinlichkeitstheoretische Beschreibung von Spielergebnissen im Fußball.

5 Entwicklung eines Prognosemodells für Spielausgänge

Fußball gilt als die weltweit populärste Sportart, und die Frage nach dem Ausgang eines Spiels beschäftigt sowohl Fans als auch Forscher:innen. Die Prognose, welche Mannschaft die nächste Weltmeisterschaft oder Champions-League gewinnen wird, löst nicht nur hitzige Diskussionen unter Anhänger:innen aus, sondern zieht selbst Gelegenheitszuschauer:innen in den Bann. Buchmacher haben diesen Umstand zu einem Geschäft gemacht: Sie nutzen komplexe Modelle, die Faktoren wie aktuelle Form, Verletzungen, historische Begegnungen und die Bedeutung des Spiels berücksichtigen, um Quoten für Sieg, Niederlage und Unentschieden zu bestimmen. So haben sich in den letzten Jahrzehnten zahlreiche hochentwickelte Prognosemodelle etabliert, die aufgrund ihrer Komplexität oft nur schwer nachvollziehbar und zugänglich sind.

Ein bekanntes Zitat der Fußballlegende Johan Cruyff verdeutlicht die scheinbare Einfachheit und zugleich die hohe Komplexität des Spiels, die eine vertiefte analytische Betrachtung erforderlich macht: *„Fußballspielen ist sehr einfach, aber Fußball auf einfache Weise zu spielen, ist das Schwierigste, was es gibt.“* Ähnlich dem Eisbergmodell wirken die beobachtbaren Aktionen auf den ersten Blick leicht verständlich, während die zugrunde liegenden taktischen und strategischen Prozesse hochkomplex sind. Daher bleibt die zuverlässige Vorhersage des Spielausgangs eine Herausforderung, da das Geschehen von zahlreichen Einflussfaktoren bestimmt wird.

Vor diesem Hintergrund besteht die zentrale Aufgabe darin, Modelle zu entwickeln, die trotz vereinfachter Annahmen belastbare Vorhersagen ermöglichen. Zwar könnte zunächst angenommen werden, dass es schwierig ist, systematische Gesetzmäßigkeiten zur Beschreibung eines so komplexen Phänomens wie eines Fußballspiels zu identifizieren. Ein wesentlicher Schritt besteht jedoch in der Definition geeigneter Beobachtungsgrößen, um zentrale Eigenschaften zu erfassen. Ziel dieser Arbeit ist es, konsequent auf subjektive Einschätzungen, persönliche Präferenzen und emotionale Bewertungen zu verzichten und stattdessen datenbasierte Größen heranzuziehen. Dabei soll ein ausgewogenes Verhältnis zwischen Modellkomplexität und Aussagekraft erreicht werden, sodass bereits mit elementaren mathematischen Mitteln fundierte Vorhersagen möglich sind.

5.1 Erste Ansätze auf Basis der Binomialverteilung

5.1.1 Relative Torverhältnisse als Modellierungsgrundlage

Ein erster vereinfachter Ansatz basiert auf dem relativen Torverhältnis der beiden Teams, wobei jeder erzielte Treffer als Bernoulli-Experiment aufgefasst wird. In Anlehnung an das Schema von Tolan [88] wird untersucht, welche Auswirkungen die im Fußball vergleichsweise geringe Anzahl an Toren – wie in Abschnitt 2.4 dargestellt – insbesondere für das unterlegene Team hat.

Zur Veranschaulichung des Zusammenhangs zwischen der Torsumme eines Spiels und der Siegwahrscheinlichkeit des schwächeren Teams werden die erzielten Tore zweier Mannschaften ins Verhältnis gesetzt, um ihre relative Spielstärke zu approximieren. Dabei sei Team A das schwächere und Team B das stärkere Team. Die Spielstärke von Team A wird durch die Erfolgswahrscheinlichkeit $p \in (0,1)$ für einen Torerfolg beschrieben, während $q = 1 - p$ die entsprechende Gegenwahrscheinlichkeit für Team B angibt. Die Gesamtanzahl der Tore in einem Spiel sei n, die Anzahl der Tore von Team A sei k. Für ungerade Werte von n treten ausschließlich die Spielausgänge Sieg oder Niederlage auf, während bei geraden Werten zusätzlich ein Unentschieden möglich ist.

Unter der Annahme, dass jedes Tor als unabhängiges Bernoulli-Experiment aufgefasst werden kann, lässt sich die Toranzahl von Team A binomialverteilt modellieren. Die Wahrscheinlichkeitsfunktion der Zufallsvariable X, die die Anzahl der von Team A erzielten Tore beschreibt, ergibt sich somit aus der Binomialverteilung mit den Parametern n und p und besitzt die Form

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Anhand konkreter Daten lässt sich dieser Ansatz illustrieren: Nach der 28. Spielrunde der Saison 2025/26 erzielten der FC St. Pauli (Team A) 25 Tore, Eintracht Frankfurt (Team B) 52 Tore [20]. Setzt man diese Werte ins Verhältnis, ergibt sich mit $\frac{52}{25} = 2{,}08$, dass die Frankfurter etwa doppelt so viele Tore erzielt haben wie der Hamburger Traditionsverein. Im Rahmen des Modells wird dieses Verhältnis dahingehend interpretiert, dass die Erfolgswahrscheinlichkeit für einen Torerfolg von Team A $p = \frac{1}{3}$ beträgt, während für Team B entsprechend $q = 1 - p = \frac{2}{3}$ gilt.

Ungerade Toranzahl

Bei ungerader Toranzahl ergeben sich für ein Spiel mit einem Tor ($n = 1$) die Gewinnwahrscheinlichkeiten:

$$P(\text{Sieg Team A}) = \frac{1}{3} \approx 33\,\% \qquad \text{und} \qquad P(\text{Sieg Team B}) = \frac{2}{3} \approx 67\,\%.$$

Für $n = 3$ Tore gewinnt Team A, wenn es zwei oder drei Tore erzielt. Unter Verwendung des Binomialkoeffizienten ergibt sich:

$$P(\text{Sieg Team A}) = \binom{3}{3} \cdot \left(\frac{1}{3}\right)^3 + \binom{3}{2} \cdot \left(\frac{1}{3}\right)^2 \cdot \left(\frac{2}{3}\right) = \frac{7}{27} \approx 26\,\%.$$

Die Gegenwahrscheinlichkeit liefert dementsprechend für Team B:

$$P(\text{Sieg Team B}) = 1 - P(\text{Sieg Team A}) = 1 - \frac{7}{27} \approx 74\,\%.$$

Mit steigender Toranzahl sinkt die Gewinnwahrscheinlichkeit des schwächeren Teams, wie in Abbildung 5.1 dargestellt. Dieser Zusammenhang zeigt sich auch für unterschiedliche Spielstärken p. Je größer der Leistungsunterschied zwischen zwei Mannschaften ist, desto wahrscheinlicher ist es, dass sich das stärkere Team in torreichen Spielen durchsetzt. Unterlegene Teams werden bei hohen Trefferzahlen daher ihrer Underdog-Rolle häufiger gerecht und verlassen das Spielfeld als Verlierer. Dies liefert zugleich eine Erklärung für die oft geringe Risikobereitschaft und die defensive Spielweise schwächerer Mannschaften, da ein offener Schlagabtausch mit einem stärkeren Gegner in der Regel mit geringeren Erfolgsaussichten verbunden ist.

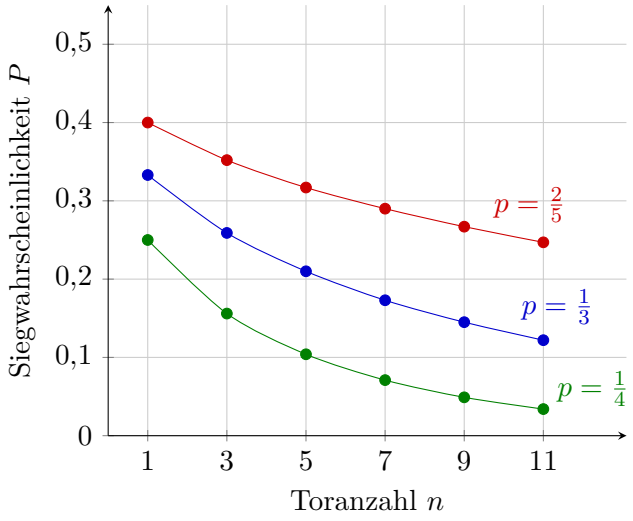

Abbildung 5.1: Siegwahrscheinlichkeit des schwächeren Teams bei *ungerader Toranzahl* (eigene Darstellung in Anlehnung an Tolan [88, S. 51])

Wie bereits bekannt ist, fallen in einem Fußballspiel im Durchschnitt etwa drei Tore. Unter dieser Annahme weist selbst ein deutlich unterlegenes Team mit einer Spielstärke von $p = \frac{1}{4}$ – wie es beispielsweise bei Begegnungen zwischen Mannschaften unterschiedlicher Ligen im Pokal vorkommen kann – noch eine Gewinnwahrscheinlichkeit von etwa 16 % auf. Vor diesem Hintergrund erscheinen sogenannte Überraschungssiege vielleicht gar nicht mehr so unerwartet, wie sie auf den ersten Blick wirken.

Gerade Toranzahl

Bei gerader Toranzahl kann zusätzlich die Wahrscheinlichkeit eines Unentschiedens berücksichtigt werden. Für $n = 2$ ergibt sich:

$$P(\text{Sieg Team A}) = \binom{2}{2} \cdot \left(\frac{1}{3}\right)^2 = \frac{1}{9} \approx 11\,\%,$$

$$P(\text{Unentschieden}) = \binom{2}{1} \cdot \left(\frac{1}{3}\right) \cdot \left(\frac{2}{3}\right) = \frac{4}{9} \approx 44\,\%,$$

$$P(\text{Sieg Team B}) = 1 - \frac{1}{9} - \frac{4}{9} = \frac{4}{9} \approx 44\,\%.$$

Für $n = 4$ gewinnt Team A bei $k = 3$ oder $k = 4$ Toren:

$$P(\text{Sieg Team A}) = \binom{4}{4} \cdot \left(\frac{1}{3}\right)^4 + \binom{4}{3} \cdot \left(\frac{1}{3}\right)^3 \cdot \left(\frac{2}{3}\right) = \frac{1}{9} \approx 11\,\%,$$

$$P(\text{Unentschieden}) = \binom{4}{2} \cdot \left(\frac{1}{3}\right)^2 \cdot \left(\frac{2}{3}\right)^2 = \frac{8}{27} \approx 30\,\%,$$

$$P(\text{Sieg Team B}) = 1 - \frac{1}{9} - \frac{8}{27} = \frac{16}{27} \approx 59\,\%.$$

Auf gleiche Weise lassen sich die Wahrscheinlichkeiten für verschiedene Torsummen n bestimmen. Da für $n = 0$ trivialerweise $P(\text{Unentschieden}) = 1$ gilt, wurde dieser Datenpunkt im zweiten Diagramm vernachlässigt. Die Abbildungen 5.2 (S. 89) und 5.3 (S. 90) veranschaulichen die Abhängigkeit von Sieg- und Unentschiedenwahrscheinlichkeiten von der Toranzahl, exemplarisch auch für andere Spielstärken p.

Fällt nach unserer Annahme in einem Spiel eine gerade Anzahl an Treffern, so kann zusätzlich ein Unentschieden auftreten. Während für niedrige Torsummen die Wahrscheinlichkeit eines Remis deutlich höher ist als die Siegwahrscheinlichkeit des schwächeren Teams, kann sich dieses Verhältnis bei größeren Torsummen und zunehmender Spielstärke des Außenseiters jedoch umkehren. Auch im Vergleich zu einer ungeraden Torsumme n liegt die Gewinnwahrscheinlichkeit von Team A auffallend niedrig. Zusammenfassend lässt sich dennoch festhalten, dass aus diesem vereinfachten Modell hervorgeht: Die Wahrscheinlichkeit, mindestens einen Punkt zu erzielen (d. h. ein Unentschieden oder einen Sieg zu erreichen), ist für das schwächere Team umso höher, je weniger Tore in einem Spiel fallen.

Nicht nur in der Theorie, sondern auch in der Praxis beobachten Millionen von Zuschauer:innen immer wieder dieses Phänomen, dass sich „David gegen Goliath“ unerwartet durchsetzt. In der Saison 2023/24 sorgte etwa der Drittligist FC Saarbrücken im DFB-Pokal für eine Sensation: In der 6. Minute der Nachspielzeit drehten sie das Spiel gegen den übermächtigen FC Bayern München und hinterließen einen großen Schock beim Rekordmeister.

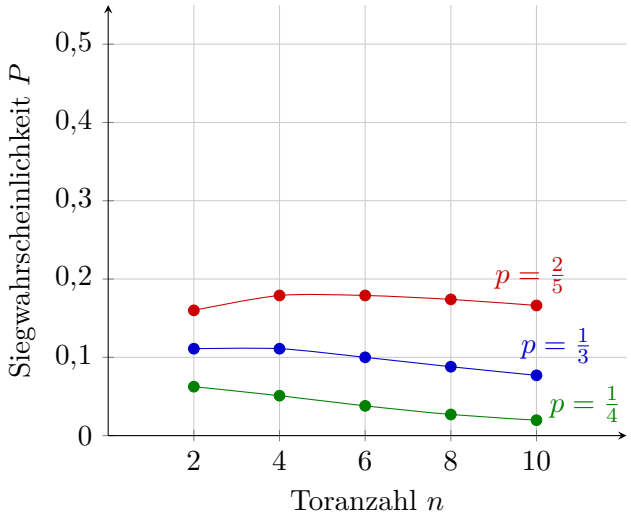

Abbildung 5.2: Siegwahrscheinlichkeit des schwächeren Teams bei *gerader Toranzahl* (eigene Darstellung in Anlehnung an Tolan [88, S. 51])

Nicht umsonst wird Fußball von Experten wie Tolan [88] als der ungerechteste Sport der Welt bezeichnet. Einen Ausschaltknopf für den Zufall gibt es eben nicht, und gerade im Fußball spielt dieser aufgrund der durchschnittlich niedrigen Trefferzahl eine besonders entscheidende Rolle. Bei Sportarten wie Handball, Basketball, Tennis oder Darts, deren Spieleigenschaften wesentlich mehr Treffer erfordern, setzt sich dagegen meist der Favorit durch. Dies liegt daran, dass glückliche Treffer des vermeintlich Schwächeren durch die Vielzahl an Möglichkeiten, Punkte zu erzielen, leichter ausgeglichen werden können. Mit zunehmender Anzahl potenzieller Treffer nimmt also die Relevanz der Zufallskomponente ab. Fußball hingegen ist so konzipiert, dass sich vergleichsweise selten wirklich aussichtsreiche Angriffe ergeben. Meist bedarf es einer besonderen Konstellation im Angriff und entsprechendem Abwehrverhalten, um den Ball erfolgreich ins Tor zu befördern. Folglich fallen im Fußball verhältnismäßig wenige Tore, sodass schon kleine Details den Unterschied im Kampf um die drei Punkte ausmachen können. Genau diese Eigenschaft ist einer der Hauptgründe, warum Fußballfans Woche für Woche in die Stadien strömen und die Spiele oft eine gewisse „Spannungsgarantie“ auf dem Platz und auf den Rängen bieten.

Aus statistischen Erhebungen der führenden Fußballligen geht hervor, dass pro Spiel und Team durchschnittlich etwa 10 bis 15 Schüsse abgegeben werden. Davon gelangen im Mittel rund 30 % bis 40 % tatsächlich auf das Tor, während der Rest entweder geblockt wird oder das Ziel verfehlt. Von diesen Abschlüssen auf das Tor führen wiederum nur etwa 25 % bis 35 % zu einem Treffer. Insgesamt resultiert somit im Durchschnitt lediglich etwa jeder zehnte Schuss in einem Tor [64, 76]. Diese Erfolgsquote wird unter anderem von Heuer [38] mit dem Werfen eines Würfels verglichen: Je öfter geworfen wird, desto näher kommt man dem erwarteten Ergebnis einer Gleichverteilung – oder im Sport dem

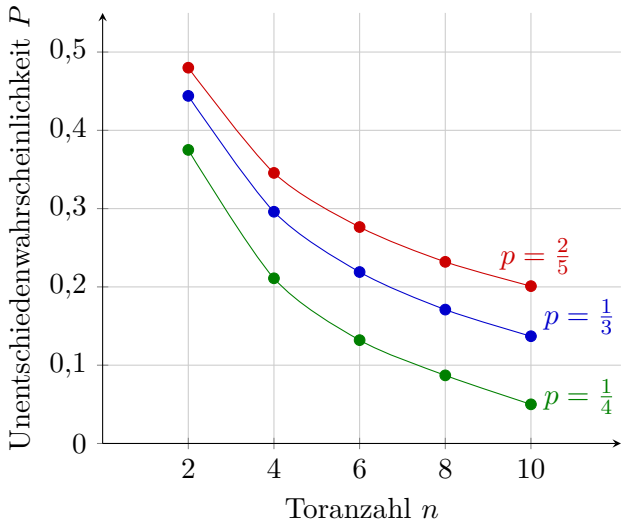

Abbildung 5.3: Wahrscheinlichkeit für ein Unentschieden bei *gerader Toranzahl* (eigene Darstellung in Anlehnung an Tolan [88, S. 51])

Sieg des Favoriten. Wer nur wenige Würfe oder Schüsse hat, ist in besonderem Maße auf Glück angewiesen.

5.1.2 Verwertung von Torschüssen als Modellierungsgrundlage

Um die Würfelanalogie weiter zu präzisieren, lässt sich das Zustandekommen von Toren durch ein vereinfachtes Zufallsexperiment modellieren. Dabei kann man sich vorstellen, dass jeder Schuss einem Wurf eines (gedachten) Würfels mit sehr vielen Seiten entspricht. Nur ein kleiner Teil dieser Seiten steht für das Ereignis „Tor“, während die überwiegende Mehrheit keinen Treffer repräsentiert. Ein Tor fällt genau dann, wenn bei einem solchen „Wurf“ – beziehungsweise Schuss – eine der wenigen Tor-Seiten realisiert wird. Die unbekannte Struktur dieses Würfels, insbesondere die Anzahl der „Tor-Seiten“, entspricht der gesuchten Trefferwahrscheinlichkeit und muss aus beobachteten Spieldaten erschlossen werden [18].

Diese Perspektive erlaubt es, grundlegende Konzepte der Wahrscheinlichkeitstheorie anschaulich zu übertragen. Interpretiert man jeden Schuss als unabhängigen Versuch mit konstanter Erfolgswahrscheinlichkeit, so ergibt sich für die Anzahl der erzielten Tore eine binomialverteilte Zufallsvariable. Selbst bei identischer Chanceneffizienz kann die tatsächlich beobachtete Toranzahl von Spiel zu Spiel erheblich variieren, insbesondere wenn – wie im Fußball üblich – die Anzahl der Versuche vergleichsweise gering ist. Daraus folgt, dass selbst unter der Annahme gleicher Schussverwertung unterschiedliche Ergebnisse keineswegs ungewöhnlich sind, wobei die Streuung maßgeblich von der Anzahl der zugrunde liegenden Schüsse abhängt [59].

Zugleich wird deutlich, dass die Schätzung der zugrunde liegenden Trefferwahrscheinlichkeit stets mit Unsicherheit behaftet ist. Werden nur wenige Beobachtungen herangezogen, kann die empirisch ermittelte Erfolgsquote deutlich vom tatsächlichen Wert abweichen. Erst mit wachsender Datenbasis stabilisiert sich diese Schätzung und nähert sich dem wahren Parameter an. Dieses Spannungsfeld zwischen zufälliger Streuung und systematischer Struktur ist zentral für die statistische Modellierung von Fußballergebnissen [38].

Vor diesem Hintergrund bietet es sich an, Torschüsse explizit als Bernoulli-Experimente mit einer bestimmten Erfolgswahrscheinlichkeit zu modellieren, die als Spielstärke p interpretiert wird. Die Verwertung von Torschüssen stellt somit eine naheliegende und zugleich theoretisch fundierte Grundlage für das nachfolgende Modell zur Beschreibung und Prognose von Spielergebnissen dar. Wie bereits im vorherigen Modell basieren die Spielstärken erneut auf empirischen Daten, wobei die Modellierung nun differenzierter erfolgt und zusätzliche, aussagekräftigere Einflussfaktoren berücksichtigt werden. Der folgende Ansatz orientiert sich am Modell von Ludwig und Oldenburg [59].

Auf Grundlage einer Auswertung der ersten 28 Spielrunden der Bundesliga-Saison 2025/26 ergibt sich, dass der FC St. Pauli (Team A) pro Spiel durchschnittlich 10,32 Torchancen in Form von Schüssen generiert und im Mittel etwa 0,89 Tore erzielt. Demgegenüber kommt Eintracht Frankfurt (Team B) im Durchschnitt auf 11,82 Schussversuche pro Spiel, von denen etwa 1,86 zu Toren führen. Auf Basis dieser Quoten werden die relativen Häufigkeiten als Trefferwahrscheinlichkeiten interpretiert, sodass sich

$$p_A = \frac{0{,}89}{10{,}32} \approx 0{,}0862 \qquad \text{und} \qquad p_B = \frac{1{,}86}{11{,}82} \approx 0{,}1574$$

ergeben. Diese dienen zur Bestimmung der Wahrscheinlichkeiten für k Treffer bei n Abschlüssen.

Da die Binomialverteilung eine natürliche Anzahl von Versuchen voraussetzt, werden die durchschnittlichen Schusszahlen entsprechend gerundet. Es sei A die Zufallsvariable, die die Anzahl der erzielten Tore von Team A beschreibt, und B die Zufallsvariable für die erzielten Treffer von Team B. Unter den getroffenen Annahmen ergeben sich für die Torwahrscheinlichkeiten der beiden Teams:

$$P_A(A = k_A) = \binom{10}{k_A} \cdot 0{,}0862^{k_A} \cdot 0{,}9138^{10-k_A},$$

$$P_B(B = k_B) = \binom{12}{k_B} \cdot 0{,}1574^{k_B} \cdot 0{,}8426^{12-k_B}.$$

Für $k_A = 1$ hat Team A nach unserer Auffassung also $n = 10$ Möglichkeiten im Spiel, diesen Treffer zu erzielen, sodass sich für die entsprechende Wahrscheinlichkeit ergibt:

$$P_A(A = 1) = \binom{10}{1} \cdot 0{,}0862 \cdot 0{,}9138^9 = 10 \cdot 0{,}0862 \cdot 0{,}4448 \approx 0{,}3830.$$

Für $k_A = 2$ beschreibt die gesuchte Wahrscheinlichkeit die Situation, dass Team A in insgesamt $n = 10$ Torschussversuchen genau zweimal erfolgreich ist. Ausgehend von diesem Modell lassen sich auch die Wahrscheinlichkeiten für weitere mögliche Toranzahlen im gleichen Kontext berechnen und interpretieren:

$$P_A(A = 0) = 0{,}9138^{10} \approx 0{,}4060,$$

$$P_A(A = 2) = \binom{10}{2} \cdot 0{,}0862^2 \cdot 0{,}9138^8 = 45 \cdot 0{,}00743 \cdot 0{,}4868 \approx 0{,}1626.$$

Für Team B ergeben sich analog:

$$P_B(B = 0) = 0{,}8426^{12} \approx 0{,}1281,$$

$$P_B(B = 1) = \binom{12}{1} \cdot 0{,}1574 \cdot 0{,}8426^{11} = 12 \cdot 0{,}1574 \cdot 0{,}1523 \approx 0{,}2871,$$

$$P_B(B = 2) = \binom{12}{2} \cdot 0{,}1574^2 \cdot 0{,}8426^{10} = 66 \cdot 0{,}02477 \cdot 0{,}1808 \approx 0{,}2950.$$

Im nächsten Schritt sind die bisher getrennt betrachteten Einzelwahrscheinlichkeiten so zu verknüpfen, dass konkrete Endstände zwischen dem FC St. Pauli und Eintracht Frankfurt modelliert werden können und darauf aufbauend Aussagen über die resultierenden Spielausgänge möglich sind. Formal lässt sich das Zustandekommen eines Spielergebnisses als zweistufiges Zufallsexperiment auffassen, bei dem zunächst die Toranzahl von Team A und anschließend jene von Team B realisiert wird. Eine zentrale Voraussetzung für diese Modellierung ist die Annahme, dass die von den beiden Teams erzielten Tore stochastisch unabhängig voneinander sind.

Diese Annahme erscheint auf den ersten Blick wenig realistisch, da der Spielverlauf im Fußball häufig durch situative Dynamiken geprägt ist. Gleichwohl lassen sich in der Praxis zahlreiche Situationen beobachten, in denen Tore scheinbar unabhängig vom bisherigen Spielgeschehen entstehen, etwa wenn eine Mannschaft unerwartet in Führung geht. Solche Beobachtungen können als Indizien für eine weitgehende Unabhängigkeit interpretiert werden [59]. Zwar könnte argumentiert werden, dass ein erzieltes Tor den weiteren Spielverlauf beeinflusst und somit auch die Wahrscheinlichkeit nachfolgender Treffer verändert, jedoch zeigen empirische Untersuchungen, dass dieser Zusammenhang nur sehr schwach ausgeprägt ist. So weist Tolan [88] eine Korrelation von lediglich 0,05 zwischen aufeinanderfolgenden Torereignissen nach, was auf einen vernachlässigbaren Einfluss hindeutet. Mit anderen Worten verändert ein Tor die Wahrscheinlichkeit eines weiteren Treffers nur in sehr geringem Ausmaß, nämlich um etwa 5 %. Eine mögliche Erklärung hierfür liegt darin, dass taktische Anpassungen beider Mannschaften einander weitgehend kompensieren. Reagiert ein Team beispielsweise auf ein Gegentor mit einer offensiveren Ausrichtung, so passt häufig auch der Gegner seine Strategie entsprechend an. Chu [15] argumentiert in diesem Zusammenhang, dass sich diese gegenläufigen Anpassungen im Mittel aufheben, sodass die Torrate im Spielverlauf annähernd konstant

bleibt. In der Folge können die Torerfolge beider Teams mit hinreichender Genauigkeit als unabhängig modelliert werden.

Nach Definition 3.5.12 können die Einzelwahrscheinlichkeiten für Treffer einfach miteinander multipliziert werden, woraus sich der Endstand zwischen Team A und Team B mit k_A zu k_B Toren wie folgt berechnen lässt:

$$\begin{aligned} P(A = k_A, B = k_B) &= P(A = k_A) \cdot P(B = k_B) \\ &= \binom{10}{k_A} \cdot 0{,}0862^{k_A} \cdot 0{,}9138^{10-k_A} \cdot \binom{12}{k_B} \cdot 0{,}1574^{k_B} \cdot 0{,}8426^{12-k_B}. \end{aligned}$$

In Tabelle 5.1 (S. 94) sind die Wahrscheinlichkeiten aller möglichen Spielausgänge festgehalten. Aufgrund der Additivität des Wahrscheinlichkeitsmaßes können die einzelnen Ereignisse kombiniert werden, um die kumulierten Wahrscheinlichkeiten für einen Sieg von Team A, ein Unentschieden oder einen Sieg von Team B zu bestimmen:

$$\begin{aligned} P(\text{Sieg Team A}) &= \sum_{k_A > k_B} P(A = k_A) \cdot P(B = k_B) \\ &= \sum_{k_B=0}^{9} \sum_{k_A=k_B+1}^{10} P(A = k_A) \cdot P(B = k_B) \approx 0{,}1524, \\ P(\text{Unentschieden}) &= \sum_{k_A=k_B} P(A = k_A) \cdot P(B = k_B) \\ &= \sum_{k=0}^{10} P(A = k) \cdot P(B = k) \approx 0{,}2179, \\ P(\text{Sieg Team B}) &= \sum_{k_B > k_A} P(A = k_A) \cdot P(B = k_B) \\ &= \sum_{k_A=0}^{10} \sum_{k_B=k_A+1}^{12} P(A = k_A) \cdot P(B = k_B) \approx 0{,}6296. \end{aligned}$$

Insgesamt verdeutlichen die in der Tabelle dargestellten Wahrscheinlichkeiten, dass sich die Wahrscheinlichkeitsmasse auf niedrige Toranzahlen konzentriert. Insbesondere weisen Ergebnisse mit null bis drei Treffern pro Team die höchsten Eintrittswahrscheinlichkeiten auf, was der typischen Torverteilung in Fußballspielen entspricht. Die Struktur der Tabelle zeigt zudem, dass die größten Wahrscheinlichkeiten entlang der Diagonalen sowie in deren unmittelbarer Umgebung liegen. Dies impliziert, dass knappe Spielausgänge – insbesondere Unentschieden oder Spiele mit geringer Tordifferenz – die wahrscheinlichsten Szenarien darstellen.

Die kumulierten Wahrscheinlichkeiten am Tabellenrand verdeutlichen darüber hinaus eine klare Überlegenheit von Team B. Während die Wahrscheinlichkeit für einen Sieg von Team A lediglich etwa 15,24 % beträgt, liegt die entsprechende Wahrscheinlichkeit für

Tabelle 5.1: Wahrscheinlichkeiten der Spielausgänge zwischen Team A und Team B

	Team A	$k_A = 0$	1	2	3	4	5	6	7	≥ 8	Summe
Team B		0,4060	0,3830	0,1626	0,0409	0,0068	0,0008	0,0001	0,0000	0,0000	1
$k_B = 0$	0,1281	5,20%	4,90%	2,08%	0,52%	0,09%	0,01%	0,00%	0,00%	0,00%	
1	0,2871	11,66%	10,99%	4,67%	1,17%	0,19%	0,02%	0,00%	0,00%	0,00%	
2	0,2950	11,98%	11,30%	4,80%	1,21%	0,20%	0,02%	0,00%	0,00%	0,00%	
3	0,1837	7,46%	7,03%	2,99%	0,75%	0,12%	0,01%	0,00%	0,00%	0,00%	
4	0,0772	3,13%	2,96%	1,25%	0,32%	0,05%	0,01%	0,00%	0,00%	0,00%	15,24%
5	0,0231	0,94%	0,88%	0,38%	0,09%	0,02%	0,00%	0,00%	0,00%	0,00%	
6	0,0050	0,20%	0,19%	0,08%	0,02%	0,00%	0,00%	0,00%	0,00%	0,00%	
7	0,0008	0,03%	0,03%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
≥ 8	0,0001	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Summe	1	62,96%									21,79%

Team B bei rund 62,96 %. Die Wahrscheinlichkeit eines Unentschiedens beläuft sich auf 21,79 % und liegt damit im erwarteten Bereich. Insgesamt bestätigt das Modell sowohl die charakteristische Torverteilung im Fußball als auch die unterschiedliche Spielstärke der beiden betrachteten Teams.

5.2 Modellierung von Toranzahlen mittels der Poisson-Verteilung

Nachdem im vorangegangenen Abschnitt die Anzahl erzielter Tore mithilfe der Binomialverteilung modelliert wurde, bietet sich im nächsten Schritt eine weiterführende Approximation an. Insbesondere in Situationen, in denen eine große Anzahl von Versuchen mit einer sehr kleinen Erfolgswahrscheinlichkeit vorliegt, lässt sich die Binomialverteilung durch die Poisson-Verteilung annähern, wie bereits in Satz 3.5.22 gezeigt wurde.

Überträgt man diese Überlegung auf den Fußball, so kann die Spieldauer von 90 Minuten als eine große Anzahl potenzieller Gelegenheiten interpretiert werden, in denen jeweils ein Tor erzielt werden kann. Formal sei $n = 90$ die Anzahl gleich langer Zeitintervalle. Die Wahrscheinlichkeit, in einem solchen Intervall ein Tor zu erzielen, sei p. Wählt man beispielsweise $p = \frac{1}{90}$, so ergibt sich im Erwartungswert ein Tor pro Spiel, während $p = \frac{2}{90}$ im Mittel zwei Tore pro Spiel liefert. Allgemein gilt für die Binomialverteilung der Erwartungswert $n \cdot p$, sodass dieser direkt der durchschnittlichen Toranzahl entspricht. Diese Modellierung knüpft unmittelbar an die zuvor eingeführte binomialverteilte Beschreibung an, interpretiert die einzelnen Versuche jedoch nicht mehr als diskrete Schüsse, sondern als zeitliche Mikrointervalle. Der Übergang zur Poisson-Verteilung erfolgt im Grenzfall $n \to \infty$ bei konstantem Produkt $\lambda = n \cdot p$. Der Parameter λ beschreibt dabei die mittlere Anzahl von Toren pro Spiel und stellt die zentrale Kenngröße der Poisson-

Verteilung dar. Die praktische Relevanz dieses Ansatzes zeigt sich in den empirischen Torstatistiken der betrachteten Teams: Für den FC St. Pauli (Team A) ergibt sich eine durchschnittliche Toranzahl von $\lambda_A = 0{,}89$, während Eintracht Frankfurt (Team B) im Mittel $\lambda_B = 1{,}86$ Tore pro Spiel erzielt. Diese Werte können direkt als Parameter der jeweiligen Poisson-Verteilungen interpretiert werden und ersetzen damit die in der Binomialverteilung separat zu bestimmenden Größen n und p [15, 41].

Neben der mathematischen Herleitung spricht auch die empirische Struktur von Torverteilungen für den Einsatz der Poisson-Verteilung. Wie aus der statistischen Häufigkeitsverteilung der Torsummen in Spielen der vergangenen fünf Saisons der deutschen Bundesliga (vgl. Tabelle 2.3 auf S. 17) hervorgeht, beginnt die Verteilung typischerweise bei $k = 0$, steigt zu einem Maximum bei $k = 2$ Toren pro Spiel an und fällt für größere Werte rasch ab. Die höchsten Ausprägungen liegen somit unmittelbar in der Nähe der mittleren Toranzahl von 3,14. Dieses charakteristische Verhalten entspricht der Form der Poisson-Verteilung, bei der die größten Wahrscheinlichkeiten im Bereich des Erwartungswertes liegen, während hohe Torzahlen nur mit geringer Wahrscheinlichkeit auftreten.

Darüber hinaus lässt sich die Eignung dieses Modells auch durch eine zeitliche Betrachtung von Torereignissen begründen. Untersuchungen von Chu [15] zeigen, dass die Zeitabstände zwischen aufeinanderfolgenden Toren näherungsweise exponentialverteilt sind. Zudem wurde von Georgii [30] gezeigt, dass sich aus einer Folge unabhängiger exponentialverteilter Zufallsvariablen – die in diesem Kontext die Zwischentorzeiten beschreiben – ein Poisson-Prozess konstruieren lässt. Daraus folgt, dass die Anzahl der Tore in einem festen Zeitintervall, wie einem Fußballspiel, poissonverteilt ist. Nach Dambeck [18] werden für die Anwendung im Fußball mehrere Annahmen getroffen: Zum einen wird vorausgesetzt, dass in hinreichend kleinen Zeitintervallen höchstens ein Tor fällt, was bei geeigneter Zerlegung der Spielzeit plausibel ist. Zum anderen wird angenommen, dass die Wahrscheinlichkeit eines Tores proportional zur Länge des betrachteten Zeitintervalls ist und dass Torereignisse unabhängig voneinander auftreten. Letztere Annahme wurde bereits im vorherigen Abschnitt diskutiert und durch empirische Befunde gestützt, sodass von einer näherungsweise konstanten Torrate im Spielverlauf ausgegangen werden kann. Die ersten beiden Bedingungen werden durch die Analyse von Chu [15] untermauert, da kein einminütiges Zeitintervall mit mehr als einem Tor beobachtet wurde und sich zeigt, dass mit zunehmender Intervalllänge auch die Anzahl der Tore steigt.

Da für eine poissonverteilte Zufallsvariable X sowohl der Erwartungswert als auch die Varianz mit dem Parameter λ übereinstimmen, also $E(X) = \lambda = V(X)$ gilt, kann an dieser Stelle überprüft werden, inwieweit sich diese Eigenschaft auch in den empirischen Daten zur Torsumme eines Bundesligaspiels widerspiegelt. Dazu werden das arithmetische Mittel $\bar{x}$ und die empirische Varianz s^2 der beobachteten Torsummen betrachtet.

Auf Grundlage der in Tabelle 2.3 zusammengefassten Häufigkeiten ergibt sich für die durchschnittliche Torsumme pro Spiel

$$\bar{x} \approx 3{,}14.$$

Da die Daten in Form einer Häufigkeitsverteilung vorliegen, ist es zweckmäßig, die empirische Varianz nicht über einzelne Beobachtungen x_i, sondern über die möglichen Ausprägungen a_j mit ihren jeweiligen Häufigkeiten h_j zu berechnen. Die Formel geht dabei in die äquivalente Darstellung

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{10} h_j (a_j - \bar{x})^2$$

über. Durch Einsetzen der Werte ergibt sich

$$\begin{aligned} s^2 = \frac{1}{1\,530 - 1} \Big(& 84(0 - 3{,}14)^2 + 173(1 - 3{,}14)^2 \\ & + 355(2 - 3{,}14)^2 + 313(3 - 3{,}14)^2 + 283(4 - 3{,}14)^2 + 168(5 - 3{,}14)^2 \\ & + 91(6 - 3{,}14)^2 + 50(7 - 3{,}14)^2 + 10(8 - 3{,}14)^2 + 3(9 - 3{,}14)^2 \Big) \approx 3{,}08. \end{aligned}$$

Damit liegen arithmetisches Mittel und empirische Varianz mit $\bar{x} \approx 3{,}14$ beziehungsweise $s^2 \approx 3{,}08$ sehr nahe beieinander. Dies entspricht einer charakteristischen Eigenschaft der Poisson-Verteilung, bei der Erwartungswert und Varianz übereinstimmen, und liefert somit einen empirischen Hinweis darauf, dass die Poisson-Verteilung die Torsumme in Fußballspielen zumindest näherungsweise angemessen beschreibt. Sie ermöglicht es folglich, die Toranzahl eines Teams allein durch den Parameter λ als Erwartungswert zu approximieren und bildet damit eine zentrale Grundlage für die weitere Modellierung von Spielergebnissen.

Im nächsten Schritt stellt sich die Frage, inwieweit die theoretisch motivierte Poisson-Verteilung die empirisch beobachtete Verteilung der Torsummen tatsächlich abbilden kann. Zu diesem Zweck wird der zuvor bestimmte empirische Mittelwert $\bar{x} \approx 3{,}14$ als Schätzer für den Erwartungswert λ verwendet und somit als einziger Parameter der Verteilung herangezogen. Die Zufallsvariable X beschreibt dabei die Torsumme pro Spiel.

Unter dieser Annahme gilt $X \sim \mathrm{P}(\lambda = 3{,}14)$ mit

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} = \frac{e^{-3{,}14} \cdot 3{,}14^k}{k!}, \quad k \in \mathbb{N}_0.$$

Für jedes k liefert diese Formel die Wahrscheinlichkeit, dass in einem Spiel genau k Tore fallen. Multipliziert man diese Wahrscheinlichkeiten mit der Gesamtanzahl der betrachteten Spiele $n = 1\,530$, so erhält man die erwarteten absoluten Häufigkeiten

$$E_k = 1\,530 \cdot P(X = k),$$

die direkt mit den empirisch beobachteten Häufigkeiten verglichen werden können. Zur besseren Interpretierbarkeit werden die erwarteten Häufigkeiten auf ganze Zahlen gerundet; diese gerundeten Werte werden im Folgenden auch in der Teststatistik verwendet.

Eine charakteristische Eigenschaft der Poisson-Verteilung besteht darin, dass ihre Werte zunächst ansteigen, bis k etwa den Erwartungswert λ erreicht, und anschließend wieder

abfallen. Genauer lässt sich zeigen, dass die Verteilung ihr Maximum im Intervall $[\lambda-1, \lambda]$ annimmt und somit im vorliegenden Fall im Bereich von $k = 2$ bis $k = 3$ liegt. Dies stimmt gut mit der empirischen Beobachtung überein, dass die häufigsten Torsummen in diesem Bereich auftreten. Eine alternative Begründung liefert die rekursive Darstellung

$$P(X = k) = \frac{\lambda}{k} \cdot P(X = k - 1),$$

aus der unmittelbar folgt, dass die Wahrscheinlichkeiten solange wachsen, wie $\frac{\lambda}{k} > 1$ gilt, und danach monoton abnehmen [11].

Die Gegenüberstellung der empirischen und theoretischen Werte erfolgt in Tabelle 5.2. Dabei werden neben den absoluten Häufigkeiten auch die relativen Häufigkeiten sowie die theoretischen Wahrscheinlichkeiten $P(X = k)$ und die daraus resultierenden erwarteten Häufigkeiten E_k berücksichtigt. Da die Wahrscheinlichkeiten für große Werte von k sehr klein sind, werden alle Beobachtungen ab $k \geq 9$ zu einer gemeinsamen Kategorie zusammengefasst.

Tabelle 5.2: Gegenüberstellung empirischer und poissonverteilter Torsummen

k	abs. Häufigkeit	rel. Häufigkeit	$P(X = k)$	E_k
0	84	0,0549	0,0433	66
1	173	0,1131	0,1359	208
2	355	0,2320	0,2134	327
3	313	0,2046	0,2233	342
4	283	0,1850	0,1753	268
5	168	0,1098	0,1101	168
6	91	0,0595	0,0576	88
7	50	0,0327	0,0258	40
8	10	0,0065	0,0101	16
≥ 9	3	0,0020	0,0051	8
Summe	1 530	1	1	1 530

Ein Vergleich der Werte zeigt, dass sich die theoretischen und empirischen Häufigkeiten insgesamt sehr ähnlich verhalten, auch wenn die Abweichungen – mit Ausnahme der Torsumme von fünf Treffern – wechselweise in beide Richtungen ausfallen. Nichtsdestotrotz wird das charakteristische Profil der Verteilung – mit einem Maximum im Bereich von zwei bis drei Toren und einem anschließenden raschen Abfall – durch das Poisson-Modell überzeugend reproduziert.

In einer idealisierten „Poisson-Fußballwelt“ ließe sich die empirische Verteilung vollständig durch die Poisson-Verteilung beschreiben. In der Realität zeigen sich zwar gewisse Abweichungen, die etwa auf taktische Einflüsse, Spielsituationen oder Abhängigkeiten

im Spielverlauf zurückgeführt werden können, dennoch wird die grundsätzliche Eignung dieses Ansatzes durch zahlreiche Studien gestützt. So zeigen zahlreiche Literaturquellen, dass die Poisson-Verteilung eine plausible Modellierung der Toranzahl liefert [11, 15, 18, 31, 38, 60, 71, 75, 78]. Auch vergleichende Analysen verschiedener Prognosemodelle kommen zu dem Ergebnis, dass Poisson-basierte Ansätze eine hohe Vorhersagequalität aufweisen und sich nur geringfügig von komplexeren Modellen unterscheiden. Dabei wird hervorgehoben, dass die Güte der Vorhersagen weniger von der konkreten Modellwahl als vielmehr von der Qualität der zugrunde liegenden Daten abhängt [27, 44]. Ebenso zeigen Ley et al. [57], dass insbesondere bivariate und unabhängige Poisson-Modelle zu den leistungsfähigsten Ansätzen in der Modellierung von Fußballergebnissen zählen.

Um die beobachtete Übereinstimmung zwischen der empirischen Verteilung und dem theoretischen Modell nicht nur qualitativ, sondern auch quantitativ zu beurteilen, bietet sich der Einsatz statistischer Testverfahren an. Insbesondere kann mithilfe des Chi-Quadrat-Anpassungstests überprüft werden, ob die Abweichungen zwischen den beobachteten und den erwarteten Häufigkeiten im Rahmen zufälliger Schwankungen liegen oder systematische Unterschiede vorliegen. Im folgenden Abschnitt wird daher untersucht, ob die Annahme einer poissonverteilten Torsumme statistisch gerechtfertigt ist.

5.3 Statistische Überprüfung der Modellgüte

In diesem Abschnitt wird die zuvor visuell motivierte Modellanpassung formal überprüft. Ziel ist es, die Übereinstimmung zwischen der empirischen Verteilung der Torsummen und der theoretischen Poisson-Verteilung statistisch zu bewerten. Zu diesem Zweck wird der Chi-Quadrat-Anpassungstest herangezogen, der eine quantitative Beurteilung der Abweichungen zwischen beobachteten und erwarteten Häufigkeiten ermöglicht.

5.3.1 Chi-Quadrat-Anpassungstest für aggregierte Daten (fünf Spielzeiten)

Da für die Durchführung des Tests bestimmte Voraussetzungen an die erwarteten Häufigkeiten gestellt werden, ist eine geeignete Klasseneinteilung erforderlich. Vor diesem Hintergrund werden die Torsummen in Tabelle 5.2 (S. 97) in die zehn Klassen $k = 0, \ldots, 9$ eingeteilt, wobei die letzte Klasse $k \geq 9$ alle Torsummen mit mindestens neun Treffern zusammenfasst.

1. Signifikanzniveau und Hypothesen: Es werde das Signifikanzniveau $\alpha = 0{,}05$ festgelegt. Die Nullhypothese lautet

$$H_0 : \text{Die Torsumme pro Spiel ist poissonverteilt mit } \lambda = 3{,}14.$$

Die Alternativhypothese H_1 besagt, dass die Torsumme nicht dieser Verteilung folgt.

2. Erwartete Häufigkeiten und Teststatistik: Seien N_k die beobachteten Häufigkeiten und E_k die unter der Nullhypothese erwarteten Häufigkeiten der jeweiligen Klassen $k = 0, \ldots, 9$. Dann gilt

$$E_k = 1\,530 \cdot P(X = k)$$

für $k = 0, \ldots, 8$, während E_9 die erwartete Häufigkeit der zusammengefassten Klasse $k \geq 9$ bezeichnet. Die Teststatistik des Chi-Quadrat-Anpassungstests lautet somit

$$D = \sum_{k=0}^{9} \frac{(N_k - E_k)^2}{E_k}.$$

3. Berechnung der Teststatistik: Aus Tabelle 5.2 ergeben sich die beobachteten und erwarteten Häufigkeiten. Einsetzen liefert

$$\begin{aligned} D = {} & \frac{(84-66)^2}{66} + \frac{(173-208)^2}{208} + \frac{(355-327)^2}{327} + \frac{(313-342)^2}{342} + \frac{(283-268)^2}{268} \\ & + \frac{(168-168)^2}{168} + \frac{(91-88)^2}{88} + \frac{(50-40)^2}{40} + \frac{(10-16)^2}{16} + \frac{(3-8)^2}{8}. \end{aligned}$$

Damit ergibt sich

$$D \approx 24{,}47.$$

4. Freiheitsgrade und Verwerfungsbereich: Die Anzahl der Klassen beträgt $j = 10$. Da der Parameter λ aus den Daten geschätzt wurde, reduziert sich die Anzahl der Freiheitsgrade um eins und es gilt

$$j - 1 - 1 = 8.$$

Für $\alpha = 0{,}05$ ergibt sich der kritische Wert

$$c_{0{,}95;8} \approx 15{,}51.$$

Der Verwerfungsbereich lautet daher

$$V = (15{,}51; \infty).$$

5. Testentscheidung und Interpretation: Da

$$D \approx 24{,}47 \in V,$$

wird die Nullhypothese auf dem Signifikanzniveau von 5 % verworfen. Die empirisch beobachtete Verteilung der Torsummen über die betrachteten fünf Bundesligasaisons weicht somit statistisch signifikant von einer Poisson-Verteilung mit $\lambda = 3{,}14$ ab. Insbesondere wurde mit einer Sicherheit von mindestens 0,95 gezeigt, dass die Poisson-Verteilung

mit $\lambda = 3{,}14$ die beobachteten Torsummen nicht adäquat beschreibt. Die festgestellten Unterschiede lassen sich folglich nicht allein durch zufällige Schwankungen erklären.

Gleichwohl ist dieses Ergebnis im Kontext der bisherigen Analyse differenziert zu betrachten. Trotz der statistisch signifikanten Abweichung zeigt die Poisson-Verteilung eine insgesamt gute Annäherung an die empirische Verteilung. In der Literatur wird zudem vielfach berichtet, dass bei vergleichbaren Untersuchungen die Nullhypothese nicht verworfen wird und die Poisson-Verteilung auch quantitativ eine zufriedenstellende Approximation der Torverteilung in Fußballspielen liefert [15, 72, 87].

5.3.2 Chi-Quadrat-Anpassungstest für die Bundesliga-Saison 2025/26

Da im weiteren Verlauf dieser Arbeit ausschließlich Daten der aktuellen Bundesliga-Saison herangezogen werden, erscheint es sinnvoll, die Modellgüte zusätzlich anhand dieser spezifischen Datengrundlage zu überprüfen. Auf diese Weise kann untersucht werden, ob die Poisson-Verteilung zumindest für jene Daten, auf denen auch die spätere Modellierung basiert, statistisch vertretbar ist.

Für die aktuelle Bundesliga-Saison 2025/26 wurden bis zum 28. Spieltag insgesamt $n = 252$ Spiele betrachtet. Die beobachteten Torsummen lauten:

$$N_0 = 11, \quad N_1 = 32, \quad N_2 = 52, \quad N_3 = 54, \quad N_4 = 50, \quad N_5 = 25,$$
$$N_6 = 19, \quad N_7 = 5, \quad N_8 = 1, \quad N_9 = 2, \quad N_{10} = 1,$$

woraus sich der empirische Mittelwert $\bar{x} \approx 3{,}21$ ergibt.

1. Signifikanzniveau und Hypothesen: Es wird erneut das Signifikanzniveau $\alpha = 0{,}05$ festgelegt. Die Nullhypothese lautet

$$H_0 : \text{Die Torsumme pro Spiel ist poissonverteilt mit } \lambda = 3{,}21.$$

Die Alternativhypothese H_1 besagt, dass die Torsumme nicht dieser Verteilung folgt.

2. Erwartete Häufigkeiten und Teststatistik: Interpretiert man den empirischen Mittelwert $\bar{x} \approx 3{,}21$ als Erwartungswert $\lambda = 3{,}21$, so können die theoretischen Wahrscheinlichkeiten der Torsumme pro Spiel der Saison 2025/26 durch

$$P(X = k) = \frac{e^{-3{,}21} \cdot 3{,}21^k}{k!}, \quad k \in \mathbb{N}_0,$$

berechnet werden. Seien N_k die beobachteten Häufigkeiten und E_k die unter der Nullhypothese erwarteten Häufigkeiten der jeweiligen Klassen $k = 0, \ldots, 7$. Dann gilt

$$E_k = 252 \cdot P(X = k)$$

für $k = 0, \ldots, 6$, während E_7 die erwartete Häufigkeit der zusammengefassten Klasse $k \geq 7$ bezeichnet. Zur besseren Übersicht werden die erwarteten Häufigkeiten auf ganze Zahlen gerundet dargestellt:

k	0	1	2	3	4	5	6	≥ 7
N_k	11	32	52	54	50	25	19	9
E_k	10	33	52	56	45	29	15	11

Es ergeben sich somit die acht Klassen $k = 0, \ldots, 7$, wobei die letzte Klasse $k = 7$ alle Torsummen mit mindestens sieben Treffern umfasst. Damit lautet die Teststatistik

$$D = \sum_{k=0}^{7} \frac{(N_k - E_k)^2}{E_k}.$$

3. Berechnung der Teststatistik: Einsetzen der beobachteten und erwarteten Häufigkeiten liefert

$$D = \frac{(11-10)^2}{10} + \frac{(32-33)^2}{33} + \frac{(52-52)^2}{52} + \frac{(54-56)^2}{56} + \frac{(50-45)^2}{45} + \frac{(25-29)^2}{29} + \frac{(19-15)^2}{15} + \frac{(9-11)^2}{11}.$$

Damit ergibt sich

$$D \approx 2{,}74.$$

4. Freiheitsgrade und Verwerfungsbereich: Die Anzahl der Klassen beträgt $j = 8$. Da der Parameter λ aus den Daten geschätzt wurde, reduziert sich die Anzahl der Freiheitsgrade um eins. Somit gilt

$$j - 1 - 1 = 6.$$

Für das Signifikanzniveau $\alpha = 0{,}05$ ergibt sich der kritische Wert

$$c_{0{,}95;6} \approx 12{,}59.$$

Der Verwerfungsbereich lautet daher

$$V = (12{,}59; \infty).$$

5. Testentscheidung und Interpretation: Da

$$D \approx 2{,}74 \notin V,$$

wird die Nullhypothese auf dem Signifikanzniveau von 5 % nicht verworfen. Die beobachteten Torsummen der aktuellen Bundesliga-Saison 2025/26 sind somit mit einer Poisson-Verteilung mit Parameter $\lambda = 3{,}21$ vereinbar. Insbesondere sprechen die vorliegenden

Daten mit einer Sicherheit von mindestens 0,95 nicht gegen die Poisson-Verteilung mit $\lambda = 3{,}21$. Die Abweichungen zwischen empirischen und theoretischen Häufigkeiten können in diesem Fall als zufällige Schwankungen interpretiert werden.

Dieses Ergebnis stützt die weitere Verwendung der Poisson-Verteilung im Rahmen der vorliegenden Arbeit zusätzlich. Während die Langzeitbetrachtung über fünf Spielzeiten hinweg noch auf statistisch signifikante Abweichungen hinweist, zeigt sich für die aktuelle Saison – und damit gerade für jene Datenbasis, auf der die späteren Modellierungen beruhen – eine gute quantitative Übereinstimmung zwischen empirischer Torverteilung und Poisson-Modell.

5.4 Modellierung von Teamstärken anhand geeigneter Leistungsindikatoren

Aufbauend auf der statistischen Überprüfung der Poisson-Verteilung stellt sich im nächsten Schritt die Frage, welche Leistungsindikatoren geeignet sind, um die Spielstärke von Teams im Rahmen eines Modells adäquat zu erfassen. Moderne Datenanalysen im Fußball ermöglichen eine detaillierte Auswertung von Spielen unter Einsatz umfangreicher technischer Verfahren. Die dabei gewonnenen Daten eröffnen vielfältige Anwendungsmöglichkeiten. Im vorliegenden Kontext steht insbesondere die Prognose zukünftiger Spielergebnisse im Fokus. Dieser Prozess lässt sich konzeptionell in zwei Schritte gliedern: Zunächst werden aus vergangenen Spielen geeignete Informationen extrahiert, um die Leistungsstärke der Teams möglichst präzise zu approximieren. Darauf aufbauend werden unter Verwendung dieser modellierten Teamstärken Vorhersagen über zukünftige Spielausgänge getroffen. Aufgrund der vor dem Spiel nur begrenzt verfügbaren Informationen ist die Einschätzung der Leistungsstärke der beteiligten Teams zwangsläufig mit Unsicherheiten behaftet, die jedoch durch eine geeignete Auswahl und Optimierung der verwendeten Daten reduziert werden können. Dabei ist von besonderem Interesse, welche Kenngrößen einen maßgeblichen Einfluss auf die zugrunde liegende Zielgröße und damit auf die Prognose des sportlichen Erfolgs haben [38].

Da Fußballspiele durch Tore entschieden werden, bietet es sich an, zunächst die Wahrscheinlichkeiten für erzielte Tore sowie für erhaltene Gegentore eines Teams zu modellieren. Aus den Torstatistiken der deutschen Bundesliga geht beispielsweise hervor, dass die TSG Hoffenheim nach den ersten 29 Spieltagen der Saison 2025/26 insgesamt 57 Tore erzielt hat [20]. Dies entspricht einem arithmetischen Mittel von

$$\bar{x} = \frac{57}{29} \approx 1{,}97$$

Toren pro Spiel. Sei X die Zufallsvariable, die die Anzahl der von Hoffenheim in einem Spiel erzielten Tore beschreibt. Unter der Annahme einer Poisson-Verteilung gilt $X \sim \mathrm{P}(1{,}97)$, sodass sich die Wahrscheinlichkeiten durch

$$P(X = k) = \frac{e^{-1{,}97} \cdot 1{,}97^k}{k!}, \quad k \in \mathbb{N}_0,$$

berechnen lassen.

Beispielhaft ergeben sich folgende Werte:

$$P(X = 0) \approx 0{,}1395, \qquad P(X = 1) \approx 0{,}2747,$$
$$P(X = 2) \approx 0{,}2706, \qquad P(X = 3) \approx 0{,}1778.$$

Daraus folgt insbesondere:

$$P(X \leq 3) = \sum_{k=0}^{3} P(X = k) \approx 0{,}8626,$$
$$P(X > 3) = 1 - \sum_{k=0}^{3} P(X = k) \approx 0{,}1374.$$

Die Ergebnisse zeigen, dass Hoffenheim mit hoher Wahrscheinlichkeit zwischen null und drei Tore pro Spiel erzielt, wobei ein oder zwei Treffer die wahrscheinlichsten Ausprägungen darstellen. Höhere Torausbeuten treten hingegen vergleichsweise selten auf. Ein Vergleich mit den empirischen Daten bestätigt die Plausibilität der Modellierung: In 25 der 29 betrachteten Spiele wurden höchstens drei Tore erzielt. Dies entspricht einem Anteil von $\frac{25}{29} \approx 86{,}21\,\%$ und stimmt somit sehr gut mit dem theoretisch bestimmten Wert überein.

Analog lässt sich die Anzahl der Gegentore modellieren. In den ersten 29 Spieltagen kassierte Hoffenheim insgesamt 43 Gegentore, was einem Mittelwert von

$$\bar{y} = \frac{43}{29} \approx 1{,}48$$

pro Spiel entspricht. Sei Y die Zufallsvariable für die Anzahl der Gegentore. Unter der Annahme $Y \sim \mathrm{P}(1{,}48)$ gilt:

$$P(Y = k) = \frac{e^{-1{,}48} \cdot 1{,}48^k}{k!}, \quad k \in \mathbb{N}_0.$$

Beispielhaft erhält man:

$$P(Y = 0) \approx 0{,}2276, \qquad P(Y = 1) \approx 0{,}3368, \qquad P(Y = 2) \approx 0{,}2492.$$

Für aggregierte Wahrscheinlichkeiten ergibt sich:

$$P(Y \leq 2) = \sum_{k=0}^{2} P(Y = k) \approx 0{,}8136,$$
$$P(Y > 2) = 1 - \sum_{k=0}^{2} P(Y = k) \approx 0{,}1864.$$

Auch hier dominieren geringe Gegentoranzahlen, während Spiele mit vielen kassierten Treffern deutlich seltener auftreten. In den betrachteten 29 Spielen lag der Anteil der Spiele mit höchstens zwei Gegentoren bei $\frac{25}{29} \approx 86{,}21\,\%$ und damit ebenfalls nahe am theoretischen Wert. Insgesamt liefert die Poisson-Verteilung somit sowohl für die erzielten Tore als auch für die erhaltenen Gegentore eine plausible Beschreibung der zugrunde liegenden Verteilungen.

Der zuvor dargestellte Ansatz ist zwar intuitiv und liefert eine erste plausible Beschreibung der Torverteilungen, stößt jedoch bei der konkreten Prognose eines Spiels zwischen zwei Teams an seine Grenzen. Die Modellierung der erzielten Tore und der kassierten Gegentore mittels separater Poisson-Verteilungen ermöglicht zunächst eine getrennte Betrachtung offensiver und defensiver Aspekte und damit eine erste Annäherung an die zugrunde liegende Spielstärke. Eine solche isolierte Betrachtung greift jedoch im Kontext eines Fußballspiels zu kurz, da die Leistung zweier Teams stets in direkter Wechselwirkung steht. Die Anzahl der erzielten Tore eines Teams hängt nicht ausschließlich von dessen Offensivstärke ab, sondern wird zugleich maßgeblich durch die Defensivqualität des Gegners beeinflusst. Entsprechend gilt dies auch umgekehrt für die kassierten Gegentore. Diese wechselseitige Abhängigkeit verdeutlicht, dass eine getrennte Analyse von Toren und Gegentoren die tatsächliche Leistungsfähigkeit eines Teams nur unvollständig abbildet. Vielmehr entsteht das Spielergebnis aus dem Zusammenspiel beider Komponenten, sodass eine integrierte Betrachtung erforderlich erscheint.

5.4.1 Die Tordifferenz als Maß der Teamstärke

Da im Rahmen des Poisson-Modells pro Team lediglich ein Parameter zur Verfügung steht, ist es sinnvoll, offensive und defensive Aspekte in einer gemeinsamen Kenngröße zu bündeln. Auch in der einschlägigen Literatur wird betont, dass zur adäquaten Beschreibung der Spielstärke sowohl erzielte als auch kassierte Tore gemeinsam berücksichtigt werden sollten. Eine naheliegende Wahl stellt die Tordifferenz dar, da sie beide Komponenten in kompakter Form vereint und somit eine aussagekräftige Charakterisierung der Teamleistung ermöglicht [18]. Empirische Untersuchungen zeigen zudem, dass die Offensive und Defensive eines Teams in etwa gleich stark zum Erfolg beitragen, da sowohl die Anzahl der erzielten Tore als auch die der kassierten Gegentore eine vergleichbare Korrelation mit der Tordifferenz aufweisen [36]. Die Tordifferenz kann somit als geeignete aggregierte Größe interpretiert werden, in der offensive und defensive Einflussfaktoren gleichgewichtig berücksichtigt sind. Darüber hinaus ist sie unmittelbar mit dem sportlichen Erfolg verknüpft, da eine positive Tordifferenz eine notwendige Voraussetzung für einen Sieg darstellt. Im Folgenden wird die Tordifferenz daher als Maß für die Leistungsstärke eines Teams verwendet. Ziel ist es, jene Leistungsindikatoren zu identifizieren, die diese Größe möglichst präzise prognostizieren können. Hierzu wird im nächsten Schritt auf Methoden der Korrelationsanalyse zurückgegriffen.

Ein zentrales Instrument stellt dabei der Korrelationskoeffizient nach Bravais-Pearson dar, der die Stärke und Richtung des linearen Zusammenhangs zwischen zwei Zufalls-

variablen misst. Nach Behrends [8] ist der empirische Korrelationskoeffizient für zwei Größen X und Y definiert durch

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

wobei vorausgesetzt wird, dass nicht alle $x_i = \bar{x}$ und nicht alle $y_i = \bar{y}$ gelten. Der Korrelationskoeffizient nimmt Werte im Intervall $[-1, 1]$ an. Dabei steht $r_{XY} = 1$ für einen perfekten positiven linearen Zusammenhang, $r_{XY} = -1$ für einen perfekten negativen Zusammenhang und $r_{XY} = 0$ für das Fehlen eines linearen Zusammenhangs; in letzterem Fall bezeichnet man die Variablen als unkorreliert [48].

Im Kontext der vorliegenden Untersuchung dient der Korrelationskoeffizient dazu, den Zusammenhang zwischen einer Kenngröße und der Tordifferenz als Zielgröße zu analysieren. Auf diese Weise lässt sich beurteilen, inwieweit einzelne Leistungsindikatoren zur Modellierung der Leistungsstärke und damit zur Vorhersage zukünftiger Spielergebnisse geeignet sind. Damit kann ein systematischer Vergleich verschiedener potenzieller Einflussgrößen hinsichtlich ihrer Prognosekraft angestellt werden. Eine hohe Prognosequalität äußert sich dabei in einer starken Korrelation zwischen dem Mittelwert der betrachteten Kenngröße in der ersten Saisonhälfte und der mittleren Tordifferenz in der zweiten Saisonhälfte [38].

Empirische Untersuchungen zeigen, dass die Tordifferenz selbst bereits eine hohe Prognosegüte aufweist. Insbesondere besitzt die Tordifferenz der ersten Saisonhälfte eine höhere Korrelation mit der Tordifferenz der zweiten Saisonhälfte als alternative Indikatoren wie die reine Anzahl erzielter Tore oder die Punkteausbeute [36]. Darüber hinaus lässt sich feststellen, dass die Leistungsstärke eines Teams, gemessen über die Tordifferenz, innerhalb einer Saison vergleichsweise stabil ist und sich nicht systematisch verändert, während über mehrere Spielzeiten hinweg deutlich größere Schwankungen auftreten können [27]. Veränderungen der Teamstärke erfolgen dabei überwiegend in der Sommerpause und deutlich seltener während der laufenden Saison oder in der Winterpause. Langfristige Analysen verdeutlichen zudem, dass die Korrelation der Teamstärken mit zunehmendem zeitlichen Abstand zwischen den betrachteten Saisons zwar abnimmt, dieser Rückgang jedoch vergleichsweise langsam erfolgt. Selbst über mehrere Jahrzehnte hinweg bleibt eine positive Korrelation bestehen, was darauf hindeutet, dass strukturelle Faktoren wie wirtschaftliche Ressourcen und institutionelle Rahmenbedingungen einen nachhaltigen Einfluss auf die sportliche Leistungsfähigkeit von Teams ausüben. Innerhalb einer Saison nimmt der Informationsgehalt der beobachteten Daten im Zeitverlauf weiter zu, sodass Schätzungen der Teamstärke mit fortschreitender Spieldauer präziser werden. Gleichzeitig bleibt die geschätzte Leistungsstärke im Saisonverlauf relativ stabil, da der Bereich plausibler Werte bereits durch Vorinformationen wie Kaderqualität oder Marktwert eingeschränkt ist. Insgesamt erweist sich die Tordifferenz somit sowohl kurz- als auch mittelfristig als eine stabile und informationsreiche Größe zur Beschreibung der Leistungsstärke eines Teams [36].

In der Literatur wird in diesem Zusammenhang gezeigt, dass neben der Tordifferenz auch

differenziertere Kenngrößen eine höhere Prognosekraft aufweisen können. Insbesondere ergibt eine Korrelationsanalyse von Heuer [36], dass die Differenz der Torchancen in der ersten Saisonhälfte stärker mit der Tordifferenz in der zweiten Saisonhälfte korreliert als die Tordifferenz selbst. Darüber hinaus wird diese Kenngröße weiter verfeinert, indem die Effizienz der Chancenverwertung berücksichtigt wird. Die daraus resultierende sogenannte effektive Torchancendifferenz, bei der ein Faktor zur durchschnittlichen Verwertungsquote integriert wird, erzielt im Vergleich die höchste Korrelation und weist somit die größte Prognosekraft auf.

Diese Ergebnisse legen nahe, dass Kenngrößen, die über reine Torstatistiken hinausgehen und die Qualität von Torchancen berücksichtigen, einen zusätzlichen Informationsgehalt für die Modellierung der Leistungsstärke besitzen. Daraus ergibt sich die Fragestellung, inwiefern derartige erweiterte Leistungsmetriken zur Verbesserung der Modellgüte beitragen können. Vor diesem Hintergrund rücken insbesondere moderne sportanalytische Kennzahlen in den Fokus, die auf umfangreichen Datenerhebungen basieren und eine differenziertere Bewertung von Spielsituationen ermöglichen. Ein zentrales Beispiel hierfür ist der Expected-Goals-Wert (xG), der die Qualität von Torchancen quantifiziert und damit eine verfeinerte Beschreibung der Teamleistung erlaubt. Auf diese Kenngröße wird im folgenden Abschnitt detailliert eingegangen.

5.4.2 Integration von sportanalytischen Metriken als Modellgrundlage

Im Zuge der zunehmenden Digitalisierung und Verfügbarkeit umfangreicher Spieldaten haben sich im Fußball neue analytische Kennzahlen etabliert, die über klassische Ergebnisstatistiken hinausgehen. Große Datenanbieter wie Opta erfassen für jedes Spiel eine Vielzahl an Ereignissen, insbesondere auch sämtliche Torabschlüsse. Diese werden hinsichtlich ihrer Qualität bewertet, indem jedem Schuss eine Wahrscheinlichkeit zugeordnet wird, mit der er zu einem Torerfolg führt [74].

Der sogenannte Expected-Goals-Wert (xG) stellt in diesem Zusammenhang ein probabilistisches Maß zur Bewertung von Torchancen dar. Er gibt die Wahrscheinlichkeit an, mit der ein bestimmter Abschluss zu einem Tor führt, basierend auf historischen Daten vergleichbarer Spielsituationen. Formal nimmt xG Werte im Intervall $[0, 1]$ an, wobei beispielsweise ein Wert von 0,1 bedeutet, dass ein entsprechender Abschluss im Mittel in 10 % der Fälle erfolgreich ist. Die Berechnung erfolgt mithilfe statistischer beziehungsweise maschineller Lernverfahren, die auf umfangreichen Datensätzen trainiert werden [93]. Dabei werden verschiedene Einflussfaktoren berücksichtigt, die die Qualität einer Torchance maßgeblich bestimmen. Hierzu zählen insbesondere die Distanz und der Winkel zum Tor, die Position des Torhüters, das Sichtfeld sowie der Defensivdruck, die Art des Abschlusses und die konkrete Spielsituation [62]. Durch die Kombination dieser Variablen wird für jede Abschlusssituation eine individuelle Torwahrscheinlichkeit geschätzt.

Ein wesentlicher Vorteil des xG-Ansatzes liegt darin, dass nicht lediglich die Anzahl der erzielten Tore betrachtet wird, sondern die Qualität der herausgespielten Chancen in die

Analyse einfließt. Da der Fußball von vergleichsweise wenigen Torereignissen lebt und Ergebnisse daher stärkeren Zufallsschwankungen unterliegen, erlaubt die Betrachtung von Torchancen eine robustere Bewertung der Teamleistung [61]. In diesem Sinne tragen Expected-Goals zu einer objektiveren Einschätzung der Spielstärke bei, indem sie die zugrunde liegende Spielweise eines Teams besser abbilden als rein ergebnisbasierte Kennzahlen [62].

Auch empirische Studien unterstreichen die hohe Aussagekraft von xG-basierten Metriken. So zeigt Heuer [37], dass sowohl Torchancen als auch Expected-Goals im Vergleich zur tatsächlichen Toranzahl einen höheren Informationsgehalt besitzen, was insbesondere auf die geringere Bedeutung zufälliger Einflüsse zurückzuführen ist. Darüber hinaus weisen Expected-Goals eine überlegene Prognosekraft für zukünftige Spielergebnisse auf und ermöglichen eine genauere Vorhersage des sportlichen Erfolgs im Vergleich zu traditionellen Kennzahlen [61]. Die Eignung von xG-Daten zur Modellierung und Prognose von Spielergebnissen wird zudem durch verschiedene modellbasierte Ansätze bestätigt. So zeigen Eggels et al. [23], dass auf xG-Daten basierende probabilistische Modelle eine gute Kalibrierung von Ergebniswahrscheinlichkeiten erreichen. Auch Erweiterungen klassischer Poisson-Modelle durch die Integration von xG-Werten führen zu leicht verbesserten Vorhersageergebnissen im Vergleich zu Modellen, die ausschließlich auf Torstatistiken beruhen [86]. Weitere Untersuchungen belegen, dass Modelle, die auf xG-Daten basieren und zusätzliche Faktoren wie den Heimvorteil berücksichtigen, eine hohe Prognosegüte aufweisen und sogar praktische Anwendungen im Bereich von Wettstrategien ermöglichen [71]. Schließlich zeigen auch empirische Analysen auf aggregierter Ebene, dass Expected-Goals zur Vorhersage langfristiger Erfolgsgrößen geeignet sind. So konnte etwa in einer Untersuchung zur englischen Premier League gezeigt werden, dass ein Großteil der Mannschaften anhand ihrer xG-Werte mit hoher Genauigkeit hinsichtlich ihrer Endplatzierung prognostiziert werden kann [62].

Insgesamt verdeutlichen diese Ergebnisse, dass Expected-Goals eine zentrale und leistungsfähige Kenngröße zur Beschreibung der Teamstärke darstellen. Dies legt nahe, die Prognosekraft der Tordifferenz und des xG-Wertes systematisch miteinander zu vergleichen. Zu diesem Zweck wird die Expected-Goals-Differenz (xGD) als Differenz aus Expected-Goals (xG) und Expected-Goals-Against (xGA) gebildet. Anschließend wird die Korrelation zwischen dieser Kenngröße in der ersten Saisonhälfte und der Tordifferenz als Maß der Leistungsstärke in der zweiten Saisonhälfte anhand der aktuellen Bundesliga-Saison bestimmt. Zum Vergleich wird analog die Korrelation zwischen der Tordifferenz der ersten und der zweiten Saisonhälfte berechnet. Auf diese Weise lässt sich beurteilen, welche der beiden Kenngrößen eine bessere Vorhersagequalität im Hinblick auf die zukünftige Leistungsstärke eines Teams aufweist.

Die zugrunde liegenden Daten wurden den offiziellen Statistiken der Bundesliga sowie xG-basierten Auswertungen entnommen [20, 91]. Betrachtet werden die ersten 17 Spieltage der ersten Saisonhälfte sowie die Spieltage 18 bis 29 der zweiten Saisonhälfte. Da die betrachteten Saisonabschnitte unterschiedlich viele Spiele umfassen, werden sämtliche Größen auf Mittelwerte pro Spiel normiert. Diese lineare Transformation beeinflusst den

Korrelationskoeffizienten nicht, gewährleistet jedoch eine bessere Vergleichbarkeit. Zur besseren Übersicht werden die mittlere Tordifferenz der ersten Saisonhälfte mit $\overline{\mathrm{TD}}_1$, die mittlere Expected-Goals-Differenz mit $\overline{\mathrm{xGD}}_1$ sowie die mittlere Tordifferenz der zweiten Saisonhälfte mit $\overline{\mathrm{TD}}_2$ bezeichnet. Die entsprechenden Werte sind in der nachfolgenden Tabelle 5.3 dargestellt.

Tabelle 5.3: Gegenüberstellung der Tordifferenz und Expected-Goals-Differenz der ersten Saisonhälfte mit der Tordifferenz der zweiten Saisonhälfte

Team	$\overline{\mathrm{TD}}_1$	$\overline{\mathrm{xGD}}_1$	$\overline{\mathrm{TD}}_2$
Bayern	3,12	2,26	2,08
Dortmund	1,00	0,65	1,17
Hoffenheim	0,88	0,35	-0,08
Leipzig	0,76	0,59	0,58
Leverkusen	0,65	0,68	0,75
Stuttgart	0,41	0,39	1,25
Frankfurt	-0,06	0,01	0,08
Freiburg	-0,12	0,05	-0,25
Union	-0,18	0,17	-1,17
M'gladbach	-0,35	-0,23	-0,67
Wolfsburg	-0,65	-0,41	-1,25
Köln	-0,24	-0,30	-0,25
Bremen	-0,88	-0,66	-0,42
HSV	-0,65	-0,40	-0,17
Augsburg	-0,88	-0,74	-0,17
St. Pauli	-0,82	-0,99	-0,92
Mainz	-0,71	-0,67	0,25
Heidenheim	-1,29	-0,87	-0,83

Auf Basis dieser Daten wird der Korrelationskoeffizient nach Bravais-Pearson zwischen den jeweiligen Kenngrößen der ersten Saisonhälfte und der Tordifferenz der zweiten Saisonhälfte berechnet. Dabei werden die 18 Teams als Beobachtungseinheiten aufgefasst. Zur Veranschaulichung der Berechnung wird exemplarisch ein Summand des Zählers betrachtet. Für den FC Bayern München ergeben sich aus Tabelle 5.3 die Werte

$$x_1 = 3{,}12, \qquad y_1 = 2{,}08.$$

Für die Mittelwerte gilt aufgrund der Struktur der Tordifferenzen

$$\bar{x}_{\mathrm{TD}_1} = 0, \qquad \bar{y}_{\mathrm{TD}_2} = 0.$$

Damit vereinfacht sich der Beitrag dieses Teams im Zähler zu

$$(x_i - \bar{x})(y_i - \bar{y}) = x_i \cdot y_i.$$

Analog werden für alle 18 Teams die entsprechenden Produkte berechnet und aufsummiert. Der Nenner wird durch die quadrierten Abweichungen der jeweiligen Variablen bestimmt. Insgesamt erhält man damit den Korrelationskoeffizienten gemäß

$$r_{XY} = \frac{\sum_{i=1}^{18}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{18}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{18}(y_i - \bar{y})^2}}.$$

Für die Expected-Goals-Differenz ergibt sich aufgrund von Rundungsabweichungen ein leicht von Null abweichender Mittelwert von

$$\bar{x}_{\mathrm{xGD}_1} \approx -0{,}12,$$

wobei diese Abweichung keinen wesentlichen Einfluss auf das Ergebnis der Korrelationsanalyse hat. Für die vorliegenden Daten ergeben sich die folgenden Korrelationskoeffizienten:

$$r_{\overline{\mathrm{TD}_1},\overline{\mathrm{TD}_2}} \approx 0{,}75,$$
$$r_{\overline{\mathrm{xGD}_1},\overline{\mathrm{TD}_2}} \approx 0{,}84.$$

Der Vergleich dieser Werte zeigt, dass die Expected-Goals-Differenz der ersten Saisonhälfte eine höhere Korrelation mit der Tordifferenz der zweiten Saisonhälfte aufweist als die Tordifferenz selbst. Dies deutet darauf hin, dass die Berücksichtigung der Qualität von Torchancen zu einer verbesserten Prognose der zukünftigen Leistungsstärke eines Teams führt. Die Ergebnisse stehen damit im Einklang mit den in der Literatur beschriebenen Befunden, wonach xG-basierte Kennzahlen einen höheren Informationsgehalt besitzen und zufällige Schwankungen in den tatsächlichen Torergebnissen teilweise ausgleichen können. Gleichzeitig zeigt die weiterhin vergleichsweise hohe Korrelation der Tordifferenz, dass auch diese Kenngröße eine solide Grundlage zur Beschreibung der Teamstärke darstellt.

Insgesamt sprechen die Ergebnisse dafür, xG-basierte Größen in die weitere Modellierung einzubeziehen, da sie einen zusätzlichen Informationsgewinn liefern und die Prognosegüte gegenüber rein torbasierten Kennzahlen verbessern. Es ist jedoch zu beachten, dass die vorliegende Analyse auf einer einzelnen Saison sowie einer vergleichsweise kleinen Stichprobe von 18 Teams basiert. Die Ergebnisse sollten daher mit entsprechender Vorsicht interpretiert werden und bedürfen zur weiteren Absicherung einer Untersuchung über mehrere Spielzeiten hinweg.

5.5 In fünf Schritten zur Spielvorhersage

In diesem Kapitel wird das entwickelte Prognosemodell exemplarisch auf ein konkretes Bundesligaspiel angewendet. Das Vorgehen erfolgt in zwei Schritten: Zunächst werden auf Basis vergangener Spieldaten sowie daraus abgeleiteter sportanalytischer Kennzahlen geeignete Größen erhoben, aus denen modellrelevante Parameter zur Beschreibung der Spielstärke bestimmt werden. Anschließend werden diese in ein Poisson-Modell

überführt, um Wahrscheinlichkeiten für konkrete Spielausgänge zu berechnen. Im Mittelpunkt stehen dabei xG-basierte Kenngrößen, da im vorangehenden Kapitel bereits begründet wurde, dass sie im Vergleich zu rein torbasierten Größen einen höheren Informationsgehalt für die Modellierung der Teamstärke besitzen. Für das hier verwendete, bewusst kompakt gehaltene Modell werden die Expected-Goals-Differenz, die Expected-Goals-Summe sowie der Heimvorteil als zentrale Einflussgrößen berücksichtigt. Ziel ist es, eine nachvollziehbare und zugleich hinreichend realitätsnahe Modellstruktur zu erhalten, die wesentliche leistungsrelevante Aspekte des Fußballspiels abbildet.

Zur Veranschaulichung des Prognoseverfahrens wird die Begegnung zwischen der TSG Hoffenheim (Team A) und Borussia Dortmund (Team B) aus der 30. Runde der Bundesliga-Saison 2025/26 betrachtet. Die verwendeten Daten stammen von der offiziellen Bundesliga-Website sowie von Plattformen mit sportanalytischen Kennzahlen [20, 64, 67, 91]. Eine zeitliche Gewichtung der erhobenen Kennzahlen wird in diesem vereinfachten Modell nicht vorgenommen. Dies erscheint im vorliegenden Kontext vertretbar, da Fischer und Heuer [27] keinen signifikanten Unterschied zwischen gleichgewichteten und stärker auf jüngere Spiele fokussierten Auswertungen feststellt, sofern ausschließlich Daten aus der aktuellen Saison betrachtet werden. Das gewählte Vorgehen orientiert sich dabei an dem Prognosemodell von Heuer [36].

Schritt 1: Expected-Goals-Summe

Im ersten Schritt wird die für das Spiel zu erwartende xG-Gesamtsumme bestimmt. In Anlehnung an Heuer et al. [39] wird diese durch

$$xGS_{A,B} = \overline{xGS}_A + \overline{xGS}_B - \overline{xGS}$$

modelliert. Dabei bezeichnet $xGS_{A,B}$ die erwartete xG-Summe des konkreten Spiels, $\overline{xGS}_A$ und $\overline{xGS}_B$ die mittleren xG-Summen pro Spiel der beiden Teams und $\overline{xGS}$ die durchschnittliche xG-Summe eines Bundesligaspiels.

Für Hoffenheim ergeben sich nach 29 Spieltagen ein xG-Wert von 47,01 und ein xGA-Wert von 44,99. Daraus folgt

$$\overline{xGS}_A = \frac{47{,}01 + 44{,}99}{29} \approx 3{,}1724.$$

Für Dortmund mit $xG = 54{,}63$ und $xGA = 32{,}87$ erhält man analog

$$\overline{xGS}_B = \frac{54{,}63 + 32{,}87}{29} \approx 3{,}0172.$$

Insgesamt wurden in den ersten 29 Runden 261 Bundesligaspiele ausgetragen, wobei sich ein kumulierter xG-Wert aller Teams von 803,71 ergibt. Somit gilt

$$\overline{xGS} = \frac{803{,}71}{261} \approx 3{,}0793.$$

Damit folgt für die zu erwartende xG-Summe des Spiels Hoffenheim gegen Dortmund

$$xGS_{A,B} \approx 3{,}1724 + 3{,}0172 - 3{,}0793 = 3{,}1103.$$

Schritt 2: Expected-Goals-Differenz

Zur Aufteilung dieser Gesamtsumme auf die beiden Teams wird im nächsten Schritt die xG-Differenz herangezogen. In Anlehnung an Heuer [36] wird sie durch

$$xGD_{A,B} = \overline{xGD}_A - \overline{xGD}_B$$

bestimmt. Dabei stehen $\overline{xGD}_A$ und $\overline{xGD}_B$ für die mittlere xG-Differenz pro Spiel der beiden Teams. Für Hoffenheim ergibt sich mit $xG = 47{,}01$ und $xGA = 44{,}99$

$$\overline{xGD}_A = \frac{47{,}01 - 44{,}99}{29} \approx 0{,}0697.$$

Für Dortmund erhält man mit $xG = 54{,}63$ und $xGA = 32{,}87$

$$\overline{xGD}_B = \frac{54{,}63 - 32{,}87}{29} \approx 0{,}7503.$$

Somit folgt

$$xGD_{A,B} \approx 0{,}0697 - 0{,}7503 = -0{,}6806.$$

Die negative xG-Differenz zeigt, dass Borussia Dortmund im direkten Vergleich eine höhere Spielstärke aufweist. Konkret bedeutet dieser Wert, dass im Mittel ein Vorteil von etwa 0,68 erwarteten Toren zugunsten von Borussia Dortmund gegenüber den Gastgebern besteht, sofern keine weiteren Einflussfaktoren berücksichtigt werden. Mit dem Gastgeber ist zugleich ein weiterer modellrelevanter Faktor angesprochen, der diese Einschätzung relativieren kann. Im Kapitel 2.7 wurde das Phänomen des Heimvorteils bereits eingeführt, welches im Folgenden in das Prognosemodell integriert wird.

Schritt 3: Heimvorteil

Da Hoffenheim im betrachteten Spiel Heimrecht besitzt, wird nun der Heimvorteil quantifiziert. In Anlehnung an Heuer und Rubner [40] wird dieser durch

$$xHV = \overline{xGD}_{\text{Heim}} - \overline{xGD}_{\text{Auswärts}}$$

erfasst. Dabei bezeichnen $\overline{xGD}_{\text{Heim}}$ und $\overline{xGD}_{\text{Auswärts}}$ die mittleren xG-Differenzen Hoffenheims in Heim- bzw. Auswärtsspielen. In den bisherigen 29 Saisonspielen absolvierte Hoffenheim 14 Heim- und 15 Auswärtsspiele. In Heimspielen lagen der mittlere xG-Wert bei 1,87 und der mittlere xGA-Wert bei 1,30, woraus

$$\overline{xGD}_{\text{Heim}} = 1{,}87 - 1{,}30 = 0{,}57$$

folgt. Auswärts ergeben sich ein mittlerer xG-Wert von 1,40 und ein mittlerer xGA-Wert von 1,50, also

$$\overline{xGD}_{\text{Auswärts}} = 1{,}40 - 1{,}50 = -0{,}10.$$

Damit erhält man für den Heimvorteil Hoffenheims

$$xHV = 0{,}57 - (-0{,}10) = 0{,}67.$$

Der resultierende Heimvorteil von $xHV = 0{,}67$ bedeutet, dass Hoffenheim in Heimspielen im Durchschnitt eine um 0,67 höhere Expected-Goals-Differenz aufweist als in Auswärtsspielen. Da sich dieser Effekt im Modell symmetrisch auf beide Teams auswirken soll, wird die Hälfte des berechneten Wertes, also $\frac{0{,}67}{2} = 0{,}335$, der zuvor bestimmten xG-Differenz zugeschlagen. Dadurch ergibt sich die heimvorteilsbereinigte xG-Differenz

$$xGD^*_{A,B} = xGD_{A,B} + \frac{xHV}{2} \approx -0{,}6806 + 0{,}335 = -0{,}3456.$$

Durch die Berücksichtigung des Heimvorteils reduziert sich der zuvor bestehende Vorteil von Borussia Dortmund deutlich, bleibt jedoch weiterhin bestehen.

Schritt 4: Modellparameter

Die zuvor berechnete xG-Summe $xGS_{A,B}$ und die angepasste xG-Differenz $xGD^*_{A,B}$ ermöglichen nun die Ableitung der erwarteten Toranzahlen beim Aufeinandertreffen der beiden Teams. Unter der Annahme zweier gleich starker Mannschaften würde die xG-Summe zunächst gleichmäßig auf beide Teams verteilt, also zu

$$\frac{xGS_{A,B}}{2} = \frac{3{,}1103}{2} = 1{,}55515.$$

Unter Berücksichtigung der xG-Differenz folgt daraus:

$$\lambda_A = \frac{xGS_{A,B}}{2} + \frac{xGD^*_{A,B}}{2} \approx 1{,}55515 + \frac{-0{,}3456}{2} = 1{,}38235,$$
$$\lambda_B = \frac{xGS_{A,B}}{2} - \frac{xGD^*_{A,B}}{2} \approx 1{,}55515 - \frac{-0{,}3456}{2} = 1{,}72795.$$

Die Werte λ_A und λ_B entsprechen den erwarteten Toranzahlen der beiden Teams im betrachteten Spiel. Dies bedeutet, dass Hoffenheim unter identischen Bedingungen im Mittel etwa 1,38 Tore erzielen würde, während für Borussia Dortmund ein Erwartungswert von etwa 1,73 Toren resultiert. Interpretativ lässt sich dies so verstehen, dass bei einer großen Anzahl hypothetischer Wiederholungen dieser Begegnung die durchschnittliche Toranzahl der beiden Teams gegen diese Werte konvergieren würde.

Im Folgenden werden diese beiden Größen als Modellparameter der Poisson-Verteilungen verwendet. Mit A sei die Zufallsvariable für die Anzahl der von Hoffenheim erzielten Tore und mit B jene für die Tore Dortmunds bezeichnet. Dann gilt:

$$P(A = k_A) = \frac{\lambda_A^{k_A} \cdot e^{-\lambda_A}}{k_A!} = \frac{1{,}38235^{k_A} \cdot e^{-1{,}38235}}{k_A!}, \quad k_A \in \mathbb{N}_0,$$

$$P(B = k_B) = \frac{\lambda_B^{k_B} \cdot e^{-\lambda_B}}{k_B!} = \frac{1{,}72795^{k_B} \cdot e^{-1{,}72795}}{k_B!}, \quad k_B \in \mathbb{N}_0.$$

Da die Unabhängigkeitsannahme im vorangegangenen Kapitel 5.1.2 bereits fachlich begründet wurde, können die beiden Einzelwahrscheinlichkeiten multipliziert werden. Für einen konkreten Endstand $k_A : k_B$ erhält man somit allgemein

$$P(A = k_A, B = k_B) = P(A = k_A) \cdot P(B = k_B)$$

und im vorliegenden Modell

$$P(A = k_A, B = k_B) = \frac{1{,}38235^{k_A} \cdot e^{-1{,}38235}}{k_A!} \cdot \frac{1{,}72795^{k_B} \cdot e^{-1{,}72795}}{k_B!}.$$

Schritt 5: Wahrscheinlichkeiten konkreter Spielausgänge

Auf Basis der Modellparameter lassen sich zunächst die Einzelwahrscheinlichkeiten für die Toranzahlen beider Teams bestimmen. Für Hoffenheim ergeben sich:

$$\begin{aligned}
&P(A = 0) \approx 0{,}2510, \quad &&P(A = 1) \approx 0{,}3470, \quad &&P(A = 2) \approx 0{,}2398,\\
&P(A = 3) \approx 0{,}1105, \quad &&P(A = 4) \approx 0{,}0382, \quad &&P(A = 5) \approx 0{,}0106,\\
&P(A = 6) \approx 0{,}0024, \quad &&P(A = 7) \approx 0{,}0005, \quad &&P(A \geq 8) \approx 0{,}0001.
\end{aligned}$$

Für Dortmund erhält man:

$$\begin{aligned}
&P(B = 0) \approx 0{,}1776, \quad &&P(B = 1) \approx 0{,}3070, \quad &&P(B = 2) \approx 0{,}2652,\\
&P(B = 3) \approx 0{,}1528, \quad &&P(B = 4) \approx 0{,}0660, \quad &&P(B = 5) \approx 0{,}0228,\\
&P(B = 6) \approx 0{,}0066, \quad &&P(B = 7) \approx 0{,}0016, \quad &&P(B \geq 8) \approx 0{,}0004.
\end{aligned}$$

Die Verteilungen der Einzelwahrscheinlichkeiten zeigen, dass für beide Teams insbesondere geringe Trefferzahlen die höchste Eintrittswahrscheinlichkeit besitzen. Für Hoffenheim ist mit $P(A = 1) \approx 0{,}3470$ ein einzelner Treffer am wahrscheinlichsten, gefolgt von keinem Tor mit $P(A = 0) \approx 0{,}2510$. Für Borussia Dortmund stellt ebenfalls ein Tor mit $P(B = 1) \approx 0{,}3070$ die wahrscheinlichste Ausprägung dar, gefolgt von zwei Treffern mit $P(B = 2) \approx 0{,}2652$.

Auffällig ist, dass sich die Wahrscheinlichkeitsmasse bei beiden Teams auf niedrige Torzahlen konzentriert, während die Verteilung für Dortmund im Vergleich zu Hoffenheim leicht zu höheren Trefferzahlen verschoben ist. Aufgrund der sehr geringen Einzelwahrscheinlichkeiten für hohe Trefferzahlen werden alle Ergebnisse mit mindestens acht Toren zu einer gemeinsamen Klasse ≥ 8 zusammengefasst, um die Darstellung übersichtlich zu halten, ohne die Aussagekraft der Verteilung wesentlich zu beeinträchtigen.

Durch Multiplikation dieser Einzelwahrscheinlichkeiten ergeben sich die Wahrscheinlichkeiten der einzelnen Endstände. Tabelle 5.4 (S. 114) zeigt die auf zwei Nachkommastellen gerundeten Prozentwerte für die jeweiligen Spielausgänge. Siege Hoffenheims sind blau, Siege Dortmunds rot und Unentschieden grün markiert.

Tabelle 5.4: Wahrscheinlichkeiten der Spielausgänge zwischen der TSG Hoffenheim und Borussia Dortmund

	Hoffenheim	$k_A = 0$	1	2	3	4	5	6	7	≥ 8	Summe
Dortmund		0,2510	0,3470	0,2398	0,1105	0,0382	0,0106	0,0024	0,0005	0,0001	1
$k_B = 0$	0,1776	4,46%	6,16%	4,26%	1,96%	0,68%	0,19%	0,04%	0,01%	0,00%	
1	0,3070	7,70%	10,65%	7,36%	3,39%	1,17%	0,32%	0,07%	0,01%	0,00%	
2	0,2652	6,66%	9,20%	6,36%	2,93%	1,01%	0,28%	0,06%	0,01%	0,00%	
3	0,1528	3,83%	5,30%	3,66%	1,69%	0,58%	0,16%	0,04%	0,01%	0,00%	
4	0,0660	1,66%	2,29%	1,58%	0,73%	0,25%	0,07%	0,02%	0,00%	0,00%	30,84%
5	0,0228	0,57%	0,79%	0,55%	0,25%	0,09%	0,02%	0,01%	0,00%	0,00%	
6	0,0066	0,16%	0,23%	0,16%	0,07%	0,03%	0,01%	0,00%	0,00%	0,00%	
7	0,0016	0,04%	0,06%	0,04%	0,02%	0,01%	0,00%	0,00%	0,00%	0,00%	
≥ 8	0,0004	0,01%	0,01%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
Summe	1	45,73%									23,43%

Die höchsten Einzelwahrscheinlichkeiten entfallen auf die Spielstände 1:1 mit 10,65 %, 1:2 mit 9,20 % und 0:1 mit 7,70 %. Insgesamt weist das Modell damit auf eine leichte Favoritenrolle Dortmunds hin. Die kumulierten Wahrscheinlichkeiten der drei Hauptausgänge lauten:

$$P(\text{Sieg Hoffenheim}) = \sum_{k_A > k_B} P(A = k_A, B = k_B) \approx 0{,}3084,$$

$$P(\text{Unentschieden}) = \sum_{k=0}^{\infty} P(A = k, B = k) \approx 0{,}2343,$$

$$P(\text{Sieg Dortmund}) = \sum_{k_A < k_B} P(A = k_A, B = k_B) \approx 0{,}4573.$$

In Prozent ausgedrückt entsprechen diese Werte einer Hoffenheimer Siegchance von 30,84 %, einer Remiswahrscheinlichkeit von 23,43 % und einer Dortmunder Siegchance von 45,73 %.

Neben konkreten Endständen lassen sich auch Wahrscheinlichkeiten für bestimmte Tordifferenzen angeben. Für einen Hoffenheimer Sieg mit genau r Toren Unterschied, also für den Fall $A = B + r$ mit $r \in \mathbb{N}$, gilt

$$P(A = B + r) = \sum_{k=0}^{\infty} P(A = k + r, B = k) = e^{-(\lambda_A + \lambda_B)} \sum_{k=0}^{\infty} \frac{\lambda_A^{k+r} \lambda_B^k}{(k+r)!\, k!}.$$

Analog erhält man für eine Niederlage Hoffenheims mit genau r Toren Unterschied

$$P(A = B - r) = \sum_{k=0}^{\infty} P(A = k, B = k + r) = e^{-(\lambda_A + \lambda_B)} \sum_{k=0}^{\infty} \frac{\lambda_A^k \lambda_B^{k+r}}{k!\, (k+r)!}.$$

Für ein Unentschieden ergibt sich als Spezialfall $r = 0$, also

$$P(A = B) = \sum_{k=0}^{\infty} P(A = k, B = k) = e^{-(\lambda_A + \lambda_B)} \sum_{k=0}^{\infty} \frac{(\lambda_A \lambda_B)^k}{(k!)^2}.$$

Im konkreten Modell mit $\lambda_A = 1{,}38235$ und $\lambda_B = 1{,}72795$ erhält man beispielsweise:

$$\begin{aligned} P(A = B) &\approx 0{,}2343, \\ P(A = B + 1) &\approx 0{,}1711, \\ P(A = B + 2) &\approx 0{,}0884, \\ P(A = B - 1) &\approx 0{,}2139, \\ P(A = B - 2) &\approx 0{,}1382. \end{aligned}$$

Damit ist ein Dortmunder Sieg mit einem Tor Differenz im Modell wahrscheinlicher als ein Hoffenheimer Sieg mit einem Tor Differenz. Zugleich zeigt sich, dass kleine Tordifferenzen deutlich häufiger auftreten als klare Siege, was gut zur typischen Torstruktur von Fußballspielen passt.

Darüber hinaus lassen sich auf Basis des Modells auch Wahrscheinlichkeiten für typische, aus dem Bereich der Sportwetten bekannte Ereignisse bestimmen. Von besonderem Interesse sind hierbei vor allem Wettkategorien wie „Beide Teams treffen“ sowie „Über-/Unter“-Wetten bezüglich der Gesamtanzahl erzielter Tore.

Die Wahrscheinlichkeit dafür, dass beide Teams mindestens ein Tor erzielen, ergibt sich zu

$$P(A \geq 1 \ \cap \ B \geq 1) = 1 - P(A = 0) - P(B = 0) + P(A = 0, B = 0).$$

Dabei werden alle Spielausgänge ausgeschlossen, in denen mindestens eines der beiden Teams kein Tor erzielt. Da das Ergebnis 0:0 in beiden Teilereignissen enthalten ist, muss es einmal wieder hinzuaddiert werden. Unter Verwendung der zuvor berechneten Einzelwahrscheinlichkeiten erhält man

$$P(\text{Beide Teams treffen}) \approx 1 - 0{,}2510 - 0{,}1776 + 0{,}0446 \approx 0{,}6160.$$

Für die Gesamtanzahl der Tore lassen sich ebenfalls entsprechende Wahrscheinlichkeiten bestimmen. So lässt sich beispielsweise die Wahrscheinlichkeit, dass mindestens drei Tore im Spiel fallen („Über 2,5 Tore“), mithilfe des Komplementärereignisses bestimmen:

$$P(A + B \geq 3) = 1 - P(A + B \leq 2) = 1 - \sum_{\substack{k_A, k_B \geq 0 \\ k_A + k_B \leq 2}} P(A = k_A, B = k_B).$$

Dabei werden alle Spielausgänge berücksichtigt, bei denen insgesamt höchstens zwei Tore erzielt werden, also 0:0, 1:0, 0:1, 2:0, 1:1 und 0:2. Daraus folgt

$$P(A + B \geq 3) \approx 0{,}6011.$$

Analog ergibt sich für die Wahrscheinlichkeit von mindestens vier Toren („Über 3,5 Tore“)

$$P(A+B \geq 4) = 1 - P(A+B \leq 3) = 1 - \sum_{\substack{k_A,k_B \geq 0 \\ k_A+k_B \leq 3}} P(A=k_A, B=k_B),$$

wobei zusätzlich alle Spielausgänge mit insgesamt drei Toren berücksichtigt werden, wie etwa 2:1, 1:2, 3:0 oder 0:3. Es ergibt sich

$$P(A+B \geq 4) \approx 0{,}3775.$$

Diese Werte verdeutlichen, dass im betrachteten Spiel mit einer relativ hohen Wahrscheinlichkeit beide Teams erfolgreich sein werden und insgesamt eine Toranzahl im Bereich des für Fußballspiele typischen Durchschnitts zu erwarten ist. Gleichzeitig nimmt die Wahrscheinlichkeit mit steigender Gesamtanzahl an Toren erwartungsgemäß ab, was der charakteristischen Struktur von Torverteilungen im Fußball entspricht.

Insgesamt zeigt das Beispiel, wie aus xG-basierten Kennzahlen, dem Heimvorteil und einem Poisson-Modell differenzierte Wahrscheinlichkeitsaussagen für einzelne Endstände, Hauptausgänge, Tordifferenzen sowie weitere wettmarktrelevante Ereignisse gewonnen werden können. Das Modell bleibt dabei bewusst einfach, bildet jedoch bereits zentrale leistungsrelevante Mechanismen ab und ermöglicht eine nachvollziehbare probabilistische Beschreibung des Spielausgangs.

5.5.1 Vergleich mit Buchmacherquoten

Zur weiteren Einordnung der Prognosegüte des entwickelten Modells bietet sich ein Vergleich mit Buchmacherquoten an. Da für Wettanbieter in der Regel insbesondere die Quoten für Sieg, Unentschieden und Niederlage öffentlich zugänglich sind, eignen sich für einen direkten Vergleich vor allem die kumulierten Wahrscheinlichkeiten dieser drei Spielausgänge. Eine solche Gegenüberstellung erscheint auch deshalb sinnvoll, weil Buchmacherquoten auf umfangreichen Datengrundlagen und komplexen Bewertungsmechanismen beruhen und damit eine naheliegende Referenz für die Beurteilung statistischer Prognosemodelle darstellen. Zudem zeigt Heuer [36], dass statistische Modelle und Wettmarktquoten hinsichtlich ihrer Vorhersagequalität grundsätzlich gut vergleichbar sind. Für die Begegnung zwischen der TSG Hoffenheim und Borussia Dortmund ergeben sich bei verschiedenen Wettanbietern die in Tabelle 5.5 (S. 117) dargestellten Quoten.

Um die Quoten mit den modellbasierten Wahrscheinlichkeiten vergleichen zu können, werden sie zunächst in implizite Wahrscheinlichkeiten umgerechnet. Dabei ist zu beachten, dass Buchmacherquoten nicht unmittelbar den tatsächlichen Eintrittswahrscheinlichkeiten entsprechen, sondern eine Gewinnmarge enthalten. Die Quoten werden von den Anbietern in der Regel auf Basis geschätzter Eintrittswahrscheinlichkeiten berechnet, deren Summe zunächst 1 ergibt und somit den sogenannten fairen Quoten entspricht. Anschließend werden diese so angepasst, dass ein Teil der Einsätze als Gewinn

Tabelle 5.5: Buchmacherquoten für die Spielausgänge Hoffenheim – Dortmund

Anbieter	Sieg Hoffenheim	Unentschieden	Sieg Dortmund
Tipico	2,65	3,90	2,40
bet365	2,55	4,00	2,50
bet-at-home	2,70	3,85	2,45
bwin	2,60	3,75	2,45

beim Buchmacher verbleibt. Dies führt dazu, dass die ausgegebenen Quoten systematisch unter den fairen Quoten liegen und die Summe der impliziten Wahrscheinlichkeiten folglich größer als 1 ist. Die Ausschüttungsquote liegt dabei im Regelfall zwischen etwa 92 % und 96 % [13, 81]. Beispielsweise würde bei einer Ausschüttungsquote von 95 % jede faire Quote mit dem Faktor 0,95 multipliziert werden, sodass die verbleibenden 5 % der Einsätze als Gewinn beim Buchmacher verbleiben.

Für einen Spielausgang mit Quote q_i ergibt sich zunächst der Kehrwert $\frac{1}{q_i}$. Da die Summe dieser Kehrwerte aufgrund des Margenaufschlags im Allgemeinen größer als 1 ist, werden die so erhaltenen Werte normiert. Für die implizite Wahrscheinlichkeit p_i gilt somit

$$p_i = \frac{\frac{1}{q_i}}{\sum_{j=1}^{3} \frac{1}{q_j}}.$$

Exemplarisch erhält man für die Tipico-Quoten zunächst:

$$\frac{1}{2{,}65} \approx 0{,}3774, \qquad \frac{1}{3{,}90} \approx 0{,}2564, \qquad \frac{1}{2{,}40} \approx 0{,}4167.$$

Da die Summe dieser Werte

$$0{,}3774 + 0{,}2564 + 0{,}4167 \approx 1{,}0505$$

beträgt, folgt nach Normierung:

$$P_{\text{Tipico}}(\text{Sieg Hoffenheim}) \approx \frac{0{,}3774}{1{,}0505} \approx 0{,}3592,$$
$$P_{\text{Tipico}}(\text{Unentschieden}) \approx \frac{0{,}2564}{1{,}0505} \approx 0{,}2441,$$
$$P_{\text{Tipico}}(\text{Sieg Dortmund}) \approx \frac{0{,}4167}{1{,}0505} \approx 0{,}3967.$$

Analog ergeben sich für die übrigen Anbieter die in Tabelle 5.6 (S. 118) aufgeführten impliziten Wahrscheinlichkeiten. Zusätzlich ist dort den Buchmacherwerten die modellbasierte Prognose gegenübergestellt.

Der Vergleich zeigt, dass sich das entwickelte Modell und die Buchmacher in der grundsätzlichen Einschätzung des Spiels nicht widersprechen: In beiden Fällen wird Borussia

Tabelle 5.6: Vergleich der modellbasierten Wahrscheinlichkeiten mit den aus den Buchmacherquoten abgeleiteten impliziten Wahrscheinlichkeiten

Quelle	Sieg Hoffenheim	Unentschieden	Sieg Dortmund
Modell	30,84 %	23,43 %	45,73 %
Tipico	35,92 %	24,41 %	39,67 %
bet365	37,63 %	23,99 %	38,38 %
bet-at-home	35,67 %	25,02 %	39,31 %
bwin	36,30 %	25,17 %	38,53 %
Mittelwert der Buchmacher	36,38 %	24,65 %	38,97 %

Dortmund als leicht favorisiert angesehen. Zugleich unterscheiden sich die konkreten Wahrscheinlichkeiten in ihrer Gewichtung. Während die Buchmacher Hoffenheim eine Siegchance von im Mittel 36,38 % zuweisen, liegt der entsprechende Modellwert mit 30,84 % deutlich darunter. Umgekehrt bewertet das Modell einen Auswärtssieg Dortmunds mit 45,73 % merklich höher als die betrachteten Wettanbieter, die hierfür im Mittel lediglich 38,97 % ansetzen. Die Remiswahrscheinlichkeit wird dagegen sehr ähnlich eingeschätzt; die Differenz zwischen Modell und dem Mittelwert der Buchmacher beträgt hier nur rund 1,22 Prozentpunkte. Dabei ist zu berücksichtigen, dass die aus den Quoten abgeleiteten Wahrscheinlichkeiten aufgrund der enthaltenen Buchmachermarge systematisch verzerrt sind und daher lediglich als Näherung an die tatsächlichen Markterwartungen interpretiert werden können.

Insgesamt spricht diese Gegenüberstellung dafür, dass das Modell die Begegnung etwas stärker zugunsten Dortmunds bewertet als die Buchmacherquoten. Das Grundmuster der Prognose bleibt jedoch vergleichbar. Damit fügt sich das Ergebnis in die Beobachtung ein, dass statistische Modelle und Wettmarktprognosen häufig ähnliche Tendenzen aufweisen, sich in der Stärke einzelner Wahrscheinlichkeitszuweisungen jedoch unterscheiden können [36].

5.5.2 Einordnung des tatsächlichen Spielausgangs

Eine abschließende Einordnung des entwickelten Prognosemodells kann exemplarisch auch im Abgleich mit dem tatsächlich eingetretenen Spielergebnis erfolgen. Dabei ist jedoch zu betonen, dass sowohl die modellbasierten Wahrscheinlichkeiten als auch die aus den Buchmacherquoten abgeleiteten Werte keine deterministischen Vorhersagen darstellen, sondern stochastische Erwartungsgrößen. Ihre Prognosegüte lässt sich daher streng genommen nicht anhand eines einzelnen Spiels beurteilen, sondern erst über eine größere Anzahl von Beobachtungen. Der Ausgang einer einzelnen Partie stellt stets nur eine Realisierung aus der Gesamtheit aller möglichen Spielverläufe und Endstände dar.

Die Begegnung zwischen der TSG Hoffenheim und Borussia Dortmund endete schließ-

lich mit einem 2:1-Heimsieg Hoffenheims. Damit trat zwar nicht der vom Modell insgesamt favorisierte Hauptausgang ein, da dieses einen Sieg Dortmunds mit 45,73 % als wahrscheinlichsten Spielausgang auswies, jedoch liegt auch das tatsächlich realisierte Ergebnis durchaus im Bereich plausibler Modellresultate. Für den konkreten Endstand 2:1 ergibt sich im Modell eine Wahrscheinlichkeit von 7,36 %. Er zählt damit nicht zu den wahrscheinlichsten Einzelresultaten, ist aber keineswegs als ungewöhnlicher oder modellfremder Ausgang zu interpretieren.

Besonders bemerkenswert ist, dass sich der Spielverlauf über weite Strecken in einer Ergebniskonstellation bewegte, die vom Modell sogar als wahrscheinlichster Einzelendstand prognostiziert wurde. So stand es lange Zeit 1:1, also genau jenes Ergebnis, dem im Modell mit 10,65 % die höchste Einzelwahrscheinlichkeit zugewiesen worden war. Zugleich zeigte sich insbesondere in der zweiten Spielhälfte, dass ein Auswärtssieg Dortmunds keineswegs unwahrscheinlich war, da die Mannschaft über längere Phasen spielbestimmend auftrat und mehrere qualitativ hochwertige Torgelegenheiten herausspielte. Erst in der achten Minute der Nachspielzeit wurde Hoffenheim nach Intervention des VAR (Video Assistant Referee) ein umstrittener Elfmeter zugesprochen, der zum entscheidenden 2:1 führte. Gerade dieser späte und spielentscheidende Moment verdeutlicht anschaulich, dass der tatsächliche Ausgang eines Fußballspiels auch von singulären, kaum vorhersagbaren Ereignissen geprägt sein kann, die sich einer präzisen Vormodellierung weitgehend entziehen.

Der Abgleich mit dem realen Spielausgang zeigt somit zweierlei: Einerseits traf die globale Tendenz des Modells in diesem konkreten Fall nicht zu, da nicht der favorisierte Dortmunder Auswärtssieg, sondern ein Hoffenheimer Heimsieg eintrat. Andererseits erwies sich die probabilistische Beschreibung des Spiels dennoch als plausibel, weil sowohl der lange Spielstand von 1:1 als auch der spätere Endstand 2:1 mit nicht vernachlässigbaren Wahrscheinlichkeiten im Modell enthalten waren. Gerade dieser Befund verdeutlicht den grundlegenden Charakter statistischer Prognosemodelle: Diese zielen nicht darauf ab, einen einzelnen Spielausgang punktgenau vorherzusagen, sondern darauf, die Menge realistischer Ergebnisse durch geeignete Wahrscheinlichkeitszuweisungen zu strukturieren. Ein Modell ist daher nicht daran zu messen, ob es den tatsächlich eingetretenen Endstand als wahrscheinlichsten Fall ausweist, sondern daran, ob sich dieser als plausibler Ausgang innerhalb der modellierten Verteilung wiederfindet. Genau dies ist im vorliegenden Beispiel gegeben.

Insgesamt bestätigt der Vergleich mit dem tatsächlichen Spielergebnis somit weniger die Treffsicherheit eines konkreten Tipps als vielmehr die grundsätzliche Eignung des entwickelten Modells zur probabilistischen Beschreibung von Fußballspielen. Zugleich wird deutlich, dass einzelne spielentscheidende Situationen – insbesondere späte und kontroverse Ereignisse wie ein VAR-induzierter Elfmeter in der Nachspielzeit – die Grenzen eines bewusst einfach gehaltenen Prämodells sichtbar machen. Gerade darin liegt jedoch kein Widerspruch zur Modellidee, sondern ein charakteristisches Merkmal des Fußballspiels selbst, dessen Ausgang trotz statistisch fundierter Beschreibung stets eine erhebliche Zufallskomponente behält.

6 Fazit und Ausblick

Ziel der vorliegenden Arbeit war die Entwicklung eines mathematisch fundierten und zugleich nachvollziehbaren Prognosemodells zur Beschreibung und Vorhersage von Spielausgängen im Fußball. Im Zentrum stand dabei die Frage, wie sich die im Fußball besonders ausgeprägte Zufallskomponente, die aus der vergleichsweise geringen Toranzahl resultiert, durch geeignete Kenngrößen systematisch erfassen und modelltheoretisch berücksichtigen lässt, um valide Wahrscheinlichkeiten für Spielausgänge abzuleiten. Vor diesem Hintergrund bestand die zentrale Herausforderung darin, ein Modell zu entwickeln, das trotz unvermeidbarer Vereinfachungen belastbare probabilistische und zugleich interpretierbare Aussagen ermöglicht.

Die schrittweise Modellentwicklung hat gezeigt, dass bereits einfache stochastische Ansätze wesentliche Strukturen des Fußballspiels erfassen können. Insbesondere wurde im Rahmen der binomialverteilten Modellierung deutlich, dass die vergleichsweise geringe Anzahl an Toren zu einer hohen Streuung möglicher Spielergebnisse führt. Dadurch ergibt sich, dass selbst deutlich unterlegene Teams eine nicht zu vernachlässigende Gewinnwahrscheinlichkeit besitzen. Diese Eigenschaft liefert eine mathematische Erklärung für die im Fußball häufig beobachteten Überraschungsergebnisse und unterstreicht die zentrale Rolle zufallsbedingter Einflüsse.

Die Einführung der Poisson-Verteilung stellte einen entscheidenden Schritt zur weiteren Präzisierung des Modells dar. Sie ergibt sich als Grenzfall der Binomialverteilung, wenn die Anzahl der Versuche stark zunimmt, während die Erfolgswahrscheinlichkeit gleichzeitig sehr klein wird, sodass das Produkt beider Größen konstant bleibt. Sowohl theoretische Überlegungen als auch empirische Analysen zeigen, dass die Toranzahl im Fußball in guter Näherung durch eine Poisson-Verteilung beschrieben werden kann. Insbesondere die Übereinstimmung von Erwartungswert und Varianz sowie die charakteristische Form der Verteilung spiegeln die empirisch beobachteten Tormuster überzeugend wider. Während sich bei langfristig aggregierten Daten statistisch signifikante Abweichungen ergeben, konnte für die aktuelle Saison eine gute Übereinstimmung festgestellt werden, was die praktische Anwendbarkeit des Modells im konkreten Prognosekontext zusätzlich stützt.

Ein wesentlicher Beitrag der Arbeit liegt in der Identifikation geeigneter Leistungsindikatoren zur Beschreibung der Teamstärke. Dabei zeigte sich zunächst, dass die Tordifferenz ein robustes und intuitiv verständliches Maß darstellt. Aufbauend darauf wurde untersucht, welche Kenngrößen sich besonders gut zur Prognose der Teamstärke eignen. Die Ergebnisse verdeutlichen, dass die Tordifferenz durch differenziertere Metriken

ergänzt werden kann. Insbesondere die Expected-Goals-Differenz (xGD) erwies sich als prognostisch überlegen, da sie die Qualität von Torchancen berücksichtigt und damit zufallsbedingte Schwankungen in den tatsächlichen Torergebnissen teilweise ausgleicht. Die durchgeführte Korrelationsanalyse bestätigt, dass xG-basierte Kennzahlen einen höheren Informationsgehalt besitzen und eine verbesserte Vorhersage zukünftiger Leistungsentwicklungen ermöglichen.

Die Kombination dieser Erkenntnisse mündete in ein konkretes Prognosemodell auf Basis unabhängiger Poisson-Verteilungen, in das zentrale Einflussgrößen wie Expected-Goals-Summe, Expected-Goals-Differenz und Heimvorteil integriert wurden. Die exemplarische Anwendung auf ein Bundesligaspiel zeigt, dass sich mit diesem Ansatz differenzierte Wahrscheinlichkeitsaussagen für Spielausgänge, konkrete Endstände und weitere Ereignisse ableiten lassen. Der Vergleich mit Buchmacherquoten verdeutlicht zudem, dass das Modell in seiner grundsätzlichen Einschätzung mit marktetablierten Prognosen übereinstimmt, auch wenn Unterschiede in der Gewichtung einzelner Wahrscheinlichkeiten bestehen. Die Einordnung des tatsächlichen Spielausgangs macht darüber hinaus die Grenzen des modellbasierten Ansatzes deutlich.

Insgesamt zeigt sich, dass die Vorhersage von Fußballspielen mit erheblichen Unsicherheiten behaftet bleibt. Die geringe Anzahl an Toren führt dazu, dass einzelne Ereignisse einen überproportional großen Einfluss auf das Spielergebnis haben. Wie Fischer und Heuer [27] hervorheben, ist der Spielausgang das Resultat weniger, kaum vorhersehbarer Ereignisse wie zufälliger Tore, strittiger Schiedsrichterentscheidungen oder spielimmanenter Effekte wie Verletzungen oder Platzverweisen. Hinzu kommen spieltagsspezifische Faktoren wie kurzfristige Ausfälle oder taktische Anpassungen, deren quantitative Erfassung im Modell nur eingeschränkt möglich ist. In diesem Sinne ist ein Fußballspiel als Folge vieler kleiner zufälliger Prozesse zu verstehen [31]. Statistische Modelle können diese Unsicherheit nicht eliminieren, sondern lediglich in Form von Wahrscheinlichkeiten strukturieren.

Vor diesem Hintergrund ist auch die Aussagekraft des entwickelten Modells kritisch zu reflektieren. Die getroffenen Annahmen, insbesondere die Unabhängigkeit von Torereignissen sowie die Konstanz der Torraten, stellen notwendige Vereinfachungen dar, die nicht in allen Spielsituationen vollständig erfüllt sind. Darüber hinaus basiert die Analyse im Wesentlichen auf Daten einer einzelnen Saison, was die Generalisierbarkeit der Ergebnisse einschränkt. Vor allem bei wettbewerbsübergreifenden Prognosen, etwa in internationalen Turnieren oder Pokalbewerben, ergeben sich zusätzliche Herausforderungen. Die geringere Anzahl an Spielen, Unterschiede in den Ligastärken sowie sich verändernde Mannschaftszusammensetzungen – insbesondere bei Weltmeisterschaften, die nur alle vier Jahre ausgetragen werden – erschweren eine verlässliche Parameterschätzung erheblich.

Ein Ansatz zur Verbesserung der Modellgüte liegt in der Erweiterung der Datenbasis und der berücksichtigten Einflussgrößen. Die Integration zusätzlicher Kenngrößen wie detaillierter Torchancenmetriken, positionsspezifischer Leistungsdaten oder weiterer spielbe-

zogener Ereignisdaten sowie eine differenzierte Gewichtung der Daten, bei der aktuellere Spiele stärker berücksichtigt werden als ältere, könnten die Modellierung weiter verfeinern. Darüber hinaus eröffnet sich ein weiterführendes Forschungsfeld im Vergleich mit komplexeren Modellansätzen, insbesondere aus dem Bereich des maschinellen Lernens. Verfahren wie neuronale Netze könnten in der Lage sein, nichtlineare Zusammenhänge besser zu erfassen und zusätzliche Informationsquellen zu integrieren. Gleichzeitig stellt sich jedoch die grundlegende Frage, inwieweit sich die Prognosequalität tatsächlich signifikant steigern lässt oder ob die inhärente Zufallskomponente des Spiels eine natürliche Grenze der Vorhersagbarkeit darstellt.

Insgesamt zeigt die vorliegende Arbeit, dass sich Fußballspiele trotz ihrer Komplexität mit vergleichsweise einfachen mathematischen Mitteln sinnvoll modellieren lassen. Die entwickelten Ansätze ermöglichen eine strukturierte und datenbasierte Beschreibung von Spielausgängen und liefern realistische Wahrscheinlichkeitsbewertungen. Gleichzeitig bleibt festzuhalten, dass auch die besten Modelle die grundlegende Unsicherheit des Spiels nicht aufheben können. Vielmehr liegt die Stärke statistischer Modelle darin, diese Unsicherheit transparent zu machen und in eine quantifizierbare Form zu überführen.

Damit erfüllt das entwickelte Modell die eingangs formulierte Zielsetzung: Es erlaubt eine fundierte probabilistische Einschätzung von Spielausgängen auf Basis geeigneter Leistungsindikatoren und verdeutlicht zugleich die Grenzen mathematischer Vorhersagbarkeit im Fußball. Gerade diese Verbindung aus Struktur und Unsicherheit ist es, die den Fußball nicht nur analytisch interessant, sondern auch für Zuschauer:innen nachhaltig faszinierend macht.

Literaturverzeichnis

[1] Adams, T. (2016, 2. Mai). *Leicester city: The greatest underdog story of all* [TNT-Sports.co.uk]. Verfügbar 16. April 2025 unter https://www.tntsports.co.uk/football/premier-league/2015-2016/leicester-city-s-premier-league-title-win-the-greatest-underdog-story-of-all_sto5521114/story.shtml

[2] Arens, T., Hettlich, F., Karpfinger, C., Kockelkorn, U., Lichtenegger, K., & Stachel, H. (2015). *Mathematik* (3. Aufl.). Springer Spektrum. https://doi.org/10.1007/978-3-642-44919-2

[3] Bach, M. (2023, 15. August). *1962: Die schwere Geburt der Bundesliga-Gründung* [NDR.de]. Verfügbar 7. August 2025 unter https://www.ndr.de/sport/fussball/1962-Die-schwere-Geburt-der-Bundesliga-Gruendung,geschichte401.html

[4] Bark, M. (2024, 12. Dezember). *Bundesliga-Trend: Meist gewinnt das Heimteam – bald nicht mehr* [Sportschau.de]. Verfügbar 8. August 2025 unter https://www.sportschau.de/fussball/bundesliga/meist-gewinnt-heimteam-bald-nicht-mehr,fussball-1476.html

[5] Bayerischer Rundfunk. (2012, 24. August). *Zeitstrahl: Über 50 Jahre Fußball-Bundesliga* [BR.de]. Verfügbar 7. August 2025 unter https://www.br.de/themen/sport/inhalt/fussball/bundesliga/50-jahre-fussball-bundesliga100.html

[6] BBC Sport. (2016, 2. Mai). *Leicester city win premier league title after tottenham draw at chelsea* [BBC.com]. Verfügbar 16. April 2025 unter https://www.bbc.com/sport/football/35988673

[7] Beckmann, N. (2022). Statistical influence of travelling distance on home advantage over 57 years in the men's german first soccer division. *German Journal of Exercise and Sport Research*, *52*(4), 657–665. https://doi.org/10.1007/s12662-021-00787-7

[8] Behrends, E. (2013). *Elementare Stochastik: Ein Lernbuch - von Studierenden mitentwickelt* (1. Aufl.). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-8348-2331-1

[9] Bendix, O. (2007). *Woher kommt der Fußball?* [Max-Planck-Gesellschaft]. Verfügbar 1. August 2025 unter https://www.ds.mpg.de/203240/11

[10] Ben-Naim, E., Vazquez, F., & Redner, S. (2007). What is the most competitive sport? *Journal of the Korean Physical Society*, *50*(1), 124–126.

[11] Berschneider, G., & Schilling, R. L. (2019). Die Poisson-Verteilung, Fußballtore und das Gesetz der kleinen Zahlen. *Der Mathematikunterricht*, *65*(6), 40–53.

[12] Biermann, C. (2009). *Die Fußball-Matrix: Auf der Suche nach dem perfekten Spiel* (1. Aufl.). Kiepenheuer & Witsch.

[13] Bola Webinformation GmbH. (n. d.). *Wettquoten Vergleich & Erklärung* [Wettbasis.com]. Verfügbar 17. April 2026 unter https://www.wettbasis.com/spezial/wettquoten

[14] Bosch, K. (2011). *Elementare Einführung in die Wahrscheinlichkeitsrechnung* (11., aktualisierte Auflage). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-8348-8331-5

[15] Chu, S. (2003). Using soccer goals to motivate the poisson process. *INFORMS Transactions on Education*, *3*(2), 64–70. https://doi.org/10.1287/ited.3.2.64

[16] Cleff, T. (2019). *Angewandte Induktive Statistik und Statistische Testverfahren: Eine computergestützte Einführung mit Excel, SPSS und Stata* (1. Aufl.). Springer Gabler. https://doi.org/10.1007/978-3-8349-6973-6

[17] Cramer, E., & Kamps, U. (2020). *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik: Eine Einführung für Studierende der Informatik, der Ingenieur- und Wirtschaftswissenschaften* (5. Aufl.). Springer Spektrum. https://doi.org/10.1007/978-3-662-60552-3

[18] Dambeck, H. (2010). Ist Fußball ein Glücksspiel? *Spektrum der Wissenschaft*, (6), 68–70. Verfügbar 27. März 2025 unter https://www.spektrum.de/magazin/ist-fussball-ein-gluecksspiel/1030089

[19] Delaney, M. (2025, 8. April). *Remembering la remontada: Barcelona 6-1 Paris Saint-Germain* [Independent.co.uk]. Verfügbar 16. April 2025 unter https://www.independent.co.uk/sport/football/european/barcelona-psg-2017-champions-league-comeback-b2729254.html

[20] DFL Deutsche Fußball Liga GmbH. (n. d.[a]). *Bundesliga: Offizielle Webseite* [Bundesliga.com]. Verfügbar 24. Juli 2025 unter https://www.bundesliga.com/de/bundesliga

[21] DFL Deutsche Fußball Liga GmbH. (n. d.[b]). *Fußball: Größen, Massen und Gewicht* [Bundesliga.com]. Verfügbar 3. Februar 2026 unter https://www.bundesliga.com/%5Bobject%20Object%5D/faq/spielbetrieb/fussball-groessen-masse-und-gewicht-22368

[22] Dilger, A., & Geyer, H. (2007). Theoretische und empirische Analyse der Drei-Punkte-Regel. *Sport und Gesellschaft*, *4*(3), 265–277. https://doi.org/10.1515/sug-2007-0304

[23] Eggels, H., van Elk, R., & Pechenizkiy, M. (2016). Explaining soccer match outcomes

with goal scoring opportunities predictive analytics. *Proceedings of the Workshop on Machine Learning and Data Mining for Sports Analytics 2016 co-located with the 2016 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2016)*, *1842*. Verfügbar 27. April 2025 unter https://ceur-ws.org/Vol-1842/paper_07.pdf

[24] Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2023). *Statistik: Der Weg zur Datenanalyse* (9. Aufl.). Springer Spektrum. https://doi.org/10.1007/978-3-662-67526-7

[25] Fédération Internationale de Football Association. (n. d.[a]). *Inside FIFA* [FIFA.com]. Verfügbar 24. Juli 2025 unter https://inside.fifa.com

[26] Fédération Internationale de Football Association. (n. d.[b]). *Video Assistant Referee Technology* [FIFA.com]. Verfügbar 4. August 2025 unter https://inside.fifa.com/innovation/standards/video-assistant-referee

[27] Fischer, M., & Heuer, A. (2025). Spielvorhersagen im Fußball: Machine Learning vs. Poisson-Ansätze. In D. Memmert (Hrsg.), *Künstliche Intelligenz und maschinelles Lernen in der Sportwissenschaft* (S. 105–117). Springer Spektrum. https://doi.org/10.1007/978-3-662-68950-9_7

[28] Frost, I. (2020). *Statistik für Wirtschaftswissenschaftler* (4. überarb. Aufl.). expert. https://doi.org/10.36198/9783838553511

[29] Fussballdaten. (n. d.). *Bundesliga: Statistiken* [Fussballdaten.de]. Verfügbar 25. August 2025 unter https://www.fussballdaten.de/bundesliga/statistik/

[30] Georgii, H.-O. (2015). *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik* (5. Auflage). De Gruyter. https://doi.org/10.1515/9783110359701

[31] Groll, A., & Schauberger, G. (2019). Statistik und Fußball. In W. Krämer & C. Weihs (Hrsg.), *Faszination Statistik* (S. 59–66). Springer Spektrum. https://doi.org/10.1007/978-3-662-60562-2_8

[32] Hartung, J. (2009). *Statistik: Lehr- und Handbuch der Angewandten Statistik* (15., überarb. und wesentlich erw. Aufl.). Oldenbourg Wissenschaftsverlag. https://doi.org/10.1524/9783486710540

[33] Heidrich, M. (2025, 16. Juni). *Datenanalyse: Warum der Heimvorteil in der Bundesliga schwindet* [NDR.de]. Verfügbar 8. August 2025 unter https://www.ndr.de/sport/fussball/Datenanalyse-Warum-der-Heimvorteil-in-der-Bundesliga-schwindet,heimvorteil218.html

[34] HEIM:SPIEL Medien GmbH & Co. KG. (n. d.). *Bundesliga 2025/2026* [weltfussball.at]. Verfügbar 7. August 2025 unter https://www.weltfussball.at/wettbewerb/bundesliga/

[35] Henze, N. (2023). *Stochastik für Einsteiger: Eine Einführung in die faszinierende*

Welt des Zufalls (14., überarbeitete und ergänzte Auflage). Springer Spektrum. https://doi.org/10.1007/978-3-662-67729-2

[36] Heuer, A. (2012). *Der perfekte Tipp: Statistik des Fußballspiels* (1. Aufl.). Wiley-VCH.

[37] Heuer, A. (2020a, 8. März). Identification of relevant performance indicators in round-robin tournaments. https://doi.org/10.48550/arXiv.2003.03774

[38] Heuer, A. (2020b). Wer wird Meister? Fußball-Vorhersagen mit statistischen Methoden. *Physik in unserer Zeit*, *51*(3), 130–137. https://doi.org/doi.org/10.1002/piuz.202001577

[39] Heuer, A., Müller, C., & Rubner, O. (2010). Soccer: is scoring goals a predictable Poissonian process? *Europhysics Letters*, *89*(3), 38007. https://doi.org/10.1209/0295-5075/89/38007

[40] Heuer, A., & Rubner, O. (2009). Fitness, chance, and myths: An objective view on soccer results. *The European Physical Journal B*, *67*(3), 445–458. https://doi.org/doi.org/10.1140/epjb/e2009-00024-8

[41] Heuer, A., & Rubner, O. (2018, 15. Oktober). How does the past of a soccer match influence its future? https://doi.org/10.48550/arXiv.1207.4471

[42] Hofbauer, F. (n. d.). *Wahrscheinlichkeitstheorie und Statistik* (Vorlesungsskriptum). Universität Wien. Wien. Verfügbar 26. März 2026 unter https://www.mat.univie.ac.at/~bruin/wkth.pdf

[43] Hofbauer, F., & Greschonig, G. (2022). *Stochastik: Eine Vorlesung für das Lehramtsstudium* (Vorlesungsskriptum). Universität Wien. Wien.

[44] Hubáček, O., Šourek, G., & Železný, F. (2022). Forty years of score-based soccer match outcome prediction: An experimental review. *IMA Journal of Management Mathematics*, *33*(1), 1–18. https://doi.org/10.1093/imaman/dpab029

[45] Kaiser, G. (2025a, 24. November). *Anzahl der Sportfans weltweit nach Sportarten 2025* [Statista.com]. Verfügbar 29. Januar 2026 unter https://de.statista.com/statistik/daten/studie/387554/umfrage/anzahl-der-sportfans-weltweit/

[46] Kaiser, G. (2025b, 28. November). *Tore pro Spiel in der 1. Fußball-Bundesliga bis 2024/2025* [Statista.com]. Verfügbar 25. Januar 2026 unter https://de.statista.com/statistik/daten/studie/1622/umfrage/bundesliga-entwicklung-der-durchschnittlich-erzielten-tore-pro-spiel/

[47] Kosfeld, R., Eckey, H. F., & Türck, M. (2016). *Deskriptive Statistik: Grundlagen - Methoden - Beispiele - Aufgaben* (6. Aufl.). Springer Gabler. https://doi.org/10.1007/978-3-658-13640-6

[48] Kosfeld, R., Eckey, H.-F., & Türck, M. (2019). *Wahrscheinlichkeitsrechnung und Induktive Statistik: Grundlagen - Methoden - Beispiele* (3. Aufl.). Springer Gabler.

https://doi.org/10.1007/978-3-658-28713-9

[49] Kovar, P., & Zart, S. (2022). Fußball. In A. Güllich & M. Krüger (Hrsg.), *Grundlagen von Sport und Sportwissenschaft: Handbuch Sport und Sportwissenschaft* (S. 603–626). Springer Spektrum. https://doi.org/10.1007/978-3-662-53404-5

[50] Kronfellner, M., Kronfellner, J., & Peschek, W. (1998). *Angewandte Mathematik: Arbeitslehrbuch 4*. öbv & hpt Hölder-Pichler-Tempsky.

[51] Kunkel, F., & Schätzle, D. (2024, 10. April). *Champions League: Als Barca gegen PSG das Unmögliche schaffte* [Sport1.de]. Verfügbar 16. April 2025 unter https://www.sport1.de/news/fussball/champions-league/2024/04/champions-league-als-barca-gegen-psg-das-unmogliche-schaffte

[52] Kunz, M. (2007). 265 million playing football. *FIFA magazine*, 10–15. Verfügbar 21. April 2026 unter https://condorperformance.com/wp-content/uploads/2020/02/emaga_9384_10704.pdf

[53] Kütting, H., & Sauer, M. J. (2011). *Elementare Stochastik: Mathematische Grundlagen und didaktische Konzepte* (F. Padberg, Hrsg.; 3. Aufl.). Springer Spektrum. https://doi.org/10.1007/978-3-8274-2760-1

[54] Lames, M. (2018). Chance involvement in goal scoring in football - an empirical approach. *German Journal of Exercise and Sport Research*, *48*(2), 278–286. https://doi.org/10.1007/s12662-018-0518-z

[55] Laplace, P.-S. (1932). *Philosophischer Versuch über die Wahrscheinlichkeit*. Akademische Verlagsgesellschaft m. b. H.

[56] Leiner, B. (2004). *Einführung in die Statistik* (9., unwes. veränd. Aufl.). Oldenbourg Wissenschaftsverlag. https://doi.org/10.1515/9783486835762

[57] Ley, C., Wiele, T. V. D., & Eetvelde, H. V. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, *19*(1), 55–73. https://doi.org/10.1177/1471082X18817650

[58] Linde, W. (2014, 11. April). *Stochastik für das Lehramt*. De Gruyter Oldenbourg. https://doi.org/10.1524/9783110362411

[59] Ludwig, M., & Oldenburg, R. (2017). Fußballergebnisse vorhersagen – mit Mathematik prognostizieren. In H. Humenberger & M. Bracke (Hrsg.), *Neue Materialien für einen realitätsbezogenen Mathematikunterricht 3* (S. 149–160). Springer Spektrum. https://doi.org/10.1007/978-3-658-11902-7_12

[60] Maozad, S. N., Noor Asyikin Mohd Razali, S., Mustapha, A., Nanthaamornphong, A., Abdul Wahab, M. H., & Razali, N. (2022). Comparative analysis for predicting football match outcomes based on poisson models, 1–4. https://doi.org/10.1109/ECTI-CON54298.2022.9795385

[61] Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving

model performance and demonstrating value. *PLoS ONE*, *18*(4), e0282295. https://doi.org/10.1371/journal.pone.0282295

[62] Mian, M. (2024). *The accuracy of expected goals in the premier league* [Thesis]. Skidmore College. https://creativematter.skidmore.edu/econ_studt_schol/164

[63] Oberhofer, H., Philippovich, T., & Winner, H. (2010). Distance matters in away games: Evidence from the german football league. *Journal of Economic Psychology*, *31*(2), 200–211. https://doi.org/10.1016/j.joep.2009.11.003

[64] Opta Analyst. (n. d.). *Bundesliga stats* [TheAnalyst.com]. Verfügbar 5. April 2026 unter https://theanalyst.com/competition/bundesliga/stats

[65] Österreichische Fußball-Bundesliga. (n. d.). *Das Ligaformat* [ÖFBL.at]. Verfügbar 5. August 2025 unter https://www.oefbl.at/de/die-liga,derspielmodus

[66] Österreichischer Fußball-Bund. (n. d.). *ÖFB* [ÖFB.at]. Verfügbar 28. Juli 2025 unter https://www.oefb.at/oefb/

[67] Over-Under Digital Inc. (n. d.). *Bundesliga: Tabelle & Statistiken* [FootyStats.org]. Verfügbar 5. April 2026 unter https://footystats.org/de/germany/bundesliga

[68] owayo GmbH. (n. d.). *Die Geschichte des Fußballs auf einen Blick* [owayo.at]. Verfügbar 19. April 2025 unter https://www.owayo.at/magazin/die-geschichte-des-fussballs-at.htm

[69] Palacios-Huerta, I. (2004). Structural changes during a century of the world's most popular sport. *Statistical Methods & Applications*, *13*(2), 241–258. https://doi.org/10.1007/s10260-004-0093-3

[70] Palacios-Huerta, I., & Garicano, L. (2014). Making the beautiful game a bit less beautiful. In I. Palacios-Huerta (Hrsg.), *Beautiful game theory. how soccer can help economics* (S. 124–150). Princeton University Press. https://doi.org/10.1515/9781400850310

[71] Partida, A., Martinez, A., Durrer, C., Gutierrez, O., & Posta, F. (2021). Modeling of football match outcomes with expected goals statistic. *Journal of Student Research*, *10*(1), 1–10. https://doi.org/10.47611/jsr.v10i1.1116

[72] Posch, L. (2011). *Wie man den Ausgang der österreichischen Fußballbundesliga berechnen kann – wahrscheinlichkeitstheoretische Betrachtungen zum Fußballspiel* [Diplomarbeit]. Universität Wien. https://doi.org/10.25365/thesis.13319

[73] Prem, K. P. (2006, 21. Mai). *Glücksspiel Fußball?* [idw-online.de]. Verfügbar 10. August 2025 unter https://idw-online.de/de/news160550

[74] Primorac, A. (2021). *Entwicklung und Erprobung eines Werkzeuges zur Quantifizierung der Spielleistung im Fußball mittels expected Goals* [Diplomarbeit]. Universität Wien.

[75] Scarf, P., & Rangel Jr., J. S. (2017). Models for outcomes of soccer matches. In J. Albert, M. E. Glickman, T. B. Swartz & R. H. Koning (Hrsg.), *Handbook of statistical methods and analyses in sports* (S. 341–354). Chapman; Hall/CRC. https://doi.org/10.1201/9781315166070

[76] Sports Reference LLC. (n. d.). *2025-2026 bundesliga stats* [FBref.com]. Verfügbar 5. April 2026 unter https://fbref.com/en/comps/20/Bundesliga-Stats

[77] Statista GmbH. (n. d.). *Fußball in Österreich - Daten & Fakten* [Statista.com]. Verfügbar 29. Juli 2025 unter https://de.statista.com/themen/2020/fussball-in-oesterreich/

[78] Stenerud, S. G. (2015). *A study on soccer prediction using goals and shots on target* [Masterarbeit]. Norwegian University of Science und Technology. Verfügbar 18. April 2025 unter https://nva.sikt.no/registration/0198ebefdf84-50825182-04fa-4423-bd50-9a328edf58c2

[79] Stickel, A., & Nufer, G. (2023). Der Einfluss steigender Zuschauerzahlen nach den COVID-19-bedingten Geisterspielen auf den Heimvorteil in der Fußball-Bundesliga. *Sciamus - Sport und Management*, (1), 1–28. https://doi.org/10.24403/jp.1310675

[80] Tappe, S. (2013). *Einführung in die Wahrscheinlichkeitstheorie* (1. Aufl.). Springer Spektrum. https://doi.org/10.1007/978-3-642-37544-6

[81] Temming, C. (n. d.). *Buchmachermargen: Die Mathematik dahinter bei Sportwetten* [Wettbasis.com]. Verfügbar 16. April 2026 unter https://www.wettbasis.com/sportwetten-news/buchmachermargen-die-mathematik-dahinter-bei-sportwetten

[82] Teves, C. (2018, 18. Juni). *Frühe Ballspiele* [planet-wissen.de]. Verfügbar 3. August 2025 unter https://www.planet-wissen.de/gesellschaft/sport/fussballgeschichte/pwiefrueheballspiele100.html

[83] Teves, C. (2020, 15. Mai). *Geschichte des Fußballs* [planet-wissen.de]. Verfügbar 1. August 2025 unter https://www.planet-wissen.de/gesellschaft/sport/fussballgeschichte/index.html

[84] The International Football Association Board. (2025). *Spielregeln 25/26*. IFAB. Verfügbar 22. April 2026 unter https://downloads.theifab.com/downloads/spielregeln-2025-26-doppelseiten?l=de

[85] The International Football Association Board. (n. d.). *Laws of the Game: Offizielle Fußballregeln* [TheIFAB.com]. Verfügbar 4. August 2025 unter https://www.theifab.com/de/

[86] Tiippana, T. (2020). *How accurately does the expected goals model reflect goalscoring and success in football?* [Bachelorarbeit]. Aalto University.

[87] Tijms, H. C. (2024). *Die faszinierende Welt der Wahrscheinlichkeitsrechnung: Sto-*

chastik in Aktion. Springer. https://doi.org/10.1007/978-3-662-69280-6

[88] Tolan, M. (2010). *So werden wir Weltmeister: Die Physik des Fußballspiels*. Piper. https://ubdata.univie.ac.at/AC07999871

[89] Toutenburg, H., & Heumann, C. (2008). *Induktive Statistik: Eine Einführung mit R und SPSS.* (4. Aufl.). Springer. https://doi.org/10.1007/978-3-540-77510-2

[90] Transfermarkt GmbH & Co. KG. (n. d.). *Bundesliga 2025/26* [Transfermarkt.de]. Verfügbar 7. August 2025 unter https://www.transfermarkt.de/bundesliga/startseite/wettbewerb/L1/saison_id/2025

[91] Understat. (n. d.). *Bundesliga 2025/26* [Understat.com]. Verfügbar 5. April 2026 unter https://understat.com/league/Bundesliga

[92] Union of European Football Associations. (n. d.). *UEFA Rankings* [UEFA.com]. Verfügbar 5. August 2025 unter https://de.uefa.com/nationalassociations/uefarankings/

[93] Whitmore, J. (2023, 8. August). *What is expected goals (xG)?* [Opta analyst]. Verfügbar 15. April 2025 unter https://theanalyst.com/2023/08/what-is-expected-goals-xg

[94] Wikimedia Foundation Inc. (n. d.[a]). *Fußball* [Wikipedia.org]. Verfügbar 19. April 2025 unter https://de.wikipedia.org/w/index.php?title=Fu%C3%9Fball&oldid=254168983

[95] Wikimedia Foundation Inc. (n. d.[b]). *Fußball-Bundesliga* [Wikipedia.org]. Verfügbar 24. Juli 2025 unter https://de.wikipedia.org/w/index.php?title=Fu%C3%9Fball-Bundesliga&oldid=257987683

[96] Wikimedia Foundation Inc. (n. d.[c]). *Fußballregeln* [Wikipedia.org]. Verfügbar 19. April 2025 unter https://de.wikipedia.org/w/index.php?title=Fu%C3%9Fballregeln&oldid=254202227

[97] Wikimedia Foundation Inc. (n. d.[d]). *Geschichte des Fußballs* [Wikipedia.org] [Page Version ID: 253226794]. Verfügbar 19. April 2025 unter https://de.wikipedia.org/w/index.php?title=Geschichte_des_Fu%C3%9Fballs&oldid=253226794

[98] Wikimedia Foundation Inc. (n. d.[e]). *History of the german football league system* [Wikipedia.org]. Verfügbar 8. August 2025 unter https://en.wikipedia.org/w/index.php?title=History_of_the_German_football_league_system&oldid=1293627765