Biometrie

Unterlagen für die Pflichtveranstaltung
Biometrie
am Fachbereich Veterinärmedizin der
Freien Universität Berlin

herausgegeben vom

Institut für Biometrie und Informationsverarbeitung
des Fachbereichs Veterinärmedizin
der

Freien Universität Berlin

Oertzenweg 19b — 14163 Berlin

Vorwort zur vierten, überarbeiteten Auflage

Nachdem wir uns lange Zeit auf geringfügig korrigierte Nachdrucke der dritten Auflage für die Pflichtveranstaltung Biometrie in der Veterinärmedizin beschränkt haben, ist diese vierte Auflage das Ergebnis der auf Datenträger gebrachten und zum Teil überarbeiteten dritten Auflage. Im wesentlichen wurden vorwiegend tiermedizinische Beispiele berücksichtigt sowie einige Textabschnitte gestrafft. Weiterhin wurden die begleitenden Übungsaufgaben im Anhang aufgenommen, dafür jedoch ein Teil alter Aufgaben weggelassen.

Für die mühevolle Übernahme auf Datenträger einschließlich der unvermeidbaren Formeln und erläuternden Grafiken danke ich besonders Herrn cand. vet. med. Klaus Roux. Für die inhaltliche und organisatorische Überarbeitung gilt mein Dank vor allem Frau Dipl.-Stat. Andrea Ochsmann, Frau PD Dr.rer.pol. Susanne Dahms und Frau Dipl.-Stat. Rose Schmitz. Weiterhin danke ich Frau Ursula Kunze für die mit Korrekturen, Organisation und Vervielfältigung zusammenhängenden Detailarbeiten.

Berlin–Düppel im Februar 2002 Prof. Dr. Hartmut Weiß

Druckdatum: 12. Februar 2007

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Was ist Biometrie ?

In direkter Übersetzung bedeutet Biometrie "Lehre vom Messen in der belebten Natur". Der Begriff "Messen" muss in dieser Definition allerdings sehr weit gefasst werden. Eine bessere Definition ist die folgende:

> "Biometrie ist die Lehre vom Schätzen und Beurteilen nicht direkt messbarer Zusammenhänge in der belebten Natur."

Zur Lösung dieser Aufgaben verwendet man statistische Methoden. Damit stellt sich eine weitere Frage:

1.2 Was ist Statistik ?

Bei dem Begriff 'Statistik' denkt man zunächst an Zahlenkolonnen und Tabellen. Fragt man nach Anwendungsgebieten, so werden Wahlprognosen, Volkszählungen, Arbeitslosenstatistiken, Wirtschaftsprognosen, Unfallstatistiken usw. genannt.

Das Wort 'Statistik' geht zurück auf das lateinische 'status', das etwa 'Zustand' oder auch 'Staat' bedeutet, und lange Zeit verstand man unter Statistik die 'Lehre von der Zustandsbeschreibung des Staates'.

Heute werden in fast allen Wissenschaften, insbesondere auch in der Medizin, für viele Zwecke statistische Verfahren eingesetzt. Dabei unterscheidet man zwei große Teilgebiete der Statistik.

1. Daten, die bei wissenschaftlichen Untersuchungen anfallen, lassen nur selten die in ihnen enthaltene Information direkt erkennen. Erst eine Aufbereitung der Daten, tabellarische Zusammenfassung, Grafiken oder die Berechnung bestimmter Kenngrößen ermöglichen einen Überblick und die Entwicklung von Arbeits- und Prüfhypothesen.
   Dieses Teilgebiet bezeichnet man als **beschreibende bzw. deskriptive Statistik**. Neuerdings haben sich auch die Begriffe explorative Statistik bzw. **exploratorische Datenanalyse (EDA)** durchgesetzt.

2. Schlüsse aus den Versuchsergebnissen, Schätzungen unbekannter Größen und das Prüfen von Hypothesen fallen dagegen in den Bereich der **schließenden bzw. induktiven Statistik**, die man auch als **konfirmatorische Datenanalyse (CDA = confirmative data analysis)** bezeichnet.

Kapitel 2

Grundbegriffe

2.1 Beispiel: Katzen im Tierheim

Hier wird ein konstruiertes Beispiel betrachtet, das für die Erklärung vieler statistischer Verfahren geeignet ist, aber nur bedingt die in der Praxis vorliegenden Sachverhalte beschreibt.

Ein Tierarzt untersucht in einem Tierheim 32 Katzen, da dort einige Fälle von Parvovirose ("Katzenseuche") aufgetreten sind.

Die dabei ermittelten Zahlen lassen sich in einer Tabelle darstellen. Diese Tabelle liefert einem Betrachter, der sich nicht für ein einzelnes Tier interessiert, sehr wenig Information. Sie enthält die Daten, so wie sie der Tierarzt ermittelt hat. Man nennt sie daher auch **Urliste** (siehe Tabelle 2.1).

2.2 Definitionen

Man verwendet zur Beschreibung einer statistischen Untersuchung bestimmte Begriffe:

Das einzelne Objekt, an dem etwas beobachtet wird, nennt man **Untersuchungseinheit**. Synonym verwendet man die Begriffe **Versuchseinheit** und **Versuchstier**. Eine Untersuchungseinheit muss nicht ein einzelnes Tier oder ein einzelner Gegenstand sein.

> Im Beispiel (2.1) ist jede einzelne Katze eine Untersuchungseinheit. In einer Untersuchung, die sich mit der Vermehrung von Katzen beschäftigt, ist es u.U. sinnvoll, eine Katzenfamilie als Untersuchungseinheit zu betrachten. Interessiert man sich für die Herzfunktion nach Verabreichung eines Medikaments, so ist die Versuchseinheit das einzelne Herz. In der Mikrobiologie kann eine Zellkultur die Versuchseinheit sein.

An den Untersuchungseinheiten werden bestimmte Eigenschaften und Charakteristika beobachtet. Diese Eigenschaften nennt man **Merkmale**.

> Bei der Untersuchung im Tierheim wurden neun Merkmale beobachtet, und zwar der Name, das Geschlecht, das Alter, das Körpergewicht, das Verhalten, die Körpertemperatur, die Leukozytenzahl, die Zellzahlen pro Gramm Darminhalt und die Diagnose der einzelnen Versuchstiere. In der Katzenfamilie ermittelt man z.B. die Anzahl der männlichen und der weiblichen Tiere in einem Wurf.

Name	**Geschlecht**	**Alter** (Jahre)	**Körpergewicht** (kg)	**Verhalten**	**Körpertemperatur** (°C)	**Leukozyten zahl** ($1/\mu l$ $= G/l$)	**Zellzahl** ($10^9\cdot$ Zellen pro g)	**Diagnose**
Mieze 1	weiblich	2.0	3.5	ruhig	38.6	8000	0.48	gesund
Peter 1	männlich	2.0	3.7	normal	38.7	9500	0.50	gesund
Mohrle	männlich	4.0	3.6	erregt	39.9	4200	0.36	gesund
Muschi	weiblich	3.5	3.4	normal	38.7	5800	0.70	gesund
Bine	weiblich	3.0	3.3	sehr ruhig	40.1	600	3.00	krank
Bonni	weiblich	4.0	3.6	normal	39.9	2500	15.00	krank
Felix 1	männlich	3.0	4.0	sehr erregt	39.3	14700	2.70	gesund
Kalle	männlich	2.5	3.5	normal	40.6	1200	12.00	krank
Minni	weiblich	2.5	3.1	erregt	38.4	10700	0.05	gesund
Trudi	weiblich	3.0	3.4	erregt	38.7	11300	0.44	gesund
Else	weiblich	3.0	3.4	sehr ruhig	37.3	400	50.00	krank
Rita	weiblich	2.0	3.0	normal	38.7	17200	2.00	krank
Max	männlich	4.0	4.3	erregt	38.6	9700	4.00	gesund
Paul	männlich	3.0	3.8	ruhig	38.5	8600	1.30	gesund
Mieze 2	weiblich	2.0	3.5	sehr erregt	38.7	7600	0.12	gesund
Kater	männlich	2.0	3.5	ruhig	39.1	21700	13.00	krank
Panther	männlich	2.5	3.3	ruhig	40.1	450	20.00	krank
Martin	männlich	1.5	3.5	sehr ruhig	38.4	9700	0.18	gesund
Susi	weiblich	3.0	2.7	normal	39.8	3200	17.00	krank
Dorle	weiblich	2.0	3.2	erregt	38.3	8700	0.54	gesund
Bodo	männlich	2.0	3.9	sehr erregt	38.5	8300	0.74	gesund
Karina	weiblich	3.5	3.1	erregt	39.6	2300	24.00	krank
Tiger	weiblich	3.0	3.9	erregt	38.6	7900	0.66	gesund
Paula	weiblich	3.0	3.3	sehr ruhig	39.0	23700	150.00	krank
Schnuk	männlich	4.5	3.7	normal	38.9	8200	0.52	gesund
Gerri	männlich	3.0	3.7	normal	38.4	6700	0.58	gesund
Gundel	weiblich	2.5	2.9	normal	40.3	1350	0.90	krank
Golo	männlich	3.5	4.1	erregt	38.4	10200	0.30	gesund
Felix 2	männlich	2.5	3.4	ruhig	38.7	5700	0.41	gesund
Peter 2	männlich	2.5	3.6	sehr ruhig	37.8	570	260.00	krank
Micki	weiblich	4.0	3.4	ruhig	41.1	800	15.00	krank
Paule	männlich	2.0	3.8	normal	39.3	3200	10.00	krank

Tab. 2.1: Urliste mit Daten von 32 Katzen

In diesem Zusammenhang nennt man die Untersuchungseinheit auch **Merkmalsträger**.

Die Beobachtung der Merkmale führt für die einzelnen Versuchseinheiten zu unterschiedlichen Ergebnissen, denn nicht alle Tiere sind gleich. Man spricht von dem für ein bestimmtes Merkmal an einer bestimmten Katze tatsächlich **beobachteten Wert (Beobachtungswert)**.

Das Körpergewicht der Katze Muschi beträgt 3.4 kg.

Für die einzelnen Merkmale können stets nur bestimmte Werte beobachtet werden.

Bei der Messung des Körpergewichts von Katzen können keine negativen Werte vorkommen, genauso unwahrscheinlich ist es, dass eine Katze 1000 kg wiegt.

> Beim Verhalten der Katze muss man vorher exakt bestimmen, wie man die verschiedenen Zustände definiert. Hat man sich für die Zustände sehr ruhig, ruhig, normal, erregt, sehr erregt entschieden, so können nur noch diese definierten Werte beobachtet werden.

Für jedes Merkmal sind also nur bestimmte Beobachtungen möglich. Diese möglichen Werte nennt man **Merkmalsausprägungen**.

> Das Körpergewicht von Katzen im Alter von 2–4 Jahren kann Werte zwischen 2.5 und 5.0 kg annehmen. Die Messgenauigkeit bei der Wägung beträgt 0.1 kg. Damit sind die möglichen Ausprägungen: 2.5 kg, 2.6 kg, 2.7 kg, 2.8 kg, ... ,4.9 kg, 5.0 kg.
>
> Beim Verhalten wurden die Zustände sehr ruhig, ruhig, normal, erregt und sehr erregt definiert. Das sind die Merkmalsausprägungen des Merkmals Verhalten in der vorliegenden Untersuchung.

2.3 Charakterisierung von Merkmalen

Das Beispiel aus Abschnitt 2.1 zeigt uns, dass Merkmale sehr unterschiedlich sein können. Die Ausprägungen sind nicht immer Zahlen, sondern oft auch Zustände. Diese Unterschiede werden herangezogen, um Merkmale in Klassen einzuteilen.

Bevor wir näher darauf eingehen, soll die Tabelle 2.1 in eine einfachere Form gebracht werden.

Häufig verwendet man für Merkmale, deren Ausprägungen Zustände sind, Codierungen (Buchstaben oder Zahlen), die den einzelnen Ausprägungen zugeordnet werden. Die Verwendung von Zahlen ist nicht zuletzt deshalb sinnvoll, weil sie für die Auswertung der Daten mit Hilfe von EDV-Anlagen am besten geeignet sind.

Tabelle 2.2 enthält die Codierungen für das Beispiel 2.1.

Merkmal	**Ausprägung**	**Codierung**
Geschlecht	männlich	1
	weiblich	0
Verhalten	sehr ruhig	1
	ruhig	2
	normal	3
	erregt	4
	sehr erregt	5
Diagnose	krank	1
	gesund	0

Tab. 2.2: Codierungen für die Merkmalsausprägungen aus Beispiel 2.1

Bei der Verwendung solcher Codes muss darauf geachtet werden, dass die Größe der Zahlen nicht bei jedem Merkmal eine Rangfolge beschreibt. Die Zuordnung der Ziffern auf die Merkmalsausprägungen erfolgt weitgehend willkürlich.

Unter Verwendung der Codierungen aus Tabelle 2.2 bekommt die Urliste des Beispiels eine neue Gestalt. Die Daten in Tabelle 2.3 nennt man auch **Rohdaten**.

An diesem Beispiel sieht man deutlich, dass Merkmale sehr unterschiedliche Eigenschaften haben.

NR	**G**	**Alt (J)**	**KGW (kg)**	**V**	**KT (°C)**	**LZ (100*G/l)**	**ZZ (10^9Z/g)**	**DIAG**
1	0	2.0	3.5	2	38.6	80.0	0.48	0
2	1	2.0	3.7	3	38.7	95.0	0.50	0
3	1	4.0	3.6	4	39.9	42.0	0.36	0
4	0	3.5	3.4	3	38.7	58.0	0.70	0
5	0	3.0	3.3	1	40.1	6.0	3.00	1
6	0	4.0	3.6	3	39.9	25.0	15.00	1
7	1	3.0	4.0	5	39.3	147.0	2.70	0
8	1	2.5	3.5	3	40.6	12.0	12.00	1
9	0	2.5	3.1	4	38.4	107.0	0.05	0
10	0	3.0	3.4	4	38.7	113.0	0.44	0
11	0	3.0	3.4	1	37.3	4.0	50.00	1
12	0	2.0	3.0	3	38.7	172.0	2.00	1
13	1	4.0	4.3	4	38.6	97.0	4.00	0
14	1	3.0	3.8	2	38.5	86.0	1.30	0
15	0	2.0	3.5	5	38.7	76.0	0.12	0
16	1	2.0	3.5	2	39.1	217.0	13.00	1
17	1	2.5	3.3	2	40.1	4.5	20.00	1
18	1	1.5	3.5	1	38.4	97.0	0.18	0
19	0	3.0	2.7	3	38.8	32.0	17.00	1
20	0	2.0	3.2	4	38.3	87.0	0.54	0
21	1	2.0	3.9	5	38.5	83.0	0.74	0
22	0	3.5	3.1	4	39.6	23.0	24.00	1
23	0	3.0	3.9	4	38.6	79.0	0.66	0
24	0	3.0	3.3	1	39.0	237.0	150.00	1
25	1	4.5	3.7	3	38.9	82.0	0.52	0
26	1	3.0	3.7	3	38.4	67.0	0.58	0
27	0	2.5	2.9	3	40.3	13.5	0.90	1
28	1	3.5	4.1	4	38.4	102.0	0.30	0
29	1	2.5	3.4	2	38.7	57.0	0.41	0
30	1	2.5	3.6	1	37.8	5.7	260.00	1
31	0	4.0	3.4	2	41.1	8.0	15.00	1
32	1	2.0	3.8	3	39.3	32.0	10.00	1

Tab. 2.3: Rohdaten

NR	Identifikationsnummer	V	Verhalten
G	Geschlecht	KT	Körpertemperatur
Alt	Alter	LZ	Leukozytenzahl
KGW	Körpergewicht	ZZ	Zellzahl
		DIAG	Diagnose

Merkmale, wie z.B. das Geschlecht, ermöglichen nur eine Einteilung der Versuchstiere in Gruppen, die untereinander als gleichwertig betrachtet werden müssen. Bei anderen Merkmalen, wie z.B. beim Verhalten, ist zwar eine Rangfolge der Merkmalsausprägungen erkennbar, jedoch sind die Unterschiede zwischen den Ausprägungen ohne genauere Information nicht interpretierbar. Für Merkmale wie das Körpergewicht beobachtet man bei den Untersuchungseinheiten Zahlen, die direkt interpretiert werden können. Größenunterschiede sind daher messbar.

In der Statistik gibt es verschiedene Möglichkeiten, Merkmale nach ihren Eigenschaften zu klassifizieren.

2.3.1 qualitativ — quantitativ

Die Klassifizierung in qualitative und quantitative Merkmale ist in den meisten Fällen relativ einfach.

Die Ausprägungen **qualitativer Merkmale** unterscheiden sich durch ihre **Art**.

Die Ausprägungen **quantitativer Merkmale** unterscheiden sich durch ihre **Größe**.

> Im Beispiel sind die Merkmale Geschlecht und Diagnose qualitativ. Auch der Name der Tiere ist ein qualitatives Merkmal. In Tabelle 2.3 wurde der Name durch eine Identifikationsnummer ersetzt. Diese Nummer ermöglicht nur die Unterscheidung der Tiere voneinander und ist deshalb qualitativ. Andere Beispiele für qualitative Merkmale sind Rasse, Haarfarbe, Beruf usw..
>
> Quantitative Merkmale sind das Körpergewicht, die Körpertemperatur, die Leukozytenzahl, das Alter und die Zellzahl.
>
> Die Zuordnung des Merkmals Verhalten in eine dieser Klassen ist schwierig. Da die Ausprägungen einer Rangfolge unterliegen, lassen sich Zuordnungen in beide Klassen begründen. Ähnliche Probleme treten bei der Einordnung von Schulnoten, Güteklassen usw. auf.

2.3.2 Skalen

Man kann für jedes Merkmal eine **Skala** festlegen, die alle möglichen Ausprägungen dieses Merkmals beinhaltet. Die Art dieser Skalen zieht man ebenfalls zur Klassifizierung von Merkmalen heran. Hier werden **drei Skalenniveaus** unterschieden. Je nach Datenqualität kann man auf die einzelnen Beobachtungswerte verschiedene Rechenoperationen anwenden. Mit steigendem Skalenniveau hat man immer größere Möglichkeiten bei der Datenauswertung.

Nominalskala

Die Nominalskala hat das niedrigste Niveau.

Beobachtet man ein **nominales Merkmal** bei mehreren Versuchseinheiten, so kann man feststellen, ob die beobachteten Werte gleich oder ungleich sind. Die möglichen Ausprägungen unterliegen keiner Rangfolge.

> Geschlecht und Diagnose sind im Beispiel 2.1 **nominalskaliert**.
>
> Andere Beispiele sind Rasse, Augenfarbe, Beruf usw.

Ordinalskala

Bei **ordinalen Merkmalen** unterliegen die möglichen Ausprägungen einer gewissen Ordnung oder Rangfolge.

Vergleicht man zwei Beobachtungswerte, so kann man wiederum feststellen, ob sie gleich oder verschieden sind. Zusätzlich kann man bei verschiedenen Werten sagen, welcher größer, besser, intensiver ist. Der Abstand zweier Beobachtungen bzw. ihrer Codierungen ist nicht interpretierbar. Daher ist es nicht sinnvoll, die Beobachtungswerte ordinaler Merkmale zu addieren oder zu subtrahieren.

Im Beispiel ist das Verhalten ein **ordinalskaliertes** Merkmal.

Die Katzen mit den Nummern 1 und 3 wurden als ruhig (=2) bzw. erregt (=4) eingestuft. Der Abstand 4-2=2 ist nicht interpretierbar.
Betrachtet man nur die Katzen Nr.7 (sehr erregt = 5) und Nr.15 (sehr ruhig = 1), so käme man, wenn man den Durchschnitt berechnet, d.h. $(5 + 1)/2 = 3$, zu der zweifellos unsinnigen Aussage: Die beiden Katzen zeigen im Durchschnitt normales Verhalten (=3).

Metrische Skala

Die möglichen Ausprägungen **metrischer Merkmale** unterliegen einer Reihenfolge, zusätzlich sind die Abstände zwischen den Werten interpretierbar. Damit sind auch Rechnungen mit den Beobachtungen sinnvoll.

Die metrische Skala wird auch **Kardinalskala** genannt.

Das Körpergewicht, das Alter, die Leukozytenzahl, die Zellzahl und die Körpertemperatur sind **metrisch skalierte** Merkmale.

Die Katze Nr.1 ist um 0.2 kg leichter als die Katze Nr.2.
Das Durchschnittsgewicht der 32 Katzen beträgt 3.503 kg.

2.3.3 diskret — stetig

Eine weitere Art der Klassifizierung von Merkmalen ist die Einteilung in stetige und diskrete Merkmale.

Ein **diskretes Merkmal** kann nur endlich viele oder höchstens abzählbar unendlich viele Ausprägungen besitzen. Es kann nur bestimmte Werte annehmen. Nominale und ordinale Merkmale sind grundsätzlich diskret.

Im Beispiel sind die Identifikationsnummer, das Verhalten, die Zellzahl und die Diagnose diskrete Merkmale.

Stetige Merkmale können theoretisch in einem bestimmten Bereich jeden beliebigen Wert annehmen.

Das Alter, das Körpergewicht und die Körpertemperatur sind stetige Merkmale.

In der Praxis kann man stetige Merkmale nur bis zu einer bestimmten Genauigkeit messen. Damit ist zwangsläufig eine Diskretisierung verbunden. Entweder sind nur 'ungenaue' Informationen verfügbar oder man muss sich an der Genauigkeit von Messgeräten orientieren. Außerdem ist in fast allen Fällen nur eine begrenzte Genauigkeit für die Untersuchung notwendig.

In dem Tierheim, in dem die im Beispiel beschriebene Untersuchung stattfand, waren nur sehr lückenhafte Informationen über das Alter verfügbar. Man hat sich auf Angaben mit halbjähriger Genauigkeit beschränkt.
Beim Körpergewicht und bei der Körpertemperatur hat man eine Stelle hinter dem Komma bestimmt.

2.3.4 Zusammenfassung

Es wurden drei Klassifizierungskriterien für Merkmale vorgestellt. Sie sollen hier zusammen mit den möglichen Rechenoperationen noch einmal in Tabelle 2.4 zusammengefasst werden.

<table>
<tr><th></th><th>Skala</th><th>mögliche Rechen-operationen für zwei Beobachtungen a und b</th><th></th></tr>
<tr><td rowspan="2">qualitativ</td><td>Nominalskala</td><td>$a = b$ oder $a \neq b$</td><td rowspan="4">diskret</td></tr>
<tr><td rowspan="2">Ordinalskala</td><td>$a = b$ oder $a \neq b$</td></tr>
<tr><td rowspan="5">quantitativ</td><td>$a < b$ oder $a > b$</td></tr>
<tr><td rowspan="4">metrische Skala</td><td>$a = b$ oder $a \neq b$</td></tr>
<tr><td>$a < b$ oder $a > b$</td><td rowspan="3">stetig</td></tr>
<tr><td>$a + b$ und $a - b$</td></tr>
<tr><td>$(a \times b$ und $a \div b)$</td></tr>
</table>

Tab. 2.4: Charakterisierung von Merkmalen

Wie Tabelle 2.4 zeigt, kommt mit steigendem Skalenniveau jeweils eine weitere Rechenoperation zu den schon vorhandenen hinzu. Umgekehrt bedeutet dies für die statistische Analyse, dass man jede Skala durch eine mit niedrigerem Niveau ersetzen kann.

Für den Übergang von der metrischen Skala zur Ordinalskala gibt es verschiedene Möglichkeiten.

2.3.4.1 Klassenbildung

Wenn man im Beispiel nicht an genauen Körpergewichtsangaben interessiert wäre, könnte man Klassen bilden; eine Möglichkeit zeigt Tabelle 2.5.

Körpergewicht	Klasse	Code
$x \leq 3.0$ kg	sehr leicht	1
3.0 kg $< x \leq 3.3$ kg	leicht	2
3.3 kg $< x \leq 3.7$ kg	mittel	3
3.7 kg $< x \leq 4.0$ kg	schwer	4
4.0 kg $< x$	sehr schwer	5

Tab. 2.5: Klassen für die Körpergewichte der Katzen

Das Körpergewicht ist ein metrisch skaliertes Merkmal. Die Abstände zwischen Beobachtungswerten sind daher interpretierbar.

Die Katze Nr.3 ist um 0.2 kg schwerer als die Katze Nr.4.

Durch Klassenbildung ist das metrisch skalierte Merkmal in ein ordinales Merkmal überführt worden. Abstände zwischen den numerischen Codes für die verschiedenen Gewichtsklassen sind nicht mehr interpretierbar. Allerdings besteht noch eine Rangfolge, mit deren Hilfe man für Katzen aus verschiedenen Gewichtsklassen feststellen kann, welche die schwerere ist.

2.3.4.2 Vergabe von Rängen

Neben der Klassenbildung gibt es noch eine andere Möglichkeit, ein metrisches Merkmal in ein ordinales zu überführen, die Vergabe von Rängen.

In Tabelle 2.6 wird nur die Teilstichprobe mit den Katzen Nr. 1 bis Nr. 12 betrachtet.

Nr.	KGW	Rang
1	3.5	7.5
2	3.7	11
3	3.6	9.5
4	3.4	5
5	3.3	3
6	3.6	9.5
7	4.0	12
8	3.5	7.5
9	3.1	2
10	3.4	5
11	3.4	5
12	3.0	1

Tab. 2.6: Rangfolge der Körpergewichte der Katzen Nr. 1 bis Nr. 12

Die Rangzahlen, die in der dritten Spalte dieser Tabelle stehen, wurden nach folgendem Prinzip vergeben:

Die leichteste Katze bekam den Rang 1, die zweitleichteste den Rang 2, usw.. Hatten mehrere Katzen das gleiche Gewicht, so wurde als Rang der Durchschnitt der Rangzahlen vergeben, die bei unterschiedlichen Beobachtungswerten vergeben worden wären.

> Beispielsweise wiegen die Katzen Nr.3 und Nr.6 beide 3.6 kg. Sie würden bei unterschiedlichem Gewicht die Ränge 9 und 10 erhalten. So bekommen beide den Rang 9.5. Die Ausprägung 3.4 kg wurde sogar dreimal beobachtet, nämlich bei den Katzen Nr.4, 10 und 11. In der Rangfolge müssten die Ränge 4, 5 und 6 vergeben werden. Alle drei Tiere erhalten den Rang 5.

Die Ränge haben nur noch ordinales Niveau. Nur aufgrund der Ränge kann man keine Aussagen über die Größe des Gewichtsunterschiedes zweier Katzen machen. Man kann nur noch feststellen, ob die Katzen verschiedene Körpergewichte haben, und, wenn dies der Fall ist, kann man erkennen, welche Katze die schwerere ist.

Entsprechend lassen sich Merkmale mit ordinalem Niveau in nominale Merkmale überführen.

2.4 Grundgesamtheit und Stichprobe

Bisher ist nichts darüber gesagt worden, wie man die Untersuchungseinheiten auswählt, von denen Daten erhoben werden sollen.

Für die reine Beschreibung von Datenmaterial ist diese Frage unwesentlich, weil man in solchen Fällen nur Aussagen über die tatsächlich beobachteten Untersuchungseinheiten machen will. Oft sieht man jedoch die deskriptive Statistik als Vorstufe der schließenden Statistik. Man ist an Verallgemeinerungen und Schlussfolgerungen interessiert. Dann erhält die Frage nach der Herkunft der Individuen, die in die Untersuchung einbezogen werden, große Bedeutung. In diesem Zusammenhang müssen einige Begriffe eingeführt werden.

Die Menge aller Untersuchungseinheiten, über die mit Hilfe einer statistischen Studie eine Aussage gemacht werden soll, nennt man **Grundgesamtheit**.

Grundgesamtheiten müssen immer sorgfältig abgegrenzt werden.

> Im Tierheim (Beispiel 2.1) interessiert man sich für den Prozentsatz der Katzen, die an Parvovirose leiden. Die Grundgesamtheit sind alle Katzen, die am Untersuchungstag im Tierheim leben.
>
> An einem bestimmten Stichtag soll eine Volkszählung in der Bundesrepublik Deutschland durchgeführt werden. Die zugehörige Grundgesamtheit sind alle Personen, die am Stichtag in der Bundesrepublik wohnen. Es gibt besondere Richtlinien für die Berücksichtigung von Ausländern und für Personen, die an mehreren Orten einen Wohnsitz haben.
>
> Das durchschnittliche Schlachtgewicht der Masthähnchen in der Bundesrepublik Deutschland soll geschätzt werden. Für diese Untersuchung besteht die Grundgesamtheit aus allen Masthähnchen, die auf den bundesdeutschen Markt kommen.
>
> Die Grundgesamtheit für eine Untersuchung über die Häufigkeit von Lungenkrebs bei Zigaretten–Rauchern sind alle Personen, die zu irgendeiner Zeit an irgendeinem beliebigen Ort Zigaretten geraucht haben oder rauchen werden.

Diese Beispiele zeigen, dass abhängig davon, welches Ziel man mit der Untersuchung verfolgt, eine **sachliche, räumliche und zeitliche Abgrenzung** notwendig ist.

> Die Untersuchung im Tierheim ist sachlich auf Katzen, räumlich auf das Tierheim und zeitlich auf den Untersuchungstag beschränkt.
>
> Eine ebenso klare Abgrenzung ist bei der Volkszählung gegeben.
>
> Dagegen müssen bei der Studie über das Schlachtgewicht von Masthähnchen alle Hähnchen, die jemals in die Bundesrepublik geliefert wurden oder in Zukunft geliefert werden, als Elemente der Grundgesamtheit betrachtet werden, wenn kein konkreter Zeitraum vorgegeben wird.
>
> Bei der Untersuchung über den Zusammenhang von Zigaretten–Rauchern und Lungenkrebs ist man an allgemeinen Aussagen interessiert. Daher muss man die Grundgesamtheit sehr weit fassen. Sie umfasst alle Raucher ohne räumliche und zeitliche Einschränkungen.

Das Ziel einer statistischen Untersuchung, d.h. die Fragen, auf die Antworten gesucht werden, sollten für die Planung der Studie die größte Bedeutung haben. Häufig muss man jedoch auch die Struktur der Grundgesamtheit berücksichtigen. Darüber hinaus sind oft das finanzielle Budget und die Zeit, die für Durchführung und Auswertung der Untersuchung zur Verfügung steht, beschränkt.

Deshalb muss geklärt werden, ob es möglich und sinnvoll ist, alle Elemente der Grundgesamtheit zu befragen bzw. zu untersuchen.

> Im Tierheim könnte man, wenn genügend Geld und Zeit zur Verfügung stehen, alle Katzen untersuchen, weil sich alle Elemente der Grundgesamtheit an einem Ort

befinden und ein Stichtag bestimmt wurde. Allerdings stellt sich die Frage, ob eine Untersuchung aller Katzen für die Beantwortung der Frage nach dem Prozentsatz der erkrankten Tiere notwendig ist.

Bei der Volkszählung ist die Grundgesamtheit ebenfalls klar umrissen und Ziel der Untersuchung ist es, Daten von allen Personen zu erheben, die in der Bundesrepublik leben.

Die Grundgesamtheit aller Masthähnchen ist sehr umfangreich. Ihr genauer Umfang ist nicht bekannt. Man kann immer nur Tiere aus bestimmten Betrieben oder Lieferungen in die Untersuchung einbeziehen.

Der Umfang der Grundgesamtheit aller Zigaretten–Raucher ist ebenfalls unbekannt. Für jede Person kann theoretisch festgestellt werden, ob sie dieser Population angehört oder nicht. Jedoch können nicht alle Personen untersucht werden, da viele bereits verstorben bzw. noch keine Raucher oder noch nicht geboren sind.

Grundgesamtheiten, die räumlich und zeitlich klar abgegrenzt sind, deren Umfang bestimmt werden kann und in denen man alle Individuen erreichen könnte, nennt man **konkret**. Dagegen bezeichnet man Grundgesamtheiten, deren Umfang unbekannt ist und deren Elemente zum Zeitpunkt der Untersuchung zum Teil gar nicht existieren, als **fiktive Grundgesamtheiten**.

Die Katzen im Tierheim und die Bevölkerung der Bundesrepublik sind, wenn man einen Stichtag vereinbart, konkrete Grundgesamtheiten. Dagegen sind die Masthähnchen und die Zigaretten–Raucher fiktive Grundgesamtheiten.

Die Beispiele dieses Abschnitts zeigen, dass man zwei Erhebungsformen unterscheidet.

Bei einer **Totalerhebung** werden Daten von allen Individuen, die zur Grundgesamtheit gehören, erhoben. Wenn dagegen nur einige Elemente befragt bzw. untersucht werden, spricht man von **Teilerhebungen** bzw. **Stichprobenerhebungen**.

Stichprobenerhebungen werden erheblich häufiger durchgeführt. Dafür gibt es viele Gründe. Der finanzielle und zeitliche Aufwand für Stichprobenerhebungen ist geringer, außerdem sind Totalerhebungen in fiktiven Grundgesamtheiten nicht möglich. Oft sind Stichprobenerhebungen auch genauer als Totalerhebungen, da man im Detail sorgfältiger arbeiten kann. Bei einigen Fragestellungen sind Totalerhebungen sinnlos. Will man z.B. die Lebensdauer einer bestimmten Serie von Kühlschränken analysieren, so läge das Ergebnis einer Totalerhebung erst dann vor, wenn alle Kühlschränke defekt sind. Dann ist aber eine Qualitätsbeurteilung für eventuelle Käufer nicht mehr interessant.

In der Stichprobentheorie gibt es unterschiedliche Verfahren für die Auswahl von Versuchseinheiten einer Studie. Die wichtigsten werden in Kapitel 3 vorgestellt.

Zunächst wird noch kurz über verschiedene Erhebungsmethoden gesprochen.

2.5 Die Erhebung von Daten

Jeder Mensch macht täglich Beobachtungen in seiner Umwelt. Nicht alles, was man wahrnimmt, kann man gleich erklären. Wenn sich bestimmte bisher ungeklärte Beobachtungen wiederholen, werden häufig Hypothesen formuliert, deren Gültigkeit dann durch intensive Beobachtungen oder durch gezielte Versuche überprüft wird.

> Vor Jahren beobachteten Ärzte, dass viele ihrer Patienten, die einen Herzinfarkt erlitten, schon vorher auffällig hohen Blutdruck hatten. Sie vermuteten einen Zusammenhang zwischen Bluthochdruck und Herzinfarkt, der aber erst nach genaueren Untersuchungen belegt werden konnte.
>
> Ein Pharmakonzern entwickelt ein neues Schädlingsbekämpfungsmittel, das sich zur Behandlung von Obstbäumen eignet. Die zuständigen Chemiker hatten den Auftrag, ein Mittel herzustellen, das ohne größere Nebenwirkungen eine bessere Wirkung hat als bisher vorhandene Präparate. Ob das neue Mittel die gestellten Anforderungen erfüllt, muss überprüft werden, bevor es allgemein eingesetzt werden kann.

Diese beiden Beispiele haben sehr unterschiedliche Strukturen. An ihnen lassen sich grundlegende Unterschiede zwischen den zwei wichtigsten Modellen für die Datenerhebung erläutern.

2.5.1 Die Beobachtungsstudie

Betrachtet man das erste Beispiel (Bluthochdruck — Herzinfarkt), so kann man genauere Information nur durch gezielte Beobachtungen bekommen. Man wählt z.B. eine Gruppe von Patienten mit Bluthochdruck aus und verfolgt, wie viele von ihnen innerhalb eines vorher festgelegten Zeitraumes einen Herzinfarkt haben. Zum Vergleich sollte man eine zweite Gruppe von Personen mit normalem Blutdruck beobachten. Beide Gruppen sollten etwa die gleiche Altersstruktur haben und auch bezüglich anderer Merkmale wie Geschlecht, Rauchverhalten, Stressbelastung usw. vergleichbar sein. Anhand der Ergebnisse der Untersuchung kann man dann prüfen, ob ein Zusammenhang zwischen Bluthochdruck und Herzinfarkt besteht.

Das Wesen einer Beobachtungsstudie ist es, dass die **Zuordnung der Individuen zu den einzelnen Gruppen** natürlich vorgegeben ist. Im obigen Beispiel kann man nicht beeinflussen, ob ein Patient Bluthochdruck hat oder nicht. Außerdem wird deutlich, dass man nur klare Aussagen machen kann, wenn die zwei Gruppen vergleichbar sind. Sind z.B. die Patienten in der Bluthochdruck–Gruppe älter als in der anderen Gruppe, so weiß man nicht, ob eine eventuell höhere Infarkt–Rate durch Bluthochdruck oder das höhere Alter zu erklären ist.

Beobachtungsstudien, wie die oben beschriebene, nennt man häufig auch **Quasi–Experimente**. Sie werden bevorzugt durchgeführt, wenn man noch sehr wenig Information über das zu untersuchende Problem hat, wenn ein Eingriff des Untersuchers die Ergebnisse beeinflussen könnte (z.B. wenn es um das Verhalten von Tieren geht) und wenn man, wie im Beispiel, keinen Einfluss auf die Untersuchungseinheiten nehmen kann. Auch **Befragungen** kann man zu den Beobachtungsstudien zählen.

2.5.2 Das Experiment

Im Gegensatz zur Beobachtungsstudie nimmt der Untersucher beim Experiment direkt Einfluss auf die Untersuchung.

Beim Vergleich des neuen Schädlingsbekämpfungsmittels mit anderen Präparaten kann er zuordnen, welches Mittel an einem Baum eingesetzt wird. **Für die Zuordnung der Versuchseinheiten zu verschiedenen Gruppen ist eine gewisse Freiheit gegeben**. Wenn Experimente durchgeführt werden, sind häufig schon Informationen vorhanden, die z.B. aus Beobachtungsstudien stammen.

Experimente nennt man oft auch Versuche. Man unterscheidet zwei Typen von Experimenten. Bei **Laborversuchen** werden möglichst alle **Störfaktoren**, wie z.B. Raumtemperatur, Luftfeuchtigkeit, Herkunft der Versuchstiere, Haltung der Tiere usw. kontrolliert. Diese Kontrollen sind möglich, weil Laborversuche in geschlossenen Räumen stattfinden und weil der Untersucher die Versuchstage selbst auswählen kann. So erreicht man, dass die Vergleichbarkeit zwischen mehreren Gruppen gewährleistet ist und dass keine äußeren Einflüsse die Versuchsergebnisse verfälschen. **Feldversuche** finden dagegen direkt in den Ställen der Bauern, auf den Feldern usw. statt. Deshalb kann man Störfaktoren wie Wetter, Ernährung, Stallhygiene usw. weniger gut oder gar nicht kontrollieren. Die Versuchseinheiten können nur aus dem vorhandenen Material ausgewählt werden. Deshalb können auch Einflussfaktoren wie z.B. das Alter der Versuchstiere, die Bodenbeschaffenheit der Felder usw. nur selten berücksichtigt werden. Allerdings kann man die Ergebnisse in der Regel besser verallgemeinern als die von Laborversuchen.

Die Erhebung von Daten sollte stets sorgfältig geplant werden. Meistens müssen bei der Planung Einflussfaktoren berücksichtigt werden, damit die Ergebnisse einer Untersuchung später interpretierbar sind. Einige Grundlagen der Versuchsplanung findet man im Abschnitt 2.5.3.

2.5.3 Biometrische Versuchsplanung

Selten werden biologische, medizinische und landwirtschaftliche Untersuchungen mit Statistikern geplant. Dies hat jedoch nicht darin seinen Grund, dass den entsprechenden Fachgebieten eine ungenügende Anzahl von Statistikern zur Verfügung steht. Vielmehr erwartet der Experimentator oft eine derartige Regelhaftigkeit seiner Messergebnisse, dass er häufig ohne Statistiker auf Grund von Beschreibungen Schlüsse aus seinen Versuchen zieht. Die Beobachtung einer unerwartet hohen Variabilität der Untersuchungsergebnisse führt häufig später doch zur statistischen Versuchsauswertung. Allerdings lassen bei der Vielzahl der Einflussmöglichkeiten auf den Versuchsablauf häufig selbst so genannte "statistisch gesicherte" Ergebnisse keine klare Interpretation auf spezielle Einflussfaktoren zu, weil beispielsweise die Tiere nicht zufällig den Versuchsgruppen zugeordnet worden sind und dadurch **Scheinwirkungen** auftreten können (Problem der nachträglichen Versuchsplanung).

Jeder Versuch kann in die drei Phasen **Planung**, **Durchführung** und **Auswertung** aufgeteilt werden. Darüber darf jedoch nicht der Zusammenhang vergessen werden, d.h. bereits bei der Planung eines Versuches ist seine Durchführung und Auswertung zu berücksichtigen.

Grundsätzlich muss die Planung vor der Durchführung eines Versuches erfolgen. Dazu gehört die klare Formulierung der Versuchsfrage mit den Teilaspekten:

1. Was will man wissen ?
2. Wie genau will man es wissen ?
3. Für welche Grundgesamtheit will man es wissen ?

Die erste Frage fordert den Experimentator auf, seine häufig komplexe Problemstellung zu konkretisieren. Dabei muss er Sachfragen durch geeignete Merkmale charakterisieren. Soll zum Beispiel die Wirkung zweier Futtermittelkombinationen verglichen werden, könnten die Gewichtszunahmen als Messgrößen herangezogen werden, über diese Merkmale werden anschließend statistische Hypothesen formuliert.

Lautet die Antwort auf die zweite Frage, dass man es so genau wie möglich wissen möchte, ergibt sich ein unendlich großer oder der Gesamtheit entsprechender Untersuchungsumfang. Es muss daher ein Kompromiss zwischen Genauigkeit der Beantwortung der Sachfrage und der Größe des Stichprobenumfanges unter Berücksichtigung der biologischen Variation und der methodischen Fehlermöglichkeiten gefunden werden.

Die dritte Frage zielt auf die Verallgemeinerungsfähigkeit der Versuchsergebnisse, die häufig in Konkurrenz zur Vergleichbarkeit steht. Wenn zur Erzielung einer guten Vergleichbarkeit z.B. nur junge, männliche Tiere aus einer bestimmten Zucht mit einem bestimmten Gewicht etc. zum Vergleich benutzt werden, so ist zwar die Vergleichbarkeit gesichert, aber die Verallgemeinerungsfähigkeit beeinträchtigt; denn aus den Versuchen geht nicht hervor, wie sich ältere oder weibliche Tiere oder solche einer anderen Zucht verhalten würden. Man zählt derartige Versuchsreihen zu den Versuchen mit einer **schmalen induktiven Basis (geringe Verallgemeinerungsfähigkeit)**.

Dieser Gesichtspunkt ist ebenfalls eng gekoppelt mit dem gewählten Versuchstyp, wobei die **Versuche unter Laborbedingungen** (Labor, Klimakammer etc.) von denjenigen **Versuchen unter praktischen Bedingungen (Feldversuche)** zu unterscheiden sind.

Die **allgemeinen Prinzipien** der biometrischen Versuchsplanung sind für biologische Untersuchungen keine anderen als für alle übrigen empirischen Wissenschaften. Sie können im wesentlichen durch die folgenden drei Punkte charakterisiert werden:

1. **Das Prinzip der Wiederholung**
   ist erforderlich, um die biologische und methodische Variabilität (Versuchsfehler) zu erfassen, ohne deren Kenntnis keine statistische Beurteilung möglich ist.

2. **Das Prinzip der zufälligen Zuordnung (Randomisierungsprinzip)**
   soll vor systematischen Verzerrungen und damit vor Fehlinterpretationen schützen. Dadurch, dass beispielsweise jedes Tier die gleiche Chance hat, einer Versuchsgruppe zugeordnet zu werden, wird garantiert, dass Ungleichheiten des Tiermaterials im Mittel ausgeglichen werden. Die zufällige Zuordnung ist die beste Methode zur Vermeidung einseitiger Einflüsse, falls keine Information über die Versuchseinheiten vorhanden ist.

3. **Das Prinzip der Reduktion des Versuchsfehlers**
   durch Einbeziehung weiterer Informationen.
   Es beinhaltet das Problem der Steigerung der Vergleichbarkeit ohne Aufgabe der Verallgemeinerungsfähigkeit. Sind Variablen und ihr Einfluss auf die untersuchten Merkmale bekannt, so lässt sich durch Berücksichtigung dieser Zusammenhänge der Versuchsfehler verringern (z.B. Einfluss des Alters auf die Körpergröße). Eine andere Möglichkeit, den Versuchsfehler zu reduzieren, bietet die so genannte **Blockbildung**, d.h. die Einbeziehung von Gruppierungsfaktoren wie z.B. Geschlecht, Rasse.

Für die biometrische Versuchsplanung hat das Prinzip der Fehlerreduktion über die Blockbildung vor allem deshalb besondere Bedeutung, weil es häufig für den Experimentator unmöglich ist, die Steigerung der Aussagekraft der Versuchsergebnisse über die Erhöhung des Versuchsumfanges, z.B. der Tierzahl, zu erreichen.

Werden an *jedem* Merkmalsträger (z.B. an jedem Tier) wiederholte Bestimmungen unter verschiedenen experimentellen Bedingungen vorgenommen, so können diese als Block betrachtet werden. Diese Blockbildung ermöglicht die Elimination der biologischen Variation zwischen den Merkmalsträgern. Es sei an vergleichende Untersuchungen hinsichtlich der Wirksamkeit von Pharmaka erinnert, bei denen z.B. die Pharmaka den Versuchstieren in zeitlichen Abständen nacheinander verabreicht werden. Dabei müssen die zeitlichen Abstände derart gewählt werden, dass eine Wirkungsbeeinflussung der Präparate ausgeschlossen ist.

Die vielfältigen Problemstellungen haben zu zahlreichen, in der Literatur dokumentierten Versuchsplänen geführt. Für den Experimentator besteht jedoch die Schwierigkeit, einen — seinem Problem adäquaten — Versuchsplan aufzufinden bzw. zu entwickeln.

Kapitel 3

Stichprobenverfahren

Eine **Stichprobe** ist eine Menge von n Individuen, die aus einer bestimmten Grundgesamtheit mit N Elementen ausgewählt wurde. Der **Stichprobenumfang** wird mit n bezeichnet. Es gilt $\mathbf{n} < \mathbf{N}$.

Für die Auswahl von Stichproben gibt es verschiedene Verfahren.

3.1 Stichprobenerhebungen mit Hilfe nicht zufälliger Auswahlverfahren

3.1.1 Willkürliche Auswahl

Wenn man Versuchseinheiten auf's Geratewohl für eine Stichprobe auswählt, bezeichnet man dies als **willkürliche Auswahl**.

> Beispiele sind die Befragung von Passanten auf der Straße, die willkürliche Auswahl einiger Versuchstiere aus einem Käfig usw.

Man spricht ebenfalls von willkürlicher Auswahl, wenn man Freiwillige in eine Stichprobe einbezieht.

Bei der willkürlichen Auswahl kann man **Verzerrungen** — so bezeichnet man Abweichungen der Ergebnisse, die man aus der Stichprobe erhält, von den tatsächlichen Verhältnissen in der Grundgesamtheit — nicht ausschließen. Solche Verzerrungen nennt man auch **systematische Fehler**.

> Befragt man Passanten auf der Straße, so erreicht man keine kranken Personen, eventuell keine Arbeitnehmer, je nach Standort und Tageszeit u.U. mehr Frauen als Männer usw...
>
> Greift man Tiere willkürlich aus einem Käfig heraus, dann ergreift man zuerst die langsameren, schwächeren Tiere.
>
> Bauern, die sich freiwillig für eine Untersuchung der hygienischen Bedingungen in ihren Ställen melden, haben wahrscheinlich sauberere Ställe als viele andere Bauern.
>
> Wenn man eine Untersuchung in der Praxis eines bestimmten Arztes durchführt, erreicht man nur Personen, die zum Patientenkreis dieses Arztes gehören und u.U. alle aus derselben gesellschaftlichen Schicht stammen.

Eine besondere Form der willkürlichen Auswahl ist die Sammlung zeitlich anfallenden Materials. Die Individuen werden in der Reihenfolge in die Stichprobe aufgenommen, in der sie zur Verfügung stehen.

> Bei einer Umfrage unter Arbeitnehmern, die in einer bestimmten Firma beschäftigt sind, wird die Befragung am Fabriktor durchgeführt. Frühmorgens um 6 Uhr wird ein Interviewer fast ausschließlich Arbeiter antreffen, dagegen erreicht er um 8 Uhr hauptsächlich Angestellte.

Wenn man eine willkürliche Auswahl durchführt, kann man bei der Datenauswertung niemals mit unverzerrten Ergebnissen rechnen. In der Regel wird eine so erhaltene Stichprobe nicht **repräsentativ** für die Grundgesamtheit sein, die man untersuchen will, d.h. sie wird kein "verkleinertes Abbild" dieser Grundgesamtheit sein. Deshalb sollte man möglichst andere Verfahren verwenden.

3.1.2 Systematische Auswahl

Eine etwas bessere, jedoch nicht optimale Auswahltechnik ist die systematische Auswahl von Versuchseinheiten. Ausgehend von einem **bestimmten Startelement** wird jedes k-te Individuum in die Stichprobe einbezogen.

> Aus der Kartei eines großen Krankenhauses oder einer Behörde wählt man jede k-te Karteikarte für die Untersuchung aus.

Man kann auch Personen mit bestimmten Geburtsdaten oder mit Nachnamen, die mit bestimmten Buchstaben beginnen, für die Stichprobe auswählen. Allerdings sollte man bei der Interpretation von Ergebnissen aus systematischen Stichproben ebenfalls sehr vorsichtig sein.

3.1.3 Quotenauswahl

Bei Befragungen werden häufig Quotenverfahren durchgeführt. Jedem Interviewer wird gesagt, wie viele Personen mit bestimmten Merkmalsausprägungen bzw. Kombinationen von Merkmalsausprägungen er befragen muss. Diese Quoten legt man möglichst so fest, dass sie für die Verhältnisse in der Grundgesamtheit repräsentativ sind.

> Die beiden Geschlechter, die verschiedenen Altersklassen, die unterschiedlichen Berufsgruppen usw. sollten in Stichprobe und Grundgesamtheit mit jeweils gleichen Anteilen vertreten sein.

Wenn man eine Quotenauswahl durchführt, kann man in der Regel mit besseren Ergebnissen rechnen als bei der rein willkürlichen Auswahl. Die Versuchseinheiten werden nur noch innerhalb der durch die Quoten gesetzten Grenzen willkürlich ausgewählt.

Wenn bei der Quotierung alle Merkmale berücksichtigt werden, die für die jeweilige Grundgesamtheit von Bedeutung sind, erhält man meistens Ergebnisse, die die Verhältnisse in der Grundgesamtheit gut beschreiben. Beispielsweise beruhen Wahlprognosen in der Bundesrepublik Deutschland auf Stichproben, die mittels Quotenauswahl erzeugt wurden. Leider sind Quotierungen von ausreichender Genauigkeit nur selten möglich, weil entweder zu wenig über die betreffende Grundgesamtheit bekannt ist oder weil der Aufwand für die Erhebung zu groß wäre. Deshalb kann auch durch die Quotenauswahl nicht garantiert werden, dass eine Stichprobe ein repräsentatives Abbild der Grundgesamtheit ist.

3.2 Stichprobenverfahren mit Hilfe zufälliger Auswahl

3.2.1 Die einfache Zufallsauswahl

Der wichtigste Grundsatz bei der zufälligen Auswahl von Individuen für eine Stichprobe ist, dass alle Elemente der Grundgesamtheit die gleiche Chance haben müssen, in die Stichprobe zu gelangen. Damit hat jede denkbare Kombination von n Individuen, die aus der Grundgesamtheit ausgewählt werden kann, die gleiche Chance, als Stichprobe zu dienen.

Eine solche **einfache Zufallsstichprobe** vom Umfang n erhält man, wenn man alle Elemente der Grundgesamtheit von l bis N durchnummeriert und dann zufällig n Elemente auswählt.

Es gibt verschiedene Möglichkeiten für die zufällige Auswahl. Bei kleinen Grundgesamtheiten kann man durch Ziehen von Losen, durch Werfen einer Münze oder durch Würfeln entscheiden, welche Individuen in die Stichprobe gelangen. Bei größeren Grundgesamtheiten verwendet man **Zufallszahlen–Tabellen**, die aus einer Serie der Ziffern $0, 1, \ldots, 9$ in zufälliger Reihenfolge bestehen.

Tabelle 3.1 ist ein Ausschnitt aus einer erheblich größeren Zufallszahlen–Tabelle. Der Übersichtlichkeit halber wurden die zufällig ermittelten Ziffern in Blöcken von je fünf Ziffern zusammengefasst. Die Durchführung einer Zufallsauswahl mit Hilfe einer Zufallszahlen–Tabelle soll an einem Beispiel erläutert werden:

> Aus einer Grundgesamtheit von 82 Ratten sollen 10 Tiere zufällig ausgewählt werden. Man nummeriert alle Tiere von l bis 82 durch. Dann wählt man aus der Tabelle eine beliebige Zeile und eine beliebige Ziffernspalte aus. Der so erhaltene Punkt gilt als Startpunkt. Hier soll der Einfachheit halber oben links begonnen werden. Man erhält die Ziffernfolge
>
> 85284 74047 13318 41800 33178 17478 47273 17680 55626 9 ...
>
> Aus dieser Folge bildet man zweistellige Zahlen:
>
> 85 28 47 40 47 13 31 84 18 00 33 17 81 74 78 47 27 31 76 ...
>
> Die Tiere mit den Nummern, die den ersten zehn so erhaltenen Zahlen entsprechen, gelangen in die Stichprobe. Zahlen, die größer sind als der Umfang der Grundgesamtheit, werden übersprungen. Erhält man mehrfach die gleiche Nummer, so wird sie nur einmal berücksichtigt.
>
> Somit wählt man die Tiere mit den Nummern 28, 47, 40, 13, 31, 18, 33, 17, 81 und 74 für die Stichprobe aus.

3.2.2 Kompliziertere Stichprobenpläne

Es gibt viele Möglichkeiten, komplizierte Stichprobenpläne zu erstellen, die auf der einfachen Zufallsauswahl beruhen.

Die systematische Auswahl jedes k-ten Elements aus einer Grundgesamtheit ist eine Zufallsstichprobe, wenn man mit einem **zufällig ausgewählten Startelement** beginnt. Solche Stichproben liefern oft gute Ergebnisse.

Man kann die Grundgesamtheit bzgl. eines bestimmten Merkmals **schichten**, d.h. man fasst Individuen mit gleichen bzw. ähnlichen Ausprägungen dieses Merkmals zu **Schichten** zusammen.

85284	74047	13318	41800	33178	17478	47273	17680	55626	98239	02693	61685
82690	74413	63976	42538	35188	10443	07245	74504	94651	97901	18277	87904
98957	69510	49031	72317	36766	30597	46887	30221	31408	69974	54113	77616
06416	37932	24658	21748	10604	21084	65441	65661	56684	92765	04557	84710
69562	33657	72657	47820	06394	58017	85811	22565	57445	73096	28967	51710
91104	91346	50171	30911	17560	32528	95711	06472	77856	18393	30492	74421
07057	12646	49343	10275	84811	17835	60717	71610	52055	90811	63851	48335
46980	00822	13643	92802	23967	11410	69977	11102	83173	94040	25592	30360
16715	21114	60598	72129	77956	06104	91572	57644	17741	65623	28714	90033
96792	79064	23033	15746	35354	87213	93536	59632	32238	21240	72150	25268
08643	69761	73969	89994	22870	79099	24128	17714	16332	86298	12158	40579
17457	03426	83884	34387	41206	53192	97483	96651	20266	17545	82355	37699
14343	67163	07248	25039	34149	34819	01940	10716	01166	99045	51278	24757
30937	53562	23779	57496	35525	61832	17868	11869	78338	19800	72138	15450
47787	58293	24719	52620	78667	56194	45425	56699	43798	05417	43520	47350
06819	06101	41028	60301	74097	52343	24064	40259	35354	94439	29222	33515
48537	62266	02920	83608	95679	78519	75440	12893	92736	09429	78085	73858
39128	23856	40064	58852	30996	47674	58852	19242	04224	92325	02903	84082
44391	71804	03455	59952	20191	43608	18900	48950	01718	76398	24349	35427
59849	77064	16065	03922	16788	58574	52730	40391	54762	85613	99038	36049
71058	76914	90319	86954	34867	03155	33925	83909	65013	65997	14828	20499
51671	34161	51290	31928	20969	32170	74941	33173	35808	24147	35765	01579
86373	75292	29271	64758	92053	33831	92018	47379	97251	02032	60166	15683
22453	59581	85638	10397	41113	82584	83710	18219	09065	48886	23428	47185
16531	31939	06728	81570	42829	26581	54770	27435	95296	41437	33656	58340
03901	56091	23920	15450	61239	41080	38214	67630	57155	99865	31230	86842
71860	50503	05690	31083	10592	12660	82607	67741	19635	04984	62793	63393
36972	89156	42356	78827	41903	63244	34760	09807	28396	56185	03106	94638
69627	18830	76201	16633	46670	88869	13053	76387	32810	32217	65359	86850
06493	30836	56788	41940	84502	30685	29685	16319	80796	45534	85302	74447
01368	91244	37842	04564	05474	04519	46041	08945	32375	23102	67537	88550
87497	60714	15110	54455	32301	84787	11478	08299	87765	60738	28192	24141
82256	70039	51746	89319	76669	72617	73742	88244	15888	28895	41557	56624
65780	61354	76418	54910	70675	29969	84615	19871	68044	20850	24795	31829
19995	48694	99657	32743	11886	67763	92433	23366	15431	54733	00378	56177
95177	45793	39381	70857	89073	43995	27617	56108	02363	14259	48524	41630
40750	79301	07511	40113	83464	73269	33890	90358	39164	36436	86890	64690
54058	54560	82327	76044	67200	32807	87825	83025	99302	61341	52646	26473
09605	26718	51052	35379	12717	39998	53819	37161	57408	54037	95160	56638
63537	59730	77554	42122	47919	67387	75396	87099	72882	34134	93967	05435
83053	71939	80978	98126	05243	19421	14814	75265	70759	44521	61201	99827
29905	13509	42347	22387	59365	40157	19012	41822	94376	70551	43229	43101
31018	27179	89231	05048	48788	78405	45815	18783	80825	20613	44729	68466
65909	25184	65144	76402	84931	43137	04287	46000	19465	47858	51725	37871
43611	72340	15275	18950	88417	29021	50683	22519	75947	46190	15679	17748

Tab. 3.1: Zufallszahlen

> Wenn man z.B. das durchschnittliche Einkommen einer Berliner Familie bestimmen will, kann man als **Schichtungskriterium** das Alter des Haushaltsvorstands verwenden. Man bildet Altersklassen und Familien, deren Haushaltsvorstände den gleichen Altersklassen angehören, bilden jeweils eine Schicht.

Innerhalb der einzelnen Schichten zieht man dann Zufallsstichproben mit bestimmtem Umfang. Mit einer solchen **geschichteten Zufallsauswahl** erzielt man häufig für die Grundgesamtheit repräsentativere Ergebnisse als mit der einfachen Zufallsauswahl.

Eine andere Strategie ist die **Klumpenauswahl**. Sie wird häufig verwendet, um den finanziellen und zeitlichen Aufwand für eine Untersuchung gering zu halten, und erfolgt oft nach geographischen Gesichtspunkten.

> Im obigen Beispiel kann man z.B. einen Berliner Stadtplan für die Auswahl benutzen. Der Plan ist in Quadrate gegliedert, die man als Klumpen betrachten kann.

Man wählt zufällig einen oder mehrere **Klumpen** aus und führt in diesen jeweils Totalerhebungen durch. Die Ergebnisse, die man aus einer mittels Klumpenauswahl erzeugten Stichprobe erhält, sind in der Regel schlechter als bei Verwendung einer einfachen Zufallsstichprobe.

In der Praxis werden sehr häufig **mehrstufige Stichprobenpläne** verwendet. Man kann z.B. eine Klumpenstichprobe ziehen und innerhalb der Klumpen eine geschichtete Zufallsauswahl durchführen.

3.3 Allgemeines zur Planung und Auswertung von Stichprobenerhebungen

Durch die zufällige Auswahl einer Stichprobe werden systematische Fehler vermieden. Sie ist diesbezüglich einer nicht zufälligen Auswahl fast immer überlegen. Allerdings muss man häufig auf zumindest teilweise nicht zufällige Auswahlverfahren zurückgreifen. Das geschieht, wie schon erwähnt, um Kosten und Zeit zu sparen.

Bei fiktiven Grundgesamtheiten ist eine Zufallsauswahl nicht möglich, da nicht alle Elemente der Grundgesamtheit erreichbar sind. Man muss Versuchseinheiten aus einer Teilpopulation für die Untersuchung auswählen.

Im Kapitel ‘Biometrische Versuchsplanung’ (Kapitel 2.5.3) wurden die einfachsten Grundsätze für die Auswahl der Versuchseinheiten und die Durchführung von Experimenten bzw. Beobachtungsstudien beschrieben.

Vor der Auswertung von Daten sollte man immer die Datenqualität überprüfen. Man sollte klären, ob systematische Verzerrungen aufgetreten sind und gegebenenfalls versuchen, sie bei der Auswertung zu berücksichtigen.

Man muss klären, ob alle ermittelten Daten tatsächlich in die Auswertung einbezogen oder ob einzelne extreme Beobachtungen (Ausreißer) nicht berücksichtigt werden sollen.

> Dies kann z.B. vorkommen, wenn die Untersuchung auf gesunde Tiere beschränkt ist und dennoch Daten von kranken Tieren ermittelt wurden.

Für diese und viele andere Probleme gibt es statistische Verfahren, die es ermöglichen, die Ergebnisse der Datenauswertung zu verbessern.

Kapitel 4

Datenaufbereitung

Aus einer Datenerhebung erhält man in der Regel sehr unübersichtliches Material, das aufbereitet werden muss, damit man sich einen Überblick verschaffen kann.

Einen ersten Eindruck vermitteln besonders gut einfache Schaubilder. Das Ziel dieses Kapitels ist es, die Konstruktion einiger Grafiken zu erläutern, die in der deskriptiven Statistik häufig verwendet werden.

Dabei werden die beiden folgenden Beispiele betrachtet.

4.1 Beispiel: kariöse Zähne bei Schulkindern

Bei $n = 100$ Schulkindern eines Jahrgangs werden die Zähne auf Karies untersucht. Für jedes Kind wird die Anzahl der kariösen Zähne registriert. Die Anzahl der kariösen Zähne ist ein diskretes metrisches Merkmal. Tabelle 4.1 enthält die Urliste.

1	0	0	3	1	5	1	2	2	0	1	0	5	2	1	0	1	0	0	4
0	1	1	3	0	1	1	1	3	1	0	1	4	2	0	3	1	1	7	2
0	2	1	3	0	0	0	0	6	1	1	2	1	0	1	0	3	0	1	3
0	5	2	1	0	2	4	0	1	1	3	0	1	2	1	1	1	1	2	2
0	3	0	1	0	1	0	0	0	5	0	4	1	2	2	7	1	3	1	5

Tab. 4.1: Anzahl kariöser Zähne bei 100 Schulkindern (Daten entnommen aus Hartung, 1998)

Der kleinste Beobachtungswert ist $x_{min} = 0$, der größte $x_{max} = 7$. Insgesamt sind acht verschiedene Merkmalsausprägungen beobachtet worden.

4.2 Beispiel: Schlachtgewicht von Masthähnchen

Aus einem Bestand von ausgeschlachteten Masthähnchen eines bestimmten einheitlichen Alters wird eine Zufallsstichprobe vom Umfang $n = 100$ gezogen. Dann wird das Schlachtgewicht der ausgewählten Tiere ermittelt. Das Schlachtgewicht ist ein stetiges metrisches Merkmal. Hier wurden allerdings die Werte nur mit einer Genauigkeit von 10 g gemessen.

Aus der Erhebung erhielt man folgende Urliste:

1480	1380	1540	1065	1450	1440	1570	1590	1560	1380
1570	1590	1455	1505	1370	1370	1675	1680	1390	1410
1490	1120	1270	1400	1590	1210	1410	1620	1480	1530
1390	1860	1340	1460	1460	1320	1230	1350	1380	1380
1520	1420	1540	1300	1550	1130	1790	1540	1410	1630
1450	1620	1490	1500	1530	1280	1360	1410	1200	1390
1450	1420	1280	1340	1490	1630	1200	1290	1310	1550
1280	1420	1590	1270	1460	1620	1270	1470	1550	1460
1450	1360	1350	1680	1360	1540	1460	1540	1380	1340
1500	1550	1410	1420	1580	1500	1230	1560	1570	1660

Tab. 4.2: Schlachtgewichte von 100 Masthähnchen in Gramm

Beim Durchsehen der Daten, findet man $x_{min} = 1065\ g$ und $x_{max} = 1860\ g$. Die Gewichte schwanken also innerhalb eines recht großen Bereiches. Daher sind viele verschiedene Merkmalsausprägungen möglich.

4.3 Der Häufigkeitsbegriff

Wenn Daten wie in den beiden Beispielen vorliegen, ist man häufig nicht an der Reihenfolge interessiert, in der die Beobachtungen gemacht wurden, sondern nur daran, wie häufig die einzelnen Ausprägungen in der Beobachtungsreihe vorkommen.

Hier soll folgende allgemeine Terminologie verwendet werden:
Betrachtet wird ein Merkmal X, das ist im Beispiel 4.1 die Anzahl der kariösen Zähne, im Beispiel 4.2 das Schlachtgewicht eines Hähnchens. Die verschiedenen Ausprägungen des Merkmals X werden mit A_1, A_2, $A_3, \ldots,$ A_k bezeichnet.

Für das Beispiel 4.1 definiert man A_1 ={0 kariöse Zähne}, A_2 ={1 kariöser Zahn}, ... A_8 ={7 kariöse Zähne}.

Nun zählt man für jede Ausprägung $A_j\,(j = 1, 2, \ldots, k)$, wie häufig sie beobachtet wurde. Wenn das Material nicht zu umfangreich ist, verschafft man sich einen ersten Überblick am besten, indem man eine Strichliste erstellt. Bei großen Datensätzen werden für diese Zwecke EDV-Anlagen und oft spezielle Programmpakete für statistische Analysen eingesetzt.

Anzahl kariöser Zähne	$H(A_j) = n_j$	$h(A_j)$	$F(A_j)$
0	30	0.30	0.30
1	34	0.34	0.64
2	14	0.14	0.78
3	10	0.10	0.88
4	4	0.04	0.92
5	5	0.05	0.97
6	1	0.01	0.98
7	2	0.02	1.00

Tab. 4.3: Häufigkeitstabelle für die Daten aus Beispiel 4.1

Definition: Die Anzahl der Versuchseinheiten, bei denen die Ausprägung A_j beobachtet wurde, nennt man **absolute Häufigkeit der Ausprägung A_j.**
Sie wird mit $H(A_j)$ bzw. n_j bezeichnet. Es gilt stets

$$\sum_{j=1}^{k} H(A_j) = \sum_{j=1}^{k} n_j = n.$$

Die absolute Häufigkeit von Kindern mit 2 kariösen Zähnen ist
H(2 kariöse Zähne)= 14.

Die absoluten Häufigkeiten hängen stark von der Gesamtzahl der Beobachtungen n ab. Daher eignen sie sich nicht zum Vergleich von Erhebungen unterschiedlichen Umfangs.

Will man Erhebungen unterschiedlichen Umfangs vergleichen, so benutzt man besser die relative Häufigkeit.

Definition: Die **relative Häufigkeit** $h(A_j)$ einer Ausprägung A_j ist definiert als

$$h(A_j) = \frac{\text{Anzahl der Versuchseinheiten mit Ausprägung } A_j}{\text{Anzahl aller Versuchseinheiten}} = \frac{1}{n} H(A_j)$$

Sie gibt für $j = 1, 2, \ldots, k$ den Anteil aller Versuchseinheiten an, bei denen das Merkmal X die Ausprägung A_j hat.

Für die relativen Häufigkeiten gilt für jeden Stichprobenumfang n

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

Die relative Häufigkeit von Kindern mit 2 kariösen Zähnen ist
h(2 kariöse Zähne)$= \frac{14}{100} = 0.14$.

Absolute und relative Häufigkeiten können für die Ausprägungen von Merkmalen jeglichen Skalenniveaus bestimmt werden.

Hat man bei ordinalen und metrischen Merkmalen die möglichen Ausprägungen der Größe nach geordnet, so kann man auch Summenhäufigkeiten betrachten:

Definition: Die **absolute Summenhäufigkeit der Ausprägung** A_j ist die Anzahl aller Versuchseinheiten, bei denen eine der Ausprägungen $A_1, A_2, \ldots, A_j$, d.h. eine Ausprägung, die $\leq A_j$ ist, beobachtet wurde. Man berechnet

$$H(A_1) + \ldots + H(A_j) = \sum_{i=1}^{j} H(A_i) \qquad \text{für } j = 1, 2, \ldots, k.$$

Definition: Entsprechend ist die **relative Summenhäufigkeit von** A_j der Anteil von Untersuchungseinheiten mit Ausprägungen $\leq A_j$:

$$h(A_1) + \ldots + h(A_j) = \sum_{i=1}^{j} h(A_i) \qquad \text{für } j = 1, 2, \ldots, k.$$

Sie wird auch mit $\mathbf{F}(\mathbf{A_j})$ bezeichnet.

Die relative Summenhäufigkeit von Kindern mit höchstens 2 kariösen Zähnen ist
F(2 kariöse Zähne) $= h(0) + h(1) + h(2) = 0.30 + 0.34 + 0.14 = 0.78$

4.4 Gruppierung von Daten

Wenn man ein Merkmal wie das Schlachtgewicht der Masthähnchen betrachtet, das sehr viele verschiedene Merkmalsausprägungen annehmen kann, wird die Häufigkeitsverteilung sehr unübersichtlich. Das ist besonders häufig bei stetigen Merkmalen der Fall, bei denen u.U. nur selten identische Beobachtungen vorkommen. Man fasst dann verschiedene Ausprägungen zu **Merkmalsklassen** zusammen. Diese Vorgehensweise wurde schon angesprochen als eine Möglichkeit, ein metrisches Merkmal in ein ordinalskaliertes Merkmal zu überführen.

Hier soll die **Klassenbildung**, die man oft auch Gruppierung nennt, noch einmal genauer erläutert werden.

Klassen bzw. Gruppen sind Teilintervalle des Intervalls, in dem alle Beobachtungswerte liegen. Eine Klasse wird eindeutig durch die Angabe von **unterer und oberer Klassengrenze** bzw. **Klassenmitte und Klassenbreite** bestimmt. Zusätzlich muss festgelegt werden, zu welcher Klasse die untere und die obere Klassengrenze gehören.

In diesem Skript wird folgende Konvention verwendet: **Zu jeder Klasse gehört die obere Klassengrenze**.

Damit ist gewährleistet, dass jeder Merkmalswert nur zu **genau einer** Größenklasse gehört. Wenn man die Klassengrenzen so wählt, dass dort keine Beobachtungswerte liegen, erspart man sich Schwierigkeiten bei der Zuordnung der Beobachtungen in die einzelnen Gruppen. Die Auswahl der Klassengrenzen bleibt demjenigen überlassen, der die Daten auswertet. Genauso ist die Anzahl der Klassen grundsätzlich beliebig. Meistens bildet man jedoch nicht mehr als 20–25 Klassen, damit die Ergebnisse und Grafiken überschaubar bleiben. Man sollte möglichst keine nach unten oder nach oben offenen Klassen — wie z.B. "1800 g und schwerer" — bilden, weil die Beobachtungshäufigkeiten in solchen Klassen grafisch nicht darstellbar sind.

Durch die Gruppierung von Daten gehen Informationen, die die Rohdaten beinhalten, verloren.

> Bei den Masthähnchen kann man aus den gruppierten Daten nicht mehr die genauen Schlachtgewichte der einzelnen Tiere ermitteln.

Je weniger Klassen es gibt, desto größer ist der Informationsverlust. Daher sollte man nicht zu breite Klassen bilden. Wenn man gleichbreite Klassen wählt, erspart man sich viel rechnerischen Aufwand. Jedoch kann es vorkommen, dass aus sachlichen Gründen Klassen verschiedener Breite besser geeignet sind.

4.5 Häufigkeitsbegriff bei gruppierten Daten

Genauso wie im Abschnitt 4.3 absolute und relative Häufigkeiten von Merkmalsausprägungen bestimmt werden, kann man **Klassenhäufigkeiten** ermitteln.

Definition: Nummeriert man die gebildeten Klassen mit $K_1, K_2, \ldots, K_k$, so erhält man mit

$$H(K_j) = \text{ (Anzahl der Beobachtungswerte in Klasse } K_j) = n_j$$

die **absolute Häufigkeit** der Klasse K_j. Die **relative Häufigkeit** ergibt sich als

$$h(K_j) = \frac{1}{n} H(K_j) = \frac{n_j}{n}$$

Entsprechend kann man auch **absolute und relative Summenhäufigkeiten** bestimmen.

Für die Daten aus Beispiel 4.2 werden in den Tabellen 4.4 und 4.5 zwei Gruppierungsvorschläge vorgestellt.

1. Gruppierungsvorschlag mit 9 Klassen:

Klasse K_j	$x_{j-1} < x \leq x_j$	Klassenmitte	$H(K_j)$	$h(K_j)$	$F(K_j)$	$f(K_j)$
K_1	$1050 < x \leq 1150$	1100	3	0.03	0.03	0.0003
K_2	$1150 < x \leq 1250$	1200	5	0.05	0.08	0.0005
K_3	$1250 < x \leq 1350$	1300	15	0.15	0.23	0.0015
K_4	$1350 < x \leq 1450$	1400	28	0.28	0.51	0.0028
K_5	$1450 < x \leq 1550$	1500	28	0.28	0.79	0.0028
K_6	$1550 < x \leq 1650$	1600	15	0.15	0.94	0.0015
K_7	$1650 < x \leq 1750$	1700	4	0.04	0.98	0.0004
K_8	$1750 < x \leq 1850$	1800	1	0.01	0.99	0.0001
K_9	$1850 < x \leq 1950$	1900	1	0.01	1.00	0.0001

Tab. 4.4: Häufigkeitstabelle für die Daten aus Beispiel 4.2 gruppiert in 9 Klassen

2. Gruppierungsvorschlag mit 17 Klassen:

Klasse K_j	$x_{j-1} < x \leq x_j$	Klassenmitte	$H(K_j)$	$h(K_j)$	$F(K_j)$	$f(K_j)$
K_1	$1050 < x \leq 1100$	1075	1	0.01	0.01	0.0002
K_2	$1100 < x \leq 1150$	1125	2	0.02	0.03	0.0004
K_3	$1150 < x \leq 1200$	1175	2	0.02	0.05	0.0004
K_4	$1200 < x \leq 1250$	1225	3	0.03	0.08	0.0006
K_5	$1250 < x \leq 1300$	1275	8	0.08	0.16	0.0016
K_6	$1300 < x \leq 1350$	1325	7	0.07	0.23	0.0014
K_7	$1350 < x \leq 1400$	1375	14	0.14	0.37	0.0028
K_8	$1400 < x \leq 1450$	1425	14	0.14	0.51	0.0028
K_9	$1450 < x \leq 1500$	1475	15	0.15	0.66	0.0030
K_{10}	$1500 < x \leq 1550$	1525	13	0.13	0.79	0.0026
K_{11}	$1550 < x \leq 1600$	1575	10	0.10	0.89	0.0020
K_{12}	$1600 < x \leq 1650$	1625	5	0.05	0.94	0.0010
K_{13}	$1650 < x \leq 1700$	1675	4	0.04	0.98	0.0008
K_{14}	$1700 < x \leq 1750$	1725	0	0.00	0.98	0.0000
K_{15}	$1750 < x \leq 1800$	1775	1	0.01	0.99	0.0002
K_{16}	$1800 < x \leq 1850$	1825	0	0.00	0.99	0.0000
K_{17}	$1850 < x \leq 1900$	1875	1	0.01	1.00	0.0002

Tab. 4.5: Häufigkeitstabelle für die Daten aus Beispiel 4.2 gruppiert in 17 Klassen

4.6 Grafische Darstellung von Häufigkeiten

Definition: Die Folgen der relativen Häufigkeiten $h(A_1)$, $h(A_2), \ldots,$ $h(A_k)$ für ein Merkmal mit den Ausprägungen $A_1, \ldots, A_k$ bzw. $h(K_1)$, $h(K_2), \ldots,$ $h(K_k)$ für ein Merkmal, das in die Klassen $K_1, \ldots, K_k$ gruppiert wurde, nennt man **(relative) Häufigkeitsverteilungen**.

Die Häufigkeitsverteilungen der Daten aus den Beispielen findet man in der dritten Spalte der Tabelle 4.3 bzw. in der jeweils fünften Spalte der Tabellen 4.4 und 4.5.

Häufigkeitsverteilungen lassen sich besonders gut übersehen, wenn man sie grafisch darstellt. Hierzu eignen sich besonders Stabdiagramme und Histogramme.

4.6.1 Das Stabdiagramm

Beim Stabdiagramm wird über jeder Ausprägung A_j bzw. x die zugehörige relative Häufigkeit $h(A_j)$ bzw. $h(x)$ als senkrechter Stab abgetragen. Stabdiagramme sind besonders für diskrete Merkmale geeignet und können auch für nominale und ordinale Merkmale verwendet werden. Bei ordinalen Merkmalen hält man sich an die Reihenfolge, die durch die Intensität der Ausprägungen vorgegeben ist.

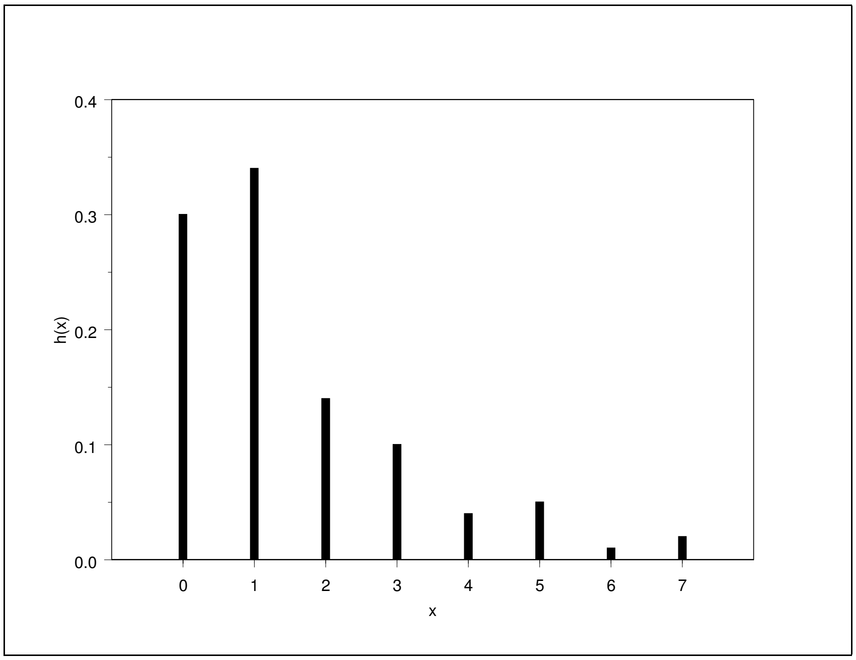


Abb. 4.1: Stabdiagramm für die Anzahl der kariösen Zähne bei 100 Schulkindern

Im Stabdiagramm trägt man in der Regel relative Häufigkeiten ab. Dadurch sind Vergleiche zwischen Diagrammen auch dann möglich, wenn die Stichprobenumfänge der jeweiligen Untersuchungen verschieden sind. Dabei sollte man allerdings stets den zugrundeliegenden Stichprobenumfang angeben.

4.6.2 Das Histogramm

Im Histogramm verwendet man anstelle der Stäbe aneinanderstoßende Rechtecke, deren **Flächen proportional zu den relativen Häufigkeiten der einzelnen Ausprägungen bzw. Klassen** sind. Histogramme verwendet man besonders bei stetigen Merkmalen und bei gruppierten Daten.

Die einzelnen Rechtecke werden über den Teilintervallen der Merkmalsachse, die den Merkmalsklassen entsprechen, gezeichnet. Ihre Höhe f an einer Stelle x der Merkmalsachse berechnet man als

$$f(x) = f(K_j) = \frac{H(K_j)}{n(x_j - x_{j-1})} = \frac{h(K_j)}{(x_j - x_{j-1})} \qquad \text{für alle } x_{j-1} < x \leq x_j$$

Dabei ist x_{j-1} die untere und x_j die obere Klassengrenze der Klasse K_j. Diese Formel geht von der Gleichsetzung der Flächen und der relativen Häufigkeiten aus. Sie kann für Histogramme mit gleich breiten und verschieden breiten Klassen verwendet werden.

Durch diese Berechnung der Rechteckhöhen erreicht man, dass die **Flächen** der einzelnen Rechtecke den relativen Häufigkeiten in den verschiedenen Klassen entsprechen. Außerdem gilt für jedes Histogramm, dass die Gesamtfläche aller Rechtecke gleich 1 ist.

Für die Schlachtgewichte der Masthähnchen erhält man mit den Gruppierungsvorschlägen aus Abschnitt 4.5 die Histogramme in den Abbildungen 4.2 und 4.3.

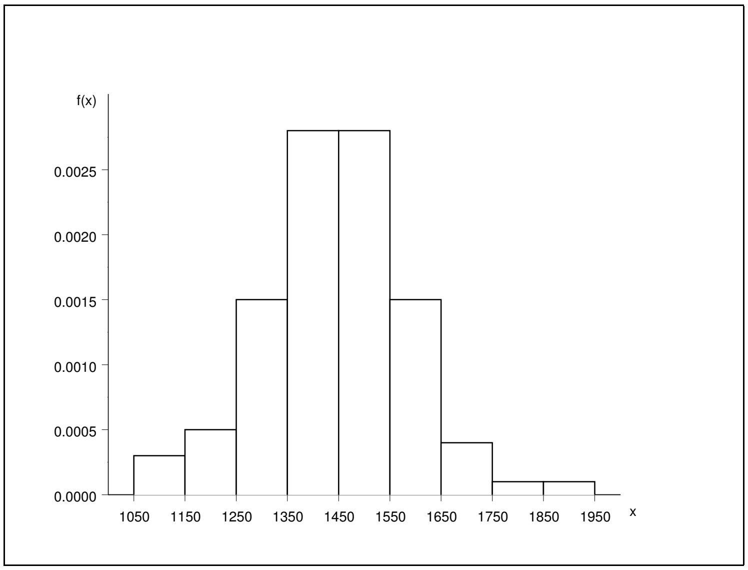


Abb. 4.2: Histogramm für die Schlachtgewichte von 100 Masthähnchen gruppiert in 9 Klassen (1. Gruppierungsvorschlag)

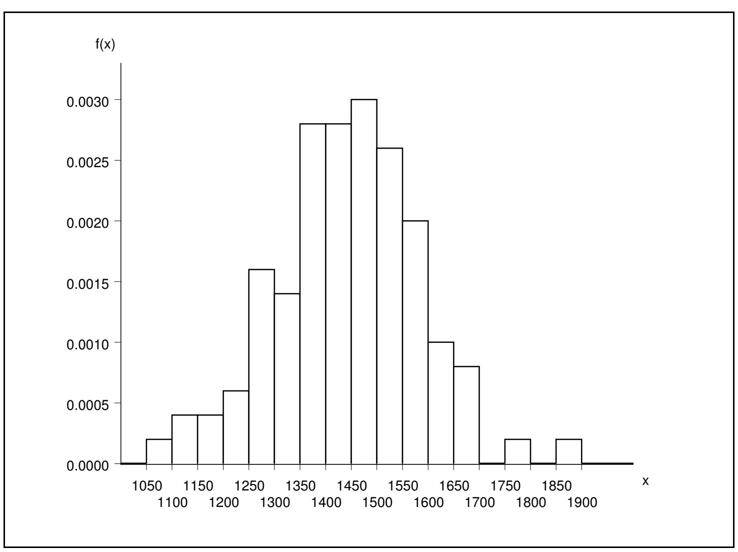


Abb. 4.3: Histogramm für die Schlachtgewichte von 100 Masthähnchen gruppiert in 17 Klassen (2. Gruppierungsvorschlag)

4.6.3 Die empirische Verteilungsfunktion

Die empirische Verteilungsfunktion ist eine Möglichkeit für die grafische Darstellung der Folge der relativen Summenhäufigkeiten $F(A_j)$ bzw. $F(K_j)$ (vgl. Spalte 4 in Tabelle 4.3 bzw. Spalte 6 in den Tabellen 4.4 und 4.5). Sie ist eine Treppenfunktion über der Achse der Merkmalsausprägungen und wir daher auch als $F(x)$ bezeichnet.

An jeder Stelle x der Merkmalsachse gibt sie an, wie groß der Anteil der Beobachtungen mit Ausprägungen $\leq x$ ist.

Die empirische Verteilungsfunktion für diskrete Merkmale ist auf jedem Intervall $(A_{j-1} \leq x < A_j)$ konstant. An der Stelle A_j springt sie um den Wert $h(A_j)$ nach oben. Es gilt $F(x) = 0$ für alle $x < A_1$ und $F(x) = 1$ für alle $x \geq A_k$. Sprungstellen liegen an allen Punkten, an denen Beobachtungen vorkommen.

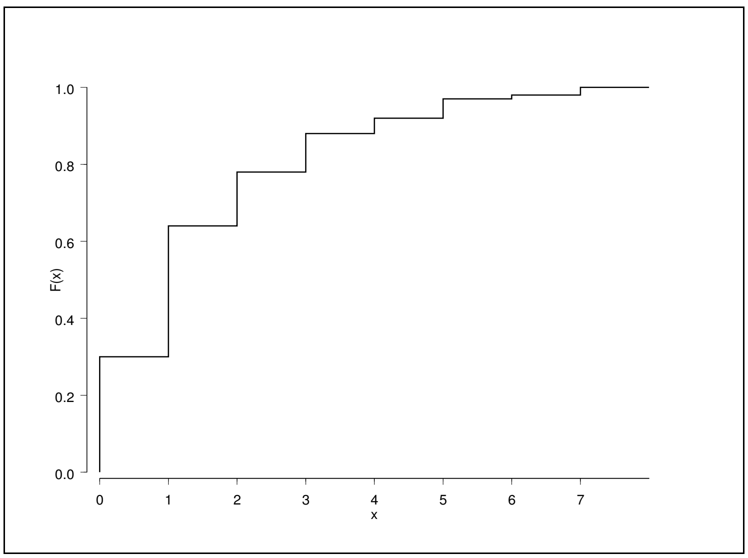


Abb. 4.4: Empirische Verteilungsfunktion der Anzahl der kariösen Zähne bei 100 Schulkindern

Im Abschnitt 4.5 wurde der Häufigkeitsbegriff bei gruppierten Daten für stetige Merkmale eingeführt. Dazu wurde für das betrachtete Merkmal eine Gruppierung bzw. Klassifizierung vorgenommen. Die Folge der relativen Summenhäufigkeiten $F(K_j)$ kann analog zur Darstellung bei diskreten Merkmalen als Treppenfunktion über den Klassen gezeichnet werden. Allerdings tritt dabei das Problem auf, geeignete Stellen für die Sprünge zu finden, d.h. an der unteren bzw. an der oberen Klassengrenze oder in der Mitte der Klasse. Deshalb wird auch für **stetige** Merkmale die empirische Verteilungsfunktion für die geordnete Folge der Originalbeobachtungen gezeichnet, zumal diese Merkmale in der Praxis auch nur diskret messbar sind (vgl. Abb. 4.5).

Mit Hilfe der empirischen Verteilungsfunktion kann man leicht den Anteil der Beobachtungen zwischen zwei Punkten x_1 und x_2 bestimmen, für die $x_1 < x \leq x_2$ gilt. $F(x_1)$ ist der Anteil der Beobachtungen, die $\leq x_1$ sind und $F(x_2)$ ist der Anteil der Beobachtungen, die $\leq x_2$ sind. Damit ist $F(x_2) - F(x_1)$ der Anteil der Beobachtungen im Intervall $(x_1 < x \leq x_2)$.

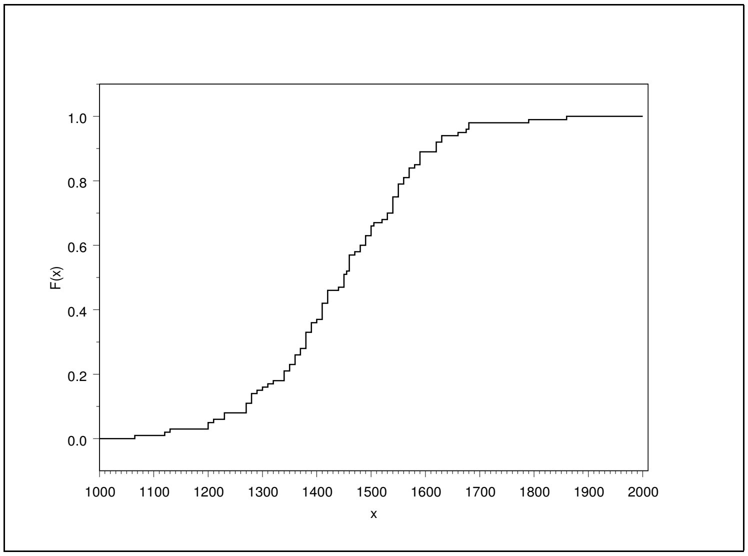


Abb. 4.5: Empirische Verteilungsfunktion der Schlachtgewichte von 100 Masthähnchen

Kapitel 5

Maßzahlen für Häufigkeitsverteilungen

Maßzahlen verwendet man, um gewisse Eigenschaften von Häufigkeitsverteilungen darzustellen. Sie sind charakteristische Größen, die einen ersten Eindruck von Lage und Gestalt einer Häufigkeitsverteilung über einer Achse mit den Merkmalsausprägungen vermitteln.

Zur Bestimmung einiger Maßzahlen benötigt man eine der Größe nach geordnete Liste der Daten. Wurden $x_1, x_2, \dots, x_n$ beobachtet, so bezeichnet man mit $x_{[1]} = x_{min}, x_{[2]}, x_{[3]}, \dots, x_{[n]} = x_{max}$ die geordneten Daten. Meistens gilt $x_1 \neq x_{[1]}$.

> Wenn man z.B. die Werte $x_1 = 7$, $x_2 = 5$, $x_3 = 1$, $x_4 = 5$, $x_5 = 2$ beobachtet, dann erhält man:
> $x_{[1]} = x_3 = 1$, $x_{[2]} = x_5 = 2$, $x_{[3]} = x_2 = 5$, $x_{[4]} = x_4 = 5$, $x_{[5]} = x_1 = 7$

Bei den Beispielen dieses Abschnitts wird auf das Katzen–Beispiel aus Kapitel 2 (Tabelle 2.1) Bezug genommen.

Verhalten	1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 5 5 5 $x_{[1]} = x_{min} = 1$ $x_{[32]} = x_{max} = 5$
Körpergewicht (kg)	2.7 2.9 3.0 3.1 3.1 3.2 3.3 3.3 3.3 3.4 3.4 3.4 3.4 3.4 3.5 3.5 3.5 3.5 3.5 3.6 3.6 3.6 3.7 3.7 3.7 3.8 3.8 3.9 3.9 4.0 4.1 4.3 $x_{[1]} = x_{min} = 2.7$ $x_{[32]} = x_{max} = 4.3$

Tab. 5.1: Geordnete Beobachtungsreihen für die Merkmale Verhalten und Körpergewicht

5.1 Lagemaße

Lage- bzw. **Lokalisationsmaße** sollen die Lage einer Häufigkeitsverteilung charakterisieren. Aus dem Datenmaterial wird ein "zentraler Wert" oder auch ein "typischer Repräsentant" ermittelt. Es stehen verschiedene Lagemaßzahlen zur Verfügung.

5.1.1 Der Modalwert x_{mod}

Als Modalwert bezeichnet man diejenige Merkmalsausprägung, die am häufigsten in der Beobachtungsreihe auftritt. Synonym werden die Begriffe **häufigster Wert**, **dichtester Wert** und **Modus** verwendet. Kommen mehrere Ausprägungen gleich häufig vor, so ist es nicht sinnvoll, den Modalwert zu bestimmen.

Je mehr verschiedene Merkmalsausprägungen möglich sind, z.B. bei hoher Messgenauigkeit, desto seltener beobachtet man identische Werte.

In solchen Fällen ist es wie beim Histogramm besser, Merkmalsklassen zu bilden, die ähnliche Ausprägungen zusammenfassen. Man definiert dann die Mitte derjenigen Klasse als Modalwert, in die die meisten Beobachtungen fallen.

Der Modalwert wird mit x_{mod} bezeichnet. Er ist die einzige Maßzahl, die für Daten jeglichen Skalenniveaus bestimmt werden kann.

Der Modalwert des Merkmals Diagnose ist die 0, d.h. die meisten Tiere sind gesund.

Für das Merkmal Verhalten wurde die 3 am häufigsten beobachtet, die meisten Tiere verhalten sich also normal.

Beim Körpergewicht ist der Modalwert nicht eindeutig. Die Ausprägungen 3.4 kg und 3.5 kg wurden je fünfmal beobachtet.

Betrachtet man die Zellzahl, so treten fast nur verschiedene Werte auf. Nur die Beobachtung 15.0×10^9 Zellen/g kommt zweimal vor. Damit ist sie der Modalwert der Einzelbeobachtungen. Die Beobachtung 15.0×10^9 Zellen/g ist jedoch kein "typischer Repräsentant" der Beobachtungsreihe. Allein im Bereich von 0.01×10^9 Zellen/g bis 1.0×10^9 Zellen/g liegen 16 der 32 Beobachtungen.

Daher ist es besser, hier Klassen zu bilden:

Klasse	Zellzahl (10^9 Zellen/g)	n_i
K_1	$0.01 < x \leq 1.0$	16
K_2	$1.00 < x \leq 5.0$	5
K_3	$5.00 < x \leq 15.0$	5
K_4	$15.00 < x \leq 50.0$	4
K_5	$50.00 < x \leq 300.0$	2

Tab. 5.2: Klassen für das Merkmal Zellzahl

Die meisten Beobachtungen fallen in die Klasse K_1. Die zugehörige Klassenmitte ist $0.505 \cdot 10^9$ Zellen/g. Damit erhält man $x_{mod} = 0.505 \cdot 10^9$ Zellen/g.

5.1.2 α-Quantile x_α

Das α-Quantil x_α einer Reihe von Beobachtungen ist der Wert x auf der Skala der Messwerte, für den gilt, dass $\alpha \times 100\%$ der Beobachtungen kleiner oder gleich x sind. Für jedes beliebige α aus dem Intervall $[0, 1]$ kann das α-Quantil bestimmt werden, sofern genügend Beobachtungen vorhanden sind.

Die Berechnung der Quantile erfolgt auf folgende Weise:

$$x_\alpha = \begin{cases} x_{[k]} & \text{falls } n \times \alpha \textbf{ keine} \text{ ganze Zahl ist;} \\ & k \text{ gibt die auf } n \times \alpha \text{ folgende ganze Zahl an} \\ \frac{1}{2}\left(x_{[k]} + x_{[k+1]}\right) & \text{falls } n \times \alpha \text{ eine ganze Zahl ist; } k = n \times \alpha \end{cases}$$

In der deskriptiven Statistik werden besonders häufig die folgenden Quantile bestimmt.

5.1.2.1 Der Median $\tilde{x} = x_{0.5}$

Der Median $x_{0.5}$ ist das 0.5-Quantil bzw. 50%-Quantil einer Verteilung und damit der Wert, für den gilt, dass jeweils die Hälfte der Beobachtungswerte $\geq$ bzw. $\leq x_{0.5}$ sind. Der Median wird auch mit $\tilde{x}$ bezeichnet und muss, je nachdem, ob die Anzahl der Beoachtungen n gerade oder ungerade ist, auf andere Art bestimmt werden:

$$\tilde{x} = x_{0.5} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{für } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n+2}{2}\right]}\right) & \text{für } n \text{ gerade} \end{cases}$$

Bei ungeradem n wählt man als Median den mittleren Wert der geordneten Daten, bei geradem n wählt man den Durchschnitt der beiden mittleren Werte.

> Im Beispiel werden 32 Katzen untersucht, d.h. $n = 32$.
> Daher wird die Formel für gerades n verwendet.
> Man berechnet $\frac{n}{2} = \frac{32}{2} = 16$, $\frac{n+2}{2} = \frac{34}{2} = 17$.
> Für die Bestimmung des Medians benötigt man also $x_{[16]}$ und $x_{[17]}$.
>
> Verhalten: $x_{[16]} = 3$, $x_{[17]} = 3$, $\tilde{x} = \frac{3+3}{2} = 3$
> Körpergewicht: $x_{[16]} = 3.5$, $x_{[17]} = 3.5$, $\tilde{x} = \frac{3.5+3.5}{2} = 3.5$kg
> Zellzahl: $x_{[16]} = 0.90$, $x_{[17]} = 1.3$, $\tilde{x} = \frac{0.9+1.3}{2} = 1.1 \times 10^9$ Zellen/g.

5.1.2.2 Die Quartile

Die Quartile zerlegen die Beobachtungswerte in vier Gruppen, die jeweils ein Viertel der Beobachtungswerte umfassen. Es gibt drei Quartile :

> das **l. Quartil** $x_{0.25}$, für das 25% der Beobachtungen $\leq x_{0.25}$ sind,
>
> das **2. Quartil** $x_{0.5}$, das gleich dem Median ist,
>
> das **3. Quartil** $x_{0.75}$, für das 75% der Beobachtungen $\leq x_{0.75}$ sind.

Das 1. Quartil wird auf Grundlage der allgemeinen Formel für Quantile folgendermaßen berechnet:

$$x_{0.25} = \begin{cases} x_{[k]} & \text{falls } n \times 0.25 \text{ keine ganze Zahl und} \\ & k \text{ die auf } n \times 0.25 \text{ folgende ganze Zahl ist} \\ \frac{1}{2}\left(x_{[n \times 0.25]} + x_{[n \times 0.25+1]}\right) & \text{falls } n \times 0.25 \text{ eine ganze Zahl ist} \end{cases}$$

Die Berechnung für das 3. Quartil erfolgt analog.

> Verhalten: $x_{0.25} = 2$, $x_{0.75} = 4$
> Körpergewicht: $x_{0.25} = 3.3$kg, $x_{0.75} = 3.7$kg
> Zellzahl: $x_{0.25} = 0.49 \times 10^9$ Zellen/g, $x_{0.75} = 14.0 \times 10^9$ Zellen/g

Grundsätzlich kann man Quantile für ordinale und metrische Daten bestimmen.

5.1.3 Das arithmetische Mittel $\overline{x}$

Das arithmetische Mittel ist der Durchschnittswert aller Beobachtungen.

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

Der Wert jeder einzelnen Beobachtung geht in die Berechnung von $\overline{x}$ ein. Damit wird für dieses Lagemaß mehr Information aus der Stichprobe berücksichtigt als für die bisher vorgestellten Maßzahlen.

Körpergewicht: $\overline{x} = 3.503$kg
Zellzahl: $\overline{x} = 18.95 \times 10^9$ Zellen/g

Das arithmetische Mittel kann nur für metrische Daten sinnvoll berechnet werden.

Für das Körpergewicht unterscheiden sich die bisher berechneten zentralen Lagemaße nur unwesentlich. Dagegen gibt es große Unterschiede bei den Maßzahlen für die Zellzahl. Das ist auf die unterschiedliche Gestalt der Häufigkeitsverteilungen zurückzuführen.

Abbildung 5.1 zeigt die einzelnen Lagemaße für die Merkmale Körpergewicht in der Stichprobe der 32 Katzen (Beispiel 2.1) mit dem zugehörigen Histogramm.

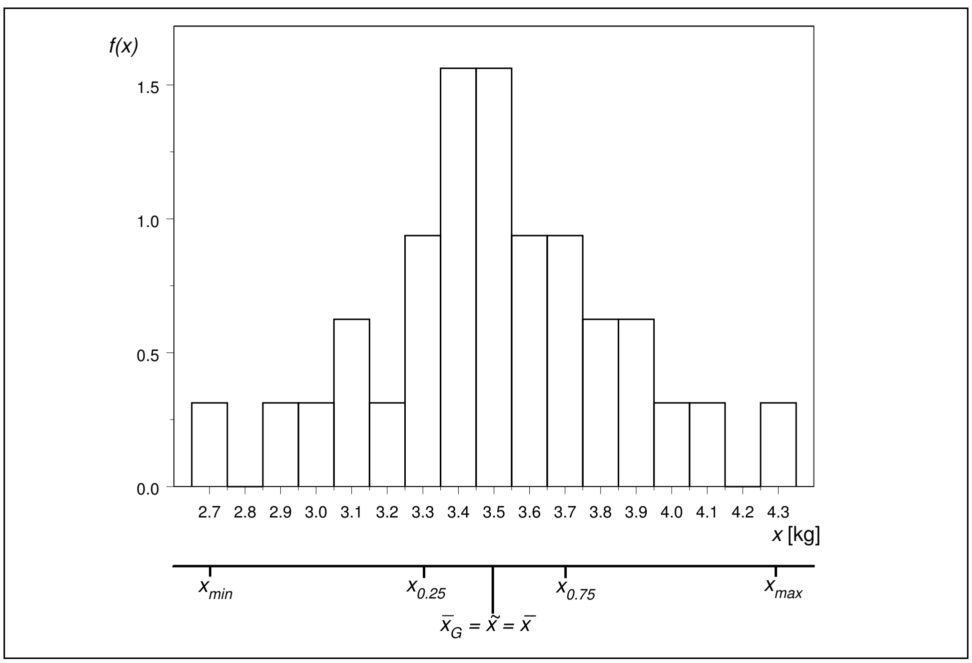


Abb. 5.1: Histogramm und Skala mit Lagemaßen für das Merkmal Körpergewicht

Das Histogramm der Körpergewichte der 32 Katzen zeigt eine annähernd symmetrische Häufigkeitsverteilung um den Wert 3.5 kg. Dagegen hat die Verteilung der Zellzahlen eine schiefe Gestalt, da "kleine" Werte erheblich häufiger beobachtet werden als "große" Werte. Man sagt, die Verteilung ist linkssteil bzw. rechtsschief (vgl. Abbildungen 5.2 bis 5.7). Derartige Häufigkeitsverteilungen kommen oft bei Merkmalen wie Zell- und Keimzahlen aber auch bei Einkommensverteilungen vor.

In solchen Fällen empfiehlt es sich, an Stelle des arithmetischen Mittels ein anderes Lagemaß zu bestimmen, da einzelne "extrem große" Beobachtungswerte das arithmetische Mittel $\overline{x}$ erheblich vergrößern.

5.1.4 Das geometrische Mittel $\overline{x}_G$

Das geometrische Mittel ist die n-te positive Wurzel aus dem Produkt aller Beobachtungen.

$$\overline{x}_G = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Es lässt sich einfacher und oft mit größerer Genauigkeit mit Hilfe der Logarithmen (hier: zur Basis 10) der Beobachtungswerte berechnen.

$$\lg \overline{x}_G = \overline{\lg x} = \frac{1}{n} \sum_{i=1}^{n} \lg x_i \Longleftrightarrow \overline{x}_G = 10^{\overline{\lg x}}$$

Bei der Berechnung des geometrischen Mittels werden ebenfalls die Werte aller Einzelbeobachtungen berücksichtigt, jedoch haben extreme Beobachtungen einen weniger starken Einfluss auf $\overline{x}_G$. Daher berechnet man im Beispiel kleinere Werte für $\overline{x}_G$ als für $\overline{x}$.

Körpergewicht: $\overline{\lg x} = 0.542$, $\overline{x}_G = 3.486$ kg
Zellzahl: $\overline{\lg x} = 0.35$, $\overline{x}_G = 10^{0.35} = 2.2 \times 10^9$ Zellen/g

Der Einfluss der Logarithmierung auf die Größen der Beobachtungen wird besonders deutlich, wenn man Tabelle 5.3 betrachtet.

x	$\lg x$
0.001	-3
0.01	-2
0.1	-1
1	0
10	1
100	2
1000	3

Tab. 5.3: Logarithmus zur Basis 10

Das geometrische Mittel ist ein Lagemaß für metrische Daten, es gilt stets $\overline{x}_G \leq \overline{x}$. Die Berechnung des geometrischen Mittels ist nur sinnvoll für Beobachtungsreihen, die ausschließlich positive Werte enthalten.

5.1.5 Allgemeine Bemerkungen

In den vorherigen Abschnitten sind verschiedene Lagemaße vorgestellt worden. Je nach Art der vorliegenden Rohdaten muss man abwägen, welche Maßzahl für das vorliegende Problem geeignet ist. Dabei muss neben dem Skalenniveau der Daten auch die Gestalt der Häufigkeitsverteilung berücksichtigt werden.

Falls kein Histogramm zur Verfügung steht, liefert die Lage der verschiedenen Lokalisationsmaße zueinander gewisse Informationen über die Gestalt einer Verteilung.

So gilt für rechtsschiefe Verteilungen stets $\overline{x} \geq x_{0.5} \geq x_{mod}$, für linksschiefe Verteilungen dagegen $\overline{x} \leq x_{0.5} \leq x_{mod}$. Für symmetrische Verteilungen gilt $\overline{x} = x_{0.5}$. Generell gilt $\overline{x}_G \leq \overline{x}$.

Man sollte sich immer überlegen, wie gut die Maßzahlen im Einzelfall interpretierbar sind. Liegt z.B. eine U-förmige Häufigkeitsverteilung vor, so kann die alleinige Angabe des Medians bzw. des arithmetischen Mittels zu Fehlinterpretationen führen.

> Im Mittelalter war die Säuglingssterblichkeit sehr hoch. Wer jedoch die ersten Jahre überlebte, konnte auch damals schon recht alt werden.
> Die Angabe "Im Mittelalter wurden die Menschen durchschnittlich 25 Jahre alt" vermittelt daher ohne weitere Angaben eine falsche Vorstellung von der damaligen Lebenserwartung.

In den Abbildungen 5.2 bis 5.7 werden einige Beispiele für Häufigkeitsverteilungen dargestellt.

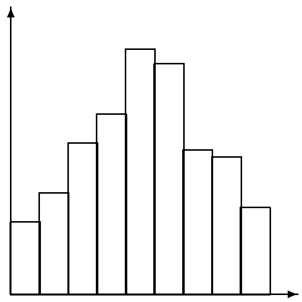


Abb. 5.2: Eingipflig

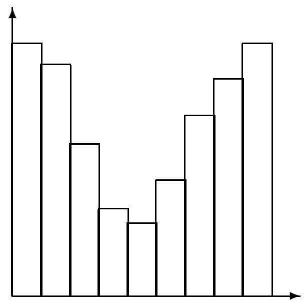


Abb. 5.3: U-förmig

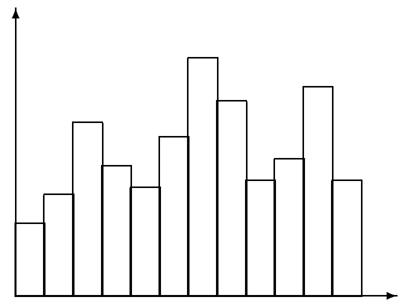


Abb. 5.4: Mehrgipflig

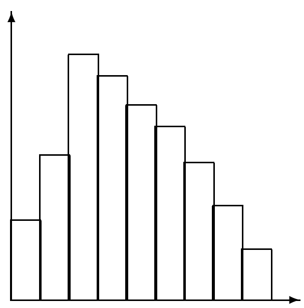


Abb. 5.5: Rechtsschief

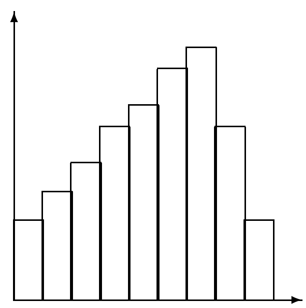


Abb. 5.6: Linksschief

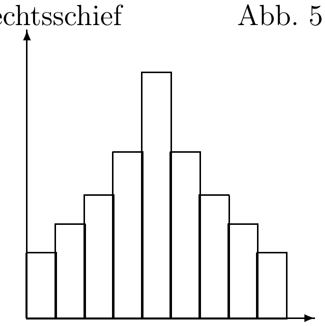


Abb. 5.7: Symmetrisch

Ein Lagemaß allein liefert sehr wenig Information über die Struktur der Rohdaten. Zwei Häufigkeitsverteilungen mit gleichem arithmetischen Mittel, gleichem Median und gleichem Modalwert können sehr verschiedene Gestalt haben, wenn die Variabilität in den beiden Datensätzen unterschiedlich ist.

5.2 Streuungsmaße

Streuungs- bzw. **Dispersionsmaße** sind Maßzahlen, die die Variabilität der Daten in der Grundgesamtheit bzw. in der Stichprobe beschreiben.

5.2.1 Die Spannweite R

Die Spannweite ist der Abstand des kleinsten Beobachtungswertes vom größten Wert. Sie wird auch **Range** genannt und daher mit R bezeichnet. Außerdem verwendet man den Begriff **Variationsbreite**.

$$R = x_{max} - x_{min}$$

Die Berechnung der Spannweite ist für metrische Daten sinnvoll.

Körpergewicht: $4.3 - 2.7 = 1.6$ kg
Zellzahl: $260 \times 10^9 - 0.05 \times 10^9 = 259.95 \times 10^9$ Zellen/g

5.2.2 Der Quartilsabstand Q

Der **Quartilsabstand** oder auch Interquartilsabstand ist der Abstand des l. Quartils vom 3. Quartil.

$$Q = x_{0.75} - x_{0.25}$$

Dieser Abstand schließt die mittleren 50% der Beobachtungen ein. Er kann für ordinale und metrische Daten bestimmt werden.

Körpergewicht: $Q = 3.7 - 3.3 = 0.4$ kg
Zellzahl: $Q = 14.0 \times 10^9 - 0.49 \times 10^9 = 13.51 \times 10^9$ Zellen/g

Selbstverständlich können auch andere **Quantilsabstände** berechnet werden.

5.2.3 Die Varianz s^2, die Standardabweichung s und der Variationskoeffizient v

Die bisher vorgestellten Streuungsmaße berücksichtigen nur einzelne Werte der Beobachtungsreihe bzw. die Abstände einiger Beobachtungen voneinander. In diesem Abschnitt sollen Maßzahlen vorgestellt werden, für deren Berechnung jeder einzelne Beobachtungswert herangezogen wird. Die Konstruktion dieser Maße wird im folgenden ausführlich erläutert.

Zunächst bestimmt man für jede Beobachtung x_i ihren Abstand zum arithmetischen Mittel $\overline{x}$. Das ist die Differenz $(x_i - \overline{x})$. Diese Differenzen sind zum Teil positiv, zum Teil negativ. Insgesamt ergibt sich

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

Die Summe der Abweichungen ist also stets gleich Null und daher als Grundbaustein eines Streuungsmaßes nicht verwendbar. Statt dessen wählt man die **Summe der Abweichungsquadrate**

$$SQ = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Alle Summanden sind positiv. Je mehr ein Beobachtungswert vom arithmetischen Mittel abweicht, desto größer ist sein Beitrag zur Summe der Abweichungsquadrate. Außerdem ist einleuchtend, dass SQ um so größer wird, je größer die Anzahl der Beobachtungen n ist.

Daher setzt man die Summe der Abweichungsquadrate ins Verhältnis zur Anzahl der Beobachtungen, indem man SQ durch $n-1$ dividiert.

Das Ergebnis

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

nennt man die **Varianz** der Beobachtungsreihe.

Auffällig ist, dass nicht durch n, sondern durch $n-1$ dividiert wird. Diese Vorgehensweise hat mathematische Gründe, die in der schließenden Statistik von großer Bedeutung sind. Sie sollen an dieser Stelle nicht erläutert werden.

Körpergewicht: $s^2 = \frac{1}{31} \sum_{i=1}^{32} (x_i - 3.503)^2 = 0.12 \text{ kg}^2$

Die Varianz ist schwer interpretierbar, da ihre Dimension das Quadrat der Dimension der Beobachtungswerte ist. Deshalb wird in der Praxis meist die **Standardabweichung** s verwendet, die die positive Wurzel aus der Varianz ist:

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

Körpergewicht: $s = \sqrt{0.12} = 0.34$ kg

Wer die Varianz oder die Standardabweichung ohne Verwendung von Computern bestimmen will, sollte die Summe der Abweichungsquadrate nicht nach der oben angegebenen Formel berechnen, sondern die folgende Umformung verwenden:

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

Daraus ergibt sich für die Varianz:

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

Varianz und Standardabweichung sind Maße für den mittleren Abstand der einzelnen Beobachtungswerte von ihrem arithmetischen Mittel $\overline{x}$. Sie sagen allerdings nur etwas über die Variabilität von Daten aus, wenn man das zugehörige $\overline{x}$ kennt.

Für eine Messreihe mit $\overline{x} = 1000$ ist eine Standardabweichung von $s = 5$ sehr klein; ist dagegen $\overline{x} = 10$, so ist dieselbe Standardabweichung groß.

In der Natur beobachtet man, dass die Variabilität von Daten oft mit der Größe der Beobachtungswerte zunimmt.

Erwachsene Labormäuse wiegen im Durchschnitt 31g mit einer Standardabweichung von $s = 2.48$g.
Junge Meerschweinchen haben ein mittleres Gewicht von 280g, die Standardabweichung der Einzelwerte beträgt 8.4g. (Daten aus Lorenz, 1997)

Die Streuungen verschiedener Beobachtungsreihen kann man nur miteinander vergleichen, wenn man die Variabilität der Daten im Verhältnis zur durchschnittlichen Größe der Beobachtungswerte betrachtet. Eine Maßzahl, die dies ermöglicht, ist der **Variationskoeffizient**.

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

Der Variationskoeffizient ist dimensionslos. Er ist nur für Beobachtungsreihen geeignet, die ausschließlich positive Werte enthalten.

Für das Körpergewicht der Labormäuse bzw. der Meerschweinchen berechnet man die beiden Variationskoeffizienten v_L und v_M:

$$v_L = \frac{2.48\text{g}}{31\text{g}} = 0.08 \qquad \text{bzw.} \qquad v_M = \frac{8.4\text{g}}{280\text{g}} = 0.03$$

Die Körpergewichte der Labormäuse haben also eine größere relative Streuung als die der Meerschweinchen.

5.3 Wie verwendet man Lage- und Streuungsmaße ?

Zur Beschreibung eines Datensatzes gibt man meist mehrere Maßzahlen an. Wie im Abschnitt 5.1.5 erläutert, liefern mehrere Lagemaße gemeinsam gewisse Informationen über die Lage und die Gestalt einer Häufigkeitsverteilung. Weitere Informationen erhält man durch die Berechnung von Streuungsmaßzahlen. Daher werden Lage- und Streuungsmaßzahlen sehr oft zusammen verwendet.

Schon bei der Vorstellung der einzelnen Maßzahlen wurde darauf hingewiesen, dass sie fast alle nicht für Datenmaterial mit beliebigem Skalenniveau geeignet sind. Tabelle 5.4 soll noch einmal einen Überblick über die Verwendbarkeit der Lage- und Streuungsmaße geben:

Skala	**Lagemaße**	**Streuungsmaße**
Nominal-skala	Modalwert	
Ordinal-skala	Modalwert	
	Median	Quartilsabstand
	Quartile	
	sonstige Quantile	sonstige Quantilsabstände
metrische Skala	Modalwert	Spannweite
	Median	Quartilsabstand
	Quartile	
	sonstige Quantile	sonstige Quantilsabstände
	arithmetisches Mittel	Standardabweichung
		Varianz
		Variationskoeffizient
	geometrisches Mittel	

Tab. 5.4: Skalen mit geeigneten Maßzahlen

Darüber hinaus ist nicht jede Kombination von Lage- und Streuungsmaßzahlen aussagekräftig.

Das arithmetische Mittel gibt man in der Regel zusammen mit der Standardabweichung an, denn die Standardabweichung berechnet man aus den Abständen der einzelnen Beobachtungen von ihrem arithmetischen Mittel. Ebenso sinnvoll sind im Zusammenhang mit dem arithmetischen Mittel die Varianz und der Variationskoeffizient. Zusätzlich ist die Angabe des größten und des kleinsten Beobachtungswertes sinnvoll.

Die verschiedenen Quantile gibt man stets gemeinsam mit Quantilsabständen an. Besonders häufig verwendet man hier Median und Quartile zusammen mit dem Quartilsabstand und der Spannweite.

5.4 Boxplots

Zur übersichtlichen grafischen Darstellung von Beobachtungsreihen verwendet man häufig Boxplots. Sie sind nach dem in Abbildung 5.8 skizzierten Schema aufgebaut.

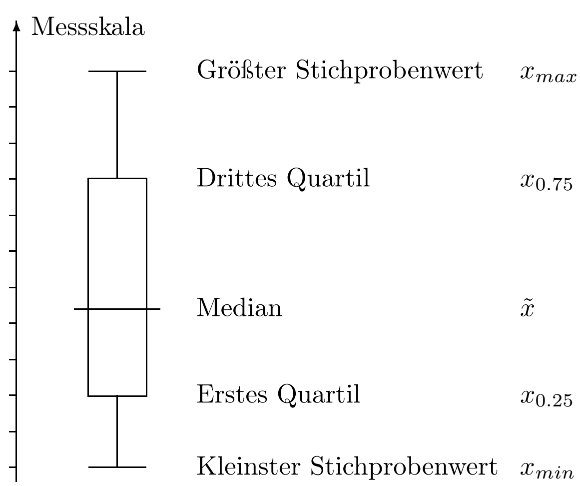


Abb. 5.8: Konstruktion des Boxplots (Quelle: Lorenz, 1997)

Größter und kleinster Beobachtungswert sowie Median und Quartile werden entlang einer Messskala abgetragen. Der Bereich zwischen erstem und drittem Quartil wird durch eine Box hervorgehoben. Innerhalb dieses Bereichs liegen die mittleren 50% der Beobachtungen.

Der Boxplot gibt außerdem Hinweise darauf, ob eine symmetrische oder eine schiefe Verteilung vorliegen könnte. Bei schiefen Verteilungen sind die Abstände des ersten und des dritten Quartils vom Median verschieden.

Boxplots sind besonders für den Vergleich von Beobachtungsreihen geeignet. Zeichnet man wie in Abbildung 5.9 mehrere Boxplots mit der gleichen Skala über- bzw. nebeneinander, so kann man Lage- und Streuungsunterschiede erkennen.

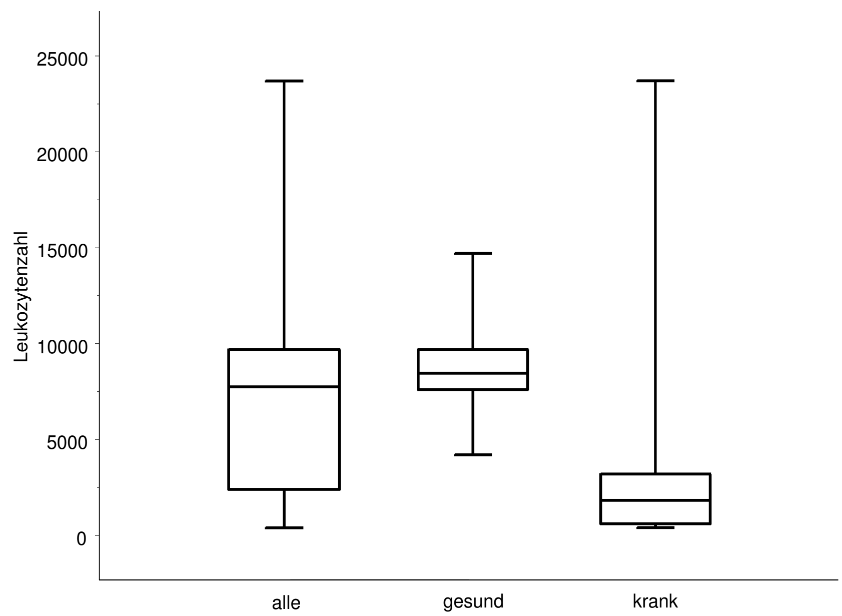


Abb. 5.9: Boxplot für den Vergleich der Leukozytenzahlen

In Abbildung 5.9 erkennt man, dass die gesunden Katzen Leukozytenwerte zwischen 5000 und 15000 haben, während mindestens 75% der kranken Katzen Werte unter 5000 aufweisen. Nur 25% der Werte der kranken Katzen liegen über ca. 3000, streuen dann aber bis fast 25000. Dadurch ist die Spannweite bei den kranken Katzen erheblich größer als bei den gesunden Katzen. Betrachtet man die Datentabelle auf S.9, wird klar, dass es sich um drei kranke Katzen mit extrem hohen Werten handelt.

Klare Schlüsse über die Schiefe der Verteilungen lassen sich aus keinem der drei Boxplots ziehen. Bei den Leukozytenwerten aller Katzen spricht zwar die Lage des Medians zu den Quartilen für eine rechtsschiefe Verteilung, das Aussehen der so genannten Whisker, die bis zum kleinsten und bis zum größten Wert reichen, gibt aber gleichzeitig den Hinweis auf eine linksschiefe Verteilung.

Kapitel 6

Wahrscheinlichkeitsrechnung

6.1 Definitionen

Definition: Ein **Zufallsvorgang** wird durch folgende Eigenschaften charakterisiert:

1. Sein Ergebnis ist nicht mit Sicherheit vorhersehbar.
2. Alle möglichen Ergebnisse sind von vornherein bekannt.
3. Der Vorgang ist tatsächlich oder gedanklich unter den gleichen Bedingungen wiederholbar.

Definition: Jedes mögliche Ergebnis eines Zufallsvorgangs heißt **Zufallsereignis**.

Definition: Jedes einzelne Ergebnis, das bei der Realisation eines Zufallsvorgangs möglich ist, nennt man **Elementarereignis**.

Beispiele:

1. Zufallsvorgang : Ziehen einer Karte aus einem Bridgeblatt
   mögliches Ereignis : Es wird Herzbube gezogen (Elementarereignis)
2. Zufallsvorgang : Werfen eines Spielwürfels
   mögliche Ereignisse: Die Zahlen 1,2,3,4,5,6 (Elementarereignisse)
3. Zufallsvorgang : Werfen zweier Würfel
   mögliche Ereignisse: Die 36 Zahlenpaare

$$\begin{array}{cccccc}
1,1 & 1,2 & 1,3 & 1,4 & 1,5 & 1,6 \\
2,1 & 2,2 & 2,3 & 2,4 & 2,5 & 2,6 \\
3,1 & 3,2 & 3,3 & 3,4 & 3,5 & 3,6 \\
4,1 & 4,2 & 4,3 & 4,4 & 4,5 & 4,6 \\
5,1 & 5,2 & 5,3 & 5,4 & 5,5 & 5,6 \\
6,1 & 6,2 & 6,3 & 6,4 & 6,5 & 6,6
\end{array}$$

6.1.1 Ereignisraum (Stichprobenraum)

Die Menge aller möglichen Ereignisse heißt Ereignisraum oder auch Stichprobenraum. Wird das Ergebnis des Zufallsvorganges durch ein einziges Merkmal bestimmt, so ist der Stichprobenraum eindimensional (Beispiele l und 2), werden zur Beschreibung eines Elementarereignisses zwei oder mehr, allgemein n Merkmale benötigt, so ist der Stichprobenraum zwei- bzw. mehrdimensional, allgemein n-dimensional (Beispiel 3: zweidimensional).

Jedes Ereignis lässt sich als Verknüpfung von Elementarereignissen darstellen.

So lässt sich in Beispiel l das Ereignis "es wird ein Bube gezogen" ableiten und folgendermaßen ausdrücken:

"Es wird ein Kreuzbube oder Pikbube oder Herzbube oder Karobube gezogen", oder abkürzend in der in der Mengenlehre gebräuchlichen Schreibweise:

$$\{\text{Kreuzbube, Pikbube, Herzbube, Karobube}\}.$$

Die Ereignisse eines Zufallsvorganges sollen mit $A,\ B,\ C\ \ldots$ oder mit $A_1,\ A_2,\ A_3,\ \ldots$ bezeichnet werden. Wir werden dabei immer voraussetzen, dass es sich um Ergebnisse desselben Zufallsvorganges handelt. Von jedem Ereignis können wir nach Realisation des Zufallsvorganges sagen, ob es eingetreten ist oder nicht. Aus dem Beispiel 2 abgeleitet, ist das Ereignis "eine gerade Zahl wird gewürfelt" eingetreten, wenn eine 2 oder eine 4 oder eine 6 gewürfelt wurde. Allgemein sprechen wir davon, dass ein Ereignis eingetreten ist, wenn eines der Elementarereignisse, aus denen es sich zusammensetzt, eingetreten ist.

6.2 Ereignisoperatoren

An Hand der Beispiele haben wir gesehen, dass sich Ereignisse miteinander verknüpfen lassen, d.h. man kann mit ihnen rechnen. Wir wollen nun zunächst die wichtigsten Operatoren für Ereignisse angeben.

6.2.1 Die Vereinigung $A \cup B$

Das Ereignis $A \cup B$ soll immer dann eingetreten sein, wenn eines der Ereignisse A oder B oder A und B gemeinsam eingetreten ist.

Regeln:

$$\begin{array}{rcll} A \cup B &=& B \cup A & \text{(Kommutativgesetz)} \\ A \cup (B \cup C) &=& (A \cup B) \cup C & \text{(Assoziativgesetz)} \end{array}$$

6.2.2 Der Schnitt $A \cap B$

Das Ereignis $A \cap B$ tritt genau dann ein, wenn sowohl A als auch B eingetreten sind.

Regeln:

$$\begin{array}{rcll} A \cap B &=& B \cap A & \text{(Kommutativgesetz)} \\ A \cap (B \cap C) &=& (A \cap B) \cap C & \text{(Assoziativgesetz)} \end{array}$$

Für die Verknüpfung von Vereinigung und Schnitt von Ereignissen gelten folgende Regeln :

$$
\begin{array}{rclll}
A \cap (B \cup C) & = & (A \cap B) \cup (A \cap C) & \text{1. Distributivgesetz} \\
A \cup (A \cap B) & = & A & \text{Einschmelzungsregel} \\
A \cup (B \cap C) & = & (A \cup B) \cap (A \cup C) & \text{2. Distributivgesetz}
\end{array}
$$

6.2.3 Das Komplementärereignis $\overline{A}$

(lies: "Komplement von A" oder kurz "A quer")

Das Nichteintreffen eines Ereignisses A fassen wir als ein neues Ereignis auf. Dem entspricht in der Logik die Negation der Aussage "A ist eingetreten".
Daraus folgt, dass das Komplementärereignis von $\overline{A}$ wieder A ist, d.h. $\overline{\overline{A}} = A$.

6.2.4 Das unmögliche Ereignis $\emptyset$ und das sichere Ereignis S

Bildet man $A \cap \overline{A}$, so erhält man ein Ereignis, das nie eintritt. Dieses Ereignis nennt man das unmögliche Ereignis und bezeichnet es mit $\emptyset$.

Betrachtet man $A \cup \overline{A}$, so erhält man ein Ereignis, das in jedem Fall eintritt; man bezeichnet es als sicheres Ereignis S. (Der Stichprobenraum entspricht dem sicheren Ereignis, deswegen wird für beide dasselbe Symbol gewählt).

Regeln:

$$
\begin{array}{rcl}
\overline{\emptyset} & = & S \\
\overline{S} & = & \emptyset \\
\overline{\overline{A}} & = & A \\
S \cup A & = & S \\
S \cap A & = & A \\
\emptyset \cup A & = & A \\
\emptyset \cap A & = & \emptyset
\end{array}
$$

6.3 Disjunkte Ereignisse und vollständiges System

Gilt für zwei Ereignisse A und B, dass $A \cap B = \emptyset$, so heißen die Ereignisse **disjunkt**.

Allgemein heißen n Ereignisse A_i, $i = 1,\ 2,\ \ldots\ ,\ n$, **paarweise disjunkt**, wenn für alle $i \neq j$ gilt, dass $A_i \cap A_j = \emptyset$. Ein System von n paarweise disjunkten Ereignissen heißt **vollständig**, wenn $A_1 \cup A_2 \cup \ldots \cup A_n = S$ ist. Man kann sich leicht überlegen, dass Elementarereignisse immer disjunkt sind.

Beispiele

Zufallsvorgang	:	Werfen eines Spielwürfels
Ereignisse	:	$A =$ Gerade Zahl
		$B =$ Zahl kleiner als 3
		$C =$ ungerade Zahl

Die Ereignisse $A,\ B,\ C$ lassen sich folgendermaßen als Verknüpfungen von Elementarereignissen schreiben:

$$\begin{array}{rcccc} A & = & \{2,4,6\} & = & \{2\} \cup \{4\} \cup \{6\} \\ B & = & \{1,2\} & = & \{1\} \cup \{2\} \\ C & = & \{1,3,5\} & = & \{1\} \cup \{3\} \cup \{5\} \end{array}$$

$$\begin{array}{rcl} A \cup B & = & \{1,2,4,6\} \\ A \cap B & = & \{2\} \\ A \cup C & = & S \\ A \cap C & = & \emptyset \\ \overline{B} & = & \{3,4,5,6\} \\ \overline{A} & = & C \end{array}$$

6.4 Wahrscheinlichkeiten und ihre Rechenregeln

Wir haben den Zufallsvorgang durch die möglichen Ereignisse charakterisiert. Dies allein reicht jedoch für eine vollständige Beschreibung nicht aus.

Dazu müssen wir noch jedem Ereignis des Stichprobenraumes eine Zahl zuordnen, die uns angibt, mit welcher Wahrscheinlichkeit dieses Ereignis bei der Realisation des Zufallsvorganges eintreten wird. Dieses Wahrscheinlichkeitsmaß soll folgende Eigenschaften haben:

1. Jedem Ereignis A aus dem Stichprobenraum ist eine zwischen 0 und 1 gelegene Zahl $P(A)$, die Wahrscheinlichkeit für das Auftreten von A, zugeordnet.

$$0 \leq P(A) \leq 1 \qquad \text{für alle } A \in S$$

2. Es seien $A_1,\ A_2,\ \ldots\ ,\ A_n$ sich paarweise ausschließende Ereignisse aus demselben Stichprobenraum. Dann ist

$$P\left(A_1 \cup A_2 \cup \ldots \cup A_n\right) = \sum_{i=1}^{n} P\left(A_i\right)$$

3. $P(S) = 1$

Bevor wir aus diesen drei Axiomen (von KOLMOGOROFF) Rechenregeln für Wahrscheinlichkeiten ableiten, wollen wir zunächst zwei Ansätze diskutieren, wie man Wahrscheinlichkeiten messen kann.

6.4.1 Gleichmöglichkeitsmodell von Laplace

Enthält der Stichprobenraum S endlich viele, gleichmögliche Elementarereignisse, dann ist die Wahrscheinlichkeit $P(A)$ für das Ereignis A gegeben durch

$$P(A) = \frac{\text{Anzahl der Elementarereignisse in } A}{\text{Anzahl der Elementarereignisse in } S}$$

oder anders ausgedrückt

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

Beispiel:
Der Zufallsvorgang ist das Werfen eines Würfels mit den Elementarereignissen $\{1\}, \{2\}, \{3\}, \{4\}$, $\{5\}$ und $\{6\}$.

Bei einem idealen Würfel hat jedes der Elementarereignisse die gleiche Chance einzutreten. Es lassen sich folglich mit Hilfe des Gleichmöglichkeitsmodells Wahrscheinlichkeiten berechnen.

$$\begin{aligned} A &= \{\text{gerade Zahl}\} \\ &= \{2,4,6\} && \Longrightarrow P(A) = \frac{3}{6} = \frac{1}{2} \\ A &= \{\text{Zahl kleiner als 5 und ungerade}\} \\ &= \{1,2,3,4\} \cap \{1,3,5\} \\ &= \{1,3\} && \Longrightarrow P(A) = \frac{2}{6} = \frac{1}{3} \end{aligned}$$

6.4.2 Statistische Definition der Wahrscheinlichkeit

Das Gleichmöglichkeitsmodell kann jedoch nicht immer angewendet werden, z.B. bei einer unendlich großen Grundgesamtheit oder bei Zufallsexperimenten, bei denen die Wahrscheinlichkeit **nicht** für alle Ereignisse gleich ist. Erfahrungsgemäß genügt das Eintreffen von Ereignissen bei den meisten Zufallsexperimenten auf lange Dauer gewissen Gesetzmäßigkeiten. Insbesondere erweist sich die relative Häufigkeit eines Ereignisses in großen Versuchsreihen als nahezu konstant.

Will beispielsweise die Polizei überprüfen, ob in einem Spielcasino asymmetrische Würfel zum Einsatz kommen, so wird sie folgendermaßen vorgehen: Die Polizei will für den asymmetrischen Würfel die Wahrscheinlichkeit bestimmen, dass mit ihm eine 6 geworfen wird. Dabei ist unser erster Gedanke, dass man Näherungswerte für diese Wahrscheinlichkeit erhält, wenn man mit dem Würfel sehr oft würfelt und auszählt, wie häufig eine 6 im Vergleich zu den anderen Zahlen geworfen wird. Würfeln wir beispielsweise 1000mal und werfen dabei 99mal eine 6, so würden wir vermuten, dass die Wahrscheinlichkeit, mit diesem Würfel eine 6 zu werfen, ungefähr bei 0.1 liegt.

Wir wollen das Ereignis "eine 6 würfeln" mit A bezeichnen, die absolute Häufigkeit, mit der dieses Ereignis eintritt, mit n_A und die relative Häufigkeit, mit der dieses Ereignis bei n Versuchen eintritt, mit $\frac{n_A}{n}$.

Im obigen Beispiel wäre also $n_A = 99$ und $\frac{n_A}{n} = 0.099$.

Wenn wir nicht nur einmal nach 1000 Würfen die relative Häufigkeit für das Werfen einer 6 berechnet hätten, sondern nach $100, \ 200, \ 500, \ \ldots$ Würfen, hätte sich das Bild in Abbildung 6.1 ergeben können :

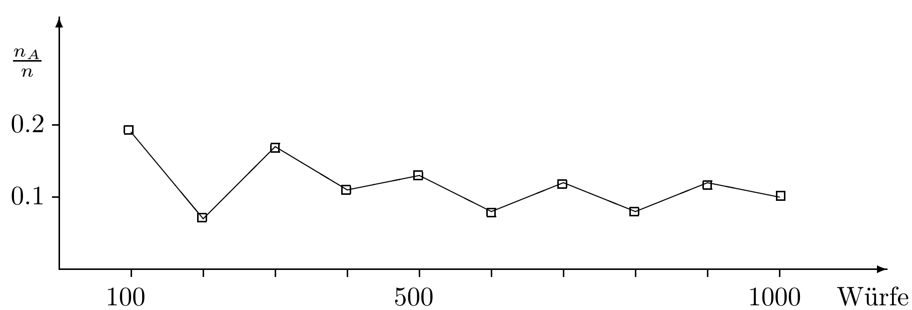


Abb. 6.1: Relative Häufigkeiten beim Würfeln einer '6'

Die relative Häufigkeit nähert sich offensichtlich einem Grenzwert. Je häufiger wir würfeln, desto stabiler wird der Verlauf der relativen Häufigkeit. Den Grenzwert wählen wir für die Wahrscheinlichkeit des Ereignisses.

Formal ausgedrückt schreibt man für diese Definition

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}.$$

6.5 Rechnen mit Wahrscheinlichkeiten

Aus den Axiomen für Wahrscheinlichkeiten lassen sich folgende Rechenregeln ableiten:

$$\begin{aligned} P(\overline{A}) &= 1 - P(A) \\ P(\emptyset) &= 1 - P(S) = 0 \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Die dritte Regel nennt man auch den **Additionssatz** der Wahrscheinlichkeitsrechnung.

Beispiel:

Zufallsvorgang	:	Aus einem Bridgespiel wird zufällig eine Karte gezogen	
Ereignisse	:	$A = \{\text{Ein Bild wird gezogen}\}$	$P(A) = 12/52$
		$B = \{\text{Eine Karokarte wird gezogen}\}$	$P(B) = 13/52$
		$C = \{\text{Ein Karobild wird gezogen}\}$	$P(C) = 3/52$

Berechnet werden soll $P(A \cup B)$.

Axiom 2 ist nicht anwendbar, da die Ereignisse A und B nicht disjunkt sind, denn

$$A \cap B = C \neq \emptyset.$$

Würden wir einfach die Wahrscheinlichkeiten $P(A)$ und $P(B)$ addieren, wäre die Wahrscheinlichkeit $P(A \cap B)$ doppelt berücksichtigt. Sie muss bei der Berechnung von $P(A \cup B)$ einmal abgezogen werden. Es ergibt sich also

$$P(A \cup B) = \frac{12}{52} + \frac{13}{52} - \frac{3}{52} = \frac{22}{52}$$

6.5.1 Bedingte Wahrscheinlichkeiten

Unter einer bedingten Wahrscheinlichkeit versteht man die Wahrscheinlichkeit des Ereignisses A unter der Bedingung, dass Ereignis B schon eingetreten ist.

$$\text{Schreibweise: } P(A \mid B)$$

$$\text{Berechnung: } P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Der Begriff soll anhand eines Beispiels erklärt werden. Die Wahrscheinlichkeit, aus der Katzenpopulation des Tierheims eine kranke Katze zu ziehen, beträgt nach dem Gleichmöglichkeitsmodell 43.75% (14 von 32 Tieren sind krank).

Der Tierarzt wählt eine Katze zufällig aus. Die Wahrscheinlichkeit, dass er eine kranke Katze auswählt, beträgt 43.75%. Nun greift aber der Tierarzt aus Bequemlichkeit nur die sehr ruhigen Katzen heraus. Wie groß ist jetzt die Wahrscheinlichkeit, dass er ein krankes Tier erwischt?

$$P(\text{krank} \mid \text{sehr ruhig}) = ?$$

Es gilt:

$$\begin{aligned}
P(\text{sehr ruhig}) &= 5/32 = 0.1563 \\
P(\text{krank und sehr ruhig}) &= 4/32 = 0.1250 \\
\Rightarrow P(\text{krank} \mid \text{sehr ruhig}) &= \frac{0.1250}{0.1563} = 0.79997 = 0.8
\end{aligned}$$

Die Wahrscheinlichkeit beträgt jetzt 80%, dass er ein krankes Tier herausgreift und liegt damit sehr viel höher als bei rein zufälliger Auswahl.

6.5.2 Unabhängige Ereignisse

Häufig ist die Wahrscheinlichkeit, dass das Ereignis A eintritt, unabhängig davon, ob das Ereignis B eingetreten ist oder nicht. Wenn dies der Fall ist, heißen A und B voneinander **unabhängige Ereignisse**. Dafür schreibt man

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

Es gilt

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

Diese Beziehung heißt **Multiplikationssatz für unabhängige Ereignisse**.

6.5.3 Satz von der totalen Wahrscheinlichkeit

Unter der Voraussetzung, dass der Ereignisraum S in paarweise disjunkte Teilmengen B_i für $i = 1, \ldots, n$ zerlegt ist, gilt für jedes Ereignis A

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$$

Beispiel:

In eine Berliner Kleintierpraxis kommen zu 30% große Hunderassen und zu 70% kleine Hunderassen. Große Hunde leiden zu 40% an Hüftgelenksdysplasie (HD), kleine Hunderassen nur zu 10%. Wie groß ist die Wahrscheinlichkeit, dass ein beliebiger Hund, der in der Praxis vorstellig wird, an Hüftgelenksdysplasie leidet?

$$\begin{aligned}
P(\text{HD}) &= P(\text{HD} \mid \text{groß}) \cdot \text{P(groß)} + \text{P(HD} \mid \text{klein)} \cdot \text{P(klein)} \\
&= 0.4 \cdot 0.3 + 0.1 \cdot 0.7 \\
&= 0.19
\end{aligned}$$

Die Wahrscheinlichkeit beträgt 19%.

6.6 Wahrscheinlichkeitsverteilungen

Der Begriff der Wahrscheinlichkeitsverteilung resultiert aus der Frage, wie sich bei einem Zufallsexperiment die Wahrscheinlichkeiten auf die verschiedenen Ereignisse verteilen, d.h. welche Wahrscheinlichkeit diese Ereignisse bei dem betreffenden Experiment jeweils besitzen.

Man unterscheidet zwei Klassen von Verteilungen, die diskreten und die stetigen. Bei Zufallsexperimenten, bei denen man zählt (Krankheitsfälle, Patienten, Diagnosen), treten diskrete Verteilungen auf. Bei Experimenten, bei denen man misst, d.h. eine kontinuierlich veränderliche Größe beobachtet (Länge, Temperatur, Milchleistung), treten stetige Verteilungen auf.

Bisher ging es darum, empirisch beobachtbare Wirklichkeit zu beschreiben. Jetzt schaffen wir uns theoretische Modelle der Grundgesamtheit, und alle Begriffe sind theoretischer Natur. Die Begriffe mit ähnlichen Namen entsprechen sich jeweils einander, wobei allerdings beide Begriffsarten prinzipiell verschieden sind. Am Beispiel: eine Grundgesamtheit besitzt eine einzige wohlbestimmte Wahrscheinlichkeitsverteilung. Die Stichproben, die man aus dieser Grundgesamtheit entnehmen kann, unterscheiden sich im allgemeinen voneinander und besitzen demzufolge im allgemeinen auch verschiedene Häufigkeitsverteilungen.

6.6.1 Zufallsvariable

In den meisten Fällen lässt sich das Ergebnis einer einzelnen Ausführung eines Zufallsexperimentes durch eine oder mehrere Zahlen kennzeichnen.

Dies gilt z.B. für das Würfeln mit einem einzelnen Würfel. Dabei erhalten wir als Ergebnis eines Wurfs eine der Zahlen $1, 2, \ldots, 6$. Dieses Ergebnis wird also durch eine variable Größe, die wir mit X bezeichnen wollen, gekennzeichnet. X ist demnach eine Funktion, die bei jedem Wurf einen der Werte $1, 2, \ldots$ oder 6 annimmt. Welchen dieser Werte sie bei einem bestimmten Wurf annimmt, das hängt 'vom Zufall' ab. Eine solche Funktion, die das Ergebnis eines Zufallsexperimentes ausdrückt, wird als **Zufallsvariable** bezeichnet. Grob gesprochen ist also eine Zufallsvariable X eine Funktion, deren Werte reelle Zahlen sind und 'vom Zufall' abhängen.

Definition: Eine Funktion X, die jedem Elementarereignis eines Stichprobenraumes S eine reelle Zahl zuordnet, heißt **Zufallsvariable**.

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

Beispiel l: Münzwurf

Der Zufallsvorgang besteht darin, dass eine Münze zweimal geworfen wird. Es können entweder Kopf (K) oder Zahl (Z) oben liegen.

Der Ereignisraum besteht aus den vier Elementarereignissen:

$$e_1 = \{KK\} \quad e_2 = \{KZ\} \quad e_3 = \{ZK\} \quad e_4 = \{ZZ\}$$

Wir definieren eine Zufallsvariable X durch die Vorschrift $x_i =$ Anzahl der 'Köpfe' in e_i. Das Ergebnis ist in der folgenden Tabelle dargestellt:

Elementarereignis in S	e_1	e_2	e_3	e_4
Wert der Zufallsvariablen X	2	1	1	0

Die Zufallsvariable X hat drei verschiedene Ausprägungen: 0, 1 und 2.
Nimmt man an, dass Kopf und Zahl mit gleicher Wahrscheinlichkeit geworfen werden, lässt sich die Wahrscheinlichkeitsfunktion der Zufallsvariablen mit Hilfe der Rechenregeln für Wahrscheinlichkeiten ableiten. Tabelle 6.1 enthält in Spalte l die Ausprägungen der Zufallsvariablen, in der 2. Spalte die Wahrscheinlichkeitsfunktion sowie in Spalte 3 die Verteilungsfunktion.

x	$P(X=x)$	$F(x)$
0	0.25	0.25
1	0.50	0.75
2	0.25	1.00
Summe	1.00	–

Tab. 6.1: Wahrscheinlichkeitsfunktion für Beispiel 1

Beispiel 2: Katzenrassen

Im Tierheim werden jeweils 3 Katzen untersucht, um die Rasse festzustellen. Jedes Tier kann der Rasse 'Europäisch Kurzhaar' (EK) oder 'Sonstige' (sonst.) angehören. Der Zufallsvorgang besteht aus der Entnahme von 3 Katzen jeweils **mit Zurücklegen**. Die Elementarereignisse sind alle möglichen Kombinationen der beiden Rassen 'Europäisch Kurzhaar' und 'Sonstige'.

Definiert man eine Zufallsvariable X über die Vorschrift

$$X = \text{Anzahl Europäisch Kurzhaar unter den 3 Tieren}$$

erhält man eine Zufallsvariable mit 4 möglichen Ausprägungen (Spalte 2 von Tabelle 6.2).

Man kann natürlich in demselben Stichprobenraum auch eine andere Zufallsvariable definieren, z.B.

$$Y = \begin{cases} 0, & \text{es ist keine Europäisch Kurzhaar–Katze unter den 3 Katzen} \\ 1, & \text{es sind ein oder mehrere Kurzhaarkatzen unter den 3 Katzen} \end{cases}$$

Die Ausprägungen von Y stehen in Spalte 3 von Tabelle 6.2.

Elementarereignis e			x	y
EK	EK	EK	3	1
EK	EK	sonst.	2	1
EK	sonst.	EK	2	1
sonst.	EK	EK	2	1
EK	sonst.	sonst.	1	1
sonst.	EK	sonst.	1	1
sonst.	sonst.	EK	1	1
sonst.	sonst.	sonst.	0	0

Tab. 6.2: Elementarereignisse für Beispiel 2

Nimmt man an, dass die Wahrscheinlichkeit, eine Europäisch Kurzhaar–Katze zu ziehen, $P(\text{EK}) = 0.9$ beträgt, und damit $P(\text{sonst.}) = 0.1$ ist, lassen sich die Wahrscheinlichkeitsfunktionen für die beiden Zufallsvariablen X und Y berechnen.

x	$P(X = x)$		$F(x)$
0	$0.1^3 =$	0.001	0.001
1	$3 \cdot 0.1^2 \cdot 0.9^1 =$	0.027	0.028
2	$3 \cdot 0.1^1 \cdot 0.9^2 =$	0.243	0.271
3	$0.9^3 =$	0.729	1.000
Summe		1.000	

y	$P(Y = y)$	$F(y)$
0	0.001	0.001
1	0.999	1.000
Summe	1.000	–

Tab. 6.3: Wahrscheinlichkeitsfunktionen für Beispiel 2

Beispiel 3: Gewichte von Ferkeln

Der Zufallsvorgang bestehe im Wiegen der Ferkel eines Aufzuchtbetriebes. Die Abbildungsvorschrift für die Zufallsvariable sei X = Gewicht des Ferkels. Die Zufallsvariable kann jeden Wert der positiven reellen Zahlen in einem bestimmten altersabhängigen Bereich annehmen.

Tritt bei der Ausführung eines bestimmten Experimentes das Ereignis ein, das einem bestimmten Zahlenwert x entspricht, so sagt man, dass die zu diesem Experiment gehörige Zufallsvariable X den Wert x angenommen habe. Wir sprechen auch kurz von dem Ereignis $X = x$ bei dem betreffenden Experiment.

Die Wahrscheinlichkeit, dass X den Wert x annimmt, bezeichnen wir mit

$$P(X = x).$$

Die Wahrscheinlichkeit des Ereignisses "X nimmt irgendeinen Wert in dem Intervall $a < X < b$ an" bezeichnen wir mit

$$P(a < X < b)$$

Entsprechend bezeichnet

$$P(X \leq c)$$

die Wahrscheinlichkeit des Ereignisses "X nimmt irgendeinen Wert an, der höchstens gleich c ist" und

$$P(X > c)$$

bezeichnet die Wahrscheinlichkeit des Ereignisses "X nimmt irgendeinen Wert an, der größer als c ist".

Die beiden zuletzt genannten Ereignisse sind für jeden beliebigen Wert von c disjunkt. Gemäß Axiom 3 von Kolmogoroff (siehe Abschnitt 6.4) gilt also

$$P(X \leq c) + P(X > c) = P(-\infty < X < \infty).$$

Aufgrund von Axiom 2 ist die Wahrscheinlichkeit auf der rechten Seite gleich 1, denn $-\infty < X < \infty$ ist ein sicheres Ereignis, da X stets irgendeinen Wert auf der Zahlengerade annehmen muss. So erhalten wir die wichtige Beziehung

$$P(X > c) = 1 - P(X \leq c) \qquad \text{für beliebiges } c \in \mathbb{R}.$$

6.6.2 Wahrscheinlichkeitsfunktion für diskrete Merkmale

Die meisten Zufallsvariablen lassen sich in zwei Klassen einteilen, nämlich in die so genannten diskreten (Beispiel 1 und 2) und die stetigen (Beispiel 3).

Eine Zufallsvariable X und ihre Verteilung heißen **diskret**, wenn folgendes gilt:

1. Die Variable X kann nur endlich viele oder abzählbar unendlich viele (reelle) Werte mit positiver Wahrscheinlichkeit annehmen; diese einzelnen Werte indizieren wir mit j, $j = 1, 2, \ldots$.

2. In jedem endlichen Intervall der reellen Zahlengeraden liegen nur endlich viele der genannten Werte. Für jedes Intervall $a < X \leq b$, das keinen solchen Wert enthält, ist die zugehörige Wahrscheinlichkeit gleich Null.

Die Wahrscheinlichkeit für das Auftreten von x_j wird mit

$$P(X = x_j) = p_j$$

bezeichnet.

Definition: Die Funktion

$$f(x) = \begin{cases} p_j & \text{für } X = x_j \qquad (j = 1, 2, \ldots) \\ 0 & \text{sonst} \end{cases}$$

heißt **Wahrscheinlichkeitsfunktion** der betreffenden Zufallsvariablen X.

Da X stets irgendeinen Wert annimmt, muss die Summe aller $f(x_j)$ gleich 1 sein. Durch die Wahrscheinlichkeitsfunktion $f(x)$ ist die "Wahrscheinlichkeitsverteilung" oder die "Verteilung" der betreffenden Zufallsvariablen vollständig bestimmt.

Bei stetigen Merkmalen ist die Wahrscheinlichkeit, einen einzelnen bestimmten Wert zu erhalten, gleich Null. Ereignisse werden daher als Intervalle dargestellt, deren Wahrscheinlichkeit, aufzutreten, als Fläche unter einer Kurve berechnet wird. Im Zusammenhang mit der im folgenden definierten Verteilungsfunktion wird darauf näher eingegangen.

6.6.3 Verteilungsfunktion

Die Verteilungsfunktion gibt die Wahrscheinlichkeit an, dass die Zufallsvariable X einen Wert annimmt, der unterhalb eines vorgegebenen Wertes x liegt und wird an der Stelle x mit

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

bezeichnet.

Aus der Verteilungsfunktion $F(x)$ lässt sich die Wahrscheinlichkeit $P(a < X \leq b)$ berechnen, da folgende grundlegende Beziehung gilt:

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

Verteilungsfunktion einer diskreten Verteilung

Ist $f(x)$ die Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariablen X, so ergibt sich die zu einem Intervall gehörige Wahrscheinlichkeit von X durch Summation. Insbesondere gilt:

$$F(x) = P(X \leq x) = \sum_{x_j \leq x} f(x_j)$$

Hierbei wird also über alle x_j summiert, die höchstens gleich der vorgegebenen Zahl x sind.

Die Verteilungsfunktion $F(x)$ einer diskreten Verteilung ist eine Treppenfunktion. An der Stelle $x = x_j$ springt sie um den Betrag $P(X = x_j)$ in die Höhe, dazwischen verläuft sie konstant.

Verteilungsfunktion einer stetigen Verteilung

Die Verteilungsfunktion lässt sich analog zum diskreten Fall definieren.

Eine Zufallsvariable X und deren Verteilung heißen stetig, wenn die zugehörige Verteilungsfunktion in Integralform dargestellt werden kann

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(v)dv.$$

Der Integrand f heißt die **Wahrscheinlichkeitsdichte** oder kurz **Dichtefunktion** der betreffenden Verteilung. Die Ableitung der Verteilungsfunktion ist die Dichtefunktion:

$$F'(x) = f(x)$$

Es gilt

$$\int_{-\infty}^{\infty} f(v)dv = 1$$

und weiterhin

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(v)dv.$$

Somit ist die Wahrscheinlichkeit des Ereignisses $a < X \leq b$ gleich der Fläche unter der Wahrscheinlichkeitsdichte $f(x)$ zwischen $x = a$ und $x = b$.

Allgemein ist $F(x)$ eine stetige, monoton steigende, nicht negative Funktion. Wegen der Stetigkeit von $F(x)$ gilt außerdem

$$P(X = a) = 0$$

Abbildung 6.2 zeigt am Beispiel der Binomialverteilung die Wahrscheinlichkeitsfunktion und die Verteilungsfunktion einer diskreten Zufallsvariablen (linke Hälfte) und am Beispiel der Normalverteilung Dichtefunktion und zugehörige Verteilungsfunktion einer stetigen Verteilung (rechte Hälfte). Diese beiden Verteilungstypen werden in Abschnitt 6.7.2 und 6.7.3 noch näher vorgestellt.

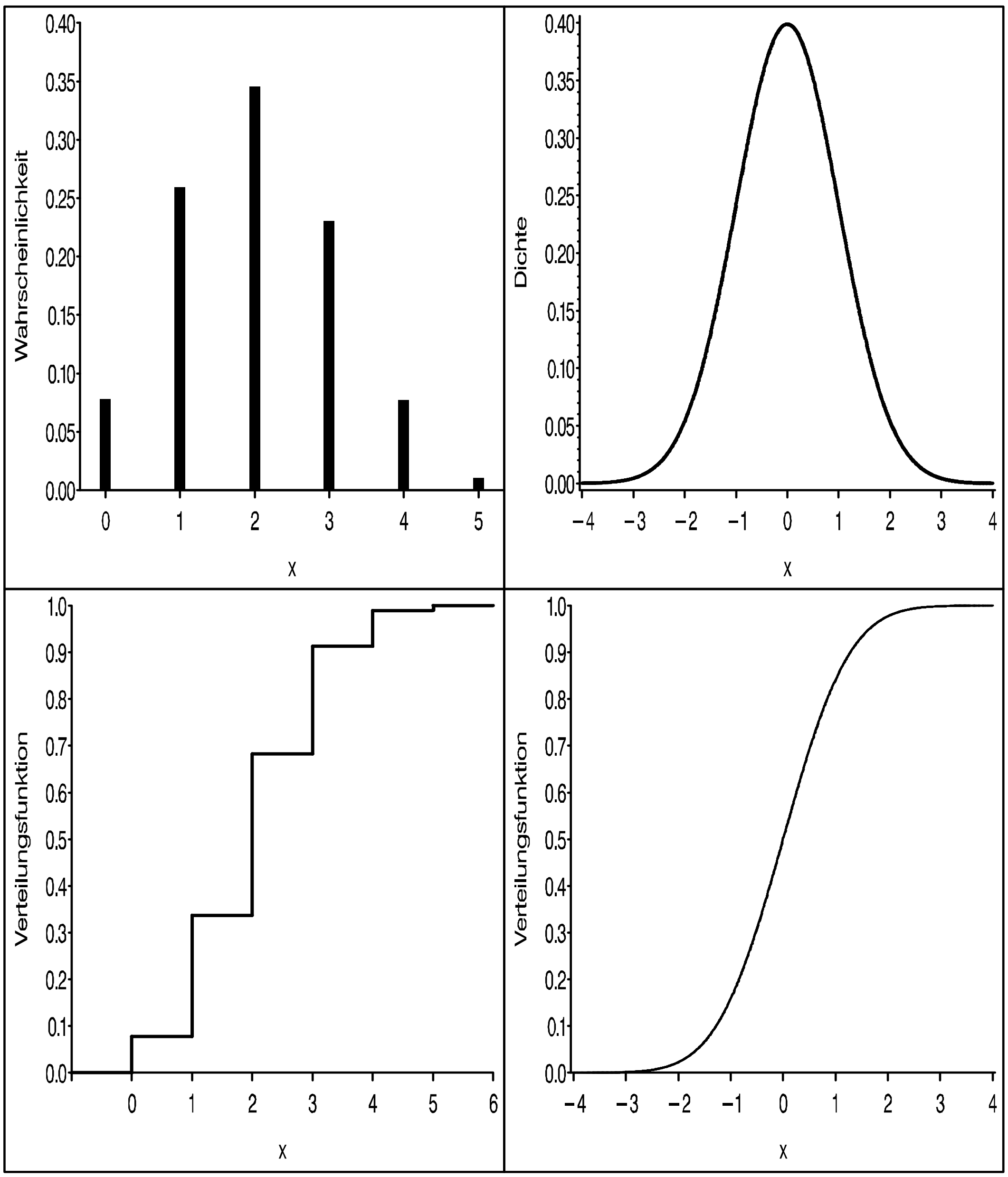


Abb. 6.2: Grafische Darstellung von diskreten und stetigen Wahrscheinlichkeitsverteilungen

6.6.4 Maßzahlen von Wahrscheinlichkeitsverteilungen

Theoretische Verteilungen wie Binomial- und Normalverteilung besitzen formal ganz ähnliche Eigenschaften wie Häufigkeitsverteilungen. Dies gilt auch für Parameter von Verteilungen, die z.B. die Lage oder die Variabilität beschreiben.

Der **Erwartungswert** einer diskreten Zufallsvariable wird analog dem arithmetischen Mittel einer empirischen Verteilung gebildet und ist eine Maßzahl für das Zentrum einer Verteilung.

Definition: Der Erwartungswert $E(X)$ einer diskreten Zufallsvariable ist definiert als

$$E(X) = x_1 p_1 + \ldots + x_k p_k = \sum_{j=1}^{k} x_j p_j.$$

Im stetigen Fall ist der Erwartungswert definiert als

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

Definition: Die **Varianz** einer diskreten Zufallsvariable ist definiert als

$$\mathrm{Var(X) = E(X - E(X))^2}.$$

Trotz der Ähnlichkeit von Erwartungswert und arithmetischem Mittel ist deren Funktion deutlich zu unterscheiden. Während das arithmetische Mittel die Lage von erhobenen Daten beschreibt, charakterisiert $E(X)$ das Verhalten eines Zufallsexperimentes.

Wir haben hier zunächst ganz allgemein das Konzept der Wahrscheinlichkeitsfunktion, Dichte- und Verteilungsfunktion beschrieben. Im folgenden werden wir einige ausgewählte Wahrscheinlichkeitsverteilungen vorstellen, die häufig bei statistischen Analysen Verwendung finden.

6.7 Beispiele für Wahrscheinlichkeitsverteilungen

6.7.1 Gleichverteilung für diskrete Merkmale

Hat jede der n Merkmalsausprägungen eines Stichprobenraums die gleiche Wahrscheinlichkeit, bei einem Zufallsvorgang einzutreten, d.h. $P(X = x_j) = \frac{1}{n}$ für $j = 1,\ 2,\ \ldots\ ,\ n$, spricht man von einer (diskreten) Gleichverteilung. Sie findet Anwendung bei der Berechnung von Wahrscheinlichkeiten beim Münzwurf, Roulette und Würfeln, außerdem wird diese Verteilung für die Randomisierung bei Versuchsplänen eingesetzt. Erwartungswert und Varianz lauten allgemein für die Gleichverteilung

$$E(X) = \frac{n+1}{2} \qquad \text{und} \qquad Var(X) = \frac{(n^2-1)}{12}$$

Beim Würfeln beträgt der Erwartungswert also $\frac{6+1}{2} = 3.5$. Bei 10maligem Würfeln erwartet man also eine Gesamtaugenzahl von 35 $(= 10 \cdot 3.5)$.

6.7.2 Binomialverteilung

Die Binomialverteilung ist die wichtigste diskrete Wahrscheinlichkeitsverteilung. Sie wird verwendet, wenn man sich für die Anzahl des Eintreffens eines bestimmten Ereignisses A bei der wiederholten Ausführung eines Zufallsexperimentes interessiert, bei dem A bei jeder Ausführung dieselbe Wahrscheinlichkeit p besitzt und sich die Ergebnisse der verschiedenen Ausführungen gegenseitig nicht beeinflussen. In Kurzform:

- Es gibt nur zwei Ergebnisse des Experimentes: A und $\overline{A}$.
- Das Zufallsexperiment wird n-mal wiederholt.
- Bei allen n Wiederholungen ist $p = P(A)$ konstant.
- Die Ergebnisse in den n Wiederholungen sind unabhängig.

Herleitung der Wahrscheinlichkeitsfunktion

Hat ein Ereignis A bei einem Zufallsexperiment die Wahrscheinlichkeit p, so ist die Wahrscheinlichkeit, dass A nicht eintrifft, gleich

$$P(\overline{A}) = q = 1 - p$$

Wir betrachten nun die Zufallsvariable

$$Y = \text{ Anzahl des Eintreffens von } A.$$

Wird das Experiment nur einmal ausgeführt, so kann Y nur die Werte 0 oder 1 annehmen, je nachdem, ob A nicht eintrifft oder eintrifft. Ein solches Experiment wird auch als 'Bernoulli–Experiment' bezeichnet. Die zugehörigen Wahrscheinlichkeiten sind

$$P(Y = 0) = q \quad \text{und} \quad P(Y = 1) = p.$$

Wird das obige Zufallsexperiment n-mal hintereinander ausgeführt, so kann man eine neue Zufallsvariable X definieren, die die n Zufallsvariablen Y aufsummiert. X kann die Werte $0, 1, \ldots, n$ annehmen. Wir fragen nach den zugehörigen Wahrscheinlichkeiten unter der Voraussetzung, dass die Ergebnisse der einzelnen Ausführungen voneinander unabhängig sind, d.h. sich gegenseitig nicht beeinflussen.

Das Ereignis

$$X = x \quad (\text{'}A \text{ tritt bei den } n \text{ Ausführungen genau } x\text{-mal ein'})$$

ist z.B. realisiert, wenn wir (in dieser Reihenfolge des Auftretens) die Einzelereignisse

$$\underbrace{A\,A\,\cdots\,A}_{x-mal}\,\underbrace{\overline{A}\,\overline{A}\,\cdots\,\overline{A}}_{(n-x)-mal}$$

erhalten. Da die Einzelereignisse unabhängig sind, hat die Realisierung des Ereignisses $X = x$ die Wahrscheinlichkeit

$$\underbrace{p \cdot p \cdot p \cdots p}_{x-mal} \cdot \underbrace{q \cdot q \cdot q \cdots q}_{(n-x)-mal} = p^x \cdot q^{n-x}$$

Jede andere Form der Realisierung des Ereignisses $X = x$ ergibt sich aus der obigen Form durch eine Änderung der Reihenfolge der Einzelereignisse. Für die Reihenfolge gibt es insgesamt $\binom{n}{x}$ Möglichkeiten.

Hierbei bezeichnet $\binom{n}{x}$ den Binomialkoeffizienten, der die Anzahl der Kombinationen angibt, von n Ereignissen x auszuwählen und durch

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \qquad \text{mit } n! = 1 \cdot 2 \cdot \ldots \cdot n$$

definiert ist (siehe auch Anhang A). Geordnet ergeben die Binomialkoeffizienten $\binom{n}{x}$ das bekannte PASCALsche Dreieck (vgl. Abbildung 6.3).

n \ x	0	1	2	3	4	5	6	7	8	9	10
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1

Abb. 6.3: Pascalsches Dreieck (bis $n = x = 10$)

Damit ergibt sich die Wahrscheinlichkeitsfunktion der Binomialverteilung als

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{für } x = 0, 1, \ldots, n.$$

Für eine binomialverteilte Zufallsvariable X mit den Parametern n und p schreibt man kurz

$$X \sim Bin(n, p)$$

Der Erwartungswert der Binomialverteilung beträgt $E(X) = np$, die Varianz ist $Var(X) = np(1-p)$. Bei $p = 0.5$ hat die Varianz den größten Wert.

Grafische Beispiele für Binomialverteilungen mit unterschiedlichem p und n sind in Abbildung 6.4 dargestellt.

Beispiel: Studenten

Es sei bekannt, dass 80% der Studenten rauchen. Man befragt $n = 5$ Studenten. Mit welcher Wahrscheinlichkeit sind alles Raucher, oder 1, 2, 3, 4 Raucher oder kein Raucher unter den Befragten?

Damit hat man folgende Angaben, die man zur Berechnung der Wahrscheinlichkeitsfunktion benötigt:

- Zufallsvariable X: Anzahl der Raucher mit Ausprägungen $x_j = 0, 1, 2, 3, 4, 5$
- $p = 0.8$ (Wahrscheinlichkeit für Raucher)
  $\Longrightarrow (1 - p) = q = 0.2$ (Wahrscheinlichkeit für Nichtraucher)
- $n = 5$.

Die Wahrscheinlichkeitsfunktion für dieses Beispiel lautet

$$P(X = x) = \binom{5}{x} 0.8^x 0.2^{5-x} \qquad \text{für } x = 0, 1, \ldots, 5$$

Die Wahrscheinlichkeit, dass genau zwei Raucher unter den 5 Studenten sind, beträgt
$P(X = 2) = 10 \cdot 0.8^2 \cdot 0.2^3 = 0.05$.

Anwendungsbereiche der Binomialverteilung

Binomialverteilungen dienen als Wahrscheinlichkeitsmodelle bei der Beurteilung von Versuchsergebnissen, die sich als relative Häufigkeit (Anteile) ausdrücken lassen. Typische Beispiele sind

- Anteil der an einer bestimmten Krankheit leidenden Patienten, die bei einem Therapieversuch nach Anwendung einer bestimmten Therapie geheilt werden.
- Anteil der Mädchengeburten auf der Neugeborenenstation eines Berliner Krankenhauses.
- Anteil der Studenten, die bei einem Lernversuch eine bestimmte Aufgabe innerhalb einer festgelegten Zeit lösen.
- Anteil der Träger der Blutgruppe A unter 5 zufällig ausgewählten Personen.
- Anteil der bei einem Wirksamkeitsversuch abgetöteten Bakterien.

Weitere häufig auftretende diskrete Wahrscheinlichkeitsverteilungen sind die **hypergeometrische Verteilung** (bei kleinen Gesamtheiten und Auswahlsätzen $\frac{n}{N} > \frac{1}{10}$ und die **Poissonverteilung** (bei seltenen Ereignissen).

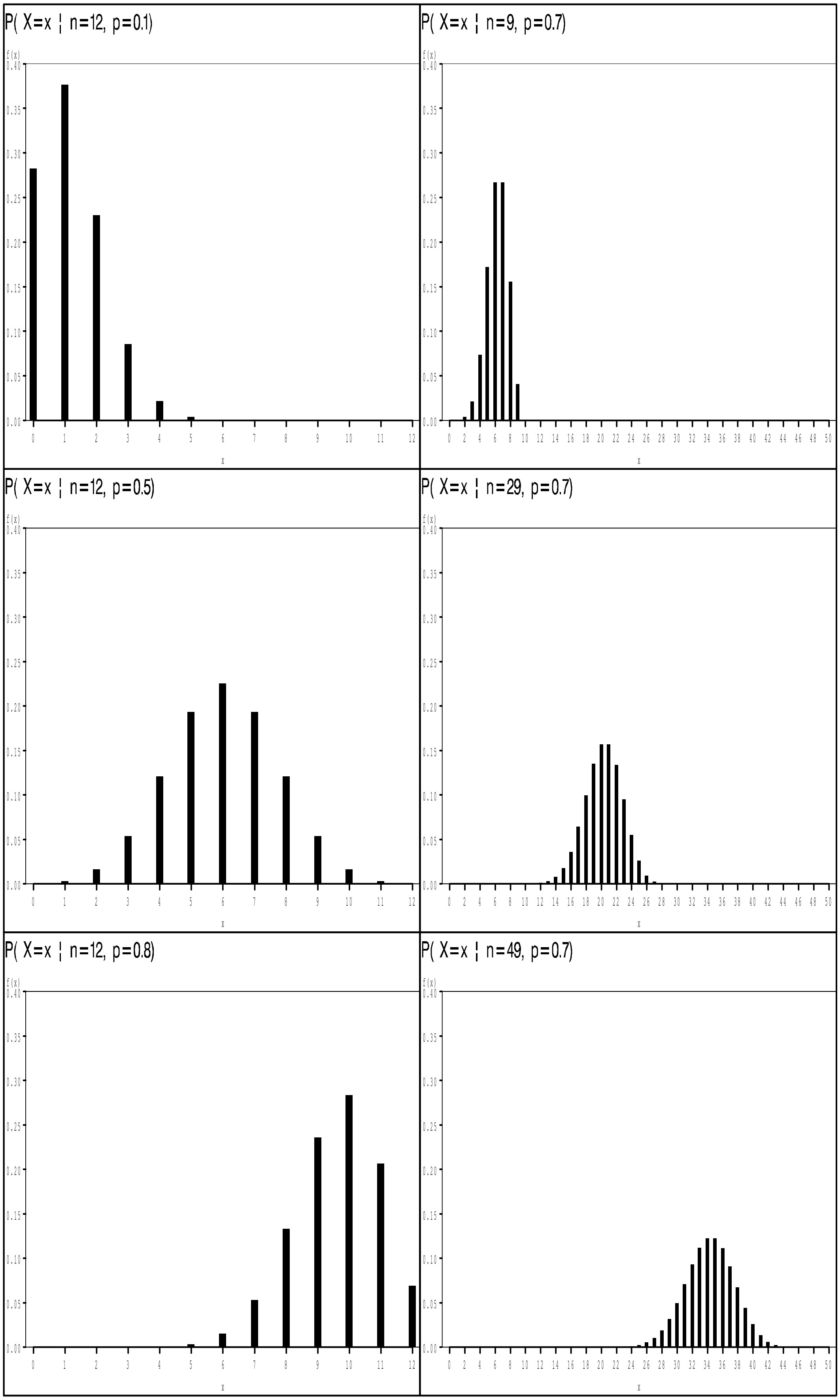


Abb. 6.4: Binomialverteilungen mit unterschiedlichem n und p

6.7.3 Normalverteilung

In der Statistik hat die **Normalverteilung** — oft **auch Gaußverteilung** genannt — aus folgenden Gründen große Bedeutung:

1. Viele Merkmale, wie Körpergröße, Körpergewicht etc. sind in der Praxis näherungsweise normalverteilt, weil unterstellt werden kann, dass zahlreiche kleine Einflüsse wie genetische und Umwelt–Einflüsse sowie Messfehler **unabhängig** und **additiv** auf die Merkmale einwirken.

2. Viele biologische Merkmale, wie Leukozytenzahl, Blutglukose, Keimzahlen, individuell letale Dosen bei toxikologischen Experimenten etc. sind in der Praxis nicht symmetrisch, sondern rechtsschief verteilt. Diese Verteilungen können oft in solche überführt werden, die approximativ der Normalverteilung entsprechen, wenn die Beobachtungswerte durch ihre Logarithmen ersetzt werden. Als Erklärung dafür wird angeführt, dass viele Einflüsse **nicht additiv**, sondern **multiplikativ** wirken. Durch Übergang zu den Logarithmen ergeben sich daraus jedoch wiederum additive Einflusskomponenten.

3. Auf Grund statistischer Sätze folgen Summen und Mittelwerte aus Zufallsstichproben bei zunehmendem Stichprobenumfang im allgemeinen näherungsweise einer Normalverteilung unabhängig davon, welche biologische Ausgangsverteilung die betrachteten Merkmale haben.

4. Die Normalverteilung ist die Grundlage der so genannten parametrischen statistischen Schätz- und Prüfverfahren.

Die **Dichtefunktion** der Normalverteilung lautet :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \quad e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{mit } \mu, x \in \mathbb{R},\, \sigma \in \mathbb{R}^+$$

mit den beiden Parametern μ und σ. Zu jedem $\mu \in \mathbb{R}$ und $\sigma \in \mathbb{R}^+$ gibt es genau eine Normalverteilung. Man sagt, die Zufallsvariable X ist normalverteilt mit Erwartungswert $E(X) = \mu$ und Varianz $Var(X) = \sigma^2$ bzw. Standardabweichung σ und schreibt dafür in Kurzform:

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

Wie man den Abbildungen 6.5 und 6.6 entnehmen kann, kennzeichnet der Erwartungswert μ die Lage der Verteilung, ohne dass sich bei fest vorgegebener Standardabweichung σ (bzw. Varianz σ^2) die Form der Dichtefunktion ändert, wenn μ variiert. Der Parameter σ bestimmt die Breite der Dichtefunktion, die wegen ihres Aussehens auch Glockenkurve genannt wird. Bei kleineren Werten wird die Form der Dichtefunktion immer steiler (enger), bei größeren Werten für σ hingegen breiter.

Man erkennt weiterhin, dass die Dichte $f(x)$ ihr Maximum an der Stelle $x = \mu$ annimmt, symmetrisch um μ ist und auch der Modalwert gleich μ ist. Aus der Symmetrie der Verteilung folgt, dass auch der Median gleich μ ist. Die Dichtefunktion hat zwei Wendepunkte, die in den Punkten $x = \mu - \sigma$ und $x = \mu + \sigma$ liegen.

Eine normalverteilte Größe X kann theoretisch alle Werte zwischen $-\infty$ und $+\infty$ annehmen, es gilt aber, dass sich die Dichte immer mehr Null annähert, je weiter die Werte an den Rändern liegen:

$$\lim_{x \to -\infty} f(x) = \lim_{x \to +\infty} f(x) = 0$$

d.h. die Wahrscheinlichkeitsmasse (die Fläche unter der Dichte) konzentriert sich symmetrisch um μ und wird in den Randbereichen (Schwänzen) immer kleiner.

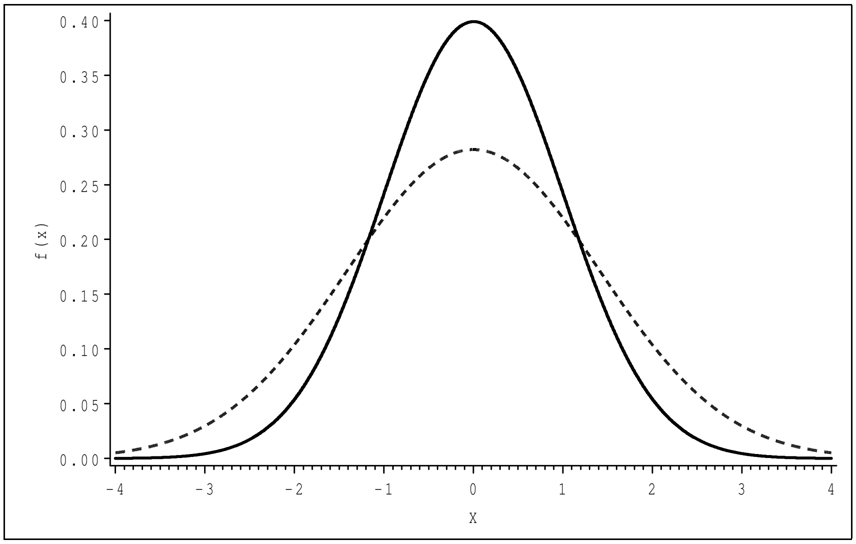


Abb. 6.5: Dichtefunktionen der Normalverteilung für Erwartungswert 0 und **Varianz 1** (durchgezogene Linie) und Erwartungswert 0 und **Varianz 2** (gestrichelte Linie)

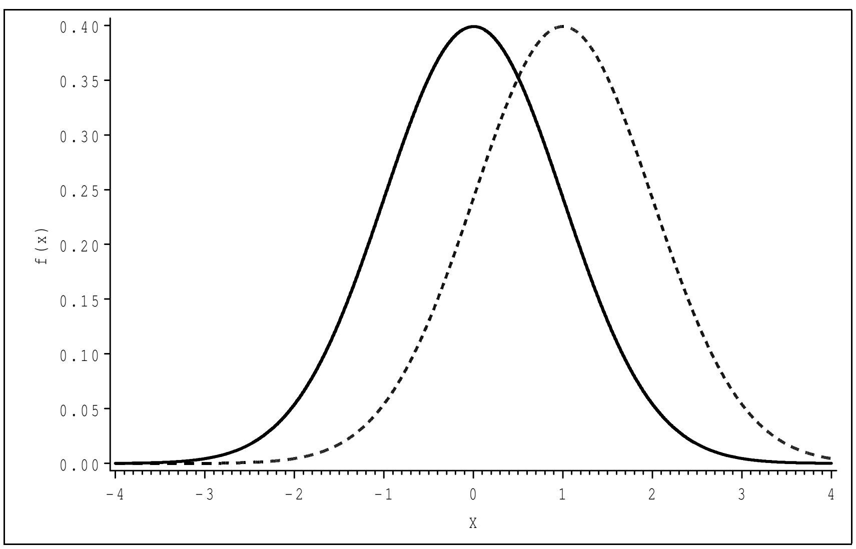


Abb. 6.6: Dichtefunktionen der Normalverteilung für **Erwartungswert 0** und Varianz 1 (durchgezogene Linie) und **Erwartungswert 1** und Varianz 1 (gestrichelte Linie)

Wahrscheinlichkeiten bestimmter Ereignisse lassen sich durch Flächen unter der Dichtefunktion veranschaulichen, dies erkennt man auch an den Abbildungen im Vorspann zu den Normalverteilungstabellen in Anhang B. Zur Bestimmung dieser Wahrscheinlichkeiten wäre es nun allerdings notwendig, für jede Kombination von μ und σ eine Funktionswerttabelle bereitzuhalten, eine sicherlich erschreckende Vorstellung. Wie im folgenden dargestellt wird, kann durch eine stets mögliche Transformation das Problem soweit reduziert werden, dass nur die im Anhang B gegebene Tabelle benötigt wird. Dazu transformiert man die mit μ und σ verteilte Zufallsgröße X über die Gleichung

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

in eine Zufallsvariable U, die eine **standardisierte Normalverteilung**, d.h. eine Verteilung mit Erwartungswert $\mu = 0$ und Standardabweichung $\sigma = 1$ besitzt. In den Tabellen im Anhang B wird die Verteilungsfunktion der Standardnormalverteilung mit $F_U(u)$ bezeichnet.

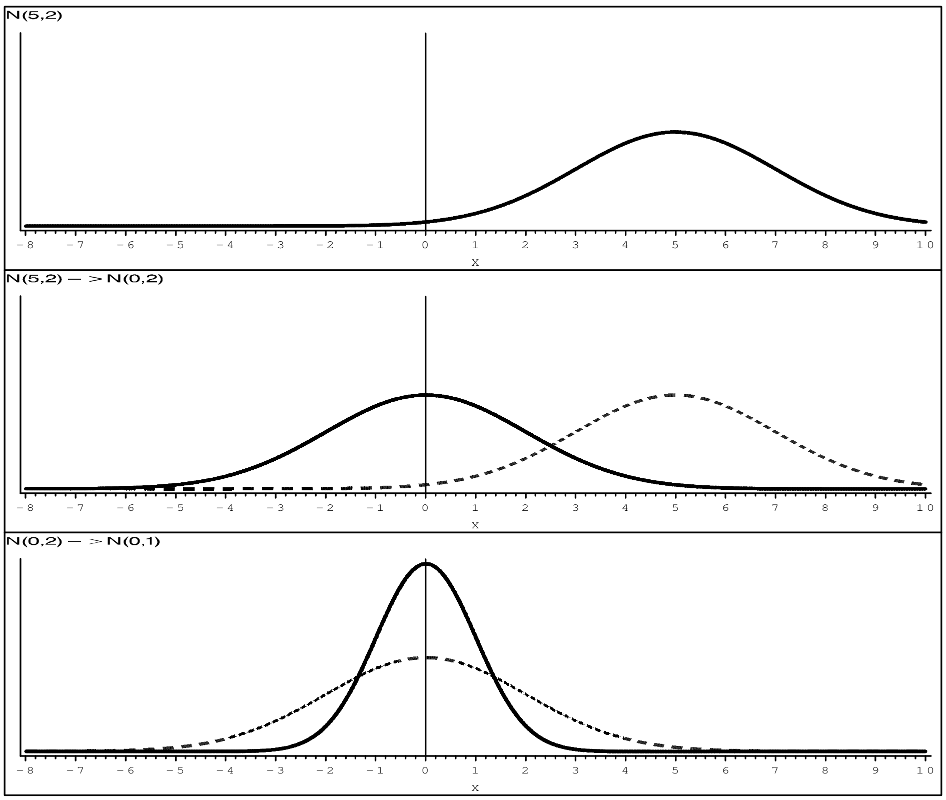


Abb. 6.7: Standardisierung der Normalverteilung $N(5, 2)$.
1. Schritt: Skalen–Verschiebung um 5 Einheiten; es ergibt sich die $N(0, 2)$.
2. Schritt: Skalen–Verzerrung mit dem Faktor 2; es entsteht die $N(0, 1)$. (Quelle: Lorenz, 1992)

Die Verteilungsfunktion ist das Integral über die Dichte, also

$$F_U(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \qquad \text{mit } y \in \mathbb{R}$$

Aus der Symmetrie der Dichtefunktion folgt generell, dass

1. $F_U(-u) = 1 - F_U(u)$ sowie $u_\alpha = -u_{1-\alpha}$

   Weiterhin kann geschrieben werden :

2. $P(x_u \leq X \leq x_o) = F(x_o) - F(x_u)$

   bezeichnet die Wahrscheinlichkeit, dass eine Realisation der normalverteilten Größe X zwischen der unteren Grenze x_u und der oberen Grenze x_o liegt.

3. Spezialfall für ein symmetrisches Intervall: $P(-u < U < +u) = D(u)$

4. $P(X \leq \mu - x) = P(X > \mu + x) \quad \Longrightarrow F(\mu - x) = 1 - F(\mu + x)$

5. $F(\mu) = 0.5$, d.h. der Median ist gleich μ.

Jede dieser Aussagen sollte sich der Leser an einer Skizze veranschaulichen; dieses Vorgehen hilft auch im weiteren sehr.

Konkret verwendet wird die angenehme Eigenschaft, mit nur einer (der standardisierten) Normalverteilungstabelle auszukommen, wie folgt:

Wäre für eine beliebige Normalverteilung der Wert $F(x_o)$ zu bestimmen, so berechnen wir

$$u = \left(\frac{x_o - \mu}{\sigma}\right)$$

und suchen $F_U(u)$ in unserer Tabelle im Anhang B, denn es gilt

$$F(x_o) = F_U(\frac{x_o - \mu}{\sigma}) = F_U(u)$$

Beispiel l: Sie kaufen ein Paket abgepackte Hähnchenschenkel mit einem Sollgewicht (Erwartungswert des Gewichtes) von 1000 g. Wie groß ist die Wahrscheinlichkeit, dass Sie ein Paket mit höchstens 800 g erwischen?

gegeben:	$\mu = 1000$, $\sigma^2 = 10000$, $x = 800$
gesucht:	$P(X \leq x)$
Lösung:	$P(X \leq x) = F(800)$
	und wegen
	$F(x) = F_U(\frac{x - \mu}{\sigma}) = F_U(u)$
	folgt
	$F(800) = F_U(\frac{800 - 1000}{100}) = F_U(u = -2) = 0.0228$

Für $u = -2$ liest man im 2. Teil der Tabelle im Anhang B (in der Ecke unten links steht $u = 2$) den Wert in der Spalte $F_U(-u)$ von 0.0228 ab.

Die Wahrscheinlichkeit beträgt 2.28%.

Beispiel 2: Masthähnchen haben ein Sollgewicht (Erwartungswert des Gewichtes) von 1475 g und eine Soll–Standardabweichung von 100 g. Um als Hähnchen der Güteklasse A in den Verkauf zu kommen, müssen sie mindestens 1250 g wiegen.

gegeben: $\mu = 1475$, $\sigma^2 = 10000$, $x = 1250$
gesucht: $P(X > x)$
Lösung: Offensichtlich ist $P(X > 1250)$ gesucht,
unsere Tabellen geben aber keine direkte Antwort auf ein solche Fragestellung.
Deshalb überlegen wir uns, dass
$P(X > x) + P(X \leq x) = 1$ ist und somit
$P(X > x) = 1 - P(X \leq x) = 1 - F(x) = F(-x)$
Man errechnet
$u = \frac{x - \mu}{\sigma} = \frac{1250 - 1475}{100} = -2.25$
und liest ab
$F_U(2.25) = 0.9878$.

Der Mäster kann davon ausgehen, dass 98.78% seines Bestandes in die Güteklasse A eingestuft werden.

Beispiel 3: Baktosil–Drops sollen im Mittel 500mg Wirkstoff enthalten. Erfahrungsgemäß ist 1% der Ware Ausschuss, da der Wirkstoffgehalt zu gering ist. Wie hoch ist der Wirkstoffgehalt (in mg) an dieser Grenze?

gegeben: $\mu = 500$, $\sigma^2 = 25$, $w = P(X \leq x_w) = 0.01$
gesucht: x_w
Lösung: Es ist also derjenige Wert x_w gesucht,
unterhalb dessen 1 % an Wahrscheinlichkeitsmasse liegt.
Mit $u_{0.01} \simeq -2.33$
(diesen Wert findet man links neben dem Wert 0.01
der mit $F_U(-u)$ überschriebenen Spalte !)
errechnet man über
$u_{0.01} = \frac{x_{0.01} - \mu}{\sigma}$
den Wert
$-2.33 = \frac{x_{0.01} - 500}{5}$
$\Rightarrow x_{0.01} = x_w \simeq 488.35$

Der Wirkstoffgehalt beträgt 488.35mg.

Abschließend wird hier noch das vielfach interessierende **zentrale Schwankungsintervall** erwähnt. Darunter versteht man ein Intervall symmetrisch um μ derart, dass

$$P\left(x_{\alpha/2} \leq X \leq x_{1-\alpha/2}\right) = 1 - \alpha$$

Zur Berechnung kann natürlich wieder die standardisierte Normalverteilung herangezogen werden, denn

$$\begin{aligned} x_{1-\alpha/2} &= \mu + u_{1-\alpha/2} \cdot \sigma \qquad \textit{bzw.} \\ x_{\alpha/2} &= \mu + u_{\alpha/2} \cdot \sigma. \end{aligned}$$

Wegen der Symmetrie der Prozentpunkte $u_{\alpha/2} = -u_{1-\alpha/2}$ bedarf es nur des Ablesens von einem Prozentpunkt, und man schreibt für die Intervallgrenzen:

$$\mu \pm u_{1-\alpha/2} \cdot \sigma$$

Nimmt $u_{1-\alpha/2}$ den Wert 1, 2 bzw. 3 an, so spricht man vom ein-, zwei- bzw. dreifachen zentralen Schwankungsintervall, und es gilt für die Wahrscheinlichkeit α:

α	$u_{1-\alpha/2}$
0.0027	3
0.0455	2
0.3173	1

Nützlich sind weiterhin die folgenden Prozentpunkte:

α	$u_{1-\alpha/2}$
0.01	2.576
0.05	1.960
0.1	1.645

Bei Einführung der Normalverteilung wurde schon angesprochen, dass viele Merkmale in der Praxis annähernd normalverteilt bzw. logarithmisch normalverteilt sind. Diese Merkmale haben oft eine sehr große Variationsbreite. Viele sind stetig. Konkrete Wahrscheinlichkeitsverteilungen sind nicht bestimmbar. Um trotzdem Methoden der schließenden Statistik anwenden zu können, arbeitet man häufig mit der Normalverteilung, nachdem man überprüft hat, dass diese Verteilungsannahme mit den in einer Stichprobe beobachteten Werten verträglich ist. Einen ersten Eindruck vermitteln oft Abbildungen, außerdem gibt es mathematische Prüfverfahren.

Betrachtet man das Beispiel 4.2 (Schlachtgewicht von 100 Masthähnchen) aus dem Kapitel Datenaufbereitung, so erkennt man aus den Abbildungen 6.8 und 6.9, dass die Normalverteilungsannahme gerechtfertigt ist.

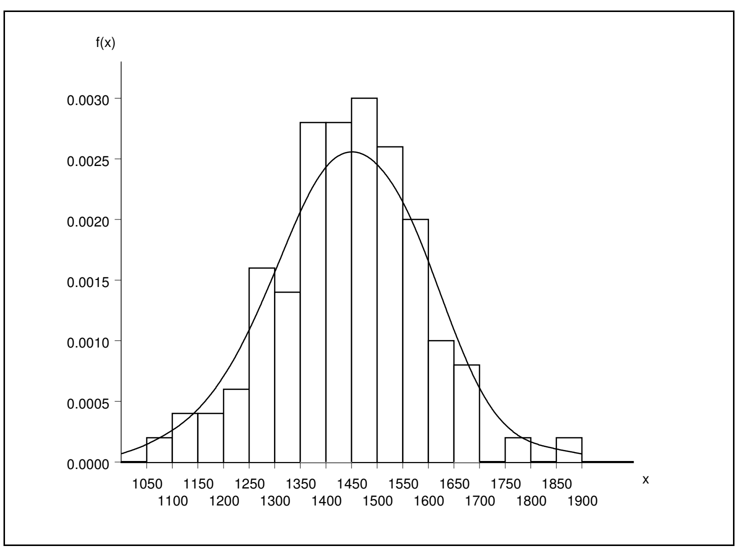


Abb. 6.8: Histogramm für die Schlachtgewichte von 100 Masthähnchen mit eingezeichneter Dichtefunktion der Normalverteilung

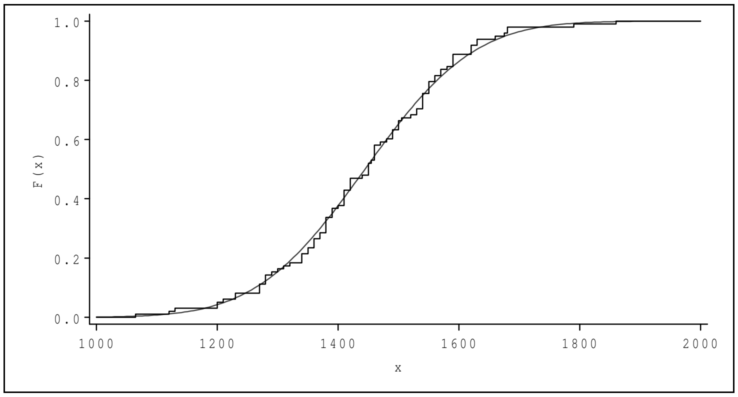


Abb. 6.9: Empirische Verteilungsfunktion der Schlachtgewichte von 100 Masthähnchen mit eingezeichneter Verteilungsfunktion der Normalverteilung

Kapitel 7

Statistische Schätzverfahren

7.1 Die Aufgaben der schließenden Statistik

Nach Meinung des großen Statistikers R. A. FISHER (1890–1962) ist es die Aufgabe des Statistikers, aufgrund empirischer Beobachtungen einen objektiven Induktionsschluss durchzuführen. Die empirischen Beobachtungen sind als Stichprobenwerte aufzufassen.

Mit Hilfe eines solchen Induktionsschlusses soll eine Aussage über die Grundgesamtheit oder deren Verteilung gewonnen werden, um aus der Situation der Ungewissheit oder der unvollständigen Information bezüglich der Grundgesamtheit soweit als möglich herauszukommen.

Bevor wir uns dem Schätzen von Parametern zuwenden, wollen wir uns im nächsten Abschnitt mit der Verteilung des Stichprobenmittels beschäftigen.

7.2 Verteilung des Stichprobenmittels und Standardfehler

Der Erwartungswert μ der Grundgesamtheit wird im allgemeinen durch das arithmetische Mittel $\overline{x}$ der Stichprobe vom Umfang n geschätzt.

Wir betrachten dazu die Schätzfunktion $\overline{X}$, die als arithmetisches Mittel der n Stichprobenvariablen X_i, $i = 1, \ldots, n$ definiert ist:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n).$$

Wie ist nun die Zufallsvariable $\overline{X}$ verteilt? Wenn die n Stichprobenvariablen X_i, $i = 1, \ \ldots \ , \ n$ alle die gleiche (μ, σ)–Normalverteilung besitzen, dann ist nach dem so genannten Additionstheorem der Normalverteilung $\overline{X}$ normalverteilt mit Erwartungswert μ und der Varianz σ^2/n bzw. der Standardabweichung $\sigma/\sqrt{n}$. Wir sagen auch, die Zufallsvariable $\overline{X}$ ist $(\mu, \sigma/\sqrt{n})$–normalverteilt:

$$\overline{X} \sim N\left(\mu_{\overline{X}} = \mu, \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\right).$$

Die Standardabweichung des Mittelwertes ($\sigma_{\overline{X}} = \sigma/\sqrt{n}$), die auch als **Standardfehler** des Mittelwertes bezeichnet wird, ist um den Faktor $1/\sqrt{n}$ kleiner als die Standardabweichung der Einzelbeobachtungen. Das ist einleuchtend, da die Mittelwerte einer Stichprobe weniger variieren als die Einzelbeobachtungen selbst. Außerdem verringert sich die Standardabweichung der Stichprobenmittelwerte mit wachsendem Stichprobenumfang.

7.3 Das Schätzen von Parametern

7.3.1 Punktschätzungen

Unter einer Punktschätzung versteht man die Angabe eines Näherungswertes für einen unbekannten Parameter einer Grundgesamtheit. Nehmen wir an, eine Grundgesamtheit oder Population sei bezüglich eines Merkmals normalverteilt. Die beiden Parameter Mittelwert μ und Varianz σ^2 seien unbekannt. Eine Schätzung gibt nun aufgrund einer Stichprobe x_1, x_2, ..., x_n **Schätzwerte** für diese unbekannten Parameter an.

Das Verfahren heißt **Punktschätzung**, weil ein bestimmter Punkt im Parameterraum ausgewählt wird. Eine Punktschätzung stellt also eine einzelne Zahl dar, die aus den Stichprobenwerten ermittelt wird.

Es liegt nahe, den Mittelwert $\overline{x}$ einer Stichprobe $x_1, x_2, \ldots, x_n$ als Punktschätzung für den unbekannten Erwartungswert μ der Grundgesamtheit zu wählen:

$$\mu \approx \overline{x}$$

Genauso kann man die empirische Varianz s^2 als Punktschätzung für die unbekannte Varianz σ^2 der Verteilung oder der Grundgesamtheit nehmen:

$$\sigma^2 \approx s^2$$

Oder wenn eine Stichprobe eine relative Häufigkeit $\frac{n_A}{n}$ für das Auftreten einer bestimmten Eigenschaft A ergibt, nehmen wir diesen Wert als Schätzwert für die unbekannte Wahrscheinlichkeit p eines binomialverteilten Merkmals in der Grundgesamtheit:

$$p \approx \frac{n_A}{n}$$

7.3.2 Vertrauensintervalle oder Intervallschätzungen

Das Stichprobenmittel $\overline{x}$ ist im Sinne der mathematischen Statistik eine unverzerrte Schätzung, die mit wachsendem Stichprobenumfang genauer wird und im Vergleich zu anderen Lokalisationsmaßen eine minimale Varianz aufweist. Wenn wir beispielsweise das arithmetische Mittel $\overline{x}$ einer Stichprobe als Schätzwert für den unbekannten Parameter μ einer Zufallsvariablen X verwenden, führen wir eine Punktschätzung durch, d.h. wir nehmen einen einzigen Punkt als Schätzwert für den Parameter. Es ist einleuchtend, dass der Schätzwert $\overline{x}$ nicht exakt mit dem unbekannten Mittelwert μ der Grundgesamtheit übereinstimmen wird. Es liegt daher die Frage nahe, wie gut solche Schätzwerte eigentlich sind. Zur Fehlerabschätzung gibt es zwei Möglichkeiten:

1. Zusätzlich zum Schätzwert gibt man die Standardabweichung des Schätzwertes an. Die Standardabweichung des Schätzwertes ergibt sich aus der Stichprobenverteilung des betreffenden Schätzers (siehe auch Abschnitt 7.2).

2. Eine etwas weitergehende Möglichkeit besteht darin, dass wir Intervalle angeben, die mit einer beliebig vorgebbaren Wahrscheinlichkeit $P = 1 - \alpha$ (z.B. $= 0.95$) den unbekannten Parameter überdecken.

Solche Intervalle werden **Vertrauensintervalle** oder **Konfidenzintervalle** genannt, die Vorgehensweise bezeichnet man als **Intervallschätzung**.

Ein solches Vertrauensintervall hängt vom Punktschätzwert, der Standardabweichung des Punktschätzers und der Vertrauenswahrscheinlichkeit ab.

Wenn wir z.B. ein Vertrauensintervall für den Erwartungswert μ betrachten, möchten wir wahrscheinlichkeitstheoretisch auf folgende Aussage hinaus:

$$P(A \leq \mu \leq B) = 1 - \alpha$$

Mit Hilfe der Stichprobenvariablen $X_1, X_2, \ldots, X_n$ sollen die beiden Grenzen A und B bestimmt werden. A und B sind als Funktionen der Stichprobenvariablen $X_1, X_2, \ldots, X_n$ anzusehen und damit selbst Zufallsvariable, also $A = A(X_1, X_2, \ldots, X_n)$ bzw. $B = B(X_1, X_2, \ldots, X_n)$. A und B heißen auch die **Vertrauensgrenzen**. Die Zahl $(1 - \alpha)$ heißt **Sicherheitswahrscheinlichkeit** oder **Vertrauenswahrscheinlichkeit** und wird bereits vor der Berechnung des Intervalls festgelegt und stellt somit eine objektive Wahrscheinlichkeit dar, die damit die Grundlage zum Vergleichen verschiedener Stichproben liefert.
$(1 - \alpha)$ ist, wie bereits gesagt, beliebig wählbar. Die Forderung $(1 - \alpha) = 1$ wäre zwar verständlich (wir wollen eben ganz genau Bescheid wissen!), aber nicht sinnvoll, weil dann unser Vertrauensintervall unendlich groß werden würde. In der Praxis übliche Werte für $(1 - \alpha)$ sind 0.90, 0.95 und 0.99.

7.4 Vertrauensintervall für den Erwartungswert einer Normalverteilung bei bekannter Varianz

Das arithmetische Mittel als Stichprobenfunktion $\overline{X}$ ist normalverteilt mit Erwartungswert μ und Standardabweichung $\sigma/\sqrt{n}$ (Standardfehler). Die standardisierte Zufallsvariable

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

ist normalverteilt mit Erwartungswert 0 und Standardabweichung 1, also standardnormalverteilt.

Damit können wir folgende Wahrscheinlichkeitsaussage aufstellen, wenn $u_{\alpha/2}$ und $u_{1-\alpha/2}$ die $(1 - \alpha)$–Grenzen der $(0, 1)$–Normalverteilung sind

$$P\left(u_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq u_{1-\alpha/2}\right) = (1 - \alpha).$$

Auflösen nach μ unter Berücksichtigung von $u_{\alpha/2} = -u_{1-\alpha/2}$ ergibt

$$P\left(\overline{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = (1 - \alpha).$$

Damit sind die Grenzen $A = \overline{X} - u_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ und $B = \overline{X} + u_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ in Abhängigkeit der Stichprobenvariablen $X_1,\ X_2,\ \ldots\ ,\ X_n$ ermittelt.

Haben wir als Realisation eine bestimmte Stichprobe $x_1,\ x_2,\ \ldots\ ,\ x_n$, dann setzen wir die Realisierung $\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$ in obige Formel ein und erhalten das $(1-\alpha)$–Vertrauensintervall:

$$\left(\overline{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{bis} \quad \overline{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Wenn wir für das theoretische $\overline{X}$ die Realisation $\overline{x}$, also die aus der konkreten Stichprobe ermittelte Zahl einsetzen, können wir die Wahrscheinlichkeitsaussage nicht mehr hinschreiben. Das zahlenmäßig jetzt bestimmte Intervall enthält μ oder enthält μ nicht. Welcher Fall vorliegt, wissen wir nicht, weil μ unbekannt ist.

Wir können aber von folgender Häufigkeitsinterpretation ausgehen:
Entnimmt man einer (μ, σ)–normalverteilten Grundgesamtheit unabhängig voneinander eine große Anzahl von Stichproben des Umfangs n und berechnet aus den Werten jeder dieser Stichproben das zugehörige Vertrauensintervall, dann überdecken erwartungsgemäß $(1-\alpha)\%$ dieser Intervalle den unbekannten Erwartungswert.

Beispiel: Masthähnchen
Beispiel 4.2 mit den 100 Masthähnchen bezieht sich auf eine Zufallsstichprobe aus einem großen Bestand. Die einzelnen Schlachtgewichte dieser Tiere wurden ermittelt und daraus $\overline{x} = 1446$g und $s = 140.8$g errechnet. Anhand dieser statistischen Maßzahlen soll auf das mittlere Schlachtgewicht des Bestandes (μ) geschlossen werden.

Im vorliegendem Beispiel ist die Standardabweichung σ im Bestand nicht bekannt und wird durch die Standardabweichung s aus der Stichprobe geschätzt. Der daraus resultierende Einfluss ist bei einem Umfang von $n = 100$ jedoch gering und wird hier vernachlässigt (exakt müsste anstelle der standardisierten Normalverteilung die so genannte standardisierte t–Verteilung für $(n-1) = 99$ Freiheitsgrade herangezogen werden).

Als Vertrauensintervall ergibt sich für $(1-\alpha) = 0.95$:

$$1446 \text{ g} - 1.96 \cdot \frac{140.8 \text{ g}}{\sqrt{100}} \quad \text{bis} \quad 1446 \text{ g} + 1.96 \cdot \frac{140.8 \text{ g}}{\sqrt{100}}$$

d.h. ab- bzw. aufgerundet

$$1418 \text{ g} \quad \text{bis} \quad 1474 \text{ g}$$

Welche Interpretation lässt nun dieses Vertrauensintervall zu ?

Würde man mehrere Stichproben durchführen und die Intervalle berechnen, so erwartet man, dass 95% dieser Intervalle den tatsächlichen Mittelwert dieses Bestandes enthalten. Daraus folgt die Aussage, dass das angegebene Intervall mit einer Chance von 0.95 zu den Intervallen gehört, die den tatsächlichen Mittelwert μ umschließen.

Für eine Vertrauenswahrscheinlichkeit von 0.99 errechnet sich ein breiteres Intervall von

$$1446 \text{ g} - 2.58 \cdot 14.08 \text{ g} \quad \text{bis} \quad 1446 \text{ g} + 2.58 \cdot 14.08 \text{ g}$$

$$\Rightarrow\ 1410 \text{ g} \quad \text{bis} \quad 1482 \text{ g}.$$

7.5 Stichprobenumfangs–Planung

Ziel einer zu planenden Stichprobenerhebung ist die Schätzung des Erwartungswertes μ einer Zufallsgröße X, für die unterstellt wird, dass sie normalverteilt ist. Als Schätzung soll ein Konfidenzintervall angegeben werden, dessen allgemeine Form in diesem Fall lautet:

$$\left[\overline{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad ; \quad \overline{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Außerdem sei angenommen, dass die Standardabweichung σ zu X bekannt ist. Unbekannt ist nur der mittlere Wert in der Grundgesamtheit.

Vorgegeben wird

1. die gewünschte Sicherheitswahrscheinlichkeit, d.h. die Wahrscheinlichkeit, mit der das aus den Stichprobenergebnissen zu berechnende Konfidenzintervall den unbekannten Erwartungswert μ überdecken soll $\Rightarrow\ 1-\alpha$,
2. die gewünschte Präzision der Schätzung, d.h. die Länge oder Breite, die das aus den Stichprobenergebnissen zu berechnende Konfidenzintervall höchstens haben soll
$$\Rightarrow\ L = 2 \cdot u_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Gesucht ist nun der dazu notwenige Stichprobenumfang n!

Die entsprechende Umstellung der Angabe für die Länge des Intervalls liefert hierfür die Formel:

$$n \geq \left(\frac{2 \cdot u_{1-\alpha/2} \cdot \sigma}{L}\right)^2$$

Beispiel: Masthähnchen
X seien die normalverteilten Gewichte von Masthähnchen nach 30 Tagen Mast. Von diesen wird angenommen, dass sie eine Standardabweichung von 120 g haben. Der Erwartungswert μ ist für eine bestimmte Herde aber unbekannt (vielleicht, weil man vermutet, dass sie aufgrund einer Erkrankung etwas zurückgeblieben sind) und soll mit Hilfe einer Stichprobenerhebung geschätzt werden.

Für die Schätzung eines Konfidenzintervalls wird zum einen vorgegeben, dass man zu 95% sicher sein will, den wahren Wert μ mit dem Intervall zu erwischen. Zum anderen soll das Intervall nicht breiter als 50 g werden.

Auf dieser Basis ist nun zu entscheiden, wie groß der Stichprobenumfang zufällig aus der Herde zu ziehender Tiere sein muss.

Aufgrund der gegebenen Informationen und der Anforderungen an die Schätzung hat man:

$\sigma = 120\text{g}$
$L \leq 50\text{g}$
$1-\alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05 \quad \Rightarrow \quad u_{1-\alpha/2} = 1.96$

Einsetzen in die Formel zur Berechnung von n liefert daher:

$$n \geq \left(\frac{2 \cdot 1.96 \cdot 120}{50}\right)^2 = 88.51$$

Es müssen also mindestens 89 Stichprobentiere gezogen werden.

Kapitel 8

Statistische Prüfverfahren

Neben dem Schätzen von Parametern (siehe Kapitel 7) ist es oft von zentralem Interesse zu überprüfen, ob bestimmte Vermutungen über einen Parameter oder eine Verteilung in der Grundgesamtheit zutreffen oder nicht. So könnte man bereits gewisse Vorstellungen über die Wirksamkeit eines Medikamentes oder über den Einfluss einer neuen Fütterungsart auf das Gewicht usw. haben.

Zwar wird die Vermutung hinsichtlich des interessierenden Merkmals in der Grundgesamtheit aufgestellt, überprüft wird sie wieder auf Stichprobenbasis. Die Überprüfung solcher Annahmen über das Verhalten des Untersuchungsmerkmals in der Grundgesamtheit fällt unter den Begriff des **statistischen Testens**. Die Regeln, die zur Überprüfung eingesetzt werden, heißen entsprechend **statistische Tests**. Um statistische Tests zur Beantwortung solcher Fragestellungen einsetzen zu können, müssen die entsprechenden Vermutungen operationalisiert und als statistisches Testproblem formuliert werden.

8.1 Formulierung von Hypothesen

Allgemein geht man dabei so vor, dass man eine Hypothese und eine Gegenhypothese formuliert. Dabei wird die **Nullhypothese** als H_0 bezeichnet, die **Gegen- oder Alternativhypothese** mit H_1. Dieses soll an zwei Beispielen verdeutlicht werden.

Beispiel 1: neues Präparat (Binomialverteilung)

Ein Pharmahersteller möchte ein neues Präparat gegen Parvovirose auf den Markt bringen. Dieses neue Präparat soll eine bessere Heilungswahrscheinlichkeit p haben als das Standardpräparat mit der Heilungswahrscheinlichkeit $p_0 = 0.7$. Um dieser Behauptung kritisch auf den Grund zu gehen, würde man für einen statistischen Test folgende Hypothese formulieren:

$$H_0 : p \leq p_0 = 0.7 \qquad \text{versus} \qquad H_1 : p > p_0 = 0.7.$$

Es handelt sich hier um eine einseitige Fragestellung. Das Verfahren wird in Abschnitt 8.5 vorgestellt.

Beispiel 2: Masthähnchen (Normalverteilung)

Der Leiter eines Schlachthofes hat den Verdacht, dass ein Mäster Masthähnchen anliefert, deren mittleres Mastendgewicht nach 34 Tagen nicht dem Sollgewicht von 1475 g entspricht. Diese Arbeitshypothese formuliert er als H_1. Der Mäster bleibt bei der Hypothese, dass seine

Hähnchen im Mittel das vorgeschriebene Sollgewicht aufweisen. Um der Sache auf den Grund zu gehen, werden folgende Hypothesen formuliert:

$$H_0 : \mu = \mu_0 = 1475\ g \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0 = 1475 g.$$

Es handelt sich hier um eine zweiseitige Fragestellung. H_0 ist in diesem Beispiel eine einfache Hypothese, H_1 eine zusammengesetzte Hypothese.

Bei den vorgestellten Tests handelt es sich um so genannte Parametertests, die Hypothesen über die Parameter einer Verteilung überprüfen, z.B. über den Erwartungswert μ einer Normalverteilung (Beispiel 2) oder den Anteil p einer Binomialverteilung (Beispiel 1). Ein drittes Beispiel zum Test auf Unabhängigkeit zwischen zwei Merkmalen wird in Kapitel 9 vorgestellt. Daneben gibt es so genannte Anpassungstests, die überprüfen, ob es sich bei den beobachteten Werten um Realisationen z.B. normalverteilter Größen handelt. Auf diese Testverfahren soll hier nicht weiter eingegangen werden.

In der Regel wird H_1 die ursprüngliche Arbeitshypothese sein. Man vermutet aus irgendwelchen Anhaltspunkten, dass Unterschiede bestehen und will diese überprüfen. H_0 ist der Standpunkt des Kritikers.

Der statistische Test entscheidet aufgrund einer Stichprobe zwischen der Nullhypothese H_0 und der Alternativhypothese H_1; dies geschieht mit Hilfe einer so genannten Stichprobenfunktion $T = T(x_1, \ldots, x_n)$, die auch **Prüfgröße**, Testgröße oder Teststatistik genannt wird. Für den Erwartungswert μ einer Normalverteilung bietet sich das arithmetische Mittel $\overline{x}$ als Prüfgröße an, beim Testen über den Parameter p einer Binomialverteilung verwendet man die relative Häufigkeit in der Stichprobe.

Ist die Abweichung nicht allzu groß, wird man die Nullhypothese beibehalten. Wenn aber die Abweichung zu groß ist und damit nicht mehr durch reinen Zufall erklärt werden kann, lehnt man die Nullhypothese ab.

Diese vage Entscheidungsvorschrift soll nun präzisiert werden, d.h. es ist konkret festzulegen, was ‘allzu groß’ bedeuten soll. Es ist zu definieren, ab welcher kritischen Abweichung c von dem in der Nullhypothese postulierten Wert diese abgelehnt werden soll. Wir verdeutlichen das Vorgehen am “Masthähnchen–Beispiel”.

Als Prüfgröße verwendet man das arithmetische Mittel $\overline{x}$ der Stichprobe. Die Grafik 8.1 zeigt die Dichtefunktion für die Zufallsvariable $\overline{X}$ bei gültiger Nullhypothese.

Man sucht also zwei kritische Werte c_1 und c_2 (bei zweiseitigen Hypothesen), für die gilt, dass jenseits dieser Grenzen die Abweichungen zwischen $\overline{x}$ und μ_0 in definierter Weise zu groß sind. Diese kritischen Werte werden so bestimmt, dass bei gültiger Nullhypothese die Prüfgröße (im “Masthähnchen–Beispiel” also $\overline{x}$) nur mit einer vorgegebenen, üblicherweise geringen Wahrscheinlichkeit in diesen Ablehnungsbereich fällt. Diese vorgegebene Wahrscheinlichkeit wollen wir mit α bezeichnen, häufig wählt man $\alpha = 5\%$, 1% oder 10%. Bei einem α von 5% riskiert man in etwa einem von 20 Fällen, dass man die Nullhypothese verwirft, obwohl sie richtig ist. Dies bedeutet, dass man eine Fehlentscheidung trifft. Darauf werden wir im nächsten Abschnitt noch näher eingehen. In Abbildung 8.1 teilt sich der Ablehnbereich auf in den Bereich links von c_1 und rechts von c_2, da man eine zweiseitige Fragestellung hat. Die Wahrscheinlichkeit α verteilt sich zu gleichen Teilen auf diese beiden Bereiche, sodass in jedem der beiden Ablehnbereiche $\alpha/2$ der Fläche unter der Dichtefunktion liegen. Formal kann man diese Forderung folgendermaßen schreiben:

$$P_{H_0}(\overline{X} < c_1 \text{ oder } \overline{X} > c_2) \leq \alpha$$

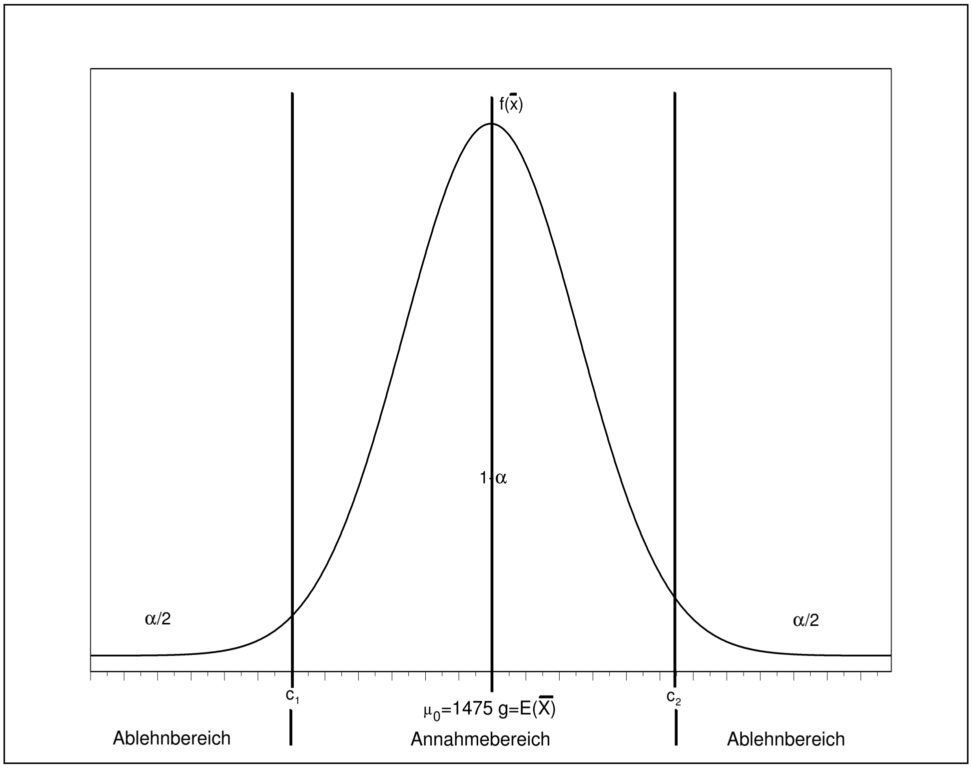


Abb. 8.1: Dichtefunktion der Normalverteilung bei Gültigkeit der Nullhypothese (“Masthähnchen–Beispiel”)

Der Index H_0 soll verdeutlichen, dass man hier die Wahrscheinlichkeitsverteilung bei Vorliegen von H_0 betrachtet.

Liegt das Stichprobenmittel $\overline{x}$ unterhalb von c_1 oder oberhalb von c_2, würde man H_0 ablehnen und die Alternativhypothese H_1 annehmen. Liegt $\overline{x}$ im Intervall zwischen c_1 und c_2, würde man H_0 beibehalten.

Wenn die Wahrscheinlichkeitsverteilung der Prüfgröße bekannt ist, kann man bei vorgegebenem α die kritischen Werte berechnen. Für das Masthähnchen–Beispiel gilt, dass $\overline{X}$ normalverteilt ist mit den Parametern μ und $\sigma/\sqrt{n}$:

$$\overline{X} \sim N(\mu,\, \sigma/\sqrt{n})$$

Das genaue Vorgehen wird in Abschnitt 8.4 erläutert.

Im vorliegenden Beispiel beträgt $\overline{x} = 1446\ g$ (bei einer bekannten Standardabweichung von $\sigma = 120\ g$ und $n = 100$). Die kritischen Werte berechnen sich zu

$$c_1 = 1451.48\ g \qquad \text{und} \qquad c_2 = 1498.52\ g$$

Der Schlachthof würde also die Lieferung ablehnen, da $\overline{x} < 1451.48\ g$ ist und man daher nicht glauben mag, dass die Herde das geforderte Sollgewicht aufweist.

8.2 Fehlentscheidungen beim Testen

Da beim statistischen Testen nur aufgrund einer Stichprobe auf die Grundgesamtheit geschlossen wird, lassen sich Fehlentscheidungen leider nicht immer vermeiden. Dabei gibt es zwei Arten, Fehler zu machen:

1. Der Test führt zur Ablehnung der Hypothese H_0, obwohl diese objektiv richtig ist. Diese falsche Entscheidung heißt **Fehler 1. Art**. Die Wahrscheinlichkeit für diesen Fehler ist die 'Irrtumswahrscheinlichkeit α'. Man lehnt die Nullhypothese ab (d.h. man entscheidet sich für die Alternative), obwohl die Nullhypothese zutrifft. α wird auch als Signifikanzniveau bezeichnet.

2. Führt der Test **nicht** zur Ablehnung der Hypothese H_0, obwohl H_1 tatsächlich richtig ist, begeht man einen so genannten **Fehler 2. Art**. Die Wahrscheinlichkeit für den Fehler 2. Art wird mit β bezeichnet.

Leider ist es nicht möglich, beide Fehlerwahrscheinlichkeiten gleichzeitig zu minimieren. Deswegen legt man üblicherweise α fest und versucht mit einem genügend großen Stichprobenumfang und Auswahl eines geeigneten Testverfahrens β so klein wie möglich zu halten. β ist i.a. eine unbekannte Größe, sodass die Entscheidung für die Nullhypothese unter Umständen mit hoher Wahrscheinlichkeit falsch sein kann. Aus diesem Grund formuliert man dann das Ergebnis des Tests vorsichtigerweise als 'Ich lehne H_0 nicht ab' oder 'H_0 kann nicht verworfen werden'.

Eine Übersicht über die Fehlerarten gibt das Schema in Tabelle 8.1.

	Wirklichkeit	
Entscheidung	H_0	H_1
für H_0	richtig $1-\alpha$	falsch β
gegen H_0	falsch α	richtig $1-\beta$

Tab. 8.1: Schema zu Entscheidungsfehlern beim statistischen Test

8.3 Allgemeiner Ablauf eines statistischen Tests

Im letzten Abschnitt wurde versucht, die Grundidee des statistischen Testens an einem Beispiel zu verdeutlichen. Zunächst wurden die Hypothesen formuliert, anschließend die Prüfgröße und deren Verteilung besprochen, woraus sich die kritischen Werte berechnen ließen. Als letztes wurden die beiden Möglichkeiten besprochen, Fehlentscheidungen zu begehen. Dem Ganzen liegt ein generelles Schema zugrunde, das auf andere Testsituationen übertragen werden kann. Es soll im folgenden kurz vorgestellt werden. In den beiden nächsten Unterkapiteln schließen sich zwei Testverfahren an, die anhand dieses **6-Punkte–Schemas** erläutert werden.

1. Formulierung der Nullhypothese H_0 und der Alternativhypothese H_1
   - einseitige oder zweiseitige Fragestellung
   - einfache oder zusammengesetzte Hypothese
   - bei einseitigen Hypothesen wird ein kritischer Wert c berechnet
   - bei zweiseitigen Hypothesen berechnet man zwei kritische Werte c_1 und c_2

2. Festlegung des Signifikanzniveaus α (Wahrscheinlichkeit für den Fehler 1. Art)

   - das Signifikanzniveau wird **vor** Durchführung des Tests festgelegt
     z. B. als $\alpha = 0.05$.

3. Festlegung des Stichprobenumfangs n

4. Formulierung der Prüfstrategie auf der Grundlage der unterstellten Wahrscheinlichkeitsverteilung

   - Auswahl einer Prüfgröße mit bekannter Wahrscheinlichkeitsverteilung
   - mit Hilfe des vorgegebenen α und der Wahrscheinlichkeitsverteilung unter H_0 können die kritischen Werte berechnet werden

5. Durchführung des Prüfverfahrens und Formulierung des Ergebnisses

   - Ziehen der Stichprobe
   - Berechnen der Prüfgröße
   - Vergleich mit den kritischen Werten
   - Ablehnung von H_0 oder Beibehaltung von H_0

6. Sachliche Interpretation

   - welche Konsequenzen werden gezogen?
   - keine Ablehnung von H_0 $\Rightarrow$ die bisherige Situation wird akzeptiert
   - Ablehnung von H_0 $\Rightarrow$ der Alternativhypothese H_1 wird Glauben geschenkt; diese Entscheidung ist mit einer Wahrscheinlichkeit von höchstens α falsch.

8.4 Test für den Erwartungswert einer Normalverteilung bei bekannter Standardabweichung (Gaußtest)

Dieser Test wurde schon in der allgemeinen Herleitung anhand des "Masthähnchen–Beispiels" eingeführt.

1. Man kann sich verschiedene Fragestellungen vorstellen, einseitig oder zweiseitig. Hier behandeln wir nur das zweiseitige Testproblem.

   $$H_0: \quad \mu = \mu_0 \qquad \text{versus} \qquad H_1: \quad \mu \neq \mu_0$$

2. Das Signifikanzniveau wird mit $\alpha = 0.05$ festgelegt.

3. Der Stichprobenumfang ist n.

4. Es liegt eine Normalverteilung mit bekannter Standardabweichung σ vor. Für die Zufallsvariable X gilt:

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

   Als Prüfgröße wird $\overline{X}$ ausgewählt. Unter H_0 gilt:

   $$\overline{X} \overset{H_0}{\sim} N(\mu_0, \sigma/\sqrt{n})$$

Es ist günstiger, anstelle von $\overline{X}$ die standardisierte Variante als Prüfgröße zu verwenden, da hier die Werte wieder vertafelt sind (siehe Anhang B):

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0, 1).$$

Für die transformierte Prüfgröße kann man jetzt schreiben:

$$P_{H_0}\left(\left|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right| > c\right) \leq \alpha$$

Unter H_0 soll die Wahrscheinlichkeit des Ablehnens höchstens α betragen. Diese Prüfgröße ist unter H_0 standardnormalverteilt. Basierend auf der Prüfgröße $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ fällt die Entscheidung für H_1 im zweiseitigen Testproblem, falls ihr Betrag $|u| > u_{1-\alpha/2}$.

Daraus ergibt sich die allgemeine Prüfstrategie:

- H_0 ablehnen, falls $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < c_1$ oder $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} > c_2$
- H_0 nicht ablehnen, falls $c_1 \leq \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \leq c_2$.

5. Für das Masthähnchen–Beispiel ergibt sich als transformierte Prüfgröße (bei $n = 100$ und $\sigma = 120\ g$ und $\alpha = 5\%$)

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = -2.417$$

   Damit ist die transformierte Prüfgröße kleiner als $c_1 = -1.96$ und man lehnt H_0 ab.

6. Aufgrund der Stichprobe wird H_0 abgelehnt, der Schlachtbetrieb kann die Lieferung des Mästers beanstanden.

8.5 Binomialtest

Die Heilungsquote von Patienten gleichen Krankheitsgrades beträgt bei Behandlung mit einem Standardpräparat $p = 0.70$. Ein neues Präparat ist entwickelt worden, von dem behauptet wird, dass es eine bessere Heilungsrate erziele. Es ist zu entscheiden, ob in Zukunft das neue Präparat anzuwenden sei.

1. Es soll untersucht werden, ob das neue Präparat besser als das Standardpräparat ist, wobei die Möglichkeit, dass es schlechter sei, außer Acht gelassen wird.
   Unter dieser Voraussetzung werden die Hypothesen für eine einseitige Prüfung wie folgt formuliert:

$$\begin{array}{lcllcl} H_0 & : & p & \leq & p_0 & = & 0.7 \\ H_1 & : & p & > & p_0 & = & 0.7 \end{array}$$

   Die Heilungsrate des Standardpräparates ist als 70% bekannt,
   p bedeutet die unbekannte Heilungsrate des neuen Präparates.

2. Die Wahrscheinlichkeit für den Fehler 1. Art wird mit $\alpha = 0.05$ festgelegt.

3. Aus untersuchungstechnischen Gründen muss der Stichprobenumfang auf $n = 20$ beschränkt werden.

4. Als Wahrscheinlichkeitsverteilung wird die Binomialverteilung zugrunde gelegt.
   Die Wahrscheinlichkeiten, bei n Patienten x Heilungserfolge zu erzielen, berechnen sich nach der Formel
   $$P(X = x \mid p,\ n) \quad = \quad \binom{n}{x} \quad p^x(1-p)^{n-x}$$
   (vgl. Abschnitt 6.7.2).

   Die Wahrscheinlichkeiten für die spezielle Binomialverteilung mit $n = 20$ unter Annahme von $p = p_0 = .70$ sind der Tabelle 8.2 zu entnehmen:

x	$P(X = x) \quad = \quad \binom{20}{x} \quad (.70)^x(.30)^{20-x}$	$P(X \leq x)$	$P(X > x)$
$0-6$	.000	.000	1.000
7	.001	.001	0.999
8	.004	.005	0.995
9	.012	.017	0.983
10	.031	.048	0.952
11	.065	.113	0.887
12	.114	.227	0.773
13	.164	.391	0.609
14	.192	.583	0.417
15	.179	.762	0.238
16	.130	.892	0.108
17	.072	.964	0.036
18	.028	.992	0.008
19	.007	.999	0.001
20	.001	1.000	0.000

   Tab. 8.2: Binomialverteilung mit $n = 20$ und $p = 0.70$

   Die Wahrscheinlichkeit, bei Gültigkeit der Nullhypothese diese abzulehnen, soll höchstens $\alpha = 0.05$ betragen. Dies kann man folgendermaßen formulieren:
   $$P_{H_0}(X > c) \leq \alpha$$
   Aus der 4. Spalte von Tabelle 8.2 kann man ablesen, dass diese Forderung gerade noch für $c = 17$ erfüllt ist, da
   $$P(X > x) = P(X > c) = P[X > 17) = 0.036 < \alpha = 0.05$$
   Der kritische Wert ist demnach $c = 17$ und es ergibt sich die folgende Prüfstrategie:

   - Werden von 20 Patienten 18 oder mehr geheilt, so verwerfe man die Prüfhypothese H_0 und nehme die Alternativhypothese H_1 als richtig an.
   - Werden 17 oder weniger Heilerfolge erzielt, so verwerfe man die Prüfhypothese H_0 **nicht**.

5. Als Grundgesamtheit werden alle Patienten mit dem gleichen Krankheitsgrad angesehen, deren genaue Zahl unbekannt ist. Von den erfassbaren Patienten werden 20 zufällig ausgewählt. Diese Auswahl wird unter Praxisbedingungen häufig nicht zufällig, sondern willkürlich erfolgen, z.B. werden die ersten 20 Patienten mit diesem Krankheitsgrad in die Untersuchung aufgenommen und mit dem neuen Präparat behandelt. Nach Feststellung des Heilungserfolges jedes einzelnen Patienten ist wie in Punkt 4 angegeben zu verfahren.

6. Wird die Prüfhypothese H_0 verworfen, so sollte in Zukunft das neue Präparat als Standardpräparat Anwendung finden.
   Im anderen Falle kann das nicht geschehen, weil die Wahrscheinlichkeit einer fehlerhaften Entscheidung größer als die vorgegebene Irrtumswahrscheinlichkeit ist. Das bedeutet jedoch nicht, dass die Heilungsrate beider Präparate gleich ist.

Bei der Annahme der Gültigkeit der Prüfhypothese, d.h. bei Verwerfen der Alternativhypothese, besteht ebenfalls die Möglichkeit, eine Fehlentscheidung zu treffen.

Die Wahrscheinlichkeit β, diesen so genannten **Fehler 2. Art** zu machen, ist jedoch nur abzuschätzen, wenn eine konkrete Alternativhypothese fixiert wird, die in dem besprochenen Prüfverfahren einen höheren Wert als 70% postulieren muss.

Als Beispiel diene die Festlegung der Alternativhypothese auf $p = p_1 = 0.9$. Die entsprechende Wahrscheinlichkeitsverteilung ist in Tabelle 8.3 dargestellt.

x	$P(X = x) = \binom{20}{x} (.90)^x (.10)^{20-x}$	$P(X \leq x$	$P(X > x)$
$0 - 12$	.000	.000	1.000
13	.002	.002	0.998
14	.009	.011	0.989
15	.032	.043	0.957
16	.090	.133	0.867
17	.190	.323	0.677
18	.285	.608	0.392
19	.270	.878	0.122
20	.122	1.000	0.000

Tab. 8.3: Binomialverteilung mit $n = 20$ und $p = 0.9$

Daraus ergibt sich bei der Anwendung der unter Punkt 4 angegebenen Strategie eine Wahrscheinlichkeit β für einen Fehler 2. Art von 0.323.

Wird aus der Kenntnis dieser hohen Fehlerwahrscheinlichkeit die Prüfstrategie in der Weise geändert, dass bei 0 bis einschließlich 16 Heilungen bei 20 Patienten die Prüfhypothese beibehalten wird und in den Fällen einer Heilung von 17, 18, 19 oder 20 Patienten die Prüfhypothese verworfen wird, bedeutet dies eine Verringerung der Wahrscheinlichkeit für einen Fehler 2. Art β auf 0.133 bei einer Erhöhung der Wahrscheinlichkeit für einen Fehler 1. Art α auf 0.108 (vgl. Abbildung 8.2).

Um neben der Wahrscheinlichkeit für einen Fehler 1. Art von $\alpha \leq 0.05$ auch für den Fehler 2. Art eine Wahrscheinlichkeit von $\beta \leq 0.05$ einhalten zu können, wäre ein Stichprobenumfang von $n = 45$ erforderlich (vgl. Abbildung 8.3).

Die Prüfstrategie würde hier wie folgt lauten:

- Behalte H_0 bei, wenn $X < 37$ Heilungen bei $n = 45$ Behandlungen registriert werden.
- Lehne H_0 $(p \leq p_0 = 0.7)$ zugunsten von H_1 $(p \geq p_1 = 0.9)$ ab, wenn $X \geq 37$ Heilungen auftreten.

Die dabei maximal auftretenden Wahrscheinlichkeiten für einen Fehler 1. Art unter H_0 und einen Fehler 2. Art unter H_1 belaufen sich auf $\alpha = 0.047$ bzw. $\beta = 0.032$.

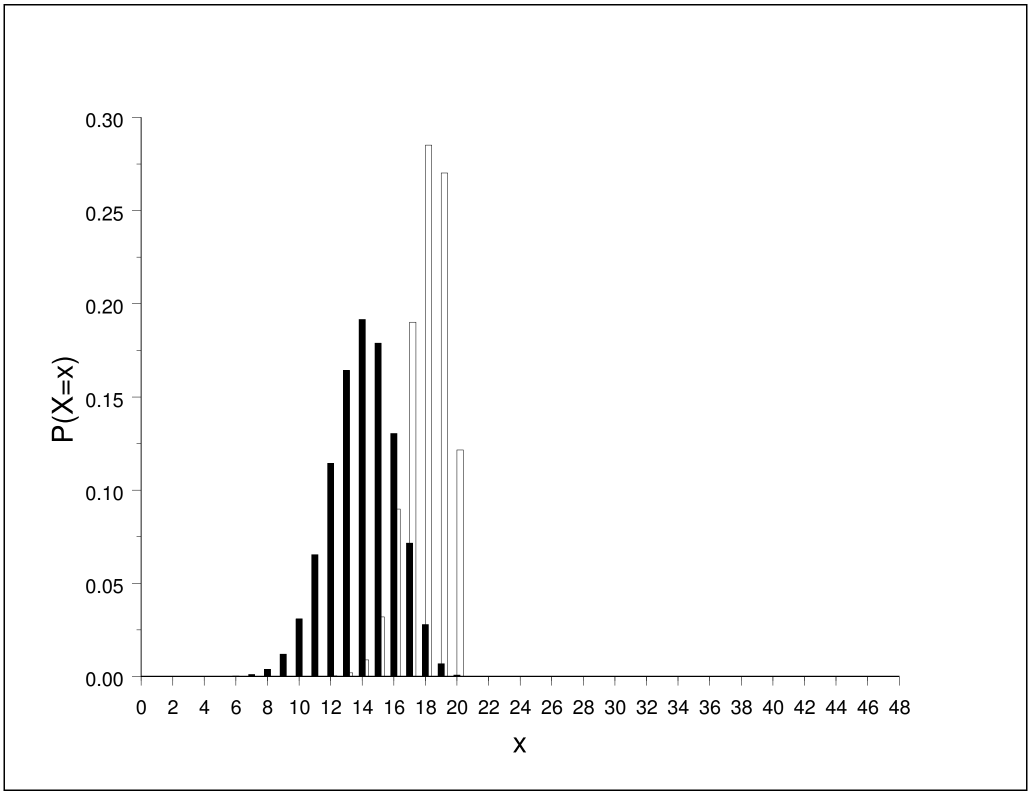


Abb. 8.2: Darstellung der Binomialverteilungen zum Test der Hypothesen $H_0 : p \leq p_0 = 0.7$ (schwarz) versus $H_1 : p \geq p_1 = 0.9$ (weiß) für $n = 20$

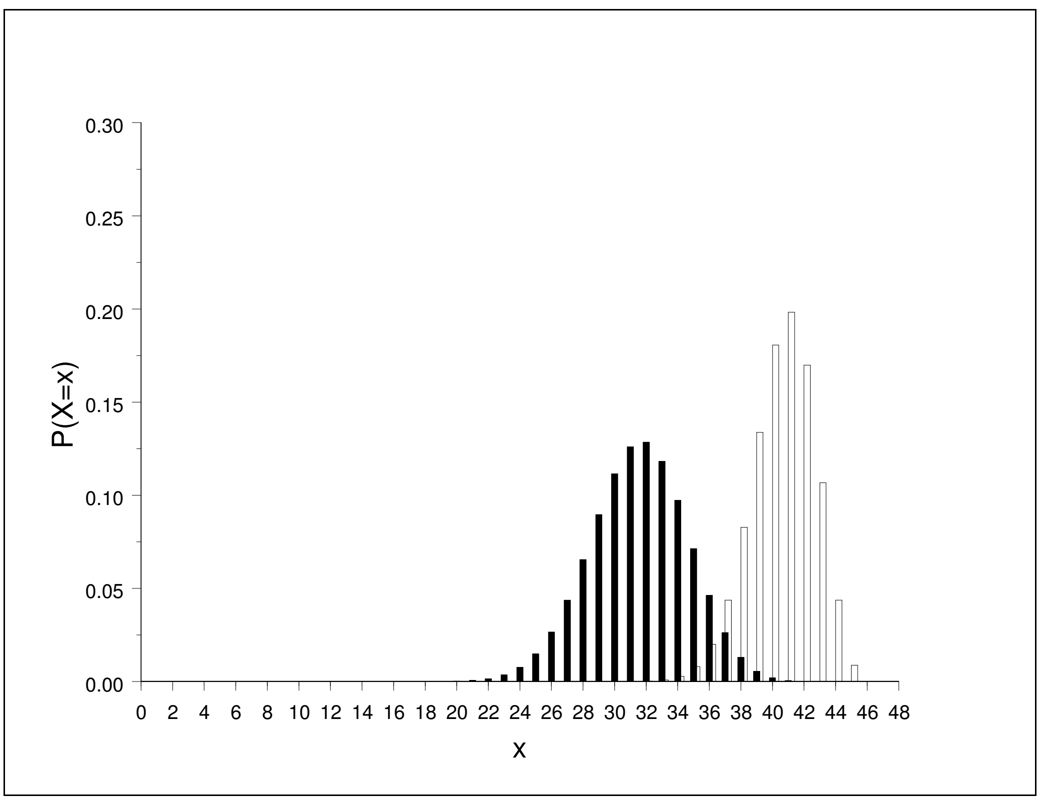


Abb. 8.3: Darstellung der Binomialverteilungen zum Test der Hypothesen $H_0 : p \leq p_0 = 0.7$ (schwarz) versus $H_1 : p \geq p_1 = 0.9$ (weiß) für $n = 45$

Kapitel 9

Der Zusammenhang zwischen zwei Merkmalen

Bis jetzt haben wir uns mit statistischen Techniken beschäftigt, die auf ein einzelnes Merkmal — in der Regel mit X bezeichnet — zugeschnitten waren. Wir haben gesehen, wie wir die Stichprobenergebnisse zusammenfassen und als Histogramm, Häufigkeitstabelle oder mit Hilfe von Lokalisations- und Streuungsmaßen darstellen können.

Diese bisher behandelten Techniken gehören zu den **univariaten Methoden**.

Im folgenden wollen wir uns den **bivariaten Methoden** zuwenden. Diese werden benötigt, wenn zwei Merkmale gemeinsam betrachtet werden. Wir bezeichnen diese Merkmale für gewöhnlich mit X und Y.

Man will oft wissen, ob zwei Größen zueinander in Beziehung stehen und, falls ja, um was für eine Beziehung es sich handelt und wie stark sie ist.

- Stehen Rauchen (X) und Blutdruck (Y) zueinander in Beziehung ?
- Hängen Sonnenflecken (X) und der Welt–Weizenertrag (Y) voneinander ab ?
- Besteht eine Assoziation zwischen dem Kreditvolumen (X) und der Inflationsrate (Y) im folgenden Jahr ?
- Lassen sich die Physikumsnoten (Y) aus den Abiturergebnissen (X) vorhersagen?

Offensichtlich kann keine dieser Fragen zufriedenstellend beantwortet werden, wenn wir jedes der Merkmale X und Y für sich allein analysieren. Beide müssen simultan betrachtet werden, um herauszufinden, wie sie zueinander in Beziehung stehen, ob tatsächlich ein rechnerischer, nicht immer auch kausaler Zusammenhang besteht.

Wie bei der Angabe von Lokalisations- und Streuungsmaßen müssen wir dabei die Art und die Skala unserer Merkmale berücksichtigen. Das heißt, es müssen z.T. unterschiedliche Methoden angewendet werden, um z.B. den Zusammenhang zwischen Haar- und Augenfarbe oder zwischen Körpergröße und Gewicht zu untersuchen.

9.1 Methoden für qualitative Merkmale

9.1.1 Die 2×2 Felder–Tafel / χ^2-Test

Wir nehmen an, dass wir die Wirkung eines Impfstoffes an einer Gruppe von 200 Katzen untersuchen wollen. Dazu wird 100 zufällig ausgewählten Katzen der Impfstoff gegeben, während die restlichen Katzen nicht geimpft werden.

Nach drei Monaten wird überprüft, wie viele Katzen der beiden Gruppen inzwischen erkrankt waren. Die Ergebnisse stellen wir in einer Kontingenztafel dar, wobei wir jeweils in einer Zeile die Zahl der erkrankten und der nicht erkrankten Katzen einer Gruppe einordnen.

	Ergebnis (Y)		
Behandlung (X)	erkrankt	nicht erkrankt	Summe
geimpft	10	90	100
nicht geimpft	34	66	100
Summe	44	156	200

Tab. 9.1: Fall A: Zahl erkrankter Katzen mit und ohne Impfung

Es interessiert natürlich, ob sich der Impfstoff bewährt, d.h. ob der Anteil an infizierten Katzen durch die Impfung gegenüber der Kontrolle vermindert ist.

Hätten wir als Ergebnis eine der folgenden 2×2 Felder–Tafeln erhalten, fiele die Entscheidung sicher leicht:

	Y		
X	ja	nein	Summe
ja	22	78	100
nein	22	78	100
Summe	44	156	200

Tab. 9.2: Fall B: kein Zusammenhang

	Y		
X	ja	nein	Summe
ja	0	100	100
nein	44	56	100
Summe	44	156	200

Tab. 9.3: Fall C: totaler Zusammenhang

Im Fall B ist kein Unterschied erkennbar, während im Fall C ein eindeutiger Zusammenhang zwischen Behandlung (X) und Erkrankung (Y) sichtbar wird. Wie sollen wir nun im Falle A entscheiden ?

Wenn wir wissen wollen, ob ein Unterschied zwischen den Gruppen, anders ausgedrückt, ob ein Zusammenhang zwischen X und Y besteht, müssen wir feststellen, wie stark die Ergebnisse von denen abweichen, die wir bei identischen Erkrankungsraten in den beiden Gruppen erwartet hätten.

Die Besetzungszahl bei gleicher Erkrankungsrate finden wir in Tabelle B. Wir berechnen sie, indem wir das Produkt aus mittlerer Erkrankungsrate (44/200) und dem Gruppenumfang (jeweils 100) bilden.

$$22 = \frac{44}{200} \cdot 100 \quad ; \quad 78 = \frac{156}{200} \cdot 100$$

Um die beobachteten mit den erwarteten Häufigkeiten zu vergleichen, berechnen wir für alle 4 Felder den Wert des Ausdruckes

$$\frac{(\text{beobachtete Häufigkeit} - \text{erwartete Häufigkeit})^2}{\text{erwartete Häufigkeit}}$$

und summieren diese 4 Werte auf. Die Summe wird als χ^2 **-Wert** bezeichnet (Sprich: chi-quadrat). In unserem Fall ergibt sich für

Tabelle A:

$$\begin{aligned} \chi^2 &= \tfrac{(10-22)^2}{22} + \tfrac{(90-78)^2}{78} + \tfrac{(34-22)^2}{22} + \tfrac{(66-78)^2}{78} \\ &= 6.545 + 1.846 + 6.545 + 1.846 \\ &= 16.782 \end{aligned}$$

Tabelle B:

$$\begin{aligned}\chi^2 &= \tfrac{(22-22)^2}{22} + \tfrac{(78-78)^2}{78} + \cdots \\ &= 0\end{aligned}$$

Tabelle C:

$$\begin{aligned}\chi^2 &= \tfrac{(0-22)^2}{22} + \tfrac{(100-78)^2}{78} + \tfrac{(44-22)^2}{22} + \tfrac{(56-78)^2}{78} \\ &= 22 + 6.205 + 22 + 6.205 \\ &= 56.41\end{aligned}$$

Wir sehen, dass ein $\chi^2 = 0$ der völligen Übereinstimmung von beobachteten und erwarteten Häufigkeiten entspricht, während sich im Extremfall totaler Abhängigkeit (Tabelle C) ein χ^2 von 56.41 ergibt.

Die Zahlen von Tabelle A ergeben ein χ^2 von 16.782, was auf einen gewissen Zusammenhang hinweist.

Grundsätzlich kann man sich bei der Vier–Felder–Tafel die Berechnung von χ^2 vereinfachen. Dazu bezeichnen wir die Besetzungszahlen unserer 2×2 Felder–Tafel wie folgt

	Y		
X	+	-	Summe
+	a	b	$a+b$
-	c	d	$c+d$
Summe	$a+c$	$b+d$	$n\,(=a+b+c+d)$

Es gilt die folgende Formel:

$$\chi^2 = \frac{(ad-bc)^2 \cdot n}{(a+c)(b+d)(a+b)(c+d)}$$

Da die Herleitung dieser Formel nur langweilige Algebra ist, verzichten wir darauf. Sie liefert aber genau dieselben Werte wie die Summen-χ^2-Formel

Im Falle A erhalten wir so:

$$\begin{aligned}\chi^2 &= \frac{((10\cdot 66)-(34\cdot 90))^2 \cdot 200}{44 \cdot 156 \cdot 100 \cdot 100} \\ &= 16.782\end{aligned}$$

Zur genaueren Beurteilung benötigen wir Methoden der schließenden Statistik. Entsprechend der in Kapitel 8 erläuterten Testidee würde zur Beurteilung eines Zusammenhanges beider Merkmale in der Grundgesamtheit wie folgt vorgegangen werden:

1. Nullhypothese (H_0): Es **besteht kein Zusammenhang** zwischen Impfung und Erkrankung, d.h. die Wahrscheinlichkeiten, dass geimpfte bzw. nicht geimpfte Katzen erkranken, sind in der Grundgesamtheit gleich (Situation in der Grundgesamtheit analog der Tabelle B).
   Alternativ- bzw. Gegenhypothese (H_1): Es **besteht ein Zusammenhang** zwischen Impfung und Erkrankung, d.h. die Wahrscheinlichkeiten, dass geimpfte bzw. nicht geimpfte Katzen erkranken, sind in der Grundgesamtheit ungleich (zweiseitige Testsituation; Impfung kann besser, aber auch schlechter sein).

2. Die Wahrscheinlichkeit für den Fehler 1. Art (Irrtumswahrscheinlichkeit α) wird mit $\alpha = 0.05$ festgelegt.

3. Der Test ist mit einer Zufallsstichprobe von $n = 200$ Katzen durchgeführt worden, wobei 100 zufällig ausgewählten Katzen der Impfstoff verabreicht worden ist.

4. Als konkrete Teststatistik wird aus der realisierten Stichprobe (vgl. Tabelle A) ein χ^2-Wert von 16.782 berechnet.

5. Bei Richtigkeit der Nullhypothese folgt die zu berechnende Teststatistik einer χ^2-Verteilung mit einem Freiheitsgrad.
   Unter Berücksichtigung von $\alpha = 0.05$ ist dann im zweiseitigen Test die Nullhypothese abzulehnen, wenn die konkrete Teststatistik den kritischen Wert $\chi^2_{1;\,0.05} = 3.84$ überschreitet, wobei das χ^2 mit der Anzahl der Freiheitsgrade und dem Signifikanzniveau α indiziert wird.

   Tabelle 9.4 zeigt einige kritische Werte für verschiedene Signifikanzniveaus.

α	$\chi^2_{1,\alpha}$
0.01	6.64
0.05	3.84
0.10	2.71

Tab. 9.4: Kritische Werte für den χ^2-Test

6. Da $16.782 > 3.84$ ist, wird die Nullhypothese "es besteht kein Zusammenhang" zugunsten der Alternativhypothese "es besteht ein Zusammenhang" zwischen Impfung und Erkrankung abgelehnt.

Die χ^2-Methode, beobachtete mit erwarteten Häufigkeiten zu vergleichen, ist nicht auf die 2×2 Felder–Tafeln beschränkt. Deshalb kann mit dieser Methode im Prinzip auch der Zusammenhang zwischen Merkmalen mit mehr als zwei Ausprägungen untersucht werden.

9.1.2 Assoziationsmaß

Will man nicht nur wissen, ob ein Zusammenhang zwischen zwei Merkmalen statistisch nachweisbar ist, sondern will man auch die Stärke des Zusammenhangs beschreiben, verwendet man bei qualitativen Merkmalen so genannte **Assoziationsmaße**. Wir wollen hier den Kontingenzkoeffizient von Pearson vorstellen, da er — genau wie die Teststatistik in Abschnitt 9.1.1 — mit Hilfe von χ^2 berechnet wird.

Der χ^2-Wert einer Kontingenztafel sagt wenig aus über die Stärke des Zusammenhanges zwischen zwei qualitativen Merkmalen . Das ist leicht einzusehen, da er bei gegebenem Verhältnis der Häufigkeiten proportional zur Gesamtzahl der Beobachtungen ist. Daher verwendet man als Assoziationsmaß den **Kontingenzkoeffizienten von Pearson**.

Die Formel lautet

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Für unabhängige Merkmale gilt $C = 0$. Da der Nenner des Bruches in der Formel für C immer größer oder gleich 1 ist, gilt stets $C < 1$. Um zu erreichen, dass bei eindeutigem Zusammenhang, bei dem zu jeder Ausprägung von Merkmal X nur *eine* Ausprägung von Y bzw. zu jeder Ausprägung von Y nur *eine* Ausprägung von X auftritt, verwendet man häufig einen korrigierten Koeffizienten. Speziell für die 4-Felder–Tafel lautet der **korrigierte Kontingenzkoeffizient**

$$C_{korr} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \sqrt{2}$$

Die Kontingenzkoeffizienten C und C_{korr} kann man für beliebige Kontingenztafeln berechnen.

Der Kontingenzkoeffizient C für das Beispiel aus Abschnitt 9.1.1 ist

- im Fall A: $C_{korr} = 0.39$
- im Fall B: $C_{korr} = 0$
- im Fall C: $C_{korr} = 0.66$
- $C_{korr} = 1$, wenn bei Nichtimpfung *alle* Katzen erkrankt wären

9.2 Methoden für quantitative Merkmale

9.2.1 Der Spearman'sche Rangkorrelationskoeffizient (r_s)

Wenn wir zwei Merkmale betrachten, die nicht nur nominales, sondern mindestens ordinales Skalenniveau aufweisen, können wir die Abhängigkeit zweier Merkmale aus der gemeinsamen Anordnung beurteilen.

Wir wollen wissen, ob Studenten mit guten Ergebnissen im 1. Leistungstest für Biometrie auch im 2. Test gut abschneiden, ob also ein Zusammenhang zwischen dem Ergebnis beim 1. Test (X) und beim 2. Test (Y) besteht.

Dazu haben wir aus dem letzten Semester eine Zufallsstichprobe von 5 Kandidaten gewählt. In der Tabelle sind die Ergebnisse zusammengestellt:

Student i	1	2	3	4	5
1. Test x_i	10	13	15	17	18
2. Test y_i	10	16	15	14	19

Die Anordnung der x- und y-Werte drücken wir durch die zugehörigen Ränge aus. (Die Rangzahlen $R(x)$ und $R(y)$ sind die Plätze, die die x- bzw. y-Werte einnehmen, wenn wir sie der Reihe nach ordnen.)

$R(x_i)$	1	2	3	4	5
$R(y_i)$	1	4	3	2	5

Wir wollen nun untersuchen, ob sich im l. und 2. Test ähnliche Ranganordnungen ergeben. Dazu bilden wir für alle $i = 1, \ldots, n$ die Differenz $d_i = R(x_i) - R(y_i)$, quadrieren und summieren auf. Der Wert für $\sum d_i^2$ ist 0, falls alle Rangzahlen Zwillinge sind, also $R(x) = R(y)$ gilt. Dann besteht zwischen den x- und den y-Werten ein strikt monotoner positiver Zusammenhang, d.h. falls die Werte für x zunehmen, wachsen immer auch die y-Werte und die Ranganordnungen der x- und y-Werte sind gleich.

Für die Rangzahlen in der obigen Tabelle ergibt sich

$$\sum_{i=1}^{5} d_i^2 = (1-1)^2 + (2-4)^2 + (3-3)^2 + (4-2)^2 + (5-5)^2 = 8$$

Maximal hätten wir für $\sum d_i^2$ den Wert

$$40 = (1-5)^2 + (2-4)^2 + (3-3)^2 + (4-2)^2 + (5-1)^2$$

bei genau umgekehrter Rangfolge der x- und y-Werte erhalten können. Dieses Maximum hängt aber noch vom Stichprobenumfang ab. Um als Abhängigkeitsmaß eine Zahl zwischen -1 und $+1$ zu erhalten, verwenden wir die Formel

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n\,(n^2-1)}$$

Dieser Rangkorrelationskoeffizient wurde zuerst von dem englischen Psychologen Charles SPEARMAN (1906) eingeführt.

Bei unseren 5 Kandidaten berechnen wir für die richtig gelösten Aufgaben im l. und 2. Leistungstest

$$r_s = 1 - \frac{6 \cdot 8}{5 \cdot (25-1)} = 0.6$$

Da es bei den Rangzahlen nicht darauf ankommt, wie gut die ursprünglichen Messwerte auf einer Geraden liegen, ist r_s zur Beschreibung sämtlicher monotonen Zusammenhänge geeignet.

Der Rangkorrelationskoeffizient versagt aber, wenn es sich um nichtmonotone Abhängigkeiten handelt, wie die Abbildung 9.1 zeigt.

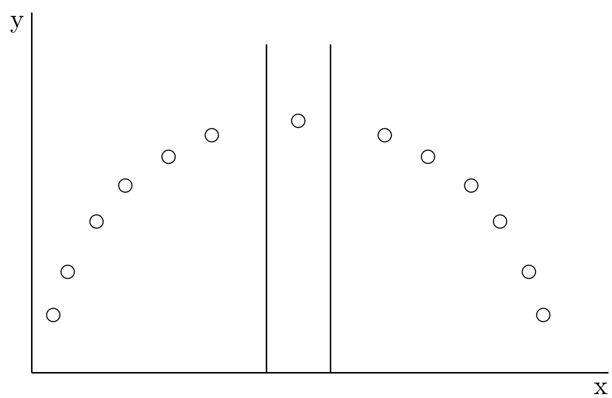


Abb. 9.1: $r_s = 0$

9.2.2 Der Pearsonsche Produkt–Moment–Korrelationskoeffizient (r)

Eine Abbildung ersetzt oft 1000 Zahlen. Dieser Satz gilt ganz besonders, wenn wir den Zusammenhang zwischen zwei quantitativen Merkmalen X und Y untersuchen wollen.

Bei der Routineuntersuchung Neugeborener werden die Körperlänge X und der Kopfumfang Y (in cm) gemessen, um Anomalien im Wachstum zu erkennen. In der folgenden Tabelle sind die Werte für eine Zufallsstichprobe vom Umfang 10 angegeben.

Merkmalsträger i	1	2	3	4	5	6	7	8	9	10
Körperlänge (cm) x_i	53	53	53	48	50	49	51	52	48	53
Kopfumfang (cm) y_i	37	38	37	35	35	35	37	36	34	36

Tab. 9.5: Körperlänge und Kopfumfang neugeborener Kinder

Aus dieser Tabelle lässt sich nicht ohne weiteres ein Zusammenhang zwischen X und Y ablesen.

Wenn wir aber unsere Wertepaare in ein Koordinatenkreuz mit X als horizontaler und Y als vertikaler Achse zeichnen, werden die Verhältnisse klarer.

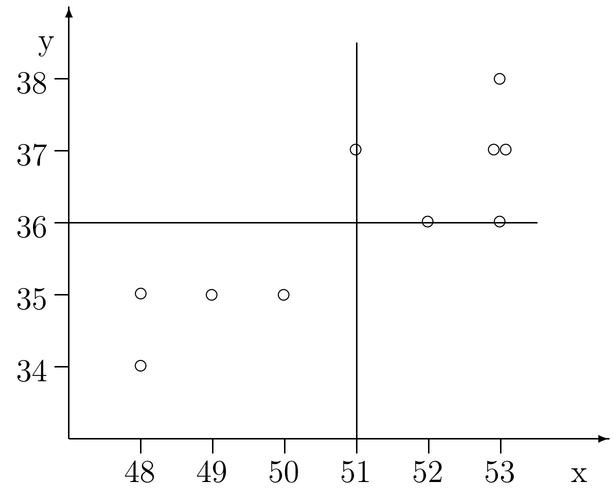


Abb. 9.2: Körperlänge (X) und Kopfumfang (Y)

Wir betrachten die 10 Punkte in Abbildung 9.2, die unseren 10 Stichprobenpaaren entsprechen. Für vier der Neugeborenen liegen sowohl Körperlänge als auch Kopfumfang unterhalb der jeweiligen Mittelwerte, für die übrigen sechs sind beide Werte größer oder gleich den Mittelwerten. So können wir allein aus der Lage unseres Punkteschwarmes schließen, dass ein gewisser Zusammenhang zwischen den untersuchten Merkmalen besteht. Der Kopfumfang nimmt offensichtlich mit wachsender Körperlänge zu. Wie können wir eine Zahl finden, die diesen Zusammenhang quantifiziert ?

Dazu erinnern wir uns an die **Stichprobenvarianz**. Die Definition lautet:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \qquad \text{für } X \quad \text{bzw.}$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y}) \qquad \text{für } Y$$

Entsprechend erhalten wir eine Größe, die die gemeinsame Variabilität unserer Merkmale misst, durch die **Stichprobenkovarianz**:

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

Sie gibt an, wie stark die Merkmale X und Y gemeinsam um ihre jeweiligen Mittelwerte schwanken.

Für unser Beispiel erhalten wir sicher eine positive Stichprobenkovarianz, denn die x- und y-Werte weichen immer in die gleiche Richtung von ihren Mittelwerten ab, sodass alle Summanden in der obigen Formel positive Vorzeichen erhalten (Merke $(-)$ mal $(-)$ gibt $(+)$).

Allerdings hätte der Gebrauch von s_{xy} als Zusammenhangsmaß den Nachteil, noch von den Maßeinheiten der Merkmale X und Y abzuhängen. Würden wir die gemessenen Strecken z.B. in mm angeben, wäre die Stichprobenkovarianz $10 \times 10 = 100$mal größer.
Deshalb dividieren wir s_{xy} durch das Produkt aus s_x und s_y, und wir erhalten als Zusammenhangsmaß

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

mit $s_x = \sqrt{s_x^2}$ und $s_y = \sqrt{s_y^2}$.

Diese Größe wurde zuerst von Karl PEARSON (1896) in obiger Form vorgeschlagen und heißt **Produkt–Moment–Korrelationskoeffizient**.

Es gibt verschiedene Formeln zur Berechnung von r, die wir hier nur kurz auflisten:

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

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Als nächstes wollen wir den Korrelationskoeffizienten für die Merkmale Körperlänge und Kopfumfang bestimmen.
Es ergeben sich $s_x = 2.11$, $s_y = 1.25$, $s_{xy} = 2.22$ und somit $r = \frac{2.22}{2.11 \cdot 1.25} = 0.84$

Es gilt **immer** $-1 \leq r \leq +1$ und für den Fall, dass r den Wert $+1$ oder -1 annimmt, liegen alle Punkte auf einer Geraden.

Deshalb ist r nur geeignet, **lineare Zusammenhänge** zu beschreiben, dabei müssen beide Merkmale quantitativ sein und ein metrisches Skalenniveau besitzen.

Handelt es sich um **nichtlineare Zusammenhänge**, wissen wir, ohne den Punkteschwarm zu betrachten, nicht, ob r klein ist, weil der Zusammenhang gering ist oder weil er stark nichtlinear ist.

Die Abbildungen 9.3 – 9.10 zeigen einige Beispiele für Korrelationen.

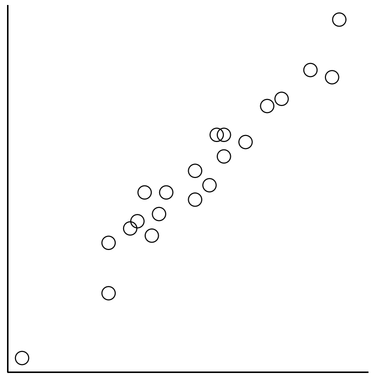


Abb. 9.3: $r = +0.98$

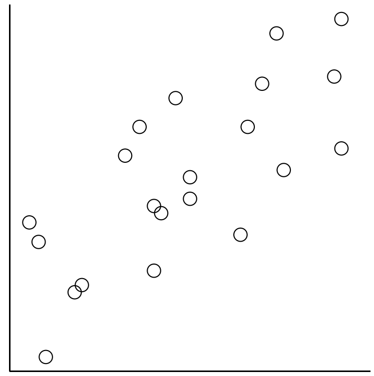


Abb. 9.4: $r = +0.78$

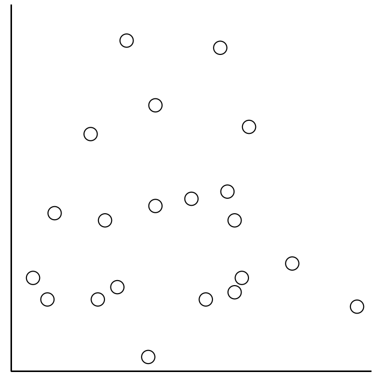


Abb. 9.5: $r = +0.04$

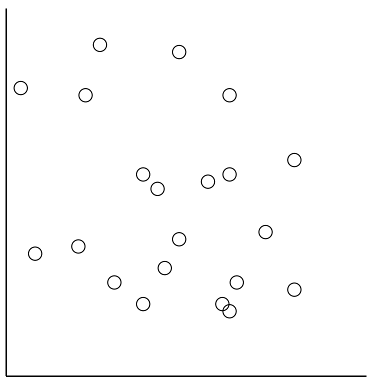


Abb. 9.6: $r = -0.37$

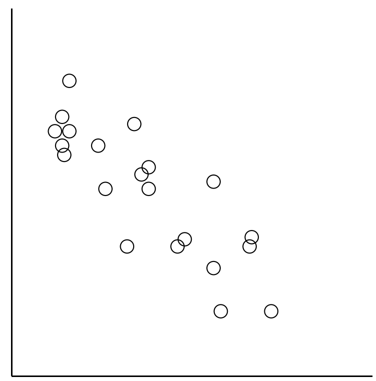


Abb. 9.7: $r = -0.92$

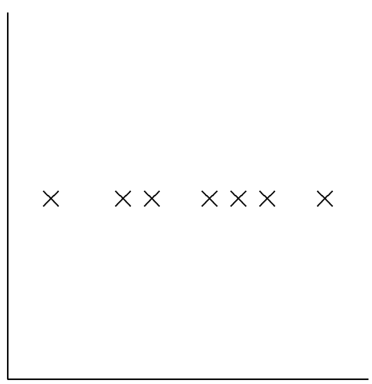


Abb. 9.8: nicht definiert

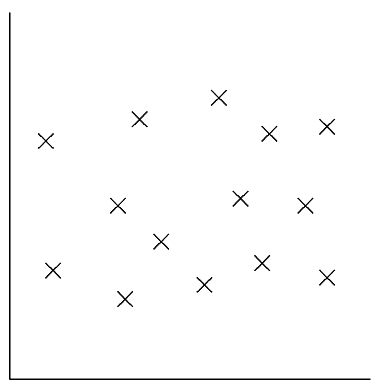


Abb. 9.9: $r = 0$

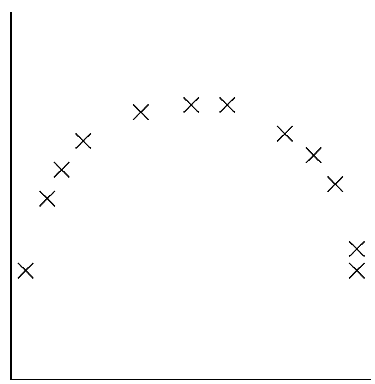


Abb. 9.10: $r = 0$

9.2.3 Warnung oder Fragen nach Ursache und Wirkung

Wir haben bisher verschiedene Methoden kennengelernt, um Zusammenhänge zwischen Merkmalen zu analysieren.

Wenn wir einen Zusammenhang zwischen zwei Merkmalen X und Y beobachten, gibt es drei verschiedene Interpretationsmöglichkeiten.

1. eine Änderung von X bewirkt eine Änderung von Y
2. eine Änderung von Y bewirkt eine Änderung von X

3. weder (1.) noch (2.) treffen zu und Änderungen von X und Y sind auf einen dritten Einflussfaktor zurückzuführen.

Die Existenz eines Zusammenhanges allein gibt uns keinen Aufschluss dafür, welche der Interpretationen die richtige ist, ein wesentlicher und oft übersehener Gesichtspunkt. Die einzige Möglichkeit für die Wahl zwischen den drei Alternativen ist die Berücksichtigung des Vorwissens.

Die Gefahr besteht nun darin, dass wir uns dessen nicht immer bewusst sind und daher nicht ausdrücklich darauf hinweisen. Gerade, wenn wir uns nicht für (1.) oder (2.) entscheiden können, geht die Suche nach dem dritten Faktor los.

(Merke : Es gibt kein Problem, das, wie komplex es auch sei, nach sorgfältiger Analyse nicht noch komplexer erscheint).

9.2.4 Die Regressionsgerade

Aus der grafischen Darstellung der gemessenen Ausprägungen zweier Merkmale an jeweils einem Merkmalsträger kann häufig ein linearer Zusammenhang für die beobachteten Werte abgelesen werden.

Bisher können wir nur beschreiben, wie stark ein solcher Zusammenhang ist. Mit dem linearen Regressionsmodell

$$Y = a + bX$$

wird versucht, aus dem beobachteten "Punkteschwarm" den zumindest im Mittel unterstellten Zusammenhang quantitativ zu beschreiben.

Dabei wird X als "erklärende" (meist fehlerfrei ermittelte) und Y als "zu erklärende" Variable aufgefasst. In vielen Problemstellungen ist es nicht eindeutig, welche Variable durch die andere zu erklären ist, sodass die X- und Y-Variable vertauscht werden können, d.h. anstelle von Y bezogen auf X betrachtet man X bezogen auf Y.

Die **"erklärende"** Variable nennt man auch **unabhängige Variable**, die **"zu erklärende"** bezeichnet man als **abhängige Variable**.

Nach der von GAUSS entwickelten **Methode der kleinsten Quadrate (MKQ)** lässt sich für den Punkteschwarm eindeutig eine Regressionsgerade aus dem möglichen Geradenbündel bestimmen, die der gestellten Bedingung genügt, dass die Summe der Quadrate der vertikalen Abstände der Punkte von der gesuchten Geraden minimal sein soll (siehe Abbildung 9.11). Diese Forderung beinhaltet eine entsprechende Festlegung der Parameter der Geraden, a und b, und lässt sich über die partiellen Ableitungen nach a und b lösen. Die Minimierung der Summe der Abweichungsquadrate folgt aus der Nullsetzung der ersten Ableitung bei positiver zweiter Ableitung. Durch Auflösen der so genannten Normalgleichungen können die Parameter der Geraden, die so genannten **Regressionskoeffizienten** $\hat{a}$ und $\hat{b}$ bestimmt werden.

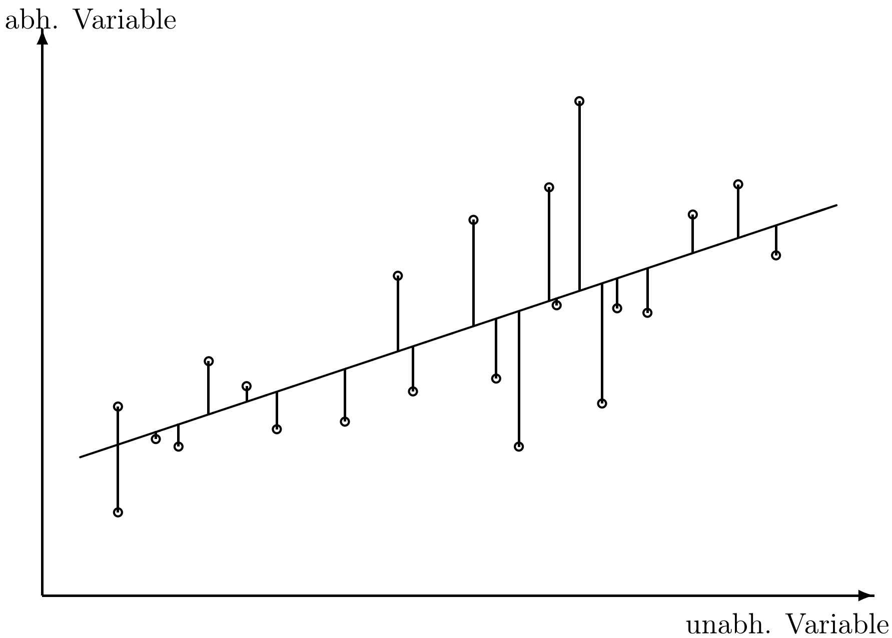


Abb. 9.11: Regressionsgerade

Den Steigungskoeffizienten der Geraden erhält man als

$$\begin{aligned}
\hat{b} &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x^2} \\
&= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{s_{xy}}{s_x^2}
\end{aligned}$$

Der Schnittpunkt mit der y-Achse ist

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

Die derart bestimmte Regressionsgerade läuft durch den **Mittelpunkt (Schwerpunkt)** des Punkteschwarmes mit den Koordinaten $\bar{x}$ und $\bar{y}$. Dabei sind $\bar{x}$ bzw. $\bar{y}$ die arithmetischen Mittelwerte der Merkmalsausprägungen x_i bzw. y_i bei den in die Analyse einbezogenen n Merkmalsträgern.

Bisher haben wir die Abhängigkeit von Y bezogen auf X mit

$$\hat{b} = \frac{s_{xy}}{s_x^2} = \hat{b}_{y/x}$$

betrachtet. Äquivalent ergeben sich die Regressionskoeffizienten für die Abhängigkeit von X bezogen auf Y:

$$\begin{aligned}
\hat{b}_{x/y} &= \frac{s_{xy}}{s_y^2} \\
\hat{a}_{x/y} &= \overline{x} - \hat{b}_{xy}\overline{y}
\end{aligned}$$

Es existiert folgende Verbindung mit dem bereits betrachteten Produkt–Moment–Korrelationskoeffizienten r, der den mittleren linearen Zusammenhang zweier Merkmale charakterisiert:

$$r^2 = \hat{b}_{y/x} \cdot \hat{b}_{x/y}$$

Nur wenn sämtliche Punkte auf einer geraden Linie liegen, sind die Regressionsgeraden x bezogen auf y sowie y bezogen auf x identisch. Daraus resultiert — wie bereits behandelt — ein Produkt–Moment–Korrelationskoeffizient von $r = |1|$. Wegen

$$\begin{aligned} r^2 &= \hat{b}_{x/y} \cdot \hat{b}_{y/x} \\ \text{gilt} \qquad \hat{b}_{x/y} &= \frac{1}{\hat{b}_{y/x}} \end{aligned}$$

Nach Durchführung der Regressionsberechnungen soll die Frage beantwortet werden, in welchem Maße sich die Variation der Variablen Y, charakterisiert durch die Varianz s_y^2 bzw. die Standardabweichung s_y, um das arithmetische Mittel $\bar{y}$ mit Hilfe der Einbeziehung der x-Werte erklären lässt.
Anstelle der Varianz s_y^2 um $\bar{y}$ verbleibt durch Berücksichtigung des Quadrates des mittleren linearen Zusammenhanges r^2 die Streuung $s_{y/x}^2$ um die Regressionsgerade $\hat{y} = \hat{a} + \hat{b}x$:

$$s_{y/x}^2 = \frac{(n-1)}{(n-2)} \cdot s_y^2 \cdot (1 - r^2)$$

$\mathbf{r^2}$ wird daher auch **Bestimmtheitsmaß** genannt und kann prozentual angegeben werden. Die Varianzreduktion wird um so geringer sein, je kleiner das Bestimmtheitsmaß ist. Ist dagegen $r^2 = 1$, so kann die Varianz s_y^2 vollständig durch Einbeziehung der Variablen X erklärt werden, d.h. die verbleibende Variation um die Regressionsgerade ist Null.

9.3 Übungsbeispiel zur Berechnung von Abhängigkeitsmaßen

Die folgende Tabelle 9.3 zeigt das Ergebnis eines konstruierten Beispiels einer Stichprobe mit den später benötigten Summierungen.

Von jedem Merkmalsträger "Mann" (i) sind die Merkmale "Alter in Jahren" (x_i) und "Systolischer Blutdruck in mmHg" (y_i) ermittelt worden. Für alle behandelten Verfahren werden die erforderlichen Rechenergebnisse vorgestellt.

i	x_i	$R(x_i)$	x_i^2	y_i	$R(y_i)$	y_i^2	$x_i y_i$	$d_i = R(x_i) - R(y_i)$	d_i^2
1	20	1.5	400	130	8.0	16900	2600	−6.5	42.25
2	20	1.5	400	120	1.5	14400	2400	0	0
3	23	3.0	529	125	4.0	15625	2875	−1.0	1.00
4	25	4.0	625	120	1.5	14400	3000	2.5	6.25
5	26	5.0	676	135	11.5	18225	3510	−6.5	42.25
6	30	6.0	900	130	8.0	16900	3900	−2.0	4.00
7	31	7.0	961	125	4.0	15625	3875	3.0	9.00
8	37	8.0	1369	130	8.0	16900	4810	0	0
9	40	9.0	1600	140	14.5	19600	5600	−5.5	30.25
10	42	10.0	1764	130	8.0	16900	5460	2.0	4.00
11	45	11.0	2025	145	17.0	21025	6525	−6.0	36.00
12	50	12.0	2500	135	11.5	18225	6750	0.5	0.25
13	53	13.0	2809	125	4.0	15625	6625	9.0	81.00
14	56	14.0	3136	150	18.5	22500	8400	−4.5	20.25
15	57	15.0	3249	140	14.5	19600	7980	0	0.25
16	60	16.0	3600	160	20.0	25600	9600	−4.5	16.00
17	61	17.0	3721	130	8.0	16900	7930	9.0	81.00
18	62	18.0	3844	140	14.5	19600	8680	3.5	12.25
19	65	19.0	4225	140	14.5	19600	9100	4.5	20.25
20	70	20.0	4900	150	18.5	22500	10500	1.5	2.25
$\sum$	873		43233	2700		366650	120120	0	408.5

Bei der grafischen Darstellung in Abbildung 9.12 als "Punkteschwarm" ist das Merkmal "Alter" auf der x-Achse und das Merkmal "Blutdruck" auf der y-Achse aufgetragen.

1. **Produkt–Moment–Korrelationskoeffizient (r):**

$$r = \frac{120120 - 873 \cdot 2700 \div 20}{\sqrt{\left(43233 - \frac{1}{20}873^2\right)}\sqrt{\left(366650 - \frac{1}{20}2700^2\right)}} = 0.682$$

2. **Spearman'scher Rangkorrelationskoeffizient (r_s):**

$$r_s = 1 - 0.307 = 0.693$$

3. **Regressionsgerade:**

$$\begin{aligned}
\bar{y} &= 135 \text{ mmHg} & \bar{x} &= 43.65 \text{ Jahre} \\
s_y^2 &= 113.15 \text{ (mmHg)}^2 & s_x^2 &= 269.82 \text{ (Jahre)}^2 \\
s_y &= 10.64 \text{ mmHg} & s_x &= 16.43 \text{ Jahre} \\
s_{xy} &= 119.21 & r^2 &= 0.4651 = 46.51\%
\end{aligned}$$

$$\begin{aligned}
\hat{b} &= \tfrac{119.21}{269.82} = 0.442 & \hat{a} &= 135 - 0.442 \cdot 43.65 = 115.71 \\
\hat{y} &= 115.71 + 0.442\,x & s_{y/x}^2 &= \tfrac{19}{18} 113.15\ (1 - 0.4651) = 63.88 \text{ (mmHg)}^2
\end{aligned}$$

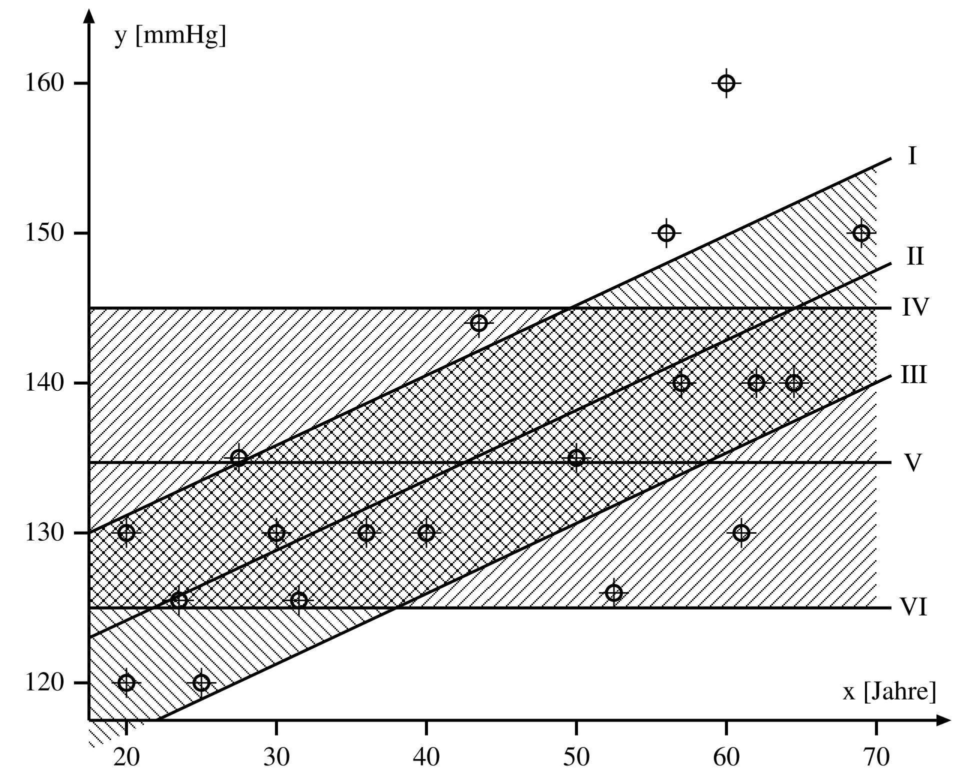
d.h. die Varianz lässt sich über die Berücksichtigung des Alters um ca. 46% reduzieren. Aus der Berechnung ergibt sich eine Standardabweichung um die Regressionsgerade

$$s_{y/x} = 8 \text{ mmHg}$$

im Gegensatz zu der Standardabweichung um den Mittelwert

$$s_y = 11 \text{ mmHg}$$

In Abbildung 9.12 ist neben dem arithmetischen Mittel der y-Werte ($\bar{y}$) zwischen den Markierungen IV und VI der Bereich eingetragen, der durch $\bar{y} \pm s_y$ gegeben ist.
Der Vergleich mit dem Bereich zwischen den Markierungen I und III der Standardabweichung der Einzelwerte um die Regressionsgerade ($\hat{y}_i \pm s_{y/x}$) zeigt die Verringerung der Variation der y-Werte, wenn eine mittlere lineare Abhängigkeit von den x-Werten vorliegt und berücksichtigt werden kann.



I	$\hat{y}_i + s_{y/x}$	IV	$\overline{y} + s_y$
II	$\hat{y}_i$	V	$\overline{y}$
III	$\hat{y}_i - s_{y/x}$	VI	$\overline{y} - s_y$

Abb. 9.12: Regressionsgeraden zum Übungsbeispiel

Literaturverzeichnis

[1] J. Dufner, U. Jensen, E. Schumacher, *Statistik mit SAS*, B.G. Teubner Verlag, Stuttgart, 1992.

[2] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, *Statistik — Der Weg zur Datenanalyse*, Springer–Verlag, 1997

[3] J. Hartung, B. Elpelt, K.-H. Klösener, *Statistik — Lehr- und Handbuch der angewandten Statistik*, 10. Auflage, Oldenbourg, München/Wien, 1998.

[4] W. Köhler, G. Schachtel, P. Voleske, *Biometrie — Einführung in die Statistik für Biologen und Agrarwissenschaftler*, Springer–Verlag, Berlin, 1984.

[5] L. Kreienbrock, S. Schach, *Biometrie — Epidemiologische Methoden*, 2. Auflage, Fischer–Verlag, Stuttgart, 1997.

[6] R. Lorenz, *Grundbegriffe der Biometrie*, 3. Auflage, Fischer–Verlag, 1997

[7] R. Schlittgen, *Einführung in die Statistik*, 6. Auflage, Oldenbourg, München/Wien, 1996

[8] L. Sachs, *Einführung in die Statistik*, 8.Auflage, Springer–Verlag, Berlin, 1997.

Anhang A

Formelsammlung

Rechenregeln für Summen

(A.1) $$\sum_{i=1}^{n} i = 1 + 2 + \ldots + n$$

(A.2) $$\sum_{i=1}^{n} c = \underbrace{c + c + \ldots + c}_{n-mal} = nc$$

(A.3) $$\sum_{i=1}^{n} x_i c = cx_1 + cx_2 + \ldots + cx_n = c \cdot \sum_{i=1}^{n} x_i$$

(A.4) $$\sum_{i=1}^{n} x_i^r = x_1^r + x_2^r + \ldots + x_n^r \qquad \text{Beachte:} \qquad \sum_{i=1}^{n} x_i^r \neq \left(\sum_{i=1}^{n} x_i\right)^r$$

(A.5) $$\sum_{i=1}^{n} (x_i \pm c) = \sum_{i=1}^{n} x_i \pm nc$$

(A.6) $$\sum_{i=1}^{n} (x_i \pm y_i) = \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i$$

(A.7) $$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \qquad \text{Beachte:} \qquad \sum_{i=1}^{n} x_i y_i \neq \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i$$

(A.8) $$\sum_{i=1}^{n} (x_i \pm y_i)^2 = \sum_{i=1}^{n} \left(x_i^2 \pm 2x_i y_i + y_i^2\right) = \sum_{i=1}^{n} x_i^2 \pm 2 \cdot \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2$$

Fakultät

Die **Fakultät** einer natürlichen Zahl k ist definiert als

$$k! \quad = \quad k \cdot (k-1) \cdot (k-2) \cdot \ldots \cdot 2 \cdot 1.$$

Es gilt:

$$1! = 1, \qquad 0! = 1.$$

Binomialkoeffizient

Der **Binomialkoeffizient** $\binom{n}{k}$ ist definiert als

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

Es gilt:

$$\binom{n}{0} = 1, \quad \binom{n}{1} = n, \quad \binom{n}{n} = 1, \quad \binom{n}{k} = 0, \text{ falls } n < k.$$

Er gibt die Anzahl der Möglichkeiten an, aus n Objekten k auszuwählen.

Produktzeichen

Das Summenzeichen $\sum$ dient zur abkürzenden Schreibweise von Summen. Etwas ähnliches gibt es auch für Produkte.

Für beliebige reelle Zahlen x_i kann man schreiben:

$$\prod_{i=1}^{n} x_i \quad = \quad x_1 \cdot x_2 \cdot \ldots \cdot x_n$$

Man liest ‘Produkt aller x_i für $i = 1$ bis n’.

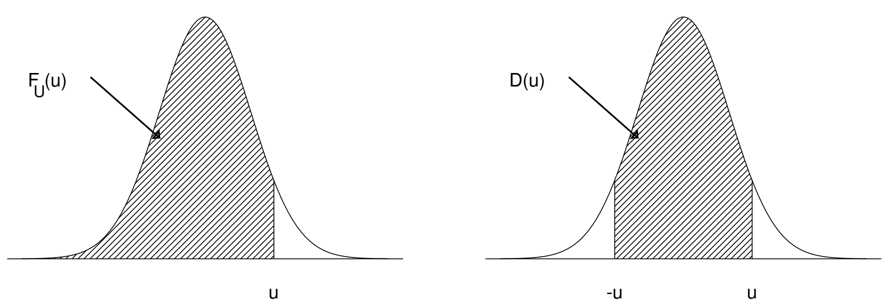
Exponentialfunktion

Anstelle von $y = \exp^x$ kann man äquivalent $y = e^x$ schreiben.

Anhang B

Normalverteilung

$$
\begin{aligned}
F_U(u) &= P(U \leq u) \\
D(u) &= F_U(u) - F_U(-u) = P(-u \leq U \leq u) \\
F_U(-u) &= 1 - F_U(u) \quad ; \quad F_U(U=0) = 0.5 \\
F_U(u_{\alpha/2}) &= \alpha/2 \\
F_U(u_{1-\alpha/2}) &= 1 - \alpha/2 \\
D(u_{1-\alpha/2}) &= F_U(u_{1-\alpha/2}) - F_U(u_{\alpha/2}) = 1 - \alpha \qquad \text{mit } u_{\alpha/2} = -u_{1-\alpha/2}
\end{aligned}
$$



u	$F_U(-u)$	$F_U(u)$	$D(u)$	u	$F_U(-u)$	$F_U(u)$	$D(u)$	u	$F_U(-u)$	$F_U(u)$	$D(u)$
	0,	0,	0,		0,	0,	0,		0,	0,	0,
0,01	4960	5040	0080	0,51	3050	6950	3899	1,01	1562	8438	6875
0,02	4920	5080	0160	0,52	3015	6985	3969	1,02	1539	8461	6923
0,03	4880	5120	0239	0,53	2981	7019	4039	1,03	1515	8485	6970
0,04	4840	5160	0319	0,54	2946	7054	4108	1,04	1492	8508	7017
0,05	4801	5199	0399	0,55	2912	7088	4177	1,05	1469	8531	7063
0,06	4761	5239	0478	0,56	2877	7123	4245	1,06	1446	8554	7109
0,07	4721	5279	0558	0,57	2843	7157	4313	1,07	1423	8577	7154
0,08	4681	5319	0638	0,58	2810	7190	4381	1,08	1401	8599	7199
0,09	4641	5359	0717	0,59	2776	7224	4448	1,09	1379	8621	7243
0,10	4602	5398	0797	0,60	2743	7257	4515	1,10	1357	8643	7287
0,11	4562	5438	0876	0,61	2709	7291	4581	1,11	1335	8665	7330
0,12	4522	5478	0955	0,62	2676	7324	4647	1,12	1314	8686	7373
0,13	4483	5517	1034	0,63	2643	7357	4713	1,13	1292	8708	7415
0,14	4443	5557	1113	0,64	2611	7389	4778	1,14	1271	8729	7457
0,15	4404	5596	1192	0,65	2578	7422	4843	1,15	1251	8749	7499
0,16	4364	5636	1271	0,66	2546	7454	4907	1,16	1230	8770	7540
0,17	4325	5675	1350	0,67	2514	7486	4971	1,17	1210	8790	7580
0,18	4286	5714	1428	0,68	2483	7517	5035	1,18	1190	8810	7620
0,19	4247	5753	1507	0,69	2451	7549	5098	1,19	1170	8830	7660
0,20	4207	5793	1585	0,70	2420	7580	5161	1,20	1151	8849	7699
0,21	4168	5832	1663	0,71	2389	7611	5223	1,21	1131	8869	7737
0,22	4129	5871	1741	0,72	2358	7642	5285	1,22	1112	8888	7775
0,23	4090	5910	1819	0,73	2327	7673	5346	1,23	1093	8907	7813
0,24	4052	5948	1897	0,74	2296	7704	5407	1,24	1075	8925	7850
0,25	4013	5987	1974	0,75	2266	7734	5467	1,25	1056	8944	7887
0,26	3974	6026	2051	0,76	2236	7764	5527	1,26	1038	8962	7923
0,27	3936	6064	2128	0,77	2206	7794	5587	1,27	1020	8980	7959
0,28	3897	6103	2205	0,78	2177	7823	5646	1,28	1003	8997	7995
0,29	3859	6141	2282	0,79	2148	7852	5705	1,29	0985	9015	8029
0,30	3821	6179	2358	0,80	2119	7881	5763	1,30	0968	9032	8064
0,31	3783	6217	2434	0,81	2090	7910	5821	1,31	0951	9049	8098
0,32	3745	6255	2510	0,82	2061	7939	5878	1,32	0934	9066	8132
0,33	3707	6293	2586	0,83	2033	7967	5935	1,33	0918	9082	8165
0,34	3669	6331	2661	0,84	2005	7995	5991	1,34	0901	9099	8198
0,35	3632	6368	2737	0,85	1977	8023	6047	1,35	0885	9115	8230
0,36	3594	6406	2812	0,86	1949	8051	6102	1,36	0869	9131	8262
0,37	3557	6443	2886	0,87	1922	8078	6157	1,37	0853	9147	8293
0,38	3520	6480	2961	0,88	1894	8106	6211	1,38	0838	9162	8324
0,39	3483	6517	3035	0,89	1867	8133	6265	1,39	0823	9177	8355
0,40	3446	6554	3108	0,90	1841	8159	6319	1,40	0808	9192	8385
0,41	3409	6591	3182	0,91	1814	8186	6372	1,41	0793	9207	8415
0,42	3372	6628	3255	0,92	1788	8212	6424	1,42	0778	9222	8444
0,43	3336	6664	3328	0,93	1762	8238	6476	1,43	0764	9236	8473
0,44	3300	6700	3401	0,94	1736	8264	6528	1,44	0749	9251	8501
0,45	3264	6736	3473	0,95	1711	8289	6579	1,45	0735	9265	8529
0,46	3228	6772	3545	0,96	1685	8315	6629	1,46	0721	9279	8557
0,47	3192	6808	3616	0,97	1660	8340	6680	1,47	0708	9292	8584
0,48	3156	6844	3688	0,98	1635	8365	6729	1,48	0694	9306	8611
0,49	3121	6879	3759	0,99	1611	8389	6778	1,49	0681	9319	8638
0,50	3085	6915	3829	1,00	1587	8413	6827	1,50	0668	9332	8664

u	F_U(-u)	F_U(u)	D(u)	u	F_U(-u)	F_U(u)	D(u)	u	F_U(-u)	F_U(u)	D(u)
	0,	0,	0,		0,	0,	0,		0,	0,	0,
1,51	0655	9345	8690	2,01	0222	9778	9556	2,51	0060	9940	9879
1,52	0643	9357	8715	2,02	0217	9783	9566	2,52	0059	9941	9883
1,53	0630	9370	8740	2,03	0212	9788	9576	2,53	0057	9943	9886
1,54	0618	9382	8764	2,04	0207	9793	9586	2,54	0055	9945	9889
1,55	0606	9394	8789	2,05	0202	9798	9596	2,55	0054	9946	9892
1,56	0594	9406	8812	2,06	0197	9803	9606	2,56	0052	9948	9895
1,57	0582	9418	8836	2,07	0192	9808	9615	2,57	0051	9949	9898
1,58	0571	9429	8859	2,08	0188	9812	9625	2,58	0049	9951	9901
1,59	0559	9441	8882	2,09	0183	9817	9634	2,59	0048	9952	9904
1,60	0548	9452	8904	2,10	0179	9821	9643	2,60	0047	9953	9907
1,61	0537	9463	8926	2,11	0174	9826	9651	2,61	0045	9955	9909
1,62	0526	9474	8948	2,12	0170	9830	9660	2,62	0044	9956	9912
1,63	0516	9484	8969	2,13	0166	9834	9668	2,63	0043	9957	9915
1,64	0505	9495	8990	2,14	0162	9838	9676	2,64	0041	9959	9917
1,65	0495	9505	9011	2,15	0158	9842	9684	2,65	0040	9960	9920
1,66	0485	9515	9031	2,16	0154	9846	9692	2,66	0039	9961	9922
1,67	0475	9525	9051	2,17	0150	9850	9700	2,67	0038	9962	9924
1,68	0465	9535	9070	2,18	0146	9854	9707	2,68	0037	9963	9926
1,69	0455	9545	9090	2,19	0143	9857	9715	2,69	0036	9964	9929
1,70	0446	9554	9109	2,20	0139	9861	9722	2,70	0035	9965	9931
1,71	0436	9564	9127	2,21	0136	9864	9729	2,71	0034	9966	9933
1,72	0427	9573	9146	2,22	0132	9868	9736	2,72	0033	9967	9935
1,73	0418	9582	9164	2,23	0129	9871	9743	2,73	0032	9968	9937
1,74	0409	9591	9181	2,24	0125	9875	9749	2,74	0031	9969	9939
1,75	0401	9599	9199	2,25	0122	9878	9756	2,75	0030	9970	9940
1,76	0392	9608	9216	2,26	0119	9881	9762	2,76	0029	9971	9942
1,77	0384	9616	9233	2,27	0116	9884	9768	2,77	0028	9972	9944
1,78	0375	9625	9249	2,28	0113	9887	9774	2,78	0027	9973	9946
1,79	0367	9633	9265	2,29	0110	9890	9780	2,79	0026	9974	9947
1,80	0359	9641	9281	2,30	0107	9893	9786	2,80	0026	9974	9949
1,81	0351	9649	9297	2,31	0104	9896	9791	2,81	0025	9975	9950
1,82	0344	9656	9312	2,32	0102	9898	9797	2,82	0024	9976	9952
1,83	0336	9664	9328	2,33	0099	9901	9802	2,83	0023	9977	9953
1,84	0329	9671	9342	2,34	0096	9904	9807	2,84	0023	9977	9955
1,85	0322	9678	9357	2,35	0094	9906	9812	2,85	0022	9978	9956
1,86	0314	9686	9371	2,36	0091	9909	9817	2,86	0021	9979	9958
1,87	0307	9693	9385	2,37	0089	9911	9822	2,87	0021	9979	9959
1,88	0301	9699	9399	2,38	0087	9913	9827	2,88	0020	9980	9960
1,89	0294	9706	9412	2,39	0084	9916	9832	2,89	0019	9981	9961
1,90	0287	9713	9426	2,40	0082	9918	9836	2,90	0019	9981	9963
1,91	0281	9719	9439	2,41	0080	9920	9840	2,91	0018	9982	9964
1,92	0274	9726	9451	2,42	0078	9922	9845	2,92	0018	9982	9965
1,93	0268	9732	9464	2,43	0075	9925	9849	2,93	0017	9983	9966
1,94	0262	9738	9476	2,44	0073	9927	9853	2,94	0016	9984	9967
1,95	0256	9744	9488	2,45	0071	9929	9857	2,95	0016	9984	9968
1,96	0250	9750	9500	2,46	0069	9931	9861	2,96	0015	9985	9969
1,97	0244	9756	9512	2,47	0068	9932	9865	2,97	0015	9985	9970
1,98	0239	9761	9523	2,48	0066	9934	9869	2,98	0014	9986	9971
1,99	0233	9767	9534	2,49	0064	9936	9872	2,99	0014	9986	9972
2,00	0228	9772	9545	2,50	0062	9938	9876	3,00	0013	9987	9973

Anhang C

Übungsaufgaben zum Kurs

C.1 Beispiel BSE

Vor einigen Jahren traten in Großbritannien die ersten Fälle der Bovinen Spongiformen Encephalopathie (BSE) auf. Dies war eine bis dahin bei Rindern unbekannte Krankheit, deren klinisches Erscheinungsbild prägnant mit dem Schlagwort 'Rinderwahnsinn' charakterisiert werden kann und die bei den betroffenen Tieren zum Tod führt.

Da die Zahl der Fälle rasch zunahm, wurde ein zentrales Meldesystem eingerichtet, um den Überblick über die Entwicklung der Epidemie zu haben und gegebenenfalls eingreifen zu können, Handelsbeschränkungen zu erlassen, etc. Alle BSE-verdächtigen Fälle müssen einer zentralen Instanz gemeldet werden, zu jedem gemeldeten Fall ist ein Fragebogen auszufüllen und die Informationen werden in eine zentrale Datenbank eingegeben, auf die dann für epidemiologische Studien zurückgegriffen werden kann.

BSE-verdächtige Tiere müssen getötet und die Gehirne histologisch bzw. pathologisch untersucht werden, um die Diagnose BSE zu bestätigen. Da das klinische Bild mit dem anderer Erkrankungen verwechselt werden kann, stellt sich zur Zeit immer noch ein gewisser Teil (10–20%) der als BSE-verdächtig gemeldeten und getöteten Tiere als negativ heraus.

Als Übungsbeispiel, an dem sich biometrische Grundbegriffe und Methoden erläutern lassen, wurde ein Datensatz aus dem BSE-Meldesystem zusammengestellt, der folgende Merkmale enthält:

Merkmal	Ausprägungen
Tiernummer	Ohrmarken–Nummer z.B. 395-45/2
Region	1 : Northern 2 : Midlands and Western 3 : Eastern 4 : South–East 5 : South–West 6 : Wales 7 : Scotland
Herdentyp	1 : Milchvieh 2 : Fleischvieh 3 : beides
Herdengröße	Anzahl erwachsener Tiere in der Herde
weitere Fälle	0 : keine 1 : ein weiterer Fall 2 : 2, 3, 4 oder 5 weitere Fälle 3 : mehr als 5 weitere Fälle
Alter	Angabe in Monaten
Gewicht	Angabe in kg
Milchleistung	Liter pro Woche im Durchschnitt
Kraftfutter	ja oder nein
BSE bestätigt	ja oder nein

Tier	Region	H-Typ	H-Größe	Fälle	Alter	Gewicht	Milch	Kr-Futter	BSE
363-08/9	1	1	209	0	70	524	29	ja	ja
112-15/9	1	1	106	1	72	455	25	nein	ja
371-44/2	5	1	90	1	54	550	32	ja	ja
400-67/2	2	1	150	0	58	497	30	ja	ja
753-34/1	7	3	127	1	68	461	33	nein	ja
221-20/8	5	1	180	2	59	571	29	ja	ja
173-16/1	5	1	290	1	75	668	30	ja	ja
480-60/8	6	1	115	0	71	593	26	ja	ja
221-20/1	5	1	150	0	75	500	26	ja	ja
560-37/5	3	1	80	1	72	650	27	nein	ja
501-05/3	6	1	34	1	89	477	36	nein	ja
420-41/1	2	1	140	2	72	549	27	ja	ja
071-09/6	1	1	32	1	58	493	30	ja	ja
560-44/1	3	1	53	0	67	556	30	nein	ja
452-73/2	1	1	71	0	67	468	29	ja	ja
174-88/2	5	2	25	1	69	613	–	ja	ja
102-06/4	1	1	74	1	58	562	26	nein	ja
360-74/4	1	1	120	0	58	531	30	nein	ja
488-45/2	6	1	66	3	66	556	35	ja	ja
153-13/2	2	1	210	1	60	575	30	ja	ja
423-25/2	2	1	118	0	65	564	32	ja	ja
141-49/6	1	1	96	0	71	523	30	ja	ja
420-59/5	2	1	305	0	66	611	35	ja	ja
681-52/7	7	1	98	0	73	588	36	nein	ja
103-14/4	1	1	70	1	70	540	26	ja	ja
334-44/8	2	1	135	0	65	595	27	nein	ja
070-71/4	1	3	68	0	69	546	–	nein	ja
050-09/9	4	1	185	1	66	446	33	nein	ja
430-61/1	5	1	107	0	59	505	34	ja	ja
364-18/1	1	3	100	3	55	544	33	ja	ja
570-47/3	3	1	112	1	56	587	37	nein	nein
064-30/4	5	1	35	0	68	676	31	nein	nein
070-13/4	1	2	65	1	51	658	–	nein	nein
070-80/4	1	1	70	0	66	531	31	nein	nein
071-22/4	1	1	50	2	66	552	31	nein	nein
081-28/1	6	1	60	0	60	636	29	nein	nein
103-53/5	1	2	53	0	59	503	–	nein	nein
110-61/9	1	1	87	1	71	652	24	nein	nein
143-02/2	1	1	128	0	69	557	25	nein	nein
333-32/4	2	1	84	1	71	596	34	nein	nein
364-70/8	1	3	125	0	66	550	29	nein	nein
370-13/1	5	1	240	1	65	641	35	nein	nein
814-88/1	7	1	102	0	53	524	29	nein	nein
845-75/5	7	1	81	0	61	626	29	nein	nein
012-28/1	2	1	119	1	65	516	31	ja	nein
334-32/4	2	1	300	1	55	496	36	ja	nein
380-49/1	4	1	90	0	65	485	30	nein	nein
521-61/4	3	1	150	0	59	514	29	ja	nein
083-21/6	6	1	100	0	49	680	30	nein	nein
086-37/1	6	1	81	1	64	568	33	ja	nein

C.2 Grundbegriffe

Aufgabe 1: Charakterisierung der Merkmale (siehe Beispiel BSE)

Sind die aufgeführten Merkmale

- quantitativ oder qualitativ?
- diskret oder stetig?
- nominal, ordinal oder metrisch?

Geben Sie für jedes dieser Merkmale einige Merkmalsausprägungen an.

Was sind jeweils die Untersuchungseinheiten?

Aufgabe 2: Stichprobenverfahren

Es soll untersucht werden, ob heutzutage bei Airedale–Terriern häufiger Hüftgelenksdysplasie auftritt als vor 30 Jahren (als schon einmal eine Untersuchung durchgeführt wurde).

- Wie sieht die Grundgesamtheit aus (räumlich, zeitlich, sachlich)?
- Wie müsste man eine rein zufällige Auswahl treffen?
- Wie könnte man eine Klumpenstichprobe und wie eine geschichtete Stichprobe ziehen?
- Welche Art von Stichprobe wäre besonders schlecht und würde das Ergebnis verfälschen?

Aufgabe 3: Stichprobe und Grundgesamtheit (Beispiel BSE)

Die 'Fälle' stammen aus dem BSE-Meldesystem und sind daraus als Stichprobe von Tieren gezogen, die 1983/84 geboren wurden.

- Wie lässt sich die Grundgesamtheit charakterisieren?
- Wie müsste man eine rein zufällige Auswahl für die Rinderpopulation in Großbritannien ziehen?
- Wie könnte man eine Klumpenstichprobe und wie eine geschichtete Stichprobe ziehen?
- Welche Art von Stichprobe wäre besonders schlecht und würde das Ergebnis verfälschen?

C.3 Häufigkeitsverteilungen

Aufgabe 4: Häufigkeitsverteilungen (HD bei Airedale–Terriern)

In der Kleintierklinik wurden 50 einjährige Airedale–Terrier auf HD untersucht. Es ergaben sich folgende absolute Häufigkeiten:

Ausprägung	absolut	relativ	abs. Summe	relativ Summe
HD-frei	10			
HD-Verdacht	20			
leichte HD	8			
mittlere HD	7			
schwere HD	5			
insgesamt	50		–	–

- Vervollständigen Sie die Tabelle!
- Stellen Sie das Ergebnis grafisch dar! (Möglichkeiten: Stabdiagramm für die relativen Häufigkeiten, empirische Verteilungsfunktion für die aufsummierten Häufigkeiten)
- Wie viel Prozent der untersuchten Airedale–Terrier sind HD-frei, bei wie viel Prozent besteht HD-Verdacht und wie viel Prozent haben höchstens leichte HD?

Aufgabe 5: Häufigkeitsverteilungen (Beispiel BSE)

- Erstellen Sie die Häufigkeitstabellen für die Merkmale "Fälle" (weitere Fälle in der Herde) und "Alter". Bilden Sie dazu für das Merkmal "Alter" geeignete Altersgruppen bzw. -klassen. Wie können Sie die von Ihnen gebildeten Altersklassen begründen?
- Stellen Sie die Häufigkeitsverteilungen für die Merkmale "Fälle" und "Alter" grafisch als Stabdiagramm und Histogramm dar. Haben Sie weitere Ideen, wie man die Verteilungen darstellen könnte?
- Wie sehen die Verteilungsfunktionen für die beiden Merkmale aus?
- Wie würden Sie die Verteilung des Merkmals "Region" darstellen? Wie sieht die Verteilungsfunktion für dieses Merkmal aus?

C.4 Lagemaße

Aufgabe 6: Lagemaße (HD bei Airedale–Terriern)

Die Untersuchung zum Auftreten von HD hat folgende Häufigkeitstabelle ergeben:

Ausprägung	absolut	relativ	abs. Summe	relativ Summe
HD-frei	10	0.20	10	0.20
HD-Verdacht	20	0.40	30	0.60
leichte HD	8	0.16	38	0.76
mittlere HD	7	0.14	45	0.90
schwere HD	5	0.10	50	1.00
insgesamt	50	1.00	–	–

Bestimmen Sie Modalwert (einfach) und Median (schon etwas schwerer)!

Aufgabe 7: Lagemaße (Alter von Professoren)

Um das Durchschnittsalter der Professoren am Fachbereich zu erforschen, wurden 5 Professoren befragt.
Hier das Ergebnis dieser Stichprobe:

Prof. Naseweiß	54 Jahre
Prof. Grasfeld	44 Jahre
Prof. Schwarz	55 Jahre
Prof. Wetterstein	64 Jahre
Prof. Ferrari	58 Jahre

Bestimmen Sie das arithmetische Mittel und den Median der Altersangaben!

Es kommt noch ein weiterer Professor in die Stichprobe: der erst 25 Jahre alte Prof. Streberlein.

Wie wirkt sich diese zusätzliche Angabe auf das arithmetische Mittel und den Median der Altersangaben aus? Welche neuen Werte erhält man?

Aufgabe 8: Lagemaße (Beispiel BSE)

Berechnen Sie für das Merkmal 'Herdengröße'

- das arithmetische Mittel
- den Median und die Quartile

und vergleichen Sie die Informationen.

C.5 Streuungsmaße

Aufgabe 9: Streuungsmaße (Alter von Professoren)

Sie erinnern sich an die Professoren?

Nochmal: Um das Durchschnittsalter der Professoren am Fachbereich zu erforschen, wurden 5 Professoren befragt. Hier das Ergebnis dieser Stichprobe:

Prof. Naseweiß	54 Jahre
Prof. Grasfeld	44 Jahre
Prof. Schwarz	55 Jahre
Prof. Wetterstein	64 Jahre
Prof. Ferrari	58 Jahre

Diesmal berechnen Sie bitte die folgenden Streuungsmaße:

Spannweite, Varianz, Standardabweichung und Variationskoeffizient.

Wieder kommt noch ein weiterer Professor in die Stichprobe: der erst 25 Jahre alte Prof. Streberlein. Wie wirkt sich diese weitere Person auf die Streuungsmaße aus und welche neuen Werte erhält man?

Aufgabe 10: Streuungsmaße (Beispiel BSE)

Berechnen Sie für das Merkmal 'Herdengröße'

- die Spannweite und den Quartilsabstand
- Varianz, Standardabweichung und Variationskoeffizient.

Welche Lage- und Streuungsmaße passen zueinander?

Aufgabe 11: Boxplots (Beispiel BSE)

Bilden Sie zwei Untergruppen anhand des Merkmals ‘BSE bestätigt’ (ja oder nein).

Bestimmen Sie für beide Gruppen die Extremwerte und Quartile für das Merkmal ‘Herdengröße’ und stellen Sie die Verteilungen als Boxplots dar.

C.6 Wahrscheinlichkeitsrechnung

Aufgabe 12: Ereignisse (Beispiel ‘robuste’ Ponies)

Eine Gruppe von 100 Ponies in Robusthaltung besteht zu gleichen Teilen aus Isländern und Norwegern, von denen jeweils die Hälfte Stuten, die andere Hälfte Wallache sind. Aus dieser Gruppe soll ein Tier für eine eingehende tierärztliche Untersuchung ausgewählt und eingefangen werden?

- Diskutieren Sie an diesem Beispiel die Begriffe Elementarereignis und Ereignisraum.
- Definieren Sie verschiedene Zufallsereignisse, die realisiert werden können.
- Definieren Sie verschiedene Zufallsereignisse, die realisiert werden können, wenn gleich 2 Tiere ausgewählt und eingefangen werden?

Aufgabe 13: Ereignisoperatoren (‘robuste’ Ponies)

Wie lassen sich folgende Zufallsereignisse als Verknüpfung von Ereignissen angeben?

- es wird ein Isländer oder eine Stute ausgewählt
- es wird eine Isländerstute ausgewählt

Sie bekommen zusätzlich auch noch die Information, dass es in jeder Rassen- und Geschlechtergruppe zu gleichen Teilen 1-, 2-, 3-, 4- und 5-jährige Tiere gibt. Wie lassen sich folgende Zufallsereignisse als Verknüpfung von Ereignissen angeben?

- es wird ein Isländer und (eine Stute oder ein 5-jähriges Pony) ausgewählt
- es wird ein Isländer oder eine Isländerstute ausgewählt
- es wird ein Isländer oder eine 5-jährige Stute ausgewählt

Formulieren Sie ein unmögliches und ein sicheres Ereignis für die Situation, dass ein Tier ausgewählt werden soll.

Aufgabe 14: Disjunkte Ereignisse und vollständiges System (‘robuste’ Ponies)

Formulieren Sie für dieses Beispiel disjunkte Ereignisse.

Betrachten Sie die folgenden Ereignisse:

Ereignis A: Isländer, Ereignis B: Norweger, Ereignis C: 5-jährig

Bestimmen Sie die Lösungsmenge für: $A \cup B$, $A \cap B$, $A \cup C$, $A \cap C$

Aufgabe 15: Berechnen von Wahrscheinlichkeiten ('robuste' Ponies)

Es soll ein Tier zufällig ausgewählt werden.
Wie groß ist die Wahrscheinlichkeit, dass

- ein Norweger untersucht wird?
- eine Stute untersucht wird?
- eine Norwegerstute untersucht wird?
- eine 2-jährige Norwegerstute untersucht wird?
- keine 2-jährige Norwegerstute untersucht wird?
- ein Norweger oder eine Stute untersucht wird?

Aufgabe 16: Bedingte Wahrscheinlichkeiten ('robuste' Ponies)

Jetzt sollen nacheinander 2 Tiere zufällig ausgewählt und eingefangen werden.

Wenn als erstes Tier bereits ein Norweger ausgewählt wurde, wie groß ist dann die Wahrscheinlichkeit, dass das zweite Tier auch ein Norweger ist?

Aufgabe 17: Wahrscheinlichkeiten bei diagnostischen Tests

Angenommen, für die Diagnose der Hundestaupe stünden zwei diagnostische Tests zur Verfügung, deren Ergebnisse unabhängig voneinander sind. Das eine Testverfahren reagiert bei einem tatsächlich an Staupe erkrankten Hund mit einer Wahrscheinlichkeit von 0.8 positiv, das andere Verfahren erkennt diesen Hund mit einer Wahrscheinlichkeit von 0.6 als erkrankt.

Wie groß ist die Wahrscheinlichkeit, die Erkrankung bei einem tatsächlich erkrankten Hund zu erkennen, wenn beide Tests angewendet werden?

C.7 Binomialverteilung

Aufgabe 18: Unabhängigkeit ('robuste' Ponies)

Es sollen wieder mal 2 Ponies zufällig ausgewählt und untersucht werden, allerdings an zwei aufeinanderfolgenden Tagen. Das am ersten Tag gefangene Tier kommt gleich nach der Untersuchung wieder in die Herde zurück.

Wie groß ist die Wahrscheinlichkeit,

- dass am zweiten Tag wieder ein Norweger untersucht wird, wenn es am ersten Tag schon ein Norweger war?
- an beiden Tagen einen Norweger zu fangen?
- dasselbe Tier zweimal zu untersuchen?

Aufgabe 19: Binomialverteilung

Angenommen, in einer Population werden Tiere unabhängig voneinander mit der gleichen Wahrscheinlichkeit von $p = 0.2$ von einer bestimmten Krankheit befallen. Ein Tierarzt untersucht nun eine zufällig aus dieser Population ausgewählte Stichprobe von $n = 4$ Tieren.

- Wie groß ist die Wahrscheinlichkeit, dass alle 4 Tiere krank sind?
- Wie groß ist die Wahrscheinlichkeit, dass mindestens ein Tier gesund ist?
- Wie groß ist die Wahrscheinlichkeit, dass alle 4 Tiere gesund sind?
- Wie groß ist die Wahrscheinlichkeit, dass mindestens ein Tier krank ist?

Aufgabe 20: Binomialverteilung (Prof. Quälgeist)

Bei Prof. Quälgeist beträgt die Durchfallquote 50% ($p = 0.5$). Heute sollen 8 Studenten ($n = 8$) von ihm geprüft werden.

- Wie groß sind die Wahrscheinlichkeiten, dass von den 8 Prüflingen überhaupt keiner, dass 1, 2, 3, 4, 5, 6, 7, oder sogar alle durchfallen?
- Mit welcher Wahrscheinlichkeit fallen höchstens 2, höchstens 3, höchstens 4, höchstens 5, höchstens 6 und höchstens 7 Studenten durch? (Anmerkung: ‘höchstens 2’ heißt, dass entweder 0, 1, oder 2 Studenten durch die Prüfung fallen.)

Fertigen Sie dazu eine Tabelle mit den verschiedenen Werte zu X: ‘Anzahl der durchgefallenen Studenten’ und den Wahrscheinlichkeiten an. Als Vorlage kann die Tabelle im Skript auf S.82 dienen.

Aufgabe 21: noch ein wenig Wahrscheinlichkeitsrechnung

Ein Therapieversuch habe die möglichen Ergebnisse “Verschlechterung”, “Zustand unverändert”, “Besserung” und “Heilung” mit gleichen Wahrscheinlichkeiten. Dann ist die Wahrscheinlichkeit dafür, dass in einer zufälligen Stichprobe von 3 Patienten bei allen 3 Patienten eine Verschlechterung eintritt, gleich ...... ?

Aufgabe 22: und noch mal Binomialverteilung (HD)

Zur Linderung ihrer Beschwerden wird bei $n = 20$ HD-belasteten Hunden eine neue Behandlungsmethode ausprobiert, die im Einzelfall mit einer Wahrscheinlichkeit von $p = 0.7$ erfolgreich sein soll.

Mit welchen Wahrscheinlichkeiten werden

- alle 20 Hunde
- wird kein Hund
- genau 10 Hunde
- genau 14 Hunde

- bis zu 12 Hunden
- weniger als 14 Hunde
- zwischen 10 und 15 Hunden
- mindestens 17 Hunde geheilt? (s. Tabelle in Abschnitt 8.5)

C.8 Normalverteilung

Aufgabe 23: Standardnormalverteilung

Machen Sie sich mit der Standardnormalverteilung (Normalverteilung mit dem Erwartungswert $\mu = 0$ und der Varianz $\sigma^2 = 1$) vertraut, indem Sie mit Hilfe der Tabelle im Anhang für eine standardnormalverteilte Zufallsvariable X folgende Fragen beantworten:

Wie groß ist die Wahrscheinlichkeit, dass X Werte

- kleiner als 1, kleiner als 2, kleiner als 3
- größer als 1, größer als 2, größer als 3
- kleiner als -1, kleiner als -2, kleiner als -3
- zwischen -1 und $+1$
- zwischen -2 und $+2$
- zwischen -3 und $+3$
- genau den Wert 1 annimmt?

Bestimmen Sie umgekehrt jeweils eine Konstante k, für die gilt, dass X mit einer Wahrscheinlichkeit von 0.8 bzw. 0.9 Werte zwischen $-k$ und $+k$ annimmt!

Aufgabe 24: Normalverteilung mit Erwartungswert μ und Standardabweichung σ

Beantworten Sie nun für eine normalverteilte Zufallsvariable X mit dem Erwartungswert μ und der Standardabweichung σ die folgenden Fragen:

Wie groß ist die Wahrscheinlichkeit, dass X Werte

- kleiner als $\mu + \sigma$, kleiner als $\mu + 2\sigma$, kleiner als $\mu + 3\sigma$
- größer als $\mu + \sigma$, größer als $\mu + 2\sigma$, größer als $\mu + 3\sigma$
- kleiner als $\mu - \sigma$, kleiner als $\mu - 2\sigma$, kleiner als $\mu - 3\sigma$
- zwischen $\mu - \sigma$ und $\mu + \sigma$
- zwischen $\mu - 2\sigma$ und $\mu + 2\sigma$
- zwischen $\mu - 3\sigma$ und $\mu + 3\sigma$
- genau den Wert $\mu + \sigma$ annimmt?

Bestimmen Sie umgekehrt jeweils eine Konstante k, für die gilt, dass X mit einer Wahrscheinlichkeit von 0.8 bzw. 0.9 Werte zwischen $\mu - k\sigma$ und $\mu + k\sigma$ annimmt!

Aufgabe 25: Geburtsmasse von Ferkeln

Für die Geburtsmasse von Ferkeln sei angenommen, dass sie normalverteilt sind mit einem Erwartungswert von μ=1300 g und einer Standardabweichung von σ=300 g.

Mit welcher Wahrscheinlichkeit sind Ferkel zu erwarten

- mit einer Geburtsmasse von mehr als 2000 g?
- mit einer Geburtsmasse zwischen 1000 g und 1500 g?
- mit einer Geburtsmasse von genau 1600 g?
- mit einer Geburtsmasse kleiner als 400 g?
- mit einer Geburtsmasse zwischen 1000 g und 1600 g, zwischen 700 g und 1900 g, bzw. zwischen 400 g und 2200 g?

Berechnen Sie die zentralen Schwankungsintervalle, in denen ein Anteil von 0.8 bzw. 0.9 der Geburtsmassen von Ferkeln zu erwarten sind.

Aufgabe 26: Gewichtszunahme

Die Gewichtszunahme (Differenz zwischen 2. und 1. Gewichtsmessung) bei einer Untersuchung an Mastschweinen sei normalverteilt mit dem Erwartungswert μ=2 kg und der Standardabweichung σ=0.6 kg.

Mit welcher Wahrscheinlichkeit nimmt ein Schwein

- zwischen 1.4 und 3.2 kg zu?
- bis zu 2 kg zu?
- zwischen 1.5 und 2.5 kg zu?
- mehr als 4 kg zu?

C.9 Schätzen

Aufgabe 27: Länge des Konfidenzintervalls

Wählen Sie diejenige Kombination aus, die das

1. kürzeste (schmalste)
2. längste (breiteste)

Konfidenzintervall liefert.

Stichprobenumfang	klein	groß	
Sicherheitswahrscheinlichkeit	90%	95%	99%
Varianz	groß	klein	

Aufgabe 28: Konfidenzintervall (Studie Mastschweine)

Die Gewichtszunahmen bei Mastschweinen seien grundsätzlich normalverteilt.

Aus einer klinischen Studie an 500 Schweinen sollen die zu erwartenden mittleren Gewichtszunahmen in verschiedenen Untergruppen geschätzt werden.

Berechnen Sie dazu 95%-Konfidenzintervalle für jede von 6 Untergruppen und interpretieren Sie die Ergebnisse.

(Anmerkung: Bei einigen Tieren fehlen in der Studie einige Angaben, sodass der Stichprobenumfang nicht immer gleich 500 ist.)

Gewichtszunahmen getrennt nach Heilung:

Heilung	n	arithm. Mittel	Standardabw. s
ja	470	1.85	0.60
nein	20	1.12	0.38

Gewichtszunahmen getrennt nach Geschlecht:

Geschlecht	n	arithm. Mittel	Standardabw. s
männlich	238	1.84	0.56
weiblich	244	1.81	0.66

Gewichtszunahmen getrennt nach Behandlung:

Präparat	n	arithm. Mittel	Standardabw. s
Baktosil	194	1.91	0.53
Hustofix	296	1.77	0.65

C.10 Statistisches Testen

Aufgabe 29: Formulieren von Hypothesen (Airedale–Terrier)

Professor Schlaumeier vermutet, dass Deutsche Schäferhunde stärker mit HD belastet sind als Airedale–Terrier. Dazu führt er eine Studie durch und möchte seine Vermutung mit Hilfe eines statistischen Tests überprüfen.

Wie sollte er Null- und Alternativhypothese formulieren?

Aufgabe 30: Fehlerarten (Airedale–Terrier)

1. Professor Schlaumeier kann beim statistischen Test die Nullhypothese nicht ablehnen. Wie wird allgemein die Wahrscheinlichkeit angegeben, dass dies eine falsche Entscheidung ist? Wie wird die Wahrscheinlichkeit angegeben, mit der er so zu einer richtigen Entscheidung kommt?

2. Professor Schlaumeier kann anhand der Studie seine Ausgangsvermutung bestätigen. Wie bezeichnet man die Wahrscheinlichkeit, mit der diese Entscheidung falsch ist? Und wie bezeichnet man die Wahrscheinlichkeit, mit der dies eine richtige Entscheidung ist?

3. In Wirklichkeit seien Schäferhunde und Airedale–Terrier gleich stark mit HD belastet. Professor Schlaumeier bestätigt mit seiner Studie jedoch seine eigene Ausgangsvermutung (siehe Aufgabe 29). Macht er einen Fehler, und wenn ja, wie wird dieser bezeichnet?

Aufgabe 31: Durchführung eines statistischen Tests (weiße Gummibärchen)

Die Tiermedizinstudentin Y.S. liebt Gummibärchen der Firma HARIBO über alles, besonders die weißen. Seit einiger Zeit hat sie jedoch den Eindruck, dass statt des früheren Anteils von ca. 30% pro Tüte sehr viel weniger weiße Gummibärchen abgepackt werden.

Bevor sie sich mit einer Beschwerde an den Geschäftsführer der Firma wendet, möchte sie sich ihrer Vermutung aber sicherer sein und mit Hilfe ihrer Kommilitoninnen einen Versuch durchführen.

- Wie muss Y.S. Null- und Alternativhypothese formulieren, damit sie bei einer möglichen Beschwerde weiß, wie sicher sie sein kann, recht zu haben?
- Formulieren Sie die Hypothesen so, dass Sie die Tabellen im Skript (S.82 oder S.83) nutzen können.
  (Tip: bei der Farbe der Gummibärchen interessiert Y.S. nur, ob sie 'weiß' oder 'farbig' sind.)

Das Signifikanzniveau legt Y.S. auf $\alpha = 0.10$ fest, der Umfang der Stichprobe soll $n = 20$ betragen. Für die Stichprobe sammelt sie von ihren Kommilitonen zufällig aus einer Tüte gezogene Gummibärchen ein. Es wird nur notiert, ob die Gummibärchen 'weiß' oder 'farbig' sind.

- Berechnen Sie den kritischen Wert, das heißt, die kritische Anzahl weißer Bärchen!
- Wie lautete der kritische Wert, wenn das Signifikanzniveau auf $\alpha = 0.2$ festgelegt wäre?
- Vergleichen Sie das Ergebnis der Stichprobe mit dem kritischen Wert. Lehnen Sie H_0 ab oder nicht?
- Welche Empfehlung geben Sie Y.S.? Welchen Fehler könnte sie machen, wenn sie Ihrer Empfehlung folgt?

C.11 Unabhängigkeitstests

Aufgabe 32: Klinische Studie Schwein

In einer Studie zur Wirksamkeit des neu entwickelten Präparates 'Baktosil' wurde für die beiden Merkmale "Heilung nach 3 Tagen" (ja/nein), und "Husten am 3. Tag" (ja/nein) an 500 Versuchsschweinen untersucht, ob es einen Zusammenhang zwischen der Behandlung (mit 'Hustofix' oder 'Baktosil') und dem jeweiligen Merkmal gibt.

Die Nullhypothese für den statistischen Test lautet: "Es gibt keinen Zusammenhang zwischen der Behandlung mit 'Hustofix' oder 'Baktosil' und dem Auftreten des jeweiligen Merkmals".

Berechnen Sie für die folgenden Kontingenztafeln jeweils den empirischen χ^2-Wert. Als Signifikanzniveau wurde $\alpha = 5\%$ festgelegt. Zu welcher Entscheidung kommen Sie hinsichtlich der Nullhypothese?

Präparat	"Heilung" ja	"Heilung" nein	Summe
Hustofix	193	7	200
Baktosil	286	14	300
Summe	479	21	500

Präparat	"Husten" ja	"Husten" nein	Summe
Hustofix	145	55	200
Baktosil	203	97	300
Summe	348	152	500

Aufgabe 33: χ^2-Test

Berechnen Sie für die folgende Vierfelder–Tafel den empirischen χ^2-Wert.

	$A+$	$A-$	Summe
$B+$	11	6	17
$B-$	8	25	33
Summe	19	31	50

Besteht ein statistisch signifikanter Zusammenhang zwischen den Merkmalen A und B? Als Irrtumswahrscheinlichkeit für den Fehler 1. Art ist $\alpha = 5\%$ gewählt worden?

Wie sähen die Vierfelder–Tafeln (bei unveränderten Randsummen) aus für den Fall

- der vollständigen Unabhängigkeit? (erwartete Häufigkeit in jeder einzelnen Zelle)
- der maximalen Abhängigkeit?

Berechnen Sie für beide Fälle den empirischen χ^2-Wert!

C.12 Korrelation und Regression

Aufgabe 34: Beurteilung der HD

Um den Grad der Übereinstimmung zwischen zwei Professoren hinsichtlich ihrer HD-Beurteilungen zu untersuchen, lässt man die Röntgenaufnahmen von 7 Hunden unabhängig von beiden Professoren beurteilen. Folgende Einstufungen auf einer Rangskala können vorgenommen werden:

HD-frei:	1
HD-Verdacht:	2
leichte HD:	3
mittlere HD:	4
schwere HD:	5

Die beiden Professoren kommen bei den 7 Hunden zu folgenden Beurteilungen:

Hund:	Prof. H.:	Prof. L.:
Wuffi	1	2
Schnuffi	3	2
Knuffi	4	5
Wurzel	5	3
Purzel	1	3
Stupsi	2	2
Tapsi	3	1

Berechnen Sie den geeigneten Korrelationskoeffizienten!

Aufgabe 35: Regression und Korrelation

Der Zusammenhang zwischen den Körpergrößen (in cm) von Müttern und denen ihrer Töchter soll untersucht werden. Dazu werden 5 Mutter–Tochter–Paare befragt:

	X: Körpergröße der Mutter	Y: Körpergröße der Tochter
Paar 1	165	174
Paar 2	164	168
Paar 3	152	154
Paar 4	170	179
Paar 5	168	166

- Berechnen Sie für die beiden Merkmale den Produkt–Moment–Korrelationskoeffizienten und den Rangkorrelationskoeffizienten nach Spearman. Berechnen Sie auch alle Hilfsgrößen (Mittelwerte, Standardabweichung, Kovarianz, Summe der Abweichungsquadrate).
- Berechnen Sie die Regressionsgerade, die die mittlere Körpergröße von Töchtern in Abhängigkeit von der Körpergröße ihrer Mütter schätzt.
- Stellen Sie die Daten in einem Punktediagramm ("Scatterplot") dar und zeichnen Sie die Regressionsgerade ein.
- Welche Körpergröße ist bei der Tochter einer 176 cm großen Mutter zu erwarten?
- Berechnen Sie das Bestimmtheitsmaß und interpretieren Sie es!

Anhang D

Übungsklausuren

Pflichtveranstaltung Biometrie **1. Leistungsnachweis SS 1998**

FB Veterinärmedizin, FU Berlin 4. 6. 1998

Aufgabe 1:

Zu den schon bekannten 5 Professoren am Fachbereich im Alter von 53, 44, 55, 64 und 58 Jahren gesellt sich auf einer Party der vor langer Zeit schon emeritierte 86-jährige Kollege Methusalem.

Wie wirkt sich das auf Modalwert, Median und arithmetisches Mittel der Altersangaben aus?

1. Der Median steigt stärker an als das arithmetische Mittel.
2. Das arithmetische Mittel steigt an, während Modalwert und Median auf jeden Fall gleich bleiben.
3. Das arithmetische Mittel steigt stärker an als der Median.
4. Der Modalwert steigt auf 86 an.
5. Die Lagemaße bleiben unverändert.

Aufgabe 2:

Ein Student, der keine Zeit hat, sich auf einen 20-Fragen–Multiple–Choice–Test vorzubereiten, beschließt, bei jeder Frage zufällig zu raten. Dabei besitzt jede Frage fünf Antwortmöglichkeiten, von denen jeweils eine die richtige ist.

Wie viele richtige Antworten sind bei so einem Studenten im Mittel zu erwarten?

1. drei
2. keine
3. sechs
4. vier
5. zwei

Aufgabe 3:

Bei einem Diagnosevergleich werden zwei Ärzte gebeten, 12 Röntgenaufnahmen anhand eines Punkteschemas von 1 bis 40 zu bewerten:

Aufnahme:	1	2	3	4	5	6	7	8	9	10	11	12
A1:	3	7	25	38	15	18	22	5	39	23	16	8
A2:	10	11	18	27	16	21	25	4	32	20	14	3

Bitte geben Sie für beide Ärzte getrennt die drei Quartile ihrer Diagnosescores an.

1. A1: $x_{0.25} = 3.5, \quad x_{0.5} = 6.5, \quad x_{0.75} = 9.5$
   A2: $x_{0.25} = 3.5, \quad x_{0.5} = 6.5, \quad x_{0.75} = 9.5$

2. A1: $x_{0.25} = 7.5, \quad x_{0.5} = 17, \quad x_{0.75} = 24$
   A2: $x_{0.25} = 10.5, \quad x_{0.5} = 17, \quad x_{0.75} = 23$

3. A1: $x_{0.25} = 8, \quad x_{0.5} = 18, \quad x_{0.75} = 25$
   A2: $x_{0.25} = 11, \quad x_{0.5} = 18, \quad x_{0.75} = 25$

4. A1: $x_{0.25} = 7, \quad x_{0.5} = 16, \quad x_{0.75} = 23$
   A2: $x_{0.25} = 10, \quad x_{0.5} = 16, \quad x_{0.75} = 21$

5. A1: $x_{0.25} = 6, \quad x_{0.5} = 17, \quad x_{0.75} = 31.5$
   A2: $x_{0.25} = 7, \quad x_{0.5} = 17, \quad x_{0.75} = 26$

Aufgabe 4:

Zu den schon bekannten 5 Professoren am Fachbereich im Alter von 53, 44, 55, 64 und 58 Jahren gesellt sich auf einer Party der vor langer Zeit schon emeritierte 86-jährige Kollege Methusalem.

Wie wirkt sich das auf die Standardabweichung s und den Variationskoeffizienten v der Altersangaben aus?

1. $s = 13.08, v = 0.218$
2. $s = 14.32, v = 0.239$
3. $s = 13.08, v = 2.85$
4. $s = 205.2, v = 3.42$
5. $s = 14.32, v = 3.42$

Aufgabe 5:

Unter der Annahme, dass Rechtshänder und Linkshänder mit der gleichen Wahrscheinlichkeit auftreten, ist die Wahrscheinlichkeit, dass von 10 zufällig ausgewählten Personen mindestens eine Person Rechtshänder ist

1. $1 - (\frac{1}{2})^{10}$
2. $1 - (\frac{1}{10})$
3. $(\frac{1}{2})^{10}$

4. doppelt so groß wie die Wahrscheinlichkeit, dass höchstens 9 Personen Rechtshänder sind.
5. $(\frac{1}{10})$

Aufgabe 6:

Bei Kälbern sei die Gewichtszunahme innerhalb einer bestimmten Zeitspanne als normalverteilt bekannt mit einer im Mittel erwarteten Gewichtszunahme von $\mu = 10$ kg und einer Varianz der Gewichtszunahmen in der betrachteten Zeitspanne von $\sigma^2 = 16$ kg^2.

Mit welcher Wahrscheinlichkeit nimmt ein Kalb mehr als 7 kg zu?

1. 0.7734
2. 0.2266
3. 0.5467
4. 0.5753
5. 1.0000

Aufgabe 7:

Zur regelmäßigen Qualitätsbeurteilung von Hamburgern in einer Fast–Food–Filiale lagen für einen Monat folgende Beschwerdegründe vor, wobei von den Kunden jeweils nur der Hauptbeschwerdegrund genannt werden konnte:

1, 2, 3, 1, 1, 1, 3, 2, 4, 1, 1, 5, 1, 2, 5, 1, 1, 1, 1, 1, 2, 3, 2, 1, 3, 1, 2, 1, 1, 2, 1, 1, 1, 1, 2

Diese Angaben haben die folgenden Bedeutungen: 1 – zu klein, 2 – zu fettig, 3 – zu viel Brot, 4 – zu salzig, 5 – zu trocken.

Welches Wertepaar $(x_{mod}, x_{0.5})$ finden Sie für den Modalwert und den Median des nominal skalierten Merkmals Hauptbeschwerdegrund?

1. $x_{mod} = 1, x_{0.5} = 3$
2. x_{mod} nicht bestimmbar, $x_{0.5} = 1$
3. $x_{mod} = 1, x_{0.5} = 1$
4. $x_{mod} = 1, x_{0.5} = 2$
5. $x_{mod} = 1$, $x_{0.5}$ nicht bestimmbar

Aufgabe 8:

Nehmen Sie einmal an, bei Rindern, die als ‘BSE-verdächtig’ gemeldet werden, ließe sich bei einem Anteil von 15% die Diagnose ‘BSE’ **nicht** bestätigen.

Wie groß wäre dann die Wahrscheinlichkeit, dass von 5 unabhängig voneinander auftretenden Verdachtsfällen **mindestens** 2 doch nicht an BSE erkrankt sind?

1. $10 \cdot 0.15^2 \cdot 0.85^3$
2. $0.15^2 \cdot 0.85^3$

3. $0.85^5 + 5 \cdot 0.15 \cdot 0.85^4$

4. $1 - 10 \cdot 0.15^2 \cdot 0.85^3$

5. $1 - (0.85^5 + 5 \cdot 0.15 \cdot 0.85^4)$

Aufgabe 9:

Für das Merkmal X:"Anzahl der Ferkel pro Wurf" wurden für $n = 100$ Würfe von Zuchtsauen zu den Werten x_i die Häufigkeiten n_i ermittelt:

x_i:	4	5	6	7	8	9	10	11	12	13	14	15	16
n_i:	1	3	0	8	11	18	20	16	11	5	4	2	1

Aus der Häufigkeitsverteilung erhält man ein arithmetisches Mittel von $\overline{x} = 10$ und eine Standardabweichung von $s = 2.25$.

Wie viele der 100 Würfe befinden sich im zweifachen Streubereich $(\overline{x} \pm 2 \cdot s)$?

1. 98

2. 93

3. 89

4. 76

5. 95

Aufgabe 10:

Welche Informationen aus einer Stichprobe vom Umfang n benötigen Sie, um für ein metrisch skaliertes Merkmal, z.B. "Schlachtgewicht von Masthähnchen (in g)", einen Boxplot zeichnen zu können?

1. $x_{[1]}, x_{0.25}, s, x_{0.75}, x_{[n]}$

2. $x_1, x_{0.25}, x_{mod}, x_{0.75}, x_n$

3. $x_{min}, x_{0.25}, x_{0.5}, x_{0.75}, x_{max}$

4. keine

5. $x_{[1]}$, Quartilsabstand, $x_{[n]}$

Aufgabe 11:

Angenommen, für die Diagnose der Hundestaupe stünden zwei diagnostische Tests zur Verfügung, deren Ergebnisse unabhängig voneinander sind. Das eine Verfahren reagiert bei einem tatsächlich an Staupe erkrankten Hund mit einer Wahrscheinlichkeit von 0.9 positiv, das andere Verfahren erkennt so einen Hund mit einer Wahrscheinlichkeit von 0.7 als erkrankt.

Wie groß ist die Wahrscheinlichkeit, dass beide Tests gleichzeitig bei einem erkrankten Hund zu einem falsch–negativen Ergebnis kommen?

1. 0.37

2. 0.97

3. 0.03
4. 0.63
5. 0.40

Aufgabe 12:

Bei der Kontrolle von $n = 4$ in kg gemessenen Gewichtswerten wurde festgestellt, dass der Maximalwert um 20 kg zu niedrig angegeben wurde.

In welcher Weise ändert die Korrektur des Maximalwertes a) die Bereichsmitte, b) den Median, c) das arithmetische Mittel und d) die Spannweite?

Sie werden um

1. a) 10 kg größer, b) 0 kg größer, c) 5 kg größer, d) 20 kg größer
2. a) 20 kg größer, b) 0 kg größer, c) 10 kg größer, d) 20 kg größer
3. a) 10 kg größer, b) 0 kg größer, c) 10 kg größer, d) 20 kg größer
4. a) 20 kg größer, b) 0 kg größer, c) 10 kg größer, d) 20 kg größer
5. a) 10 kg größer, b) 5 kg größer, c) 5 kg größer, d) 20 kg größer

Aufgabe 13:

In einer Studie an 100 dreijährigen Schäferhunden ergab sich zu dem Auftreten von Hüftgelenksdysplasie folgende Häufigkeitstabelle:

A_i	$H(A_i)$	$h(A_i)$	$F(A_i)$
HD-frei	10	0.10	0.10
HD-Verdacht	30	0.30	0.40
leichte HD	35	0.35	0.75
mittlere HD	15	0.15	0.90
schwere HD	10	0.10	1.00

Geben Sie Modalwert, Median und arithmetisches Mittel an.

1. Modalwert: leichte HD, Median und arithmetisches Mittel: lassen sich nicht angeben
2. Modalwert: leichte HD, Median: leichte HD, arithmetisches Mittel: lässt sich nicht angeben
3. Modalwert: 10, Median: 15, arithmetisches Mittel: 20
4. Modalwert: HD-Verdacht, Median: HD-Verdacht, arithmetisches Mittel: lässt sich nicht angeben
5. Modalwert: 35, Median: 15, arithmetisches Mittel: lässt sich nicht angeben

Aufgabe 14:

Zwei Studenten versuchen unabhängig voneinander dieselbe Statistik–Aufgabe zu lösen, wobei jeder mit einer Wahrscheinlichkeit von 0.6 die richtige Lösung erarbeitet.

Wie groß ist die Wahrscheinlichkeit, dass wenigstens einer der Studenten das richtige Ergebnis findet?

1. 0.42
2. 0.96
3. 0.68
4. 0.84
5. 1.20

Aufgabe 15:

Zur regelmäßigen Qualitätsbeurteilung von Hamburgern in einer Fast–Food–Filiale lagen für einen Monat folgende Beschwerdegründe vor, wobei von den Kunden jeweils nur der Hauptbeschwerdegrund genannt werden konnte:

1, 2, 3, 1, 1, 1, 3, 2, 4, 1, 1, 5, 1, 2, 5, 1, 1, 1, 1, 1, 2, 3, 2, 1, 3, 1, 2, 1, 1, 2, 1, 1, 1, 1, 2

Diese Angaben haben die folgenden Bedeutungen: 1 – zu klein, 2 – zu fettig, 3 – zu viel Brot, 4 – zu salzig, 5 – zu trocken.

Mit welcher relativen Häufigkeit der Kundenbeschwerden (in %) waren die Hamburger in jenem Monat entweder zu klein oder zu fettig?

1. 58%
2. 99%
3. 94%
4. 80%
5. 75%

Aufgabe 16:

Wenn eine Lieferung mit einer Wahrscheinlichkeit von 0.1 unvollständig ist und, unabhängig davon, mit einer Wahrscheinlichkeit von 0.2 verspätet eintrifft, wie groß ist dann die Wahrscheinlichkeit, dass die Lieferung vollständig und termingerecht eintrifft?

1. 1.10
2. 0.30
3. 0.92
4. 0.68
5. 0.72

Aufgabe 17:

Der Milchfettgehalt X von Kühen einer Population sei eine normalverteilte Zufallsgröße mit $\mu = 3.7\%$ und $\sigma = 0.3\%$. Um einen züchterischen Fortschritt zu erreichen, sollen die Tiere mit niedrigen Leistungen geschlachtet werden, sodass nur 50% der Population zur Zucht selektiert werden.

Unter welcher Schranke für den Milchfettgehalt muss ein Tier liegen, das geschlachtet werden soll?

1. 3.70%
2. 3.39%
3. 3.82%
4. 3.10%
5. 3.63%

Aufgabe 18:

Ein Boxplot ist eine geeignete Möglichkeit der grafischen Darstellung von

1. Kennzahlen für nominale Merkmale
2. relativen Häufigkeiten
3. kumulierten Häufigkeiten
4. verschiedenen Streuungsmaßen
5. Quartilen sowie Minimum und Maximum

Aufgabe 19:

Bei Bauer Meier ist davon auszugehen, dass am Tag der Schlachtung noch 50% seiner Schweine Antibiotika–Rückstände aufweisen, die über dem zulässigen Grenzwert liegen. Die verwendete Nachweistechnik ist aber leider nur in der Lage, ein Überschreiten des Grenzwertes mit einer Wahrscheinlichkeit von 0.8 zu entdecken. Wenn der Grenzwert nicht überschritten ist, wird dies dagegen sicher erkannt, mit falsch–positiven Ergebnissen ist nicht zu rechnen.

Mit welcher Wahrscheinlichkeit wird bei Anwendung dieses Verfahrens eine Grenzwertüberschreitung angezeigt, wenn ein zufällig aus Meiers Lieferung gezogenes Schwein untersucht wird?

1. 0.5
2. 0.4
3. 0.3
4. 0.8
5. 1.0

Aufgabe 20:

Wie viel Prozent der Beobachtungen liegen bei einem normalverteilten Merkmal im Bereich von $(\mu - 2 \cdot \sigma)$ bis $(\mu + \sigma)$?

1. 93.32%
2. 95.45%
3. 86.64%
4. 68.27%
5. 81.86%

Pflichtveranstaltung Biometrie **1. Leistungsnachweis SS 1999**

FB Veterinärmedizin, FU Berlin 6. 6. 1999

Aufgabe 1:

In einer Schulklasse wurden die Zähne von 10 Kindern auf Kariesbefall untersucht. Die Anzahl der mit Karies befallenen Zähne für die einzelnen Kinder ist:

Kind-Nr.	1	2	3	4	5	6	7	8	9	10
Anzahl kariöser Zähne	0	1	1	2	2	1	0	4	1	1

Der Median der Daten ist

1. 5.5
2. 1.3
3. 1.5
4. 1
5. 4

Aufgabe 2:

Mit Hilfe eines Histogrammes werden nach Definition des Skriptes

1. relative Häufigkeiten in Wahrscheinlichkeiten umgerechnet.
2. qualitative Merkmale grafisch flächenproportional dargestellt.
3. die absoluten Klassenhäufigkeiten klassierter stetiger Merkmale direkt ablesbar grafisch dargestellt.
4. abhängige Ereignisse grafisch dargestellt.
5. klassierte stetige Merkmale grafisch so dargestellt, dass die relativen Klassenhäufigkeiten als flächenproportionale Rechtecke erscheinen.

Aufgabe 3:

Die Wahrscheinlichkeit, dass eine bestimmte Behandlung erfolgreich ist, sei 0.8. Die Behandlungserfolge bei verschiedenen Patienten seien unabhängig voneinander. Dann ist die Wahrscheinlichkeit, dass von

A) 5 Behandlungen alle erfolgreich sind: 0.8^5

B) 2 Behandlungen genau eine erfolgreich ist: $2 \cdot 0.8 \cdot 0.2$

C) 2 Behandlungen genau eine nicht erfolgreich ist: $2 \cdot 0.2$

D) 5 Behandlungen nur genau eine einzige erfolgreich ist: $5 \cdot 0.8 \cdot 0.2^4$

Welche der folgenden Lösungen trifft zu?

1. A, B und C sind richtig.

2. A, B und D sind richtig.

3. nur C ist richtig.

4. nur C und D sind richtig.

5. nur A und D sind richtig.

Aufgabe 4:

Da Sie nicht gerne beim Würfeln verlieren, haben Sie sich einen Würfel präpariert, der anstelle der Eins und der Zwei jeweils sechs Augen zeigt.

Wie sieht nun die Menge der Elementarereignisse aus?

1. $\{3\}$ $\{4\}$ $\{5\}$ $\{6\}$ $\{6\}$ $\{6\}$

2. $\{3\}$ $\{4\}$ $\{5\}$ $\{6\}$

3. der Würfel ist gefälscht, es gibt keine Elementarereignisse

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

5. $\{6\}$ $\{6\}$ $\{3\}$ $\{4\}$ $\{5\}$ $\{6\}$

Aufgabe 5:

Welche Aussage trifft **nicht** zu?

Der arithmetische Mittelwert $\overline{x}$ ist

1. Ausreißer–empfindlich.

2. stets größer als der Median.

3. bei symmetrisch verteilten, metrisch skalierten Merkmalen das Lagemaß mit dem höchsten Informationsgehalt, verglichen mit dem Modalwert oder dem Median.

4. als Lagemaß nur bei metrisch skalierten Merkmalen sinnvoll.

5. ein Lagemaß.

Aufgabe 6:

Bei einer Standardnormalverteilung liegen 5% aller Messwerte unterhalb des Wertes -1.645. Dann liegen 95% der Messwerte unterhalb von

1. 1.645

2. 1.96

3. -0.645

4. 0.95

5. 0.05

Aufgabe 7:

Der Heilungserfolg einer Therapie betrage unter den betroffenen Patienten $p = 0.8$. Wie groß ist die Wahrscheinlichkeit, dass von 25 zufällig herausgegriffenen Patienten mindestens 24 geheilt werden?

1. 0.8^{24}
2. $0.8^{25} + 0.8^{24}$
3. $0.8^{25} + 25 \cdot 0.8^{24} \cdot 0.2$
4. $25 \cdot 0.8^{24} \cdot 0.2$
5. $0.8^{25} + 0.8^{24} \cdot 0.2$

Aufgabe 8:

Die Gewichte von Ferkeln im Alter von 2 Monaten seien normalverteilt mit einem Erwartungswert von $\mu = 12$ kg und einer Standardabweichung von $\sigma = 2$ kg. Wie viel Prozent der Ferkelgewichte kann man im Bereich zwischen 10 kg und 14 kg erwarten?

1. 85%
2. 50%
3. 95.45%
4. 99.73%
5. 68.27%

Aufgabe 9:

Welche Merkmalstypen und beschreibenden Maßzahlen passen gut zusammen?

1. Median und Quartilsbreite zu einem metrisch skalierten, quantitativen Merkmal.
2. Arithmetischer Mittelwert und Variationskoeffizient zu einem nominal skalierten Merkmal.
3. Median und Quartilsbreite zu einem nominal skalierten Merkmal.
4. Arithmetischer Mittelwert und Varianz zu einem ordinal skalierten Merkmal.
5. Die Angabe des geometrischen Mittels und der Anzahl der verschiedenen Ausprägungen zu einem metrisch skalierten, qualitativen Merkmal.

Aufgabe 10:

Anhand der empirischen Verteilungsfunktion kann man ablesen:

1. arithmetischen Mittelwert und Median
2. sämtliche Lagemaße
3. den Variationskoeffizienten

4. Median, oberes und unteres Quartil
5. arithmetischen Mittelwert und Standardabweichung

Aufgabe 11:

Bei einer größeren Untersuchung von $n = 400$ Kühen wurde die Milchleistung dieser Kühe erhoben. Dabei wurde festgestellt, dass bei 40% der Kühe die Milchleistung zwischen 30 Litern (inklusive) und 35 Litern (exklusive) lag. Die absolute Häufigkeit der Kühe, die in diese Milchleistungsklasse fallen, war

1. 16
2. 40
3. 200
4. 120
5. 160

Aufgabe 12:

Das Merkmal Hinken (H) bei einem Hund kann ein Hinweis auf das Vorliegen einer Hüftgelenksdysplasie (HD) sein.

Insgesamt kommt die Diagnose "HD" bei Hunden mit einer Wahrscheinlichkeit von $P(\text{HD}) = 0.06$ vor. In der Hundegesamtheit kommt Hinken mit einer Häufigkeit von 30% vor, also ist $P(\text{H}) = 0.3$. Und 65% der Hunde mit HD hinken, demnach ist $P(\text{H}|\text{HD}) = 0.65$.

Gesucht ist die Wahrscheinlichkeit, dass bei auftretendem Hinken tatsächlich eine HD vorliegt, also $P(\text{HD}|\text{H})$.

1. 0.2
2. 0.13
3. 0.039
4. 0.46
5. 0.78

Aufgabe 13:

Eine Beobachtungsreihe soll aus 17 unterschiedlichen, der Größe nach geordneten Messwerten $x_{[1]}, \ldots, x_{[17]}$ bestehen. Der Median dieser Beobachtungen ist definiert als

1. $x_{[8.5]}$
2. $x_{[9]}$
3. $0.5 \cdot (x_{[8]} + x_{[9]})$
4. $x_{[8]}$
5. $0.5 \cdot (x_8 + x_9)$

Aufgabe 14:

Bei einer Vorsorgeuntersuchung waren 15% der untersuchten Personen herzkrank und 15% lungenkrank, 80% hatten keine der beiden Krankheiten.

Wie hoch war der Anteil der untersuchten Personen, die sowohl herz- als auch lungenkrank waren?

1. 10%
2. 5%
3. 15%
4. 20%
5. 2.25%

Aufgabe 15:

Bei einer Infektionskrankheit von Schäferhunden verlaufen 40% der Fälle **klinisch unauffällig**. Wie groß ist die Wahrscheinlichkeit, dass von zwei infizierten Schäferhunden mindestens einer **klinisch auffällig** erkrankt?

1. $2 \cdot 0.6 \cdot 0.4$
2. $1 - 0.4^2$
3. 0.4^2
4. 0.6^2
5. $1 - 0.6^2$

Aufgabe 16:

Bei dem BSE-Beispiel im Skript (S.107) werden außer der Tieridentifikation neun verschiedene Parameter erhoben. Ordnen Sie den folgenden Begriffen (A) bis (D) die entsprechende Angabe aus (1) bis (5) zu.

Begriff: (A) Beobachtungseinheit, (B) Grundgesamtheit, (C) Merkmal, (D) Ausprägung
Angabe: (1) Tier 560-37/5, (2) 27 Liter, (3) Milchleistung, (4) BSE-verdächtige Kühe in Großbritannien, (5) BSE

1. $(A) - (1)$, $(C) - (2)$, $(D) - (3)$, $(B) - (4)$
2. $(C) - (1)$, $(D) - (2)$, $(A) - (3)$, $(B) - (4)$
3. $(A) - (1)$, $(D) - (2)$, $(C) - (3)$, $(B) - (4)$
4. $(A) - (1)$, $(D) - (2)$, $(C) - (3)$, $(B) - (5)$
5. $(D) - (1)$, $(C) - (2)$, $(A) - (3)$, $(B) - (5)$

Aufgabe 17:

Zu den schon bekannten 5 Professoren am Fachbereich im Alter von 53, 44, 55, 64 und 58 Jahren gesellt sich auf einer Party der vor langer Zeit schon emeritierte 86-jährige Kollege Methusalem.

Wie wirkt sich das auf die statistischen Maßzahlen aus?

1. Median und arithmetischer Mittelwert bleiben unverändert
2. der Median bleibt unverändert, der arithmetische Mittelwert steigt an
3. die Quartile bleiben unverändert und damit auch der Quartilsabstand
4. Varianz und Spannweite steigen an
5. Minimum und Median steigen an

Aufgabe 18:

Für das Merkmal X:"Anzahl der Ferkel pro Wurf" wurden für $n = 100$ Würfe von Zuchtsauen zu den Werten x_i die Häufigkeiten n_i ermittelt:

x_i:	4	5	6	7	8	9	10	11	12	13	14	15	16
n_i:	1	3	0	8	11	18	20	16	11	5	4	2	1

Geben Sie das 1. Quartil, das 2. Quartil (Median), das 3. Quartil und den Quartilsabstand dieser Verteilung an.

1. $x_{0.25} = 9$, $x_{0.5} = 10$, $x_{0.75} = 11$ und $Q = 2$
2. $x_{0.25} = 10$, $x_{0.5} = 11$, $x_{0.75} = 12$ und $Q = 2$
3. $x_{0.25} = 8$, $x_{0.5} = 9$, $x_{0.75} = 10$ und $Q = 2$
4. $x_{0.25} = 9$, $x_{0.5} = 10$, $x_{0.75} = 11$ und $Q = 12$
5. $x_{0.25} = 9$, $x_{0.5} = 10$, $x_{0.75} = 11$ und $Q = 3$

Aufgabe 19:

In einem Patientengut betrage der Anteil der Lungenkranken 60%, der Anteil der Herzkranken 50% und der Anteil derjenigen, die an beiden Krankheiten leiden, 30%.

Der Anteil der Kranken, die **nur genau eine** der beiden Krankheiten haben, beträgt

1. 30%
2. 60%
3. 20%
4. 80%
5. 50%

Aufgabe 20:

Die Gewichte von Ferkeln im Alter von 20 Tagen seien normalverteilt mit einem Erwartungswert von $\mu = 2500$g und einer Standardabweichung von $\sigma = 600$g. Wie groß ist die Wahrscheinlichkeit, dass ein zufällig gezogenes Ferkel schwerer als 2500 g ist?

1. 0.25
2. 0.05
3. 0
4. 0.75
5. 0.5

Pflichtveranstaltung Biometrie **2. Leistungsnachweis SS 1998**

FB Veterinärmedizin, FU Berlin 16. 7. 1998

Aufgabe 1:

Zwei beobachtete Merkmale mit jeweils zwei Ausprägungen werden auf ihre statistische Unabhängigkeit hin überprüft. Aus dem folgenden Untersuchungsergebnis

	$Y=0$	$Y=1$	$\sum$
$X=0$	25	10	
$X=1$	25	10	
$\sum$			

errechnet man ein empirisches χ^2 von

1. 14.7
2. 0
3. 20.3
4. 5.3
5. 9.9

Aufgabe 2:

Die Massezunahme X in einer festgelegten Zeitspanne bei Mastschweinen sei eine normalverteilte Zufallsvariable mit der bekannten Standardabweichung $\sigma = 0.5$ kg. Aus einer Stichprobe von $n = 100$ Mastschweinen bestimmte man die mittlere Massezunahme $\overline{x} = 1.5$ kg. Welches 0.95-Konfidenzintervall kann man für die unbekannte mittlere Massezunahme μ berechnen?

1. $[1.30; 1.70]$
2. $[1.38; 1.62]$
3. $[1.45; 1.55]$
4. $[1.40; 1.60]$
5. $[1.42; 1.58]$

Aufgabe 3:

In einer umfangreichen Studie wurde der Zusammenhang zwischen dem Alter (X, angegeben in Jahren) und dem Blutdruck (Y) bei Frauen untersucht. Aus den Studiendaten ergab sich ein Produkt–Moment–Korrelationskoeffizient von $r = 0.89$. Um die Beziehung zwischen den beiden Merkmalen zu beschreiben, wurde die folgende Regressionsgerade für den Blutdruck in Abhängigkeit vom Alter der Frau geschätzt:

$\hat{y} = 80 + 1.5 \cdot x$

Welche der folgenden Aussagen ist **richtig**?

1. Bei 30-jährigen Frauen erwartet man im Mittel einen Blutdruckwert von 125.
2. Alle 30-jährigen Frauen haben einen Blutdruckwert von 110.

3. Jede 30-jährige Frau hat einen Blutdruckwert von genau 81.5.
4. Bei 40-jährigen Frauen erwartet man im Mittel einen Blutdruckwert von 120.
5. Jede Frau von 30 Jahren hat einen Blutdruckwert von genau 125.

Aufgabe 4:

Das Merkmal X sei in der Grundgesamtheit normalverteilt mit bekannter Varianz σ^2. In einer Simulationsstudie sollen mit insgesamt 120 aus dieser Grundgesamtheit simulierten Stichproben des Umfangs n ebenso viele 0.95-Konfidenzintervalle für den Erwartungswert μ berechnet werden. Wie viele solcher 0.95-Konfidenzintervalle darf man erwarten, die μ nicht überdecken?

1. 8
2. 5
3. 3
4. 0
5. 6

Aufgabe 5:

Eine nach Rasse, Alter und Ort abgegrenzte Tierpopulation gelte grundsätzlich als gesund, wenn wenigstens 80% der Tiere gesund sind. Die Beurteilung des Gesundheitsstatus so einer Tierpopulation kann mit Hilfe eines statistischen Tests erfolgen, mit dem die Nullhypothese geprüft wird, die Population sei gesund. Der Wahrscheinlichkeit, diese Population irrtümlich als gesund einzustufen, entspricht dann

1. die Irrtumswahrscheinlichkeit α.
2. die Wahrscheinlichkeit β für den Fehler 2. Art.
3. keine der genannten Wahrscheinlichkeiten.
4. die Sicherheitswahrscheinlichkeit $(1-\alpha)$.
5. die Wahrscheinlichkeit α für den Fehler 1. Art.

Aufgabe 6:

Die Herzmasse von Schlachtschweinen sei normalverteilt mit einem Erwartungswert von $\mu =$ 450g und einer Standardabweichung von $\sigma = 20$g. Wie ist dann das arithmetische Mittel $\overline{X}$ aus Stichproben von jeweils 25 zufällig ausgewählten Schweineherzen verteilt?

Es ist

1. immer genau 450 g.
2. ist normalverteilt mit dem Erwartungswert 450 g und einer Standardabweichung von 4g.
3. binomialverteilt mit $n = 25$ und $p = 0.5$, also mit einem Erwartungswert von 12.5 g und einer Standardabweichung von $\sqrt{6.25}$.

4. normalverteilt mit dem Erwartungswert 450 g und einer Standardabweichung von 0.8 g.
5. normalverteilt mit dem Erwartungswert 450 g und einer Standardabweichung von 20 g.

Aufgabe 7:

Welche Aussage über Vertrauensintervalle zur Schätzung des unbekannten Erwartungswertes eines normalverteilten Merkmals ist **richtig**?

1. Bei einer vierfachen Varianz gegenüber der Ausgangssituation wird das Vertrauensintervall doppelt so breit, sofern alle übrigen Vorgaben konstant bleiben.
2. Mit abnehmender Vertrauenswahrscheinlichkeit verbreitert sich das zugehörige Vertrauensintervall.
3. Mit zunehmender Vertrauenswahrscheinlichkeit verkürzt sich das zugehörige Vertrauensintervall.
4. Mit zunehmendem Stichprobenumfang verbreitert sich das zugehörige Vertrauensintervall.
5. Mit abnehmendem Stichprobenumfang verkürzt sich das zugehörige Vertrauensintervall.

Aufgabe 8:

Wenn die Kovarianz zweier Merkmale gleich dem Produkt der beiden Standardabweichungen ist, handelt es sich um

1. eine vollständige positive lineare Korrelation.
2. eine schwache negative lineare Korrelation.
3. eine schwache positive lineare Korrelation.
4. keinen linearen Zusammenhang.
5. eine vollständige negative lineare Korrelation.

Aufgabe 9:

Aus einer Lieferung von Futtermittelpaketen soll eine Stichprobe gezogen werden, um abzuschätzen, welches mittlere Gewicht μ die Pakete der Lieferung haben. Man kann unterstellen, dass die Gewichte normalverteilt sind und eine Standardabweichung von $\sigma = 5$g haben. Wie viele Pakete muss man mindestens für die Stichprobe ziehen, damit das 95%-Konfidenzintervall für den Erwartungswert μ nicht breiter als 4 g wird?

1. 27 Pakete
2. 7 Pakete
3. 25 Pakete
4. 5 Pakete
5. 16 Pakete

Aufgabe 10:

Bei einer Untersuchung zum Zusammenhang von zwei Krankheitssymptomen A und B stellt man bei 60 Erkrankten fest, dass insgesamt 40 das Symptom A zeigen und insgesamt 30 das Symptom B, wobei 25 Erkrankte beide Symptome haben.

Mit einem χ^2-Test soll geprüft werden, ob die beiden Symptome unabhängig voneinander auftreten. Wie sieht das Ergebnis dieses Tests bei einer Irrtumswahrscheinlichkeit von $\alpha = 0.05$ für einen Fehler 1. Art aus und wie begründen Sie diese Entscheidung? (Hinweis: $\chi^2_{1;0.05} = 3.84$)

1. Die Teststatistik ist kleiner als der kritische Wert 3.84, deshalb kann die Unabhängigkeitshypothese nicht abgelehnt werden.
2. Die Teststatistik ist kleiner als der kritische Wert 3.84, deshalb wird die Unabhängigkeitshypothese verworfen.
3. Die Teststatistik ist größer als Null, deshalb wird die Unabhängigkeitshypothese verworfen.
4. Die Teststatistik ist größer als der kritische Wert 3.84, deshalb wird die Unabhängigkeitshypothese verworfen.
5. Die Teststatistik ist nicht identisch mit dem kritischen Wert 3.84, deshalb wird die Unabhängigkeitshypothese verworfen.

Aufgabe 11:

Bei 30 Merkmalsträgern wurden die Ausprägungen zweier verschiedener Merkmale gemessen. Die Berechnung der Korrelationskoeffizienten ergab für den Spearmanschen Rangkorrelationskoeffizienten $r_S = -0.98$ und für den Produkt–Moment–Korrelationskoeffizienten $r = -0.68$. Welche Aussage lässt sich auf dieser Grundlage über den Zusammenhang der beiden Merkmale machen?

1. Das Ergebnis weist auf einen nichtlinear ansteigenden Zusammenhang hin.
2. Das Ergebnis weist darauf hin, dass weder ein linearer noch ein nichtlinearer Zusammenhang besteht.
3. Das Ergebnis weist auf einen linear abfallenden Zusammenhang hin.
4. Das Ergebnis weist auf einen linear ansteigenden Zusammenhang hin.
5. Das Ergebnis weist auf einen nichtlinear abfallenden Zusammenhang hin.

Aufgabe 12:

Nehmen Sie einmal an, bei Rindern im Alter von 5 Jahren seien die Gewichte normalverteilt mit einer Standardabweichung von $\sigma = 40$ kg. Für eine Gruppe von 100 Tieren in diesem Alter erhält man ein mittleres Gewicht von $\overline{x} = 587$ kg.

Bitte geben Sie für solche Tiere das 90%-Konfidenzintervall zur Schätzung des erwarteten mittleren Gewichtes an.

1. $[579.160; 594.840]$

2. $[586.216; 587.784]$
3. $[586.342; 587.568]$
4. $[576.696; 597.304]$
5. $[580.420; 593.580]$

Aufgabe 13:

Der Pianist Ludwig B., verstorben im Jahre 1976, war berühmt für seine explosive und mitreißende Spielweise. Seine Gattin Ludmilla gab nach seinem Tode eine Biographie über ihren Mann heraus, die schnell ein Erfolg auf dem Buchmarkt wurde.

Hier ein kleiner Auszug aus dem Buch, in dem Ludmilla B. unter anderem für einige öffentliche Konzerte ihres Gatten die Anzahl der zerstörten Klaviertasten und die Länge des anschließenden Applauses verzeichnet hat:

Datum des Konzerts	20.2.57	3.7.60	5.5.65	1.2.72	1.1.74
Anzahl X der zerstörten Tasten	7	4	11	5	3
Dauer Y des Applauses in Minuten	16	10	24	12	8

Schätzen Sie die Koeffizienten $\hat{a}$ und $\hat{b}$ der linearen Regressionsbeziehung von Y auf X. Wie lautet die geschätzte Gerade?

1. $\hat{y} = 2 + 2 \cdot x$
2. $\hat{y} = 14 - 2 \cdot x$
3. $\hat{y} = -2 + 8 \cdot x$
4. $\hat{y} = -23.92 + 6.32 \cdot x$
5. $\hat{y} = 12 + 2 \cdot x$

Aufgabe 14:

Um die HD-Beurteilungen zweier Tierärzte zu überprüfen, werden $n = 6$ Hunde von beiden Ärzten untersucht. Die Stärke der HD wird mit Hilfe der folgenden Skala beurteilt:

HD-frei: 1, HD-Verdacht: 2, leichte HD: 3, mittlere HD: 4, schwere HD: 5

Die folgende Tabelle zeigt die Beurteilungen der beiden Ärzte:

Hund:	Stromer	Benji	Tapsi	Kaspar	Jenny	Hasso
Tierarzt A:	2	4	1	5	3	3
Tierarzt B:	1	3	2	4	4	3

Welche der folgenden Möglichkeiten gibt hierzu den passenden Korrelationskoeffizienten an?

1. $r = 0.40$
2. $r_S = 0.76$
3. $r = 0.60$
4. $r_S = 0.24$
5. $r_S = 0.86$

Aufgabe 15:

Beim Prüfen statistischer Hypothesen bezeichnet α die Wahrscheinlichkeit für einen Fehler 1. Art und β die Wahrscheinlichkeit für einen Fehler 2. Art. Welche Aussage ist **falsch**?

1. α ist die Wahrscheinlichkeit, dass die Nullhypothese nicht zutrifft, obwohl sie angenommen wurde.
2. $1-\beta$ ist die Wahrscheinlichkeit, mit der man die Nullhypothese ablehnt, wenn sie falsch ist.
3. $1-\alpha$ bezeichnet die Wahrscheinlichkeit, mit der man zu keiner Ablehnung der Nullhypothese kommt, wenn diese zutrifft.
4. β ist die Wahrscheinlichkeit, mit der man zu keiner Ablehnung der Nullhypothese kommt, obwohl die Alternativhypothese gilt.
5. $1-\beta$ ist die Wahrscheinlichkeit für das Verwerfen der Nullhypothese, wenn die Alternativhypothese gilt.

Aufgabe 16:

Zu einem statistischen Test werden die Prüf- und Alternativhypothese sowie die Wahrscheinlichkeit für den Fehler 1. Art festgelegt. Die Wahrscheinlichkeit für den Fehler 2. Art nimmt zu, wenn im Vergleich zum ersten Planungsansatz

1. die Teststatistik (Prüfgröße) weniger streut.
2. der Stichprobenumfang erhöht wird.
3. der Stichprobenumfang reduziert wird.
4. der Stichprobenumfang verdoppelt wird.
5. der Stichprobenumfang unverändert bleibt.

Aufgabe 17:

Bei der Verkostung mehrerer Sorten eines Lebensmittels vergeben 2 Verkoster jeweils eine Geschmacksnote. Die Bewertungsskala reicht von 0 bis 10. Ein Ergebnis einer solchen Verkostung zeigt folgende Tabelle:

Sorte	1	2	3	4	5
1. Verkoster	3	4	6	2	8
2. Verkoster	4	3	6	5	7

Welcher der nachstehenden Werte für den Rangkorrelationskoeffizienten r_S ist der richtige?

1. $r_S = 0.50$
2. $r_S = 0.70$
3. $r_S = 0.55$
4. $r_S = 0.60$
5. $r_S = 0.65$

Aufgabe 18:

Welche der folgenden Aussagen ist **falsch**?

1. Die Wahrscheinlichkeit für den Fehler 2. Art lässt sich verringern, wenn man den Stichprobenumfang erhöht.
2. Wenn man die Nullhypothese verwirft, obwohl sie richtig ist, begeht man den Fehler 1. Art.
3. Wenn man die Nullhypothese nicht verwirft, obwohl sie zutrifft, begeht man den Fehler 2. Art.
4. Die Wahrscheinlichkeit für den Fehler 1. Art wird mit α bezeichnet und vor Versuchsbeginn festgelegt.
5. Es ist nicht möglich, α und β gleichzeitig zu minimieren.

Aufgabe 19:

Bei dem bisher verwendeten Standardpräparat für eine bestimmte Indikation muss man leider mit einer Wahrscheinlichkeit von 0.9 mit unerwünschten Nebenwirkungen rechnen. Eine Firma hat nun ein neues Präparat entwickelt, bei dem die Wahrscheinlichkeit, dass solche Nebenwirkungen auftreten, deutlich geringer sein soll.

Um diese Behauptung der Firma kritisch zu prüfen, soll mit dem neuen Präparat eine Studie an 20 Probanden durchgeführt werden. Zu prüfen ist die Hypothese $H_0 : p \geq 0.9$, wobei man bereit ist, eine Irrtumswahrscheinlichkeit für einen Fehler 1. Art von bis zu 5% in Kauf zu nehmen.

Bei wie vielen Patienten dürfen die Nebenwirkungen **höchstens** auftreten, damit man die Nullhypothese ablehnen kann? Und mit welcher Wahrscheinlichkeit β für einen Fehler 2. Art wird man die Nullhypothese nicht ablehnen können, wenn Nebenwirkungen bei dem neuen Präparat tatsächlich nur mit einer Wahrscheinlichkeit von 0.7 auftreten? (Skript S.82 und S.83)

1. H_0 kann nicht abgelehnt werden, da schon die Wahrscheinlichkeit, dass bei 20 Patienten eine Nebenwirkung auftritt, größer als 5% ist.
2. höchstens 18 Patienten; $\beta = 0.323$
3. höchstens 16 Patienten; $\beta = 0.043$
4. höchstens 15 Patienten; $\beta = 0.238$
5. höchstens 15 Patienten; $\beta = 0.762$

Aufgabe 20:

Ein 95%-Vertrauensintervall für den Mittelwert eines normalverteilten Merkmals mit bekannter Standardabweichung der Einzelwerte $\sigma = 20$ hat eine Breite von $2 \cdot 1.96$. Welcher Stichprobenumfang wurde hier berücksichtigt?

1. $n = 196$
2. $n = 100$

3. $n = 400$
4. $n = 20$
5. $n = 1$

Pflichtveranstaltung Biometrie **2. Leistungsnachweis SS 1999**

FB Veterinärmedizin, FU Berlin 22. 7. 1999

Aufgabe 1:

Das Schlachtgewicht einer Entenrasse sei als normalverteilt bekannt mit einer Standardabweichung von $\sigma = 60$g. Aus einer Stichprobe von $n = 16$ Tieren wurde der unbekannte Erwartungswert μ der Gewichte durch einen arithmetischen Mittelwert von $\overline{x} = 2386$g geschätzt. Bestimmen Sie das zugehörige 99%-Konfidenzintervall für μ.

1. $2386 \pm 2.576 \cdot \frac{60}{\sqrt{16}}$
2. $2386 \pm 1.96 \cdot \frac{60}{\sqrt{16}}$
3. $2386 \pm 1.96 \cdot \frac{60}{16}$
4. $2386 \pm 2.576 \cdot \frac{60}{16}$
5. $2386 \pm 2.576 \cdot 60$

Aufgabe 2:

Es soll ein Konfidenzintervall für den Erwartungswert einer Normalverteilung bei bekannter Standardabweichung σ berechnet werden. Der Stichprobenumfang sei n. Welche der fünf folgenden Möglichkeiten ergibt das **breiteste** Konfidenzintervall?

1. Ein 95%-Konfidenzintervall mit $n = 100$ und $\sigma = 100$.
2. Ein 99%-Konfidenzintervall mit $n = 10$ und $\sigma = 10$.
3. Ein 90%-Konfidenzintervall mit $n = 5$ und $\sigma^2 = 100$.
4. Ein 99%-Konfidenzintervall mit $n = 5$ und $\sigma = 10$.
5. Ein 95%-Konfidenzintervall mit $n = 100$ und $\sigma^2 = 100$.

Aufgabe 3:

Welche der folgenden Aussagen trifft **nicht** zu?

1. Die Regressionsgerade einer linearen Regression geht immer durch den Schwerpunkt $(\overline{x}, \overline{y})$.
2. r^2 ist immer größer oder gleich Null.
3. Die Berechnung eines Produkt–Moment–Korrelationskoeffizienten setzt voraus, dass die Daten metrisch skaliert sind.
4. Für ordinalskalierte Daten kann der Spearman'sche Rangkorrelationskoeffizient berechnet werden.
5. r ist immer größer oder gleich Null.

Aufgabe 4:

Um zu untersuchen, ob das Auftreten einer Krankheit in Rinderherden mit der Größe der Herde zusammenhängt, wurden für 5 Herden die beiden Merkmale "Herdengröße" X und "Anzahl erkrankter Tiere" Y erhoben:

x_i	160	140	100	190	150
y_i	13	17	10	18	15

Es wurde bereits berechnet, dass eine Kovarianz von $s_{xy} = 81.5$ vorliegt, und dass die Standardabweichungen $s_x = 32.7$ für X sowie $s_y = 3.2$ für Y betragen. Geben Sie bitte den Produkt–Moment–Korrelationskoeffizienten r und den Rangkorrelationskoeffizienten r_s an.

1. $r = 0.993, r_s = 0.6$
2. $r = 0.007, r_s = 0.6$
3. $r = 0.779, r_s = 0.4$
4. $r = 0.779, r_s = 0.6$
5. $r = 0.007, r_s = 0.4$

Aufgabe 5:

Welche Aussage trifft **nicht** zu?

Für die Standardnormalverteilung gilt:

1. der Erwartungswert ist 0.
2. die Varianz ist 1.
3. die Standardabweichung ist 1.
4. die Spannweite ist 2.
5. der Median ist 0.

Aufgabe 6:

Wenn die Kovarianz zweier Merkmale dem negativen Produkt der beiden Standardabweichungen entspricht, dann handelt es sich um

1. keinen linearen Zusammenhang.
2. eine vollständige negative lineare Korrelation.
3. eine vollständige positive lineare Korrelation.
4. eine schwache positive lineare Korrelation.
5. eine schwache negative lineare Korrelation.

Aufgabe 7:

Um den Erwartungswert der Schlachtgewichte einer Entenrasse zu schätzen, wollen Sie auf Basis einer Stichprobenauswertung ein 95%-Konfidenzintervall angeben. Dabei unterstellen Sie, dass diese Gewichte normalverteilt sind und eine Standardabweichung von $\sigma = 60$g haben. Wie groß muss Ihre Stichprobe mindestens sein, wenn die Länge des Konfidenzintervalls nicht mehr als 30 g betragen soll?

1. mindestens 7
2. mindestens 16
3. mindestens 36
4. mindestens 62
5. mindestens 8

Aufgabe 8:

In einem Versuch werden 100 kranke Tiere zufällig in zwei Behandlungsgruppen aufgeteilt. Eine Gruppe wird mit dem Präparat A, die andere mit dem Präparat B behandelt. Den Heilungserfolg gibt die folgende Tabelle wieder:

	Heilung	keine Heilung	Summe
Präparat A	12	38	50
Präparat B	18	32	50
Summe	30	70	100

Der kritische Wert der χ^2-Verteilung für $\alpha = 0.05$ ist 3.84. Welche der folgenden Aussagen lässt sich bei dieser Irrtumswahrscheinlichkeit treffen?

1. Die Hypothese, Heilungserfolg und Art der Behandlung seien unabhängig, wird abgelehnt.
2. Die Hypothese, der Heilungserfolg sei abhängig vom eingesetzten Präparat, kann nicht verworfen werden.
3. Die Wahrscheinlichkeit, dass der Heilungserfolg vom Präparat abhängt, beträgt 95%.
4. Die Hypothese, der Heilungserfolg sei vom eingesetzten Präparat unabhängig, lässt sich nicht ablehnen.
5. Die Hypothese, der Heilungserfolg hänge nicht vom Präparat ab, wird verworfen.

Aufgabe 9:

Wie groß ist die Wahrscheinlichkeit, dass eine Beobachtung zwischen dem 1. Quartil und dem Median der entsprechenden Wahrscheinlichkeitsverteilung liegt?

1. 1.25
2. 0
3. $1 - 0.25$
4. 0.5
5. 0.25

Aufgabe 10:

Liegt eine Prüfgröße im Annahmebereich der Prüfgrößenverteilung, so formuliert man:

1. die Nullhypothese kann nicht verworfen werden.

2. die Alternativhypothese ist falsch.

3. die Alternativhypothese lässt sich mit Sicherheit ausschließen.

4. die Nullhypothese ist bewiesen.

5. es ist keine Testentscheidung möglich.

Aufgabe 11:

Das Wachstum von Bakterien auf Nährböden kann durch die Zugabe von Antibiotika gehemmt werden (Hemmstofftest). Folgende Daten wurden bei einer Versuchsreihe zu einem Antibiotikum ermittelt:

Dosis x_i (mg/ml)	5	10	15	20
Hemmung y_i(%)	25	35	55	85

$s_x^2 = 125/3$, $s_y^2 = 700$, $s_{xy} = 500/3$

Bestimmen Sie die Regressionsgerade $\hat{y} = \hat{a} + \hat{b}\,x$.

1. $\hat{y} = 20 + 2 \cdot x$

2. $\hat{y} = 0 + 4 \cdot x$

3. $\hat{y} = 25 + 2 \cdot x$

4. $\hat{y} = 15 + 4 \cdot x$

5. $\hat{y} = 0 + 2 \cdot x$

Aufgabe 12:

Ein statistischer Test dient

1. zum Prüfen der Nullhypothese.

2. zum Schätzen eines Parameters.

3. zur Berechnung der Wahrscheinlichkeit der Alternativhypothese.

4. zur Berechnung der Wahrscheinlichkeit der Nullhypothese.

5. zur Ermittlung des Betrages einer Differenz.

Aufgabe 13:

Wie groß ist bei Unterstellung einer Normalverteilung die Wahrscheinlichkeit, dass ein Beobachtungswert innerhalb des Schwankungsintervalls $\mu \pm 3 \cdot \sigma$ auftritt?

1. 0.9987

2. 0.9973

3. 0.5

4. 0

5. 0.0013

Aufgabe 14:

Ein Vertrauensintervall mit einer Vertrauenswahrscheinlichkeit von $1 - \alpha = 0.99$ für den Mittelwert eines normalverteilten Merkmals mit bekannter Standardabweichung der Einzelwerte $\sigma = 20$ hat eine Breite von $2 \cdot 2.576$.

Welcher Stichprobenumfang wurde hier berücksichtigt?

1. $n = 100$
2. $n = 1$
3. $n = 196$
4. $n = 400$
5. $n = 20$

Aufgabe 15:

Um die HD(Hüftgelenksdysplasie)–Beurteilungen zweier Tierärzte zu überprüfen, werden $n = 8$ Hunde von beiden Ärzten untersucht. Die Stärke der HD wird mit Hilfe der folgenden Skala beurteilt:

HD-frei: 1, HD-Verdacht: 2, leichte HD: 3, mittlere HD: 4, schwere HD: 5

Die folgende Tabelle zeigt die Beurteilungen der beiden Ärzte:

Hund:	1	2	3	4	5	6	7	8
Tierarzt A:	1	3	2	1	5	4	1	2
Tierarzt B:	2	1	3	1	5	4	2	3

Welche der folgenden Möglichkeiten gibt hierzu den passenden Korrelationskoeffizienten an (r_S: Rang–Korrelationskoeffizient nach Spearman, r: Produkt–Moment–Korrelationskoeffizient nach Pearson)?

1. $r_S = 0.679$
2. $r_S = 0.321$
3. $r_S = 0.905$
4. $r = 1.589$
5. $r = 0.750$

Aufgabe 16:

Welches der folgenden Intervalle umfasst alle Zahlenwerte, die der Steigungskoeffizient der Regressionsgeraden, b, annehmen kann?

1. $[-\infty, \infty]$
2. $[0, 1]$
3. $[0, \infty]$
4. $[1, \infty]$

5. $[-1, 1]$

Aufgabe 17:

Ein bestimmtes Lehrfach in der Tiermedizin ist mit einem Leistungstest abzuschließen. Die Erfolgsquote von Studenten, die sich einzeln darauf vorbereiten, liegt bei $p = 0.7$. Von 20 Studenten ist bekannt, dass sie sich in Gruppenarbeit darauf vorbereitet haben. Wie viele dieser Studenten müssten den Test mindestens bestanden haben, um bei einer Irrtumswahrscheinlichkeit (für den Fehler 1. Art) von $\alpha = 0.2$ (bei einseitiger Fragestellung) eine höhere Erfolgsquote der Gruppenarbeit zu bestätigen? Als Lösungshilfe sei auf die Tabelle in dem Skript S. 82 verwiesen.

1. 14
2. 17
3. 11
4. 16
5. 10

Aufgabe 18:

Welche Aussage trifft zu?

Die Wahrscheinlichkeit α für den Fehler 1. Art muss bei jedem statistischen Test grundsätzlich

1. aus den Daten berechnet werden.
2. bei der Versuchsplanung vorher festgelegt werden.
3. nach der Durchführung des Tests festgelegt werden.
4. spätestens nach der Berechnung der Prüfgröße gewählt werden.
5. so angepasst werden, dass man die gewünschte Entscheidung trifft.

Aufgabe 19:

In der Tabelle ist die Anzahl der Studenten angegeben, die bei den Professoren X.Y. und V.W. die Prüfung zum ersten Staatsexamen bestanden haben bzw. durchgefallen sind. Man teste bei einer Irrtumswahrscheinlichkeit von $\alpha = 0.05$ die Nullhypothese, dass die Anteile der Studenten, die durchgefallen sind, bei den zwei Professoren gleich groß sind (die kritische Grenze liegt bei $\chi^2_{1;0.05} = 3.84$).

	X.Y.	V.W.	
bestanden	45	20	65
nicht bestanden	5	5	10
	50	25	75

1. $\chi^2 = 1.4423$, die Nullhypothese kann abgelehnt werden.
2. $\chi^2 = 0$, die Nullhypothese kann nicht abgelehnt werden.

3. $\chi^2 = 0.0115$, die Nullhypothese kann nicht abgelehnt werden.
4. $\chi^2 = 1.4423$, die Nullhypothese kann nicht abgelehnt werden.
5. $\chi^2 = 0.0115$, die Nullhypothese kann abgelehnt werden.

Aufgabe 20:

Die grafische Darstellung, die man in der linearen Regressionsanalyse benutzt, um die Daten darzustellen, ist ein(e)

1. Punktwolke
2. Flussdiagramm
3. Stabdiagramm
4. Histogramm
5. Polygonzug