1 Aussagen und Häufigkeiten

Grundgesamtheit: räumlich und zeitlich abgegrenzte Menge von Untersuchungseinheiten

Merkmale oder Variable: Eigenschaften der Untersuchungseinheiten

Ausprägungen: Werte der Merkmale

Datenliste: Folge von Untersuchungseinheiten

Umfang: Anzahl der Elemente einer Datenliste

Ziel einer statistischen Untersuchung: Aussagen über die Untersuchungsobjekte

(1.3) Definition *Es sei A eine Aussage über ein Merkmal der Untersuchungsobjekte. Unter der absoluten Häufigkeit $h(A)$ versteht man die Anzahl der Untersuchungsobjekte, für die die Aussage A zutrifft. Unter der relativen Häufigkeit $f(A)$ versteht man den Anteil (Prozentsatz) der Untersuchungsobjekte, für die die Aussage A zutrifft.*

$$f(A) = \frac{h(A)}{n}$$

Sind A und B zwei beliebige Aussagen, so bedeuten:

$A \cup B$:	A oder B trifft zu.
	Mindestens eine der beiden Aussagen A, B trifft zu.
$A \cap B$:	A und B treffen zu.
	Beide Aussagen A und B treffen gleichzeitig zu.
A':	A trifft nicht zu.
	Das Gegenteil von A trifft zu.
$A \subseteq B$:	Wenn A, dann B.
	B ist eine umfassendere Aussage als A.

$A = \emptyset$:	A ist unmöglich.
	Die Aussage A trifft niemals zu.
$A = \Omega$:	A ist sicher.
	Die Aussage A trifft stets zu.

$$h(\emptyset) = 0, \quad f(\emptyset) = 0$$
$$h(\Omega) = n, \quad f(\Omega) = 1$$

Es seien A und B zwei beliebige Aussagen:

- Wenn $A \cap B = \emptyset$, so sind die Aussagen unvereinbar, die Aussagen können nicht gleichzeitig zutreffen, sie schließen einander aus.

- Wenn $A \cup B = \Omega$, so trifft stets mindestens eine der Aussagen A oder B zu. Die Aussagen A und B schöpfen gemeinsam alle Möglichkeiten aus.

Monotoniegesetz:

$$A \subseteq B \Rightarrow h(A) \leq h(B) \text{ und } f(A) \leq f(B)$$

Additionsgesetz:

$$A \cap B = \emptyset \Rightarrow \begin{cases} h(A \cup B) = h(A) + h(B), \\ f(A \cup B) = f(A) + f(B) \end{cases}$$

Siebformel:

$$h(A \cup B) = h(A) + h(B) - h(A \cap B),$$
$$f(A \cup B) = f(A) + f(B) - f(A \cap B)$$

2 Ereignisse und Wahrscheinlichkeiten

Stochastik : Zufallsexperimente

(2.2) Definition *Unter einem Zufallsexperiment versteht man ein grundsätzlich wiederholbares Experiment mit mehreren möglichen Ergebnissen. Die Versuchsergebnisse sind nicht vorhersagbar, sondern wechseln zufällig von Versuchswiederholung zu Versuchswiederholung.*

Zufallsexperimente können nur statistisch beschrieben werden.

Ereignisse: Aussagen über die Versuchsergebnisse

- Wenn die Aussage A für ein Versuchsergebnis zutrifft, so sagt man, daß das Ereignis A eingetreten ist oder beobachtet worden ist.
- Wenn A nicht zutrifft, dann sagt man, das Ereignis A ist nicht eingetreten.

Ist das Zufallsexperiment unter identischen Versuchsbedingungen beliebig oft reproduzierbar, so kann es statistisch ausgewertet werden, dh. die Versuchsergebnisse werden einer Datenanalyse unterworfen.

Statistische Gesetzmäßigkeiten:

Die relative Häufigkeiten von Ereignissen scheinen mit wachsendem Datenumfang einem festen Wert zuzustreben.

(2.3) Beispiel: Münzwurf

Münze wird n–mal geworfen.

Häufigkeit des Ereignisses $A =$„Die Zahlseite liegt oben"

$n =$	$h(A)$	$f(A)$	$\lvert f(A) - \frac{1}{2} \rvert$
10	3	$0,3$	$0,2$
100	47	$0,47$	$0,03$
500	254	$0,508$	$0,008$
1000	488	$0,488$	$0,012$
5000	2453	$0,4906$	$0,0094$

Die relativen Häufigkeiten konvergieren anscheinend gegen den Wert 0,5.

(2.4) Beispiel: Würfelwurf

Zwei Würfel werden geworfen.

Häufigkeit des Ereignisses $A =$„Die Augensumme ist mindestens 10"

n	$h(A)$	$f(A)$
10	0	0
100	19	$0,19$
500	80	$0,16$
1000	170	$0,17$
2000	349	$0,1745$

Auch hier scheinen die relativen Häufgkeiten gegen einen Grenzwert zu konvergieren.

(2.5) Empirisches Gesetz der grossen Zahl:
Wird ein Zufallsexperiment unter identischen Bedingungen wiederholt, und zwar so, daß die einzelnen Versuchsergebnisse einander nicht beeinflussen können, dann konvergieren die relativen Häufigkeiten mit wachsender Anzahl der Versuchswiederholungen gegen einen Grenzwert:

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

Der Grenzwert p hängt vom jeweiligen Ereignis A ab, daher schreibt man $p = P(A)$.

Empirisches Gesetz der großen Zahl:

Die langfristige durchschnittliche Häufigkeit ist als naturgesetzartige Eigenschaft des Zufallsexperiments ansehbar.

(2.6) Definition *Unter der Wahrscheinlichkeit $P(A)$ eines Ereignisses A versteht man den Grenzwert der relativen Häufigkeiten $f_n(A)$.*

Wahrscheinlichkeiten sind nichts anderes als idealisierte relative Häufigkeiten.

Daher gelten für Wahrscheinlichkeiten gleiche Rechengesetze:

(1)	$0 \leq P(A) \leq 1$	
(2)	$P(\emptyset) = 0,\ P(\Omega) = 1$	
(3)	$A \subseteq B \Rightarrow P(A) \leq P(B)$	**Monotoniegesetz**
(4)	$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$	**Additionsgesetz**
(5)	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	**Siebformel**

Zwei Wege, um eine Wahrscheinlichkeit zu bestimmen:

Statistische Methoden: Die relative Häufigkeit $f(A)$ des Ereignisses ist ein Schätzer für die unbekannte Wahrscheinlichkeit.

Mathematische Methoden: Mit mathematischen Methoden werden die exakten Werte von Wahrscheinlichkeiten berechnet.

Methode von LAPLACE

Seien $A_1, A_2, \ldots, A_m$ Ereignisse, die eine Zerlegung der Ereignismenge bilden:

$$P(A_1) + P(A_2) + \cdots + P(A_m) = P(\Omega) = 1.$$

Die Ereignisse seien gleichwahrscheinlich:

$$P(A_1) = P(A_2) = \ldots = P(A_m) = \frac{1}{m}.$$

(2.7) Beispiel Werfen einer Münze

(2.8) Beispiel Werfen eines Würfels

Ziehungsexperimente

Die Grundgesamtheit bestehe aus N Untersuchungsobjekten.

Eigenschaft A besitzt die relative Häufigkeit (den Anteil) p

Totalerhebung, Stichprobenerhebung

Mikrozensus, Volkszählung, Inventur

Grundgesamtheit N mit M Untersuchungsobjekten mit Eigenschaft A:

$$p = \frac{M}{N}$$

Gesucht ist die Wahrscheinlichkeit $P(A)$, daß bei einer zufälligen Ziehung ein Untersuchungsobjekt mit der Eigenschaft A gezogen wird.

Da jedes einzelne Untersuchungsobjekt die Wahrscheinlichkeit $\frac{1}{N}$ hat, folgt:

$$P(A) = \frac{M}{N}$$

(2.11) Anwendung: Meinungsumfrage

Die politische Partei A kann 40 % der Wählerstimmen auf sich vereinigen. Bei einer Meinungsumfrage werden zufällig ausgewählte Personen befragt. Wie groß ist die Wahrscheinlichkeit, daß eine befragte Person Wähler der Partei A ist?

Der Anteil der Wähler der Partei A ist $p = 0,4$. Daher beträgt die Wahrscheinlichkeit, daß eine zufällig befragte wahlberechtigte Person Wähler der Partei A ist, gerade 0,4.

3 Statistik einer relativen Häufigkeit

Es sei A ein Ereignis bei einem Zufallsexperiment mit $P(A) = p$.

Stichprobe: Versuchsergebnisse bei n-maligem Wiederholen des Zufallsexperimentes

Stichprobenumfang: Umfang n der Daten

Relative Häufigkeit: $\hat{p} = f_n(A)$, ist ein Schätzer der Wahrscheinlichkeit p

Prognoseintervalle

Wie groß ist die zufällige Schwankung der relativen Häufigkeit $\hat{p}$ um die Wahrscheinlichkeit p?

Die Schwankung $\hat{p} - p$ ist abhängig vom

- Stichprobenumfang n
- Wert der Wahrscheinlichkeit p

Die durchschnittliche Größe der Zufallsschwankungen ist proportional zur Standardabweichung:

$$SD := \sqrt{\frac{p(1-p)}{n}}.$$

Faustregel:

- Mit etwa 67 % Sicherheit betragen Zufallsschwankungen nicht mehr als eine Standardabweichung: $|f_n(A) - p| \leq SD$.
- Mit etwa 95 % Sicherheit betragen Zufallsschwankungen nicht mehr als zwei Standardabweichungen: $|f_n(A) - p| \leq 2\,SD$.
- Mit etwa 99,5 % Sicherheit betragen Zufallsschwankungen nicht mehr als drei Standardabweichungen: $|f_n(A) - p| \leq 3\,SD$.

Statistische Sicherheit:

Berechnet man die relativen Häufigkeiten $f_n(A)$ in sehr vielen, voneinander unabhängigen Stichproben, so erfüllt der als Sicherheit angegebene Prozentsatz von Stichproben die entsprechende Ungleichung:

$$|\hat{p} - p| \leq cSD \iff p - cSD \leq \hat{p} \leq p + cSD$$

(3.4) Aufgabe

$p = 0,2$; Stichprobe vom Umfang $n = 200$

$$SD = \sqrt{\frac{0,2 \cdot 0,8}{200}} = 0,0283$$

Prognoseintervall:

$$0,1434 = 0,2 - 2 \cdot 0,0283 \leq \hat{p} \leq 0,2 + 2 \cdot 0,0283 = 0,256$$

(3.5) Aufgabe

$p = 0,4$; Stichprobenumfang $n = 1000$

$$SD = \sqrt{\frac{0,4 \cdot 0,6}{1000}} = 0,0155$$

Prognoseintervall:

$$0,369 = 0,4 - 2 \cdot 0,0155 \leq \hat{p} \leq 0,4 + 2 \cdot 0,0155 = 0,431$$

Länge: Genauigkeit eines Prognoseintervalles

Wahl des Wertes c beeinflußt: Genauigkeit und Sicherheit

Standardabweichung wird beeinflußt von

- Wahrscheinlichkeit p
- Stichprobenumfang n

Die Prognoseintervalle für $f_n(A)$ sind umso genauer, je näher die Wahrscheinlichkeit p an 0 oder 1 liegt.

(3.6) $\sqrt{n}$–Gesetz:

Die statistische Genauigkeit eines Prognoseintervalls steigt proportional zur Wurzel aus dem Stichprobenumfang.

Konfidenzintervalle

Sei p eine unbekannte Wahrscheinlichkeit, deren relative Häufigkeit $\hat{p}$ beobachtet wird.

Ungleichung:

$$\hat{p} - c\,SD \leq p \leq \hat{p} + c\,SD$$

Nachteil: Berechnung von SD

(3.8) Definition *Unter einem Konfidenzintervall für eine unbekannte Wahrscheinlichkeit p versteht man ein Überdeckungsintervall $p_1 \leq p \leq p_2$ für p, dessen Grenzen p_1 und p_2 wohl von den Daten, aber nicht von der unbekannten Wahrscheinlichkeit p abhängen.*

Exakte Methode:

Lösungen der quadratischen Gleichung:

$$(p - \hat{p})^2 = \frac{c^2}{n} p(1-p).$$

(3.11) Aufgabe

Stichprobe der 50 ausgewählten Bewerberinnen. Relative Häufigkeit: $\hat{p} = f_n(A) = \frac{14}{50} = 0,28$

$$(0,28 - p)^2 = \frac{2^2}{50} p(1-p) = 0,08p(1-p)$$

$p_1 = 0,1730$; $p_2 = 0,4196$; Konfidenzintervall:

$$0,17 \leq p \leq 0,42$$

(3.12) AUFGABE: HOCHRECHNUNG

Um einen Fischbestand unbekannter Größe zu messen, hat man 300 Fische gefangen, sie markiert und wieder ausgesetzt. Nach einiger Zeit wurden 500 Fische gefangen, von denen sich 113 als markiert herausstellen.

$p = \frac{300}{N}$; $n = 500$; $\hat{p} = \frac{113}{500} = 0,226$

$$(0,226 - p)^2 = 2^2 \frac{p(1-p)}{500} = 0,008p(1-p)$$

$p_1 = 0,190851$; $p_2 = 0,265497$

$$0,190851 \leq \frac{300}{N} \leq 0,265497$$

$$1129,95 \leq N \leq 1571,90$$

Robuste Methode:

Formel:

$$p_{1,2} = \hat{p} \pm \frac{c}{2\sqrt{n}}$$

(3.13) AUFGABE

$$0,28 - 2 \cdot \frac{1}{2\sqrt{50}} \leq p \leq 0,28 + 2 \cdot \frac{1}{2\sqrt{50}}$$

also

$$0,14 \leq p \leq 0,42$$

Bootstrapmethode:

$$p_{1,2} = \hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(3.14) AUFGABE: MARKTFORSCHUNG

Wieviele Hausfrauen muß man befragen, um den Bekanntheitsgrad einer Waschmittelsorte mit der Genauigkeit $\pm 2\,\%$ bestimmen zu können ?

$$\hat{p} - \frac{c}{2\sqrt{n}} \leq p \leq \hat{p} + \frac{c}{2\sqrt{n}}$$

$$\frac{2}{2\sqrt{n}} \leq 0,02 \quad \Rightarrow \quad n \geq 2500$$

TESTPROBLEME

Es sollen zwischen zwei alternativen Aussagen über p eine Entscheidung getroffen werden.

(3.17) ANWENDUNG: MARKTFORSCHUNG

Es ist bekannt, daß in einer Stadt mindestens 60 % der Konsumenten das Produkt A dem Produkt B vorziehen. Nach einer Werbekampagne für das Produkt B erklären 80 von 160 befragten Konsumenten, sie würden das Produkt B vorziehen. Ist damit nachgewiesen, daß die Werbekampagne wirksam war ?

p: Anteil der B–Konsumenten nach der Werbekampagne

$p > 0,4 =: p_0$

$\hat{p} = 0,5$; Stichprobe vom Umfang $n = 160$

(3.18) Definition *Ein statistischer Test über eine unbekannte Wahrscheinlichkeit p ist ein Prüfverfahren, das zwischen zwei Aussagen der Form*

Nullhypothese: $p = p_0$

Alternative: $p \neq p_0$

über die unbekannte Wahrscheinlichkeit p entscheidet. Die Entscheidung wird auf Grund empirischer Daten getroffen.

Standardscores von relativen Häufigkeiten

$$\frac{\hat{p} - p}{SD}$$

Faustregel:

- Mit etwa 67 % Sicherheit liegt ein Standardscore zwischen -1 und $+1$.
- Mit etwa 95 % Sicherheit liegt ein Standardscore zwischen -2 und $+2$.
- Mit etwa 99,5 % Sicherheit liegt ein Standardscore zwischen -3 und $+3$.

Prüfverfahren:

Testgröße T beruht auf der Unterstellung der Nullhypothese $p = p_0$:

$$T = \frac{\hat{p} - p_0}{SD} \quad \text{mit} \quad SD = \sqrt{\frac{p_0(1-p_0)}{n}}$$

$-2 \leq \frac{\hat{p} - p_0}{SD} \leq 2$: Das Ergebnis ist nicht signifikant.
Keine Entscheidung.

$\frac{\hat{p} - p_0}{SD} > 2$: Das Ergebnis ist signifikant.
Entscheidung: $p > p_0$.

$\frac{\hat{p} - p_0}{SD} < -2$: Das Ergebnis ist signifikant.
Entscheidung: $p < p_0$.

(3.20) Aufgabe: Marktforschung

Der Wert der Testgröße beträgt

$$\frac{\hat{p} - p_0}{SD} = \frac{0,5 - 0,4}{\sqrt{\frac{0,4 \cdot 0,6}{160}}} = 2,58.$$

Dieser Wert ist signifikant. Wirksamkeit der Werbekampagne ist nachgewiesen.

Testtheorie

Statistische Tests:

- Aus den Daten wird eine Testgröße T berechnet.
- Es wird ein Annahmebereich der Testgröße T festgelegt, der folgende Eigenschaft hat: Falls die Nullhypothese zutrifft, liegt die Testgröße mit hoher Wahrscheinlichkeit (=Signifikanzniveau) innerhalb des Annahmebereiches.
- Die Grenzen des Annahmebereiches heißen kritische Werte. Überschreitet die Testgröße einen kritischen Wert, dann liegt ein signifikantes Ergebnis vor, welches dazu führt, daß die Hypothese verworfen wird.

Fehlentscheidungen:

Fehler 1.Art: Die Hypothese wird verworfen, obwohl sie zutrifft.

Fehler 2.Art: Die Hypothese wird beibehalten, obwohl sie nicht zutrifft.

	H wird nicht verworfen	H wird verworfen
H trifft zu	Entscheidung richtig	Fehlentscheidung 1.Art
H trifft nicht zu	Fehlentscheidung 2.Art	Entscheidung richtig

(3.22) Definition *Unter dem Signifikanzniveau versteht man die Sicherheit eines Tests, mit der sich der Fehler 1.Art vermeiden läßt.*

Durch die Wahl der kritischen Werte ist das Signifikanzniveau kontrollierbar, und daher ist der Fehler 1.Art selten.

Das Verwerfen der Nullhypothese ist ein statistischer Beweis dafür, daß sie tatsächlich falsch ist.

(3.23) Definition *Unter der Trennschärfe versteht man die Sicherheit eines Tests, mit der sich der Fehler 2.Art vermeiden läßt.*

- Eine Erhöhung des Stichprobenumfangs n unter Beibehaltung des Signifikanzniveaus führt zu einer Erhöhung der Trennschärfe.
- Eine Erhöhung des Signifikanzniveaus unter Beibehaltung des Stichprobenumfangs führt zu einer Senkung der Trennschärfe.
- Wenn $|p - p_0|$ groß ist, dann ist auch die Trennschärfe des Tests groß.

Man kann nicht davon ausgehen, daß der Fehler 2.Art selten ist. Deshalb darf das Beibehalten der Nullhypothese nicht als statistischer Beweis der Nullhypothese interpretiert werden.

4 Der Vergleich von relativen Häufigkeiten

Zwei von einander unabhängige Zufallssexperimente: Ereignis A_1 bzw. A_2 mit $P(A_1) = p$ und $P(A_2) = q$

Frage nach dem Unterschied $p - q$

(4.1) Anwendung : Verkehrsstatistik

	Schweden	Ausländer
Schwere Unfälle im Monat vor der Neuordnung:	512	261
Schwere Unfälle im Monat nach der Neuordnung:	510	189

A_1 Ereignis: „Schwerer Unfall im Monat vor ...",
A_2 Ereignis: „Schwerer Unfall im Monat nach ...".

Konfidenzintervalle

Standardabweichung der Differenz zweier relativer Häufigkeiten $\hat{p}$ und $\hat{q}$:

$$SD := \sqrt{SD_1^2 + SD_2^2} = \sqrt{\frac{p(1-p)}{n_1} + \frac{q(1-q)}{n_2}}$$

Prognoseintervalle für die Differenz $\hat{p} - \hat{q}$:

$$p - q - c\,SD \leq \hat{p} - \hat{q} \leq p - q + c\,SD$$

Konstruktion von Konfidenzintervallen

Robuste Methode:

$$p - q = \hat{p} - \hat{q} \pm \frac{c}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Bootstrapmethode:

$$p - q = \hat{p} - \hat{q} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{q}(1-\hat{q})}{n_2}}$$

(4.4) Aufgabe

$\hat{p} = \frac{261}{512+261} = 0,338$; $\hat{q} = \frac{189}{510+189} = 0,270$; $\hat{p} - \hat{q} = 0,068$

$$SD_{max} = \frac{1}{2}\sqrt{\frac{1}{773} + \frac{1}{699}} = 0,026$$

und

$$\widehat{SD} = \sqrt{\frac{0,338(1-0,338)}{773} + \frac{0,27(1-0,27)}{699}} = 0,024$$

Robuste Methode:

$$0,017 = 0,068 - 2 \cdot 0,026 \leq p - q \leq 0,068 + 2 \cdot 0,026 = 0,117$$

Bootstrapmethode:

$$0,021 = 0,068 - 2 \cdot 0,024 \leq p - q \leq 0,068 + 2 \cdot 0,024 = 0,113$$

Testprobleme

Entscheidung zugunsten der Aussage $p \neq q$

(4.7) Definition *Ein statistischer Test über den Unterschied zwischen zwei Wahrscheinlichkeiten p und q im Rahmen eines Zweistichprobenproblems ist ein Prüfverfahren, das zwischen den Aussagen*

Nullhypothese: $p = q$

Alternative: $p \neq q$

entscheidet. Die Entscheidung wird auf Grund empirischer Daten getroffen, bei denen $f(A)$ und $f(B)$ aus unabhängigen Stichproben gewonnen werden.

Prüfverfahren:

Testgröße bei Unterstellung der Nullhypothese $p = q$:

$$T = \frac{\hat{p} - \hat{q}}{SD} \quad \text{wobei} \quad SD = \sqrt{p_0(1-p_0)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Wie ist aber dabei der unterstellte gemeinsame Wert p_0 zu wählen ?

$$\hat{p}_0 = \frac{n_1\hat{p} + n_2\hat{q}}{n_1 + n_2}$$

Es gibt drei Möglichkeiten:

$-2 \leq \frac{\hat{p} - \hat{q}}{SD} \leq 2$: Das Ergebnis ist nicht signifikant.
Es ist keine Entscheidung zugunsten von $p \neq q$ möglich.

$\frac{\hat{p} - \hat{q}}{SD} > 2$: Das Ergebnis ist signifikant.
Entscheidung zugunsten der Aussage $p > q$.

$\frac{\hat{p} - \hat{q}}{SD} < -2$: Das Ergebnis ist signifikant.
Entscheidung zugunsten der Aussage $p < q$.

(4.8) Aufgabe

In einer Erhebung an Schulkindern wurde untersucht, ob sie mit der rechten oder mit der linken Hand schreiben.

Ergebnis:

	linke Hand	Gesamt
Knaben	991	12 629
Mädchen	1478	25 045

Kann daraus geschlossen werden, daß der Anteil der Linkshänder bei Knaben und Mädchen unterschiedlich ist ?

$$\hat{p}_0 = \frac{991 + 1478}{12629 + 25045} = 0,0655$$

$$\widehat{SD} = \sqrt{0,0655(1-0,0655)}\sqrt{\frac{1}{12629} + \frac{1}{25045}} = 0,0027$$

$$\frac{\hat{p} - \hat{q}}{\widehat{SD}} = \frac{0,0785 - 0,059}{0,0027} = 7,22$$

Daher schließen wir auf $p > q$.

5 Qualitative Merkmale

(5.1) Definition *Eine Eigenschaft heißt ein quantitatives Merkmal, wenn seine Ausprägungen Ergebnisse eines Zähl– oder Meßvorgangs sind. Jede andere Eigenschaft nennt man ein qualitatives Merkmal.*

(5.4) Definition *Unter einer Kodierung oder Skalierung eines Merkmals versteht man eine Abbildung der Ausprägungen des Merkmals in die Menge der reellen Zahlen. Der Skalentyp einer Kodierung ist die Gesamtheit jener Eigenschaften der Zahlen, die eine Eigenschaft der Merkmalsausprägungen abbilden.*

Es gibt im wesentlichen die folgenden Skalentypen:

Nominalskala: Das kodierte Merkmal hat keine Eigenschaften, welche sich in der Kodierung niederschlagen.

Ordinalskala: Die Ausprägungen des Merkmals besitzen eine natürliche Anordnung, welche durch die Anordnung der Codewerte ausgedrückt wird.

Intervallskala: Die Ausprägungen des Merkmals besitzen eine Anordnung und Distanzen, welche durch die Anordnung und die Abstände der Codewerte ausgedrückt werden.

Deskriptive Statistik von qualitativen Merkmalen

Mögliche Ausprägungen $A_1, A_2, \ldots, A_m$: vollständiges System von alternativen Eigenschaften

Alternativ: Die Eigenschaften schließen einander paarweise aus.

Vollständig: Die Eigenschaften erfassen alle Möglichkeiten.

(5.7) Definition *Unter der Häufigkeitsverteilung oder empirischen Verteilung eines qualitativen Merkmals versteht man die Liste der absoluten Häufigkeiten $h(A_1), h(A_2), \ldots, h(A_m)$ bzw. der relativen Häufigkeiten $f(A_1), f(A_2), \ldots, f(A_m)$.*

Häufigkeitsverteilungen: Tabellen oder Diagramme

Häufigkeitstabelle

Ausprägung	abs. Hfk.	rel. Hfk.
A_1	$h(A_1)$	$f(A_1)$
A_2	$h(A_2)$	$f(A_2)$
$\vdots$	$\vdots$	$\vdots$
A_m	$h(A_m)$	$f(A_m)$
Summe	n	1

Stabdiagramm: Die Ausprägungen werden durch Stäbe unterschiedlicher Länge dargestellt. Die Häufigkeiten sind zu den Stablängen proportional.

Sektorendiagramm: Die Ausprägungen werden durch Sektoren eines Kreises dargestellt. Die Häufigkeiten sind zu den Sektorenwinkeln proportional.

Endliche stochastische Modelle

(6.1) Beispiel

Beim Münzwurf bilden die Ereignisse $A_1 = \{\text{Zahl}\}$ und $A_2 = \{\text{Wappen}\}$ ein vollständiges System alternativer Möglichkeiten. Das Merkmal „Bildseite", welches beim Münzwurf beobachtet wird, ist ein zufälliges qualitatives Merkmal mit den Ausprägungen A_1 und A_2.

(6.2) Beispiel

Das Merkmal Videofilm der DEMO–Daten ist ein zufälliges Merkmal, da seine Daten aus einem Zufallsexperiment (Reaktion auf die Präsentation eines Videofilms) stammen.

(6.3) Definition *Unter einem endlichen Zufallsexperiment versteht man ein Zufallsexperiment mit endlich vielen alternativen Ergebnissen.*

Alternative Ergebnisse $A_1, A_2, \ldots, A_m$ mit Wahrscheinlichkeiten

$$p_1 = P(A_1), p_2 = P(A_2), \ldots, p_m = P(A_m).$$

(6.4) Definition *Unter der Wahrscheinlichkeitsverteilung eines endlichen Zufallsexperiments mit m alternativen Ergebnissen $A_1, A_2, \ldots, A_m$ versteht man die Liste der Wahrscheinlichkeiten $p_1, p_2, \ldots, p_m$.*

Eigenschaften:

$$0 \le p_i \le 1 \text{ für } i = 1, 2, \ldots, m,$$
$$p_1 + p_2 + \cdots + p_m = 1$$

Darstellungsmöglichkeit: Tabelle oder Diagramm

(6.5) DEFINITION *Ein endliches Zufallsexperiment, dessen alternative Ergebnisse $A_1, A_2, \ldots, A_m$ gleichwahrscheinlich sind, dh.*

$$p_1 = p_2 = \cdots = p_m = \frac{1}{m},$$

heißt eine LAPLACE–Experiment.

Die Wahrscheinlichkeitsverteilung ist in diesem Fall eine gleichmäßige Verteilung.

Wahrscheinlichkeit zusammengesetzter Ereignisse:

$$B = A_{i_1} \cup A_{i_2} \cup \ldots \cup A_{i_k}$$

$$P(B) = \frac{k}{m} = \frac{\text{„Anzahl der günstigen Fälle”}}{\text{„Anzahl der möglichen Fälle”}}$$

ZIEHUNGSEXPERIMENTE

Menge mit N Gegenständen $a_1, a_2, \ldots, a_N$

Zufällige Stichprobe vom Umfang n, wobei alle Gegenstände die gleiche Chance haben.

Ziehungsmöglichkeiten: Ziehen mit Zurücklegen, Ziehen ohne Zurücklegen.

Ziehen mit Zurücklegen

N^n mögliche Stichproben

Wahrscheinlichkeit einer bestimmten Stichprobenfolge:

$$p = \frac{1}{N^n}$$

Ziehen ohne Zurücklegen

$N(N-1)(N-2)\cdots(N-n+1)$ mögliche Stichproben.

Wahrscheinlichkeit einer bestimmten Stichprobefolge:

$$p = \frac{1}{N(N-1)(N-2)\cdots(N-n+1)}$$

Wahrscheinlichkeit einer bestimmten Stichprobenmenge

$n(n-1)(n-2)\cdots 2\cdot 1$ mögliche Reihenfolgen

$$p = \frac{n(n-1)(n-2)\cdots 2\cdot 1}{N(N-1)(N-2)\cdots(N-n+1)}$$

(6.6) AUFGABE

Wahrscheinlichkeit für einen Haupttreffer beim Lotto „6 aus 49” ?

$$p = \frac{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{49\cdot 48\cdot 47\cdot 46\cdot 45\cdot 44} = 7,15\cdot 10^{-8}$$

(6.7) AUFGABE

Rubbelkarten: 11 Feldern, davon 3 Gewinnfelder

Wahrscheinlichkeit eines Haupttreffers:

$$p = \frac{3\cdot 2\cdot 1}{11\cdot 10\cdot 9} = \frac{1}{165}$$

7 Empirische Prüfung von Modellen

(7.2) Definition *Ein statistischer Test einer Hypothese über eine endliche Wahrscheinlichkeitsverteilung* $(p_1, p_2, \ldots, p_m)$ *ist ein Prüfverfahren, das zwischen den Aussagen*

Nullhypothese: $(p_1, p_2, \ldots, p_m) = (p_{01}, p_{02}, \ldots, p_{0m})$

Alternative: $(p_1, p_2, \ldots, p_m) \neq (p_{01}, p_{02}, \ldots, p_{0m})$

entscheidet. Die Entscheidung wird auf Grund empirischer Daten getroffen.

Ziel: Prüfverfahren zur Beurteilung der Vereinbarkeit der empirische Verteilung (Häufigkeitsverteilung) mit der hypothetischen Wahrscheinlichkeitsverteilung

Maßzahl für das Ausmaß der Zufallsschwankung:

(7.4) Definition *Es sei* $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)$ *eine Häufigkeitsverteilung, die aus einem Zufallsexperiment mit der Wahrscheinlichkeitsverteilung* $(p_{01}, p_{02}, \ldots, p_{0m})$ *stammt. Definiert man*

$$z_i = \sqrt{n}\frac{\hat{p}_i - p_{0i}}{\sqrt{p_{0i}}} \text{ für } i = 1, 2, \ldots, m,$$

so nennt man die Liste $(z_1, z_2, \ldots, z_m)$ *die standardisierte Häufigkeitsverteilung.*

Praktische Datenanalyse:

A_1	p_1	$\hat{p}_1$	$\hat{p}_1 - p_1$	z_1
A_2	p_2	$\hat{p}_2$	$\hat{p}_2 - p_2$	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_m	p_m	$\hat{p}_m$	$\hat{p}_m - p_m$	z_m

Faustregel:
Der Maximalwert der standardisierten Häufigkeitsverteilung bei Gültigkeit der Hypothese ist mit hinreichender statistischer Sicherheit dem Betrage nach ≤ 3.

Die Chiquadrat–Methode

(7.7) Definition *Unter der Chiquadrat–Größe (für die Prüfung eines stochastischen Modells) versteht man*

$$\chi^2 = \sum_{i=1}^{m} n\frac{(\hat{p}_i - p_{0i})^2}{p_{0i}},$$

das ist die Quadratsumme der Komponenten der standardisierten Häufigkeitsverteilung.

Prüfverfahren:

Falls die Chiquadrat–Größe $(m-1) + 3\sqrt{m-1}$ übersteigt, dann wird die Hypothese verworfen.

(7.8) Aufgabe

Beurteilen Sie, ob die empirische Verteilung der Sternbilder in den DEMO–Daten mit der Hypothese einer gleichmäßigen Verteilung vereinbar ist.

Chiquadrat–Größe: $\chi^2 = 4,88$. Da $m = 12$, ist der kritische Wert $c = 11 + 3 \cdot \sqrt{11} = 20,9$.

(7.9) Aufgabe

Beurteilen Sie, ob die empirische Verteilung der Religionsbekenntnisse in den DEMO–Daten mit der Hypothese einer gleichmäßigen Verteilung vereinbar ist.

Chiquadrat–Größe: $\chi^2 = 17,8$, $m = 5$; $c = 4 + 3 \cdot \sqrt{4} = 10$

8 Der Vergleich von empirischen Verteilungen

Empirische Verteilungen von zwei qualitativen Merkmalen liegen vor.

Merkmal 1: Ausprägungen $A_1, A_2, \ldots, A_m$ mit Wahrscheinlichkeiten $P(A_1) = p_1,\ P(A_2) = p_2, \ldots,\ P(A_m) = p_m$ und relativen Häufigkeiten $f(A_1) = \hat{p}_1,\ f(A_2) = \hat{p}_2, \ldots, f(A_m) = \hat{p}_m$

Merkmal 2: Ausprägungen $B_1, B_2, \ldots, B_m$ mit Wahrscheinlichkeiten $P(B_1) = q_1,\ P(B_2) = q_2, \ldots,\ P(B_m) = q_m$ und relativen Häufigkeiten $f(B_1) = \hat{q}_1,\ f(B_2) = \hat{q}_2, \ldots, f(B_m) = \hat{q}_m$

(8.1) Beispiel: Umfragen über das Wahlverhalten

1944 Präsidenten-Wahl	Erstes Interview	Zweites oder späteres Interview	Gesamt
Roosevelt	138	217	355
Dewey	124	200	324
ohne Stimmabgabe	90	142	232
andere, oder zu jung	39	78	117
G e s a m t	391	637	1028

(8.3) Definition *Ein statistischer Test über den Unterschied zwischen zwei Wahrscheinlichkeitsverteilungen $(p_1, p_2, \ldots, p_m)$ und $(q_1, q_2, \ldots, q_m)$ ist ein Prüfverfahren, das zwischen den Aussagen*

Nullhypothese: $(p_1, p_2, \ldots, p_m) = (q_1, q_2, \ldots, q_m)$

Alternative: $(p_1, p_2, \ldots, p_m) \neq (q_1, q_2, \ldots, q_m)$

entscheidet. Die Entscheidung wird auf Grund empirischer Daten getroffen.

Das Prüfverfahren beruht auf standardisierten Differenzen der relativen Häufigkeiten.

Die hypothetischen gemeinsamen Werte werden mit $p_{0i} := p_i = q_i$ und ihre Schätzer mit

$$\hat{p}_{0i} = \frac{n_1 \hat{p}_i + n_2 \hat{q}_i}{n_1 + n_2}$$

bezeichnet.

(8.5) Definition *Es seien $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)$ und $(\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_m)$ zwei Häufigkeitsverteilungen, die aus unabhängigen Zufallsexperimenten stammen. Die Liste der Größen*

$$z_i = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\hat{p}_i - \hat{q}_i}{\sqrt{\hat{p}_{0i}}}$$

heißt die standardisierte Verteilung der Häufigkeitsdifferenzen.

(8.6) Aufgabe

Standardisierte Verteilung der Differenzen im Beispiel (8.1):

i	$\hat{p}_i$	$\hat{q}_i$	$\hat{p}_{0i}$	z_i
1	0,35	0,34	0,35	0,33
2	0,32	0,31	0,32	0,09
3	0,23	0,22	0,23	0,24
4	0,1	0,12	0,11	−1,05

Beurteilung des Schwankungsbereichs der standardisierten Häufigkeitsdifferenzen:

Faustregel:
Der Maximalwert der standardisierten Häufigkeitsdifferenzen bei Gültigkeit der Hypothese ist mit hinreichender statistischer Sicherheit dem Betrage nach ≤ 3.

Die Chiquadrat–Methode

(8.8) Definition *Unter der Chiquadrat–Größe (für den Vergleich zweier empirischer Verteilungen) versteht man*

$$\chi^2 = \sum_{i=1}^{m} \frac{n_1 n_2}{n_1 + n_2} \frac{(\hat{p}_i - \hat{q}_i)^2}{\hat{p}_{0i}},$$

das ist die Quadratsumme der Liste von standardisierten Häufigkeitsdifferenzen.

Die Anzahl der Freiheitsgrade dieser Chiquadrat–Größe beträgt $df = m - 1$.

PRÜFVERFAHREN:

Falls die Chiquadrat–Größe den kritischen Wert $(m-1)+3\sqrt{m-1}$ übersteigt, wird die Hypothese verworfen.

(8.9) AUFGABE

Überprüfen Sie im Beispiel (8.1), ob die empirischen Verteilungen signifikant voneinander abweichen.

Chiquadrat–Größe: $\chi^2 = 1,26$, kritischer Wert: $c = 3+3\sqrt{3}$

9 BEDINGTE WAHRSCHEINLICHKEITEN

Zwei qualitative Merkmale mit jeweils zwei Ausprägungen A, A' bzw. B, B'.

Kombinationen der Ausprägungen:

$$A \cap B,\ A \cap B',\ A' \cap B,\ A' \cap B'$$

Kontingenztafel (Vierfeldertafel):

	B	B'	
A	$h(A \cap B)$	$h(A \cap B')$	$h(A)$
A'	$h(A' \cap B)$	$h(A' \cap B')$	$h(A')$
	$h(B)$	$h(B')$	

Gewöhnliche relative Häufigkeiten: $f(A \cap B) = \dfrac{h(A \cap B)}{n}$

Bedingte relative Häufigkeiten: $f(A|B) = \dfrac{h(A \cap B)}{h(B)} = \dfrac{f(A \cap B)}{f(B)}$

„bedingte relative Häufigkeit von A unter der Bedingung B"

(9.1) AUFGABE

Von 1000 Verkehrsunfällen waren 280 mit tödlichem Ausgang. Davon ereigneten sich 80 bei einer Geschwindigkeit von mehr als 150 km/h. Insgesamt ereigneten sich 900 Verkehrsunfälle bei einer niedrigeren Geschwindigkeit.

A =:„Unfall endet tödlich",
B =:„Unfall ereignet sich bei mehr als 150 km/h".

	B	B'	
A	80	200	280
A'	20	700	720
	100	900	1000

$$f(A|B) := \frac{h(A \cap B)}{h(B)} = 0,8 \quad \text{und} \quad f(A|B') := \frac{h(A \cap B')}{h(B')} = 0,22$$

Empirisches Gesetz der großen Zahl:

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

(9.3) Definition *Die bedingte Wahrscheinlichkeit $P(A|B)$ von A unter B ist die Wahrscheinlichkeit von A, gemessen an Versuchen, bei denen B eintritt:*

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

Eine bedingte Wahrscheinlichkeit ist nur dort sinnvoll, wo $P(B) \neq 0$.

Produktformel: $P(A \cap B) = P(A|B)P(B)$

(9.4) Aufgabe

Ein Wurf mit zwei Würfeln ergibt eine Augensumme ≥ 10. Wie groß ist die Wahrscheinlichkeit, daß dabei mindestens eine 6 auftritt ?

S: Augensumme, A: ,,mindestens eine Sechs".

$$P(A|S \geq 10) = \frac{P(A \cap (S \geq 10))}{P(S \geq 10)}$$

$$P(A \cap (S \geq 10)) = \frac{5}{36} \quad \text{und} \quad P(S \geq 10) = \frac{6}{36},$$

und daher

$$P(A|S \geq 10) = \frac{5}{6}.$$

(9.5) Aufgabe

Ein Unternehmen produziert zwei Sorten von Produkten. Vier Prozent aller Produkte sind Ausschuß. Von den einwandfreien Produkten gehören 75 % zur Sorte 1. Wie groß ist die Wahrscheinlichkeit, daß ein zufällig ausgewähltes Produkt zur Sorte 1 gehört und einwandfrei ist ?

A ,,Produkt gehört zur Sorte 1"

B ,,Produkt ist einwandfrei"

$P(B') = 0,04;\ P(A|B) = 0,75$

$$P(A \cap B) = P(A|B)P(B) = 0,75 \cdot 0,96 = 0,72$$

(9.6) Anwendung: **Qualitätskontrolle**

Ein Konsument bezieht Glühbirnen von drei Herstellern A, B und C. Je 25 % der Glühbirnen stammen von den Herstellern A und B, der Rest stammt vom Hersteller C. Die vom Konsumenten verlangte Mindestqualität einer Glühbirne bestehe darin, daß sie eine Lebensdauer von 300 Stunden besitzt. Der Glühbirnen des Herstellers A erfüllen diese Anforderung zu 90 %, die des Herstellers B zu 70 % und die des Herstellers C zu 50 %.

Wie groß ist die Wahrscheinlichkeit dafür, daß eine zufällig ausgewählte Glühbirne, die den Anforderungen nicht entspricht, vom Hersteller A (bzw. B, C) stammt ?

$P(A) = 0.25$; $P(B) = 0.25$; $P(C) = 0.50$

L: die Glühbirne besitzt die erforderliche Lebensdauer

$P(L|A) = 0.9$; $P(L|B) = 0.7$; $P(L|C) = 0.5$

Gesucht: $P(A|L')$, $P(B|L')$, $P(C|L')$

Tabelle:

	A	B	C	
L	0.225	0.175	0.25	0.65
L'	0.025	0.075	0.25	0.35
	0.25	0.25	0.50	1

$P(A|L') = 0.0714$, $P(B|L') = 0.214$, $P(C|L') = 0.714$

(9.7) Formel für die inverse Wahrscheinlichkeit:

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

(9.8) Formel für die totale Wahrscheinlichkeit:
Es sei $(B_1, B_2, \ldots, B_m)$ eine Zerlegung. Dann gilt:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_m)P(B_m)$$

(9.9) Formel von BAYES:
Es sei $(B_1, B_2, \ldots, B_m)$ eine Zerlegung.

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_m)P(B_m)} \text{ für } i = 1, 2, \ldots, m.$$

Entscheidungsprobleme

(9.10) Anwendung: Labormedizin

Labortest:

E_+:„Der Patient leidet an der Krankheit"

E_-:„Der Patient leidet nicht an der Krankheit"

$P(E_+|K_+) = 0,95$

$P(E_-|K_-) = 0,80$

Verläßlichkeit des Labortests: Fehlerwahrscheinlichkeiten

$$P(E_-|K_+) = 1 - P(E_+|K_+) = 0,05$$
$$P(E_+|K_-) = 1 - P(E_-|K_-) = 0,2$$

Aus der Sicht des Patienten: a posteriori Wahrscheinlichkeiten

$P(K_+|E_+) =$? Wieviele der als krank eingestuften Untersuchungspersonen sind tatsächlich krank ?

$P(K_-|E_-) =$? Wieviele der als gesund eingestuften Untersuchungspersonen sind tatsächlich gesund ?

$$P(K_+|E_+) = \frac{P(E_+|K_+)P(K_+)}{P(E_+|K_+)P(K_+) + P(E_+|K_-)P(K_-)}$$

$$P(K_-|E_-) = \frac{P(E_-|K_-)P(K_-)}{P(E_-|K_+)P(K_+) + P(E_-|K_-)P(K_-)}$$

Es müssen die a priori Wahrscheinlichkeiten $P(K_+)$ und $P(K_-)$ der mögliche Zustände des Patienten bekannt sein.

$P(K_+) = 0,7$:

$$P(K_+|E_+) = \frac{0,95 \cdot 0,7}{0,95 \cdot 0,7 + 0,2 \cdot 0,3} = 0,91$$

$$P(K_-|E_-) = \frac{0,8 \cdot 0,3}{0,05 \cdot 0,7 + 0,8 \cdot 0,3} = 0,87$$

$P(K_+) = 0,05$:

$$P(K_+|E_+) = \frac{0,95 \cdot 0,05}{0,95 \cdot 0,05 + 0,2 \cdot 0,95} = 0,2$$

$$P(K_-|E_-) = \frac{0,8 \cdot 0,95}{0,05 \cdot 0,05 + 0,8 \cdot 0,95} = 0,997$$

Binäre Entscheidungsprobleme:

Z_1 und Z_2: mögliche Zustände

E_1: Entscheidung zugunsten von Z_1

E_2: Entscheidung zugunsten von Z_2

	E_1	E_2
Z_1	richtig	falsch
Z_2	falsch	richtig

	E_1	E_2
Z_1	$Z_1 \cap E_1$	$Z_1 \cap E_2$
Z_2	$Z_2 \cap E_1$	$Z_2 \cap E_2$

$Z_1 \cap E_2$ heißt Fehler 1.Art

$Z_2 \cap E_1$ heißt Fehler 2.Art

(9.11) Definition *Unter den Fehlerwahrscheinlichkeiten eines binären Entscheidungsproblems versteht man die bedingten Wahrscheinlichkeiten $P(E_2|Z_1)$ und $P(E_1|Z_2)$.*

Verläßlichkeit von Einzelentscheidungen:

Faustregel:

- Die Entscheidung E_1 gilt als verläßlich, wenn $P(E_1|Z_2)$ wesentlich kleiner ist als $P(E_1|Z_1)$: Die Entscheidung E_1 wird unter Z_2 wesentlich seltener getroffen als unter Z_1.
- Die Entscheidung E_2 gilt als verläßlich, wenn $P(E_2|Z_1)$ wesentlich kleiner ist als $P(E_2|Z_2)$: Die Entscheidung E_2 wird unter Z_1 wesentlich seltener getroffen als unter Z_2.

(9.12) Anwendung: Qualitätskontrolle

Produkt mit den Zuständen Z_1=„tauglich" und Z_2=„mangelhaft".

$P(E_2|Z_1)$: Produzentenrisiko

$P(E_1|Z_2)$: Konsumentenrisiko

(9.13) Anwendung: Test einer Hypothese

Hypothese: Z_1=„richtig" und Z_2=„falsch"

Signifikanzniveau: $1 - P(E_2|Z_1)$

Trennschärfe: $1 - P(E_1|Z_2)$

(9.14) Definition *Die bedingten Wahrscheinlichkeiten $P(Z_1|E_1)$ und $P(Z_2|E_2)$ heißen a posteriori Wahrscheinlichkeiten, weil durch sie die Beurteilung von Einzelentscheidungen im nachhinein (a posteriori) möglich ist.*

$$P(Z_1|E_1) = \frac{P(E_1|Z_1)P(Z_1)}{P(E_1)} \qquad P(Z_2|E_2) = \frac{P(E_2|Z_2)P(Z_2)}{P(E_2)}$$

Es werden die Größen $P(Z_1)$ und $P(Z_2)$ benötigt.

(9.15) Definition *Die Wahrscheinlichkeiten $P(Z_1)$ und $P(Z_2)$ der einzelnen Zustände heißen a priori Wahrscheinlichkeiten, denn sie geben an, mit welchen Häufigkeiten der Zustände Z_1 und Z_2 man von vornherein (a priori) rechnen muß.*

10 Gekoppelte Ereignisse

Ereignisse A und B:

„B begünstigt A", wenn $P(A|B) > P(A)$.

„A begünstigt B", wenn $P(B|A) > P(B)$.

Äquivalent mit: $P(A \cap B) > P(A)P(B)$

(10.1) Definition *Zwei Ereignisse A und B begünstigen einander oder sind positiv gekoppelt, wenn $P(A \cap B) > P(A)P(B)$.*

(10.2) Definition *Zwei Ereignisse A und B behindern einander oder sind negativ gekoppelt, wenn $P(A \cap B) < P(A)P(B)$.*

Das Gegenteil von Koppelung heißt Unabhängigkeit.

(10.3) Definition *Zwei Ereignisse A und B heißen gekoppelt oder stochastisch abhängig, wenn $P(A \cap B) \neq P(A)P(B)$. Sie heißen stochastisch unabhängig, wenn $P(A \cap B) = P(A)P(B)$.*

(10.4) Aufgabe

In einer technischen Untersuchung werden an PKWs folgende Merkmale erhoben:

R: Der PKW weist Rostschäden auf.
S: Der PKW besitzt eine Hohlraumversiegelung.

$P(R) = 0,37$; $P(S) = 0,71$; $P(R \cap S) = 0,11$

$P(R \cap S) = 0,11 < P(R)P(S) = 0,2626$

Koppelung zwischen zwei Ereignissen A und B: Vergleich der Vierfeldertafeln.

	B	B'	
A	$P(A \cap B)$	$P(A \cap B')$	$P(A)$
A'	$P(A' \cap B)$	$P(A' \cap B)$	$P(A')$
	$P(B)$	$P(B')$	

	B	B'	
A	$P(A)P(B)$	$P(A)P(B')$	$P(A)$
A'	$P(A')P(B)$	$P(A')P(B)$	$P(A')$
	$P(B)$	$P(B')$	

Differenzen der Tabelleneinträge:

Positive Koppelung

	B	B'	
A	$+$	$-$	
A'	$-$	$+$	

Negative Koppelung

	B	B'	
A	$-$	$+$	
A'	$+$	$-$	

(10.6) Aufgabe

Wahrscheinlichkeiten

tatsächlich

	S	S'	
R	$0,11$	$0,26$	$0,37$
R'	$0,60$	$0,03$	$0,63$
	$0,71$	$0,29$	

bei Unabhängigkeit

	S	S'	
R	$0,2627$	$0,1073$	$0,37$
R'	$0,4473$	$0,1827$	$0,63$
	$0,71$	$0,29$	

Differenzen

	S	S'	
R	$-0,1527$	$0,1527$	0
R'	$0,1527$	$-0,1527$	0
	0	0	

Vierfelderkorrelation:

$$\rho(A,B) = \frac{P(A \cap B) - P(A)P(B)}{\sqrt{P(A)P(A')P(B)P(B')}}$$

Die Vierfelderkorrelation hat folgende Eigenschaften:

- $-1 \leq \rho(A,B) \leq 1$.
- stochastisch unabhängig: $\rho(A,B) = 0$
- positiv gekoppelt: $\rho(A,B) > 0$
- negativ gekoppelt: $\rho(A,B) < 0$

(10.8) Aufgabe

$\rho(R,S) = -0,697$

Interpretation von Koppelungen

Zwischen den Ereignissen A und B besteht eine kausale Beziehung: Ursache und Wirkung

(10.9) Beispiel

Ist die negative Koppelung kausal erklärbar ?

(10.10) Beispiel

F: „Die Person ist farbenblind."
H: „Die Person arbeitet in ihrem Privathaushalt aktiv mit."

	F	F'	
H	$0,00725$	$0,49275$	$0,5$
H'	$0,04775$	$0,45225$	$0,5$
	$0,055$	$0,945$	

$\rho = -0.1776$

Kausale Interpretation:

1. Haushaltsarbeit schützt vor Farbenblindheit.
2. Farbenblindheit ist bei Haushaltsarbeit hinderlich.

Versteckter Faktor: „Geschlecht"

Männer

	F	F'	
H	$0,005$	$0,045$	$0,05$
H'	$0,095$	$0,855$	$0,95$
$'$	$0,10$	$0,90$	

Frauen

	F	F'	
H	$0,0095$	$0,9405$	$0,95$
H'	$0,0005$	$0,0495$	5
	$0,01$	$0,99$	

In einzelnen Tabellen herrscht Unabhängigkeit, aber bei Überlagerung starke Koppelung.

(10.12) Definition *Eine stochastische Koppelung von Ereignissen heißt Scheinkoppelung, wenn sie ausschließlich durch eine versteckten Faktor bewirkt wird.*

- Eine bestehende Koppelung ist kein Beweis für einen kausalen Zusammenhang.
- Kausale Interpretationen sind immer nur als vorläufige Theorien anzusehen, da ein versteckter Faktor noch nicht entdeckt sein könnte.

Das Kontingenzproblem

Kontingenztafel (Vierfeldertafel):

	B	B'	
A	$f(A \cap B)$	$f(A \cap B')$	$f(A)$
A'	$f(A' \cap B)$	$f(A' \cap B')$	$f(A')$
	$f(B)$	$f(B')$	

(10.14) DEFINITION *Unter einem Test für das Kontingenzproblem versteht man ein Verfahren, das eine Entscheidung zwischen den Aussagen*

Nullhypothese: *A und B sind stochastisch unabhängig*

Alternative: *A und B sind gekoppelt*

herbeiführt. Die Entscheidung wird auf Grund von empirischen Daten getroffen.

Empirische Vierfelderkorrelation:

$$\hat{\rho} = r = \frac{f(A \cap B) - f(A)f(B)}{\sqrt{f(A)f(A')f(B)f(B')}}.$$

Prüfverfahren:

$-2 \leq \sqrt{n}r \leq 2$: Das Ergebnis ist nicht signifikant.
Keine Entscheidung zugunsten der Nullhypothese.

$\sqrt{n}r < -2$: Das Ergebnis ist signifikant.
Entscheidung zugunsten negativer Koppelung.

$\sqrt{n}r > 2$: Das Ergebnis ist signifikant.
Entscheidung zugunsten positiver Koppelung.

(10.15) AUFGABE Hypothese der Unabhängigkeit für die Merkmale Flugangst und Schulbildung.

	$FF = 0$	$FF = 1$	
$EE = 0$	55	10	65
$EE = 1$	14	21	35
	69	31	100

$r = 0,4601$; $\sqrt{n}r = 4,601$

DAS SYMMETRIEPROBLEM

Kontingenztafel (Vierfeldertafel):

	B	B'	
A	$f(A \cap B)$	$f(A \cap B')$	$f(A)$
A'	$f(A' \cap B)$	$f(A' \cap B')$	$f(A')$
	$f(B)$	$f(B')$	

(10.16) Definition *Ein Test für den Vergleich zweier Wahrscheinlichkeiten im Rahmen eines Symmetrieproblems ist ein Verfahren, welches eine Entscheidung zwischen den Aussagen*

Nullhypothese: $P(A) = P(B)$

Alternative: $P(A) \neq P(B)$

herbeiführt. Die Entscheidung wird auf Grund von empirischen Daten getroffen, bei denen $f(A)$ und $f(B)$ aus einer Stichprobe gewonnen werden.

Die Prüfmethode beruht darauf, nur solche Daten in die Analyse einzubeziehen, bei denen genau eines der beiden Ereignisse eintritt.

$$C = (A \cap B') \cup (A' \cap B)$$

	B	B'	
A	$f(A \cap B)$	$f(A \cap B')$	$f(A)$
A'	$f(A' \cap B)$	$f(A' \cap B')$	$f(A')$
	$f(B)$	$f(B')$	

Prüfverfahren:

Das Prüfverfahren beruht auf dem Standardscore der relativen Häufigkeit $f(A|C)$ und ist somit ein Spezialfall des Tests einer Hypothese über eine Wahrscheinlichkeit.

(10.18) Aufgabe

Wahrscheinlichkeiten der Ereignisse ,,Hochschulreife", ,,Flugangst":

	$FF = 0$	$FF = 1$	
$EE = 0$	55	10	65
$EE = 1$	14	21	35
	69	31	100

$$f(A|C) = \frac{14}{14+10} = \frac{14}{24}$$

$$\frac{f(A|C) - 0,5}{\sqrt{\frac{0,5 \cdot 0,5}{n}}} = 0,816$$

11 Gekoppelte Merkmale

Zwei qualitative Merkmale: $(A_1, A_2, \ldots, A_r)$ und $(B_1, B_2, \ldots, B_s)$

Kontingenztafel:

Häufigkeitstabelle der Kombinationen $A_i \cap B_j$:

	B_1	B_2	$\ldots$	B_s	
A_1	$f(A_1 \cap B_1)$	$f(A_1 \cap B_2)$	$\ldots$	$f(A_1 \cap B_s)$	$f(A_1)$
A_2	$f(A_2 \cap B_1)$	$f(A_2 \cap B_2)$	$\ldots$	$f(A_2 \cap B_s)$	$f(A_2)$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
A_r	$f(A_r \cap B_1)$	$f(A_r \cap B_2)$	$\ldots$	$f(A_r \cap B_s)$	$f(A_r)$
	$f(B_1)$	$f(B_2)$	$\ldots$	$f(B_s)$	

(11.1) Aufgabe

Kontingenztafeln der Merkmale Bekenntnis und Videofilm:

	$VI=1$	$VI=2$	$VI=3$	$VI=4$	
$CO=1$	0.04	0.13	0.10	0.06	0.33
$CO=2$	0.06	0.10	0.06	0.05	0.27
$CO=3$	0.02	0.04	0.07	0.00	0.13
$CO=4$	0.07	0.04	0.00	0.04	0.15
$CO=5$	0.06	0.02	0.02	0.02	0.12
	0.25	0.33	0.25	0.17	1.00

(11.2) Definition *Zwei stochastische Merkmale mit den alternativen Ausprägungen $(A_1, A_2, \ldots, A_r)$ und $(B_1, B_2, \ldots, B_s)$ heißen stochastisch unabhängig, wenn*

$$P(A_i \cap B_j) = P(A_i)P(B_j) \text{ für alle Kombinationen } A_i \cap B_j.$$

Ist das nicht der Fall, dann sind die beiden Merkmale gekoppelt.

(11.3) Definition *Unter einem Test für das Kontingenzproblem versteht man ein Verfahren, das eine Entscheidung zwischen den Aussagen*

Nullhypothese: *Die Merkmale sind stochastisch unabhängig.*

Alternative: *Die Merkmale sind gekoppelt.*

herbeiführt. Die Entscheidung wird auf Grund empirischer Daten getroffen.

Indifferenztafel:

	B_1	B_2	$\ldots$	B_s	
A_1	$f(A_1)f(B_1)$	$f(A_1)f(B_2)$	$\ldots$	$f(A_1)f(B_s)$	$f(A_1)$
A_2	$f(A_2)f(B_1)$	$f(A_2)f(B_2)$	$\ldots$	$f(A_2)f(B_s)$	$f(A_2)$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
A_r	$f(A_r)f(B_1)$	$f(A_r)f(B_2)$	$\ldots$	$f(A_r)f(B_s)$	$f(A_r)$
	$f(B_1)$	$f(B_2)$	$\ldots$	$f(B_s)$	

Standardisierung der Differenzen der Tabelleneinträge:

$$\sqrt{n}r^*(A_i, B_j) = \sqrt{n}\frac{f(A_i \cap B_j) - f(A_i)f(B_j)}{\sqrt{f(A_i)f(B_j)}}$$

Struktur der Koppelung: Vorzeichenmuster der Differenzen

Signifikanz der Koppelung: Bei stochastischer Unabhängigkeit ist der Maximalwert der standardisierten Differenzen dem Betrage nach ≤ 3.

(11.4) Aufgabe

Koppelung der Merkmale Bekenntnis und Videofilm:

	$VI=1$	$VI=2$	$VI=3$	$VI=4$
$CO=1$	0.0825	0.1089	0.0825	0.0561
$CO=2$	0.0675	0.0891	0.0675	0.0459
$CO=3$	0.0325	0.0429	0.0325	0.0221
$CO=4$	0.0375	0.0495	0.0375	0.0255
$CO=5$	0.0300	0.0396	0.0300	0.0204

	$VI=1$	$VI=2$	$VI=3$	$VI=4$
$CO=1$	−1.4797	0.6394	0.6093	0.1647
$CO=2$	−0.2887	0.3652	−0.2887	0.1914
$CO=3$	−0.6934	−0.1400	2.0801	−1.4866
$CO=4$	1.6783	−0.4270	−1.9365	0.9080
$CO=5$	1.7321	−0.9849	−0.5774	−0.0280

Die Chiquadratmethode

(11.5) Definition *Unter der Chiquadrat–Größe (für das Kontingenzproblem) versteht man die Quadratsumme*

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} n\, r^*(A_i, B_j)^2.$$

der standardisierten Differenzen zwischen Kontingenztafel und Indifferenztafel.

$df = (r-1)(s-1)$

Prüfverfahren:

Wenn die Chiquadrat–Größe den kritischen Wert $df + 3\sqrt{df}$ übersteigt, dann wird die Nullhypothese verworfen.

(11.6) Aufgabe

Signifikanz der Koppelung bei den Merkmalen Bekenntnis und Videofilm:

Chiquadratgröße: $\chi^2 = 22,2482$ mit $df = 3 \cdot 4 = 12$

kritischer Wert: $12 + 3\sqrt{12} = 22,39$

12 Zufallsgrössen

(12.1) Definition *Unter einer Zufallsgröße versteht man ein zufälliges quantitatives Merkmal.*

Einfache Zufallsgröße: Endlich viele verschiedene Werte.

Diskrete Zufallsgröße: Ganzzahlige Werte.

(12.2) Beispiele

1. Wurf eines Würfels, Augenzahl
2. Wurf von zwei Würfeln, Augensumme
3. Eine Maschine produziert Werkstücke. Geometrische oder physikalische Eigenschaften der Werkstücke.
4. Ertrag einer landwirtschaftlich genutzten Fläche.
5. Simulation von Zufallszahlen.
6. Ziehen von Stichproben.

Man kann mit Zufallsgrößen rechnen. Ein algebraischer Ausdruck, bestehend aus Zufallsgrößen, definiert eine neue Zufallsgröße.

X eine einfache Zufallsgröße mit den möglichen Werten $a_1 < a_2 < \ldots < a_k$.

$(X = a_i)$: „Die Zufallsgröße X nimmt den Wert a_i an".

(12.3) Definition *Es sei X eine einfache Zufallsgröße mit den möglichen Werten $a_1 < a_2 < \ldots < a_k$ und den Wahrscheinlichkeiten*

$$p_1 = P(X = a_1),\ p_2 = P(X = a_2),\ \ldots,\ p_k = P(X = a_k).$$

Unter der Wahrscheinlichkeitsverteilung der Zufallsgröße X versteht man die Liste

$$(a_1, p_1),\ (a_2, p_2),\ \ldots,\ (a_k, p_k)$$

der Paare, bestehend aus Werten und Wahrscheinlichkeiten.

Man kann die Wahrscheinlichkeitsverteilung einer einfachen Zufallsgröße als Tabelle oder als Diagramm darstellen.

(12.4) Beispiel

Gegeben sind die Wahrscheinlichkeiten $P(A) = 0,2$, $P(B) = 0,3$ und $P(C) = 0,1$.

Unter Berücksichtigung der Unabhängigkeit der Ereignisse A, B, C und des Additionsgesetzes erhalten wir:

$$\begin{aligned} P(X = 0) &= P(A' \cap B' \cap C') = P(A')P(B')P(C') \\ &= 0,8 \cdot 0,7 \cdot 0,9 = 0,504 \end{aligned}$$

$$\begin{aligned} P(X = 1) &= P(A \cap B' \cap C') + P(A' \cap B \cap C') + P(A' \cap B' \cap C) \\ &= P(A)P(B')P(C') + P(A')P(B)P(C') + P(A')P(B')P(C) \\ &= 0,2 \cdot 0,7 \cdot 0,9 + 0,8 \cdot 0,3 \cdot 0,9 + 0,8 \cdot 0,7 \cdot 0,1 = 0,398 \end{aligned}$$

$$P(X=2) = P(A\cap B\cap C') + P(A\cap B'\cap C) + P(A'\cap B\cap C)$$
$$= P(A)P(B)P(C') + P(A)P(B')P(C) + P(A')P(B)P(C)$$

$$P(X=3) = P(A\cap B\cap C) = P(A)P(B)P(C)$$

(12.5) Aufgabe

a_i	$P(X_1=a_i)$
10000	0,3
0	0,7

a_i	$P(X_2=a_i)$
20000	0,4
−2000	0,6

a_i	$P(X_3=a_i)$
4000	0,95
−15000	0,05

$X_1=$	$X_2=$	$X_3=$	$X_1+X_2+X_3=$	$P(X_1+X_2+X_3=\ldots)$
10000	20000	4000	34000	$0,3\cdot 0,4\cdot 0,95 = 0,114$
10000	20000	−15000	15000	$0,3\cdot 0,4\cdot 0,05 = 0,006$
10000	−2000	4000	12000	$0,3\cdot 0,6\cdot 0,95 = 0,171$
10000	−2000	−15000	−7000	$0,3\cdot 0,6\cdot 0,05 = 0,009$
0	20000	4000	24000	$0,7\cdot 0,4\cdot 0,95 = 0,266$
0	20000	−15000	5000	$0,7\cdot 0,4\cdot 0,05 = 0,014$
0	−2000	4000	2000	$0,7\cdot 0,6\cdot 0,95 = 0,399$
0	−2000	−15000	−17000	$0,7\cdot 0,6\cdot 0,05 = 0,021$
				$= 1,000$

(12.6) Aufgabe

Ein Wertpapier im Wert von 3000 GE wird mit 8 % jährlich verzinst. Seine Fälligkeit wird am Ende jedes Jahres ausgelost, wobei die Wahrscheinlichkeit für die Fälligkeit 0,2 beträgt. Am Ende des dritten Jahres ist das Wertpapier spätestens endgültig fällig. Bestimmen Sie die Wahrscheinlichkeitsverteilung der Auszahlungshöhe.

	a_i	$P(X=a_i)=$
A_1	$3000\cdot 1,08$	$0,2$
$A_1'\cap A_2$	$3000\cdot (1,08)^2$	$0,8\cdot 0,2$
$A_1'\cap A_2'$	$3000\cdot (1,08)^3$	$0,8^2$

13 Unabhängige Versuchswiederholungen

- Das Zufallsexperiment besteht aus Versuchen mit jeweils zwei alternativen Ergebnissen A und A'.
- Die Versuche werden n–mal wiederholt.
- Die Versuchswiederholungen sind unabhängig voneinander und erfolgen unter gleichen Versuchsbedingungen.

Die absolute Häufigkeit $h_n(A)$ ist in diesem Fall eine Zufallsgröße.

(13.1) Anwendung: Qualitätssicherung

Die automatische Produktion eines Werkstücks hat die mögliche Ergebnisse $A =$„Werkstück ist brauchbar" und $A' =$„Werkstück ist mangelhaft". Es sei bekannt, daß $P(A') = 0.03$. Wie groß sind die Wahrscheinlichkeiten $P(h_{10}(A) = 0)$, $P(h_{10}(A) = 1)$ usw. ?

(13.2) Anwendung: Statistikklausur

Bei einer Statistik–Klausur werden 20 Fragen gestellt, und für jede Frage werden 5 alternative Anworten angeboten. Um die Klausur zu bestehen, müssen mindestens 10 Fragen richtig beantwortet werden. Ein Student, der gar keine Ahnung hat, beantwortet die Fragen zufällig. Wie groß ist die Wahrscheinlichkeit, daß er die Klausur besteht ? Welche Trefferwahrscheinlichkeit muß ein Student besitzen, damit er eine 50 %–ige Erfolgswahrscheinlichkeit hat ?

Es sei nun $n = 2$.

A_1 : A beim ersten Versuch

A_2 : A beim zweiten Versuch

$$P(A_1 \cap A_2) = P(A_1)P(A_2) = p^2$$
$$P(A_1 \cap A_2') = P(A_1)P(A_2') = p(1-p)$$
$$P(A_1' \cap A_2) = P(A_1')P(A_2) = (1-p)p$$
$$P(A_1' \cap A_2') = P(A_1')P(A_2') = (1-p)^2$$

$$\begin{aligned} P(h_2(A) = 0) &= P(A_1' \cap A_2') &&= (1-p)^2 \\ P(h_2(A) = 1) &= P(A_1 \cap A_2') + P(A_1' \cap A_2) &&= 2p(1-p) \\ P(h_2(A) = 2) &= P(A_1 \cap A_2) &&= p^2 \end{aligned}$$

(13.5) Definition: **Binomialverteilung**

$$P(h_n(A) = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, 2 \ldots, n.$$

(13.6) Beispiel

Vereinfachen Sie die Formel für die Wahrscheinlichkeiten der Binomialverteilung im Fall $p = \frac{1}{2}$.

Lösung: $P(h_n(A) = k) = \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = \binom{n}{k} \frac{1}{2^n}$

(13.7) Beispiel Beantworten Sie die Frage im Beispiel (13.1).

Lösung:

$$P(h_{10}(A) = 0) = \binom{10}{0} 0,03^0\, 0,97^{10} = 0,97^{10} = 0,7374.$$

$$P(h_{10}(A) = 1) = \binom{10}{1} 0,03^1\, 0,97^9 = 10 \cdot 0,03 \cdot 0,97^9 = 0,228.$$

(13.8) Beispiel

Beantworten Sie die Fragen im Beispiel (13.2).

Lösung: Die Wahrscheinlichkeit, die Klausur zu bestehen, beträgt

$$P(h_{20}(A) \geq 10) = \sum_{k=10}^{20} \binom{20}{k} \left(\frac{1}{5}\right)^k \left(\frac{4}{5}\right)^{20-k}$$

$$= 1 - \sum_{k=0}^{9} \binom{20}{k} \left(\frac{1}{5}\right)^k \left(\frac{4}{5}\right)^{20-k} = 1 - 0,9974 = 0,0026 = \frac{1}{384}.$$

(13.9) Beispiel

Die Wahrscheinlichkeit eines schweren Unfalls betrage bei einem technischen Verfahren 1:10000 im Laufe eines Jahres. Wie groß ist die Wahrscheinlichkeit dafür, daß beim Betrieb von 30 Anlagen im Laufe von 10 Jahren der Unfall mindestens einmal auftritt ?

Lösung:

$$P(h_{300}(A) \geq 1) = 1 - P(h_{300}(A) = 0)$$

$$= 1 - \binom{300}{0} 0,0001^0\, 0,9999^{300} = 1 = 0,9999^{300} = 0,02956.$$

(13.10) Beispiel

Es sei bekannt, daß 0.05% einer Bevölkerungsgruppe jährlich durch einen gewissen Unfall getötet wird. Bei einer Versicherung sind 10000 Personen dieser Gruppe gegen den Unfall versichert. Wie groß ist die Wahrscheinlichkeit, daß in einem gegebenen Jahr mehr als drei dieser Versicherten durch diesen Unfall umkommen?

Lösung:

$$P(h_{10000}(A) > 3) = 1 - P(h_{10000}(A) \leq 3)$$

$$= 1 - \binom{10000}{0} 0,0005^0\, 0,9995^{10000} - \binom{10000}{1} 0,0005^1\, 0,9995^{9999}$$

$$- \binom{10000}{2} 0,0005^2\, 0,9995^{9998} - \binom{10000}{3} 0,0005^3\, 0,9995^{9997}$$

$$= 1 - 0,00673 - 0,03366 - 0,08419 - 0,14037 = 0,73505.$$

14 Das Ziehen von Stichproben

- Es liegt eine endliche Grundgesamtheit mit N Objekten vor.
- Die Objekte in der Grundgesamtheit sind Träger eines Merkmals mit zwei alternativen Ausprägungen A und A'.
- Es gibt M Objekte in der Grundgesamtheit, die die Eigenschaft A besitzen, dh. der Anteil der Objekte mit der Eigenschaft A ist $p = \frac{M}{N}$.
- Aus der Grundgesamtheit wird eine Stichprobe von n Objekten gezogen.

(14.1) Anwendung: **Glückspiel**

Bei einer Tombola werden Lose in einer Urne angeboten. Lose können Treffer oder Nieten sein. Enthält die Urne N Lose, von denen M Treffer sind, so beträgt, wie wir aus Abschnitt 2 wissen, die Wahrscheinlichkeit, bei der Ziehung eines Loses einen Treffer zu ziehen, $P(h_1(A)) = p = \frac{M}{N}$. Ein Teilnehmer kauft n Lose. Wie groß sind seine Chancen, eine bestimmte Anzahl von Treffern zu erzielen ?

(14.2) Anwendung: **Qualitätskontrolle**

Eine Lieferung von N Produkten einhält M mangelhafte Stücke. Um den Anteil der mangelhaften Stücke zu überprüfen, wird eine Stichprobe vom Umfang n gezogen. Wie groß sind die Wahrscheinlichkeiten, dabei eine bestimmte Anzahl mangelhafter Produkte zu entdecken ?

Ziehen mit Zurücklegen

Die Objekte werden nacheinander zufällig gezogen. Nach jeder Ziehung wird das eben gezogene Objekt wieder in die Urne zurückgelegt.

Binomialverteilung:

$$P(h_n(A) = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \left(\frac{M}{N}\right)^k \left(1 - \frac{M}{N}\right)^{n-k}.$$

Ziehen ohne Zurücklegen

Die Objekte werden nacheinander zufällig gezogen. Nach jeder Ziehung wird das eben gezogene Objekt nicht mehr in die Urne zurückgelegt.

Stichprobe vom Umfang $n = 2$:

$$P(h_2(A) = 0) = \frac{N-M}{N} \cdot \frac{N-M-1}{N-1} = \frac{(N-M)(N-M-1)}{N(N-1)}$$

$$P(h_2(A) = 1) = \frac{M}{N} \cdot \frac{N-M}{N-1} + \frac{N-M}{N} \cdot \frac{M}{N-1} = 2\frac{M(N-M)}{N(N-1)}$$

...

(14.4) Definition: **Hypergeometrische Verteilung**

$$P(h_n(A) = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \text{ für } k = 0, 1, 2, \ldots, n.$$

(14.5) Aufgabe: **Lotterie**

Wie groß ist die Wahrscheinlichkeit, beim Lotto „6 aus 49" genau 4 „Richtige" zu haben ?

Lösung: Es ist $N = 49$, $M = 6$, $n = 6$ und $k = 4$. Daher ist

$$P(h_6(A) = 4) = \frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}} = 0,0009686.$$

(14.6) Aufgabe: Qualitätssicherung

Ein Lebensmittelhändler kauft bei einem Bauern erstmalig 250 Eier. Um im Hinblick auf weitere Käufe deren Qualität zu prüfen, zerschlägt er 10 Eier und untersucht, ob sie verdorben sind. Wie groß ist die Wahrscheinlichkeit, daß genau 1 bzw. genau 2 der untersuchten Eier verdorben sind, wenn 25 der 250 Eier verdorben sind.

Lösung: Es ist $N = 250$ und $M = 25$. Wir berechnen

$$P(h_{10}(A) = 1) = \frac{\binom{25}{1}\binom{225}{9}}{\binom{250}{10}} = 0,3954$$

$$P(h_{10}(A) = 2) = \frac{\binom{25}{2}\binom{43}{8}}{\binom{250}{10}} = 0,1968.$$

(14.7) Aufgabe: **Qualitätssicherung**

Lieferungen eines Produktionsbetriebs, bestehend aus Serien zu je 100 Stück, werden vom Empfänger kontrolliert. Es werden Stichproben vom Umfang 5 gezogen, und die Serie wird zurückgewiesen, wenn davon ein Stück mangelhaft ist. Wie groß ist die Wahrscheinlichkeit, daß mit diesem Verfahren eine Serie mit 5 % Ausschuß zurückgewiesen wird ?

Lösung: Wenn die Serie von $N = 100$ Stück 5 % Ausschuß enthält, so beträgt $M = 5$. Dann gilt

$$P(h_5(A) \geq 1) = 1 - P(h_5(A) = 0) = 1 - \frac{\binom{5}{0}\binom{95}{5}}{\binom{100}{5}} = 1 - 0,7696 = 0,2304.$$

Wenn die Kennzahlen der endlichen Grundgesamtheit N und M im Vergleich zum Stichprobenumfang n sehr groß sind, dann stimmt die Hypergeometrische Verteilung annähernd mit der Binomialverteilung überein.

Wir fassen zusammen:

Die Binomialverteilung ist der Grenzfall der Hypergeometrischen Verteilung für eine unendlich große Grundgesamtheit.

15 Lage und Streuung

Der Mittelwert ist der Durchschnittswert:

Will man alle Daten der Datenliste durch eine konstante Zahl c ersetzen und zwar so, daß die Summe der Daten unverändert bleibt, dann muß $c = \overline{x}$ sein.

Der Mittelwert ist der Gleichgewichtspunkt des Stabdiagramms:

Unterstützt man das Stabdiagramm an jener Stelle, an der sich der Mittelwert befindet, dann ist das Stabdiagramm im Gleichgewicht.

Der Mittelwert ist der Ausgleichswert:

Sucht man eine Zahl c, deren Abweichungen $x_i - c$ von den Daten die Summe 0 haben, dann muß $c = \overline{x}$ sein.

Der Mittelwert ist die am besten an die Daten angepaßte Zahl im Sinn des Prinzips der kleinsten Quadrate:

Sucht man eine Zahle c, für die die Summe der quadrierten Abstände („Abweichungsquadratsumme")

$$(x_1 - c)^2 + (x_2 - c)^2 + \cdots + (x_n - c)^2$$

möglichst klein ist, dann muß $c = \overline{x}$ sein.

Werden die Daten einer linearen Transformation $y_i = a\,x_i + b$ unterworfen, so verändert sich das arithmetische Mittel in gleicher Weise: $\overline{y} = a\,\overline{x} + b$.

Werden die Daten einer linearen Datentransformation $y_i = ax_i + b$ unterworfen, so verändert sich die Varianz gemäß $s_y^2 = a^2 s_x^2$ und die Standardabweichung $s_y = |a| s_x$.

Streuungszerlegung einer Datenliste

Wir zerlegen die Abweichungen in zwei Teile:

$$(x_i - a) = (\overline{x} - a) + (x_i - \overline{x}).$$

SS^*	$=$	$\sum_{i=1}^{n}(\overline{x}-a)^2 = n(\overline{x}-a)^2$:
	$=$	Ausmaß der systematischen Abweichung der Daten von a
SS_R	$=$	$\sum_{i=1}^{n}(x_i-\overline{x})^2 = ns_x^2$:
	$=$	Ausmaß der inneren Streuung der Daten
SS_T	$=$	$\sum_{i=1}^{n}(x_i-a)^2$:
	$=$	Totale Abweichung der Daten von a

Die Größen SS^*, SS_R und SS_T heißen Komponenten der Streuung.

(15.10) Satz von der Streuungszerlegung:

$$SS_T = SS^* + SS_R$$

Die totale Abweichung SS_T ist die Summe aus der systematischen Abweichung SS^* und der inneren Streuung SS_R der Daten.

Der Anteil $\frac{SS^*}{SS_T}$ der systematischen Abweichung an der Gesamtstreuung heißt das Bestimmtheitsmaß.

(15.12) Aufgabe: **Verkehrspsychologie**

Wie stark unterscheiden sich die Schätzwerte von der tatsächlichen Geschwindigkeit ?

Lösung: Der Vergleichswert ist $a = 60$. Wir haben $n = 24$ Daten mit dem Mittelwert $\overline{x} = 53,04$ und der Varianz $s_x^2 = 83,96$.

$$\begin{aligned} SS^* &= n(\overline{x}-a)^2 = 24(53,04-60)^2 = 1162,6 \\ SS_R &= ns_x^2 = 2015 \\ SS_T &= SS^* + SS_R = 1162,6 + 2015 = 3177,6 \end{aligned}$$

Daraus ergibt sich als Bestimmtheitsmaß

$$\frac{SS^*}{SS_R} = \frac{1162,6}{3177,6} = 0,366.$$

(15.14) Aufgabe: **Qualitätssicherung**

Eine Maschine, die Werkstücke mit einer Größe von $50\,cm$ herstellen soll, produziert eine Serie von 200 Stück mit einem Mittelwert von $58\,cm$ und einer Varianz von $36\,cm^2$. Ist die Maschine eher dejustiert oder ungenau ?

Lösung: Der Vergleichswert beträgt $a = 50$. Wir haben $n = 200$ Daten mit dem Mittelwert $\overline{x} = 58$ und der Varianz $s_x^2 = 36$. Wir berechnen die Komponenten der Streuung:

$$\begin{aligned} SS^* &= n(\overline{x}-a)^2 = 200(58-50)^2 = 12800 \\ SS_R &= ns_x^2 = 200 \cdot 36 = 7200 \\ SS_T &= SS^* + SS_R = 12800 + 7200 = 20000 \end{aligned}$$

Daraus ergibt sich als Bestimmtheitsmaß

$$\frac{SS^*}{SS_R} = \frac{12800}{20000} = 0,64.$$

Standard–Scores einer Datenliste

$$z = \frac{x - \overline{x}}{s}.$$

Der Standard–Score von x ist die Differenz zwischen x und $\overline{x}$ gemessen in Vielfachen der Standardabweichung.

Standard–Scores haben folgende Eigenschaften:

Standard–Scores haben den Mittelwert 0 und die Varianz 1. Das heißt, sie enthalten keine Information mehr über Lage und Streuung des ursprünglichen Datensatzes.

Kennt man Mittelwert und Varianz der ursprünglichen Daten, dann kann man aus den Standard–Scores die Originaldaten zurückgewinnen:

$$x_i = \overline{x} + z_i s_x$$

Grundsätzlich sind beliebig große Standard–Scores denkbar. Allerdings ist die Häufigkeit von extremen Standard–Scores beschränkt.

(15.16) **Ungleichung von TSCHEBYSCHEFF** *Es sei $x_1, x_2, \ldots, x_n$ eine Datenliste und z bezeichne einen beliebigen Standard–Score der Datenliste. Dann ist*

$$f_n(|z| \geq c) \leq \frac{1}{c^2}.$$

Das heißt, daß die relative Häufigkeit von Standard-Scores mit $|z_i| \geq c$ höchstens $1/c^2$ betragen kann.

(15.17) Aufgabe

Suchen Sie in den DEMO–Daten Personen mit ungewöhnlichem Alter.

Lösung: Das Alter der Männer hat den Mittelwert $\overline{x} = 34,9$ und die Standardabweichung $s = 5,3$. Auffällig sind Daten, deren Standard–Score dem Betrag nach größer als 2 ist. Der Bereich der ,,normalen" Daten wird eingegrenzt durch

$$x = 34,9 \pm 2 \cdot 5,3 = \begin{cases} 45,5 \\ 24,3 \end{cases}$$

Das Alter der Frauen hat den Mittelwert $\overline{x} = 25,3$ und die Standardabweichung $s = 2,75$. Auffällig sind Daten, deren Standard–Score dem Betrag nach größer als 2 ist. Der Bereich der ,,normalen" Daten wird eingegrenzt durch

$$x = 25,3 \pm 2 \cdot 2,75 = \begin{cases} 30,8 \\ 19,8 \end{cases}$$

16 Der Erwartungswert einer Zufallsgrösse

Begriff des Erwartungswerts: Langfristiger Durchschnitt der Werte der Zufallsgröße.

(16.1) **Empirisches Gesetz der großen Zahl**
Es sei X eine einfache Zufallsgröße. Wird das Zufallsexperiment unter identischen Bedingungen wiederholt, und zwar so, daß sich die einzelnen Versuchsergebnisse nicht beeinflussen können, dann konvergieren die Mittelwerte der Realisationen $x_1, x_2, \ldots$ von X gegen einen Grenzwert:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i = \mu.$$

Der Grenzwert μ heißt Erwartungswert und hängt von der Zufallsgröße X ab: $\mu = E(X)$.

Einfache Zufallsgröße: $E(X) = a_1 p_1 + a_2 p_2 + \cdots + a_m p_m$

(16.3) Aufgabe Berechnen Sie den Erwartungswert der Augenzahl beim Werfen eines Würfels:

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6}.$$

(16.4) Aufgabe Berechnen Sie die Erwartungswerte der Zufallsgrößen aus den Beispielen (12.4) und (12.6):

$$E(X) = 0 \cdot 0,504 + 1 \cdot 0,398 + 2 \cdot 0,092 + 3 \cdot 0,006 = 0,6.$$

$$E(X) = 3240 \cdot 0,2 + 3499 \cdot 0,16 + 3779 \cdot 0,64 = 3626,4.$$

(16.6) *Es seien X und Y zwei Zufallsgrößen und α, β reelle Zahlen. Dann gilt*

$$E(\alpha X + \beta) = \alpha E(X) + \beta$$
$$E(X + Y) = E(X) + E(Y)$$

(16.7) Aufgabe

Berechnen Sie im Beispiel (12.5) den Erwartungswert der Summe der Erträge.

$$\begin{aligned} E(X_1) &= 10000 \cdot 0,3 + 0 \cdot 0,7 = 3000 \\ E(X_2) &= 20000 \cdot 0,4 + (-2000) \cdot 0,6 = 6800 \\ E(X_3) &= 4000 \cdot 0,95 + (-15000) \cdot 0,05 = 3050 \end{aligned}$$

$$E(X_1 + X_2 + X_3) = 3000 + 6800 + 3050 = 12850.$$

Indikatorgrößen:

$$X_A = \begin{cases} 1, & \text{wenn } A \text{ eintritt,} \\ 0, & \text{wenn } A \text{ nicht eintritt} \end{cases}$$

Der Erwartungswert von X_A beträgt $E(X_A) = P(A)$.

Der Erwartungswert einer Indikatorgröße ist gleich der Wahrscheinlichkeit jenes Ereignisses, dessen Eintreten sie anzeigt.

(16.9) Anwendung: Faire Glückspiele

Ein Glückspiel heißt fair, wenn der Erwartungswert des Gewinns G mit dem Einsatz übereinstimmt.

Lotto „6 aus 49":

$$\text{Einsatz} = E(G) = E(\text{ Gewinnhöhe } \cdot X_A) = \text{ Gewinnhöhe } \cdot P(A)$$

$$\text{Gewinnhöhe} = 10 \cdot \frac{1}{P(A)} = 10 \cdot \frac{49 \cdot 48 \cdots 45 \cdot 44}{6 \cdot 5 \cdots 2 \cdot 1} = 139\,838\,160$$

Versicherungsmathematik

Erlebensversicherung: Zusage, nach Ablauf eines Zeitraums von n Jahren ein Kapital K auszuzahlen, falls der Versicherungsnehmer zu diesem Zeitpunkt noch lebt.

A: Ereignis, daß der Versicherungsnehmer die n Jahre der Wartezeit überlebt.
Auszahlung: $K \cdot X_A$
Barwert der Auszahlung: $B = Kv^n X_A$

Risiko der Versicherung: Erwartungswert $R = E(B)$ des Barwertes der Auszahlung.
q: Sterbewahrscheinlichkeit, $P(A) = (1-q)^n$

$$\text{Risiko} = R = E(B) = Kv^n(1-q)^n.$$

Äquivalenzprinzip: Prämie = Risiko

(16.10) Aufgabe

Berechnen Sie die Risikoprämie für einen 40–jährigen Österreicher und eine 40–jährige Österreicherin, die eine Erlebensversicherung über ein Kapital von 100 000 GE, auszuzahlen nach 10 Jahren, abschließen wollen.

Männer 3%;

$$R_m = 100000 \left(\frac{1}{1,03}\right)^{10} (1-0,003)^{10} = 72207.$$

Männer 12%:

$$R_m = 100000 \left(\frac{1}{1,12}\right)^{10} (1-0,003)^{10} = 31244,34.$$

Unabhängige Versuchswiederholungen

Ein Versuch mit den alternativen Ergebnissen A und A' wird n–mal unabhängig wiederholt. Die absolute Häufigkeit $h_n(A)$ ist dann eine Zufallsgröße mit einer Binomialverteilung.

(16.11) *Der Erwartungswert einer Zufallsgröße $h_n(A)$, deren Verteilung eine Binomialverteilung ist, beträgt $E(h_n(A) = np$*

(16.12) Aufgabe

Bei einer Statistikklausur werden 20 Fragen gestellt, und für jede Frage werden 5 alternative Antworten angeboten. Wie groß ist der Erwartungswert für die Anzahl der richtigen Antworten eines völlig unvorbereiteten Studenten ?

Da $n = 20$ und $p = 0,2$, beträgt der Erwartungswert $E(h_{20}(A)) = np = 4$.

Ziehungsexperimente

Eine Grundgesamtheit enthält N Objekte, die ein Merkmal mit den alternativen Ausprägungen A und A' tragen. Die Anzahl der Objekte mit der Ausprägung A betrage M. Es wird eine Stichprobe vom Umfang n ohne Zurücklegen gezogen. Die absolute Häufigkeit $h_n(A)$ ist dann eine Zufallsgröße mit einer hypergeometrischen Verteilung.

$E(h_n(A)) = np = n\frac{M}{N}$.

(16.14) Aufgabe

Beim Lotto „6 aus 49" beteiligen sich 1 Million Personen, von denen jeder durchschnittlich 10 Tips abgibt. Mit wievielen Haupttreffern muß durchschnittlich gerechnet werden ?

$$E(h_n(A)) = nP(A) = 10\,000\,000\frac{6 \cdot 5 \cdots 2 \cdot 1}{49 \cdot 48 \cdots 45 \cdot 44} = 0,715.$$

17 Die Varianz einer Zufallsgrösse

(17.1) Definition *Es sei X eine Zufallsgröße und μ ihr Erwartungswert. Falls die Zufallsgröße $(X-\mu)^2$ einen Erwartungswert besitzt, dann heißt dieser Erwartungswert $E\Big((X-\mu)^2\Big)$ die Varianz der Zufallsgröße. Die Varianz wird mit dem Symbol $V(X)$ bezeichnet.*

Die Wurzel $\sqrt{V(X)}$ wird als Standardabweichung von X bezeichnet.

Die Varianz einer Zufallsgröße ist die mathematische Fassung der Vorstellung eines langfristigen Durchschnittswerts der quadratischen Abweichungen vom Erwartungswert.

(17.2) *Eine einfache diskrete Zufallsgröße X besitze die möglichen Werte $a_1, a_2, \ldots, a_m$, die sie mit den Wahrscheinlichkeiten $p_1, p_2, \ldots, p_m$ annimmt. Dann beträgt die Varianz dieser Zufallsgröße*

$$V(X) = (a_1-\mu)^2p_1 + (a_2-\mu)^2p_2 + \cdots + (a_m-\mu)^2p_m.$$

(17.3) *Es sei X eine Zufallsgröße mit dem Erwartungswert μ und der Varianz $V(X)$. Dann gilt*

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

(17.4) AUFGABE Berechnen Sie die Varianz der Augenzahl beim Werfen eines Würfels.

$$E(X^2) = \frac{1}{6}(1+4+9+16+25+36) = 15,1667.$$

$$V(X) = E(X^2) - \mu^2 = 15,1667 - (3,5)^2 = 2,91667.$$

(17.5) AUFGABE Berechnen Sie Varianz der Zufallsgröße aus dem Beispiel (12.4).

$$E(X^2) = 0,2 \cdot (3240)^2 + 0,16 \cdot (3499)^2 + 0,64 \cdot (3779)^2 = 13\,198\,138,4.$$

$$V(X) = E(X^2) - \mu^2 = 13\,198\,138,4 - (3\,626,4)^2 = 47\,361,44.$$

(17.6) *Es sei X eine Zufallsgröße mit der Varianz $V(X)$. Dann gilt*

$$V(aX+b) = a^2 V(X) \quad \text{für } a,b \in \mathbb{R}.$$

(17.7) *Sind X und Y unabhängige Zufallsgrößen, dann gilt*

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

(17.8) AUFGABE

Berechnen Sie Varianz der Zufallsgröße aus dem Beispiel (12.5).

Zunächst berechnen wir die Varianzen der Zufallsgrößen X_1, X_2 und X_3. Wir erhalten

$$E(X_1^2) = 10000^2 \cdot 0,3 + 0 \cdot 0,7 = 30\,000\,000 \qquad V(X_1) = 21\,000\,000$$
$$E(X_2^2) = 20000^2 \cdot 0,4 + 2000^2 \cdot 0,6 = 162\,000\,000 \qquad V(V_2) = 116\,160\,000$$
$$E(X_3^2) = 4000^2 \cdot 0,95 + 15000^2 \cdot 0,05 = 26\,450\,000 \qquad V(X_3) = 17\,147\,500.$$

Da die Zufallsgrößen unabhängig sind, folgt daraus

$$V(X_1+X_2+X_3) = V(X_1) + V(X_2) + V(X_3) = 154\,307\,500.$$

(17.9) *Es sei A ein Ereignis und X_A seine Indikatorgröße. Die Varianz von X_A beträgt*

$$V(X_A) = P(A)(1-P(A)).$$

(17.10) AUFGABE

Berechnen Sie Varianz der Zufallsgröße aus dem Beispiel (12.4).

$$V(X_A) = 0,2 \cdot 0,8 = 0,16$$
$$V(X_B) = 0,3 \cdot 0,7 = 0,21$$
$$V(X_C) = 0,1 \cdot 0,9 = 0,09$$
$$V(X) = V(X_A) + V(X_B) + V(X_C) = 0,46.$$

Unabhängige Versuchswiederholungen

(17.11) *Die Varianz einer Zufallsgröße $h_n(A)$ mit einer Binomialverteilung beträgt*

$$V(h_n(A)) = np(1-p)$$

Varianz der relativen Häufigkeit:

$$V(f_n(A)) = V(\frac{1}{n}h_n(A)) = \frac{1}{n^2}V(h_n(A)) = \frac{P(A)(1-P(A))}{n}.$$

$$SD = \sqrt{\frac{P(A)(1-P(A))}{n}}.$$

18 Statistik eines Mittelwerts

Es sei X eine Zufallsgröße mit dem Erwarungswert $E(X) = \mu$. Eine Stichprobe von Realisationen $x_1, x_2, \ldots, x_n$ dieser Zufallsgröße führt zum Mittelwert $\overline{x}$. Wie stark kann $\overline{x}$ bei gegebenem μ schwanken ?

Prognoseintervalle

(18.1) Anwendung: Fertigungskontrolle

Eine Maschine produziert Werkstücke, deren Ausmaß X den Erwartungswert $\mu = 120$ und die Varianz $\sigma^2 = 78$ besitzt. Zur Kontrolle werden die Durchschnittswerte von X für Serien von je 100 Stück kontrolliert. In welcher Bandbreite sind die Schwankungen dieser Durchschnittswerte zu erwarten ?

(18.2) Definition *Unter der Standardabweichung des Mittelwerts aus n unabhängigen Realisationen von X versteht man die Größe*

$$SD = \frac{\sigma}{\sqrt{n}}.$$

Es sei X eine Zufallsgröße mit dem Erwartungswert mu und der Varianz σ^2. Ein Zufallsexperiment bestehe darin, n unabhängige Realisationen $X_1, X_2, \ldots, X_n$ der Zufallsgröße X zu beobachten. Der Mittelwert

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

ist eine Zufallsgröße mit den Eigenschaften

$$E(\overline{X}) = \mu \quad \text{und} \quad V(\overline{X}) = \frac{\sigma^2}{n}.$$

Praktisch gilt für Mittelwerte sogar die gleiche Faustregel wie für relative Häufigkeiten.

Faustregel:

- Mit etwa 67 % Sicherheit betragen Zufallsschwankungen nicht mehr als eine Standardabweichung: $|\overline{x} - \mu| \leq SD$.
- Mit etwa 95 % Sicherheit betragen Zufallsschwankungen nicht mehr als zwei Standardabweichungen: $|\overline{x} - \mu| \leq 2\,SD$.
- Mit etwa 99,5 % Sicherheit betragen Zufallsschwankungen nicht mehr als drei Standardabweichungen: $|\overline{x} - \mu| \leq 3\,SD$.

(18.6) Aufgabe Berechnen Sie Prognoseintervalle im Beispiel (18.1).

$$120 - c\frac{\sqrt{78}}{10} \leq \overline{x} \leq 120 + c\frac{\sqrt{78}}{10}.$$

$c = 2$:

$$\overline{x} = 120 \pm 2 \cdot 0,88 = 120 \pm 1,6.$$

(18.7) Aufgabe Eine Schulklasse mit 28 Schülern wird hinsichtlich ihrer Körpergröße untersucht. Der Mittelwert beträgt $\mu = 155\,cm$ und die Varianz ist $\sigma^2 = 144\,cm^2$. Anläßlich einer Veranstaltung wird eine Gruppe von 5 Schülern zufällig ausgewählt. Mit welcher mittleren Körpergröße ist bei dieser Gruppe zu rechnen ?

$$155 - c \cdot \frac{12}{\sqrt{5}}\sqrt{\frac{28-5}{28-1}} \leq \overline{x} \leq 155 + c \cdot \frac{12}{\sqrt{5}}\sqrt{\frac{28-5}{28-1}},$$

Die Schätzung der Varianz

Unbekannte Varianz σ^2: Bootstrapmethode.

(18.8) Definition *Es sei $x_1, x_2, \ldots, x_n$ eine Datenliste. Dann nennt man die Größe*

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

die Stichprobenvarianz der Datenliste.

Die Stichprobenvarianz s_{n-1}^2 ist ein unverfälschter Schätzer der Varianz σ^2.

Konfidenzintervalle

Der Erwartungswert μ sei unbekannt. Mit Hilfe eines beobachteten Mittelwertes $\overline{x}$ soll der Erwartungswert μ geschätzt werden:

$$\overline{x} - c\,SD \leq \mu \leq \overline{x} + c\,SD$$

Unter einem Konfidenzintervall für einen unbekannten Erwartungswert μ versteht man ein Überdeckungsintervall für μ, dessen Grenzen wohl von den Daten, aber nicht von den unbekannten Größen μ und σ^2 abhängen.

Bootstrapmethode:

$$SD = \frac{\sigma}{\sqrt{n}} \approx \widehat{SD} := \frac{s_{n-1}}{\sqrt{n}}$$

$$\overline{x} - c\frac{s_{n-1}}{\sqrt{n}} \leq \mu \leq \overline{x} + c\frac{s_{n-1}}{\sqrt{n}}$$

(18.13) Aufgabe Die DEMO–Daten bilden eine Stichprobe aus der Gesamtheit aller Bewerber um das außerirdische Siedlungsprojekt. Schätzen Sie die durchschnittliche Körpergröße getrennt nach Geschlechtern.

Männer: $\overline{x} = 166,44$, $s = 10,84$

$$s_{n-1} = \sqrt{\frac{50}{49}}\,10,84 = 10,95.$$

$$\mu = 166,44 \pm 2 \cdot \frac{10,95}{\sqrt{50}} = 166,44 \pm 3,1.$$

Testprobleme

(18.14) Definition *Ein statistischer Test über einen unbekannten Erwartungswert μ ist ein Prüfverfahren, das zwischen den Aussagen*

$$\textbf{Nullhypothese:} \quad \mu = \mu_0$$
$$\textbf{Alternative:} \quad \mu \neq \mu_0$$

über den unbekannten Erwartungswert μ entscheidet. Die Entscheidung wird auf Grund empirischer Daten getroffen.

(18.15) Definition *Unter dem Standard–Score eines Mittelwerts versteht man die Größe*

$$\frac{\overline{x} - \mu}{SD} = \sqrt{n}\frac{\overline{x} - \mu}{\sigma}.$$

Faustregel für die Beurteilung von Standard–Scores:

- Mit etwa 67 % Sicherheit liegt ein Standardscore zwischen –1 und +1.
- Mit etwa 95 % Sicherheit liegt ein Standardscore zwischen –2 und +2.
- Mit etwa 99,5 % Sicherheit liegt ein Standardscore zwischen –3 ud +3.

Prüfverfahren: $\widehat{SD} = \frac{s_{n-1}}{\sqrt{n}}$.

$$-2 \leq \frac{\overline{x} - \mu_0}{\widehat{SD}} \leq 2: \quad \text{Das Ergebnis ist nicht signifikant.}$$
$$\frac{\overline{x} - \mu_0}{\widehat{SD}} > 2: \quad \text{Das Ergebnis ist signifikant.}$$
$$\frac{\overline{x} - \mu_0}{\widehat{SD}} < -2: \quad \text{Das Ergebnis ist signifikant.}$$

(18.16) Aufgabe

Liegt beim außerirdischen Siedlungsprojekt die durchschnittliche Intelligenz der Bewerber über oder unter dem Durchschnitt der Weltbevölkerung ?

$\overline{x} = 107,45$, $s = 11,99$. Die Stichprobenstandardabweichung beträgt somit $s_{n-1} = \sqrt{\frac{100}{99}}11,99 = 12,05$.

$$SD \approx \widehat{SD} = \frac{12,05}{\sqrt{100}} = 1,205.$$

Testgröße:

$$\frac{107,45 - 100}{1,205} = \frac{7,45}{1,205} > 2.$$

(18.17) Aufgabe

Unterscheiden sich im Beispiel (15.12) die geschätzten Geschwindigkeiten signifikant von der wahren Geschwindigkeit ?

Von $n = 24$ Versuchspersonen beträgt $\overline{x} = 53,04$ und $s = 9,16$. Daraus ergibt sich $s_{n-1} = 9,36$ und $\widehat{SD} = 1,91$. Der Wert der Testgröße beträgt

$$\frac{53,04 - 60}{1,91} = \frac{-6,96}{1,91} < -2.$$

Varianzanalyse

Die ANOVA–Tabelle ist ein Analyseschema mit folgendem Aufbau:

	SS	df	MSS
$*$	SS^*	1	MSS^*
R	SS_R	$n-1$	MSS_R
	SS_T	n	

Erste Spalte: Quadratsummen (Sums of Squares).

Zweite Spalte: Anzahl der Freiheitsgrade (degrees of freedom).

Dritte Spalte: Mittlere Quadratsummen.

Signifikanzproblem: F–Größe

$$F = \frac{MSS^*}{MSS_R} = \left(\sqrt{n}\frac{\overline{x} - \mu_0}{s_{n-1}}\right)^2.$$

Die Hypothese $\mu = \mu_0$ wird verworfen, wenn die F–Größe einen kritischen Wert c übersteigt, der im vorliegenden Fall als Faustregel mit $c = 4$ festgelegt werden kann.

Relevanzproblem: Bestimmtheitsmaß

$$\frac{SS^*}{SS_T}.$$

(18.18) Aufgabe

Führen Sie den Test im Beispiel (15.12) anhand der ANOVA–Tabelle durch.

Die ANOVA–Tabelle hat die Form

	SS	df	MSS
$*$	$1162,6$	1	$1162,6$
R	$2015,0$	23	$87,6$
	$3177,6$	24	

Daher beträgt die F–Größe

$$F = \frac{1162,6}{87,6} = 13,27.$$

19 Der Vergleich von zwei Mittelwerten

Es seien X und Y zwei unabhängige Zufallsgrößen mit den Erwartungswerten $E(X) = \mu_1$, $E(Y) = \mu_2$, und den Varianzen $V(X) = \sigma_1^2$, $V(Y) = \sigma_2^2$.

Das Testproblem

(19.1) Definition *Ein statistischer Test über den Unterschied zwischen zwei unbekannten Erwartungswerten μ_1 und μ_2 von unabhängigen Zufallsgrößen ist ein Prüfverfahren, welches zwischen den Aussagen*

Nullhypothese: $\mu_1 = \mu_2$

Alternative: $\mu_1 \neq \mu_2$

entscheidet. Die Entscheidung erfolgt auf Grund empirischer Daten.

Der Grundgedanke jedes Prüfverfahrens für dieses Testproblem besteht darin, die Differenz $\overline{x} - \overline{y}$ als Basis für die Konstruktion einer Prüfgröße zu verwenden.

(19.2) Definition *Unter der Standardabweichung der Mittelwertsdifferenz $\overline{x} - \overline{y}$ aus zwei unabhängigen Stichproben versteht man die Größe*

$$SD = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

Prüfverfahren:

Das Prüfverfahren beruht auf der Testgröße

$$\frac{\overline{x} - \overline{y}}{\widehat{SD}}, \text{ wobei } \widehat{SD} = \sqrt{\frac{s_{x,n-1}^2}{n_1} + \frac{s_{y,n-1}^2}{n_2}}$$

Es gibt insgesamt drei Möglichkeiten:

$-2 \leq \dfrac{\overline{x} - \overline{y}}{\widehat{SD}} \leq 2$: Das Ergebnis ist nicht signifikant.

$\dfrac{\overline{x} - \overline{y}}{\widehat{SD}} > 2$: Das Ergebnis ist signifikant.

$\dfrac{\overline{x} - \overline{y}}{\widehat{SD}} < -2$: Das Ergebnis ist signifikant.

(19.4) Aufgabe

Bestehen in den DEMO–Daten beim Merkmal Intelligenz signifikante Unterschiede zwischen Männern und Frauen ?

Wir haben zwei Stichproben mit $n_1 = 50$ und $n_2 = 50$. Es gilt $\overline{x} = 104,72$ und $s_{x,n-1} = 12,387$, sowie $\overline{y} = 110,36$ und $s_{y,n-1} = 11,133$.

$$\widehat{SD}_{\overline{x}-\overline{y}} = \sqrt{\frac{12,387^2}{50} + \frac{11,133^2}{50}} = 2,355.$$

Als Wert der Testgröße erhalten wir

$$\frac{104,72 - 110,36}{2,355} = -2,39.$$

(19.5) Aufgabe

Bestehen in den DEMO–Daten beim Merkmal Körpergröße signifikante Unterschiede zwischen Männern mit und ohne Flugangst ?

Wir haben zwei Stichproben mit $n_1 = 18$ und $n_2 = 32$. Es gilt $\overline{x} = 164,37$ und $s_{x,n-1} = 12,678$, sowie $\overline{y} = 167,61$ und $s_{y,n-1} = 9,87$.

$$\widehat{SD}_{\overline{x}-\overline{y}} = \sqrt{\frac{12,678^2}{18} + \frac{9,87^2}{32}} = 3,46.$$

Als Wert der Testgröße erhalten wir

$$\frac{164,37 - 167,61}{3,46} = -0,936.$$

Varianzanalyse

Falls $\sigma_1^2 = \sigma_2^2$, dann ist ein genaueres Verfahren zur Beurteilung des Unterschieds $\mu_1 - \mu_2$ möglich.

Anwendungen: Medizin, Landwirtschaft, Marktforschung

Der Vergleich der Mittelwerte erfolgt unter den angegebenen Voraussetzungen mit Hilfe einer Varianzanalyse.

Der Grundgedanke der Varianzanalyse besteht darin, die Daten der beiden Stichproben mit dem Gesamtmittelwert aller Daten zu vergleichen.

Der Gesamtmittelwert beträgt

$$m = \frac{n_1\overline{x} + n_2\overline{y}}{n_1 + n_2}.$$

Man kann die Abweichung von Einzeldaten vom Gesamtmittelwert in zwei Teile zerlegen:

$$(x_i - m) = (\overline{x} - m) + (x_i - \overline{x}),$$
$$(y_i - m) = (\overline{y} - m) + (y_i - \overline{y}).$$
$$(\overline{x} - m) = \frac{n_2}{n_1 + n_2}(\overline{x} - \overline{y}), \quad (\overline{y} - m) = \frac{n_1}{n_1 + n_2}(\overline{y} - \overline{x}).$$

$$SS_{ZW} = \sum_{i=1}^{n_1}(\overline{x} - m)^2 + \sum_{i=1}^{n_2}(\overline{y} - m)^2$$
$$= n_1(\overline{x} - m)^2 + n_2(\overline{y} - m)^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{x} - \overline{y})^2$$

$$SS_{IN} = \sum_{i=1}^{n_1}(x_i - \overline{x})^2 + \sum_{i=1}^{n_2}(y_i - \overline{y})^2 = n_1 s_x^2 + n_2 s_y^2$$

$$SS_T = \sum_{i=1}^{n_1}(x_i - m)^2 + \sum_{i=1}^{n_2}(y_i - m)^2$$

(19.9) Satz von der Streuungszerlegung

Es gilt

$$SS_T = SS_{ZW} + SS_{IN}.$$

Die Gesamtstreuung SS_T ist die Summe aus dem systematischen Unterschied SS_{ZW} zwischen den Stichproben und der inneren Streuung SS_{IN} der einzelnen Stichproben.

ANOVA–Tabelle:

	SS	df	MSS
ZW	SS_{ZW}	1	MSS_{ZW}
IN	SS_{IN}	n_1+n_2-2	MSS_{IN}
	SS_T	n_1+n_2-1	

Signifikanzproblem: F–Größe

$$F = \frac{MSS_{ZW}}{MSS_{IN}}.$$

Relevanzproblem: Bestimmtheitsmaß

$$\frac{SS_{ZW}}{SS_T}.$$

(19.10) Aufgabe Führen Sie mit den DEMO–Daten beim Merkmal Intelligenz eine Varianzanalyse zwischen Männern und Frauen durch.

Die ANOVA–Tabelle lautet

	SS	df	MSS
ZW	$795,24$	1	$795,24$
IN	$13\,591,60$	98	$138,69$
	$14\,386,84$	99	

F–Größe $F = 5,7339$, Bestimmtheitsmaß $0,05851$

20 Empirische Korrelation

Die Untersuchungsobjekte seien Träger von zwei quantitativen Merkmalen X und Y. Für jedes Untersuchungsobjekt erhalten wir als Beobachtung ein Datenpaar (x, y), wobei x die beobachtete Ausprägung des Merkmals X ist und y die beobachtete Ausprägung des Merkmals Y.

Datenliste: Liste von Datenpaaren (x_i, y_i), $i = 1, 2, \ldots, n$.

Wechselseitige Beziehungen: Streudiagramm.

Das Streudiagramm

Das Streudiagramm ist ein Punktediagramm, das die Datenpaare (x_i, y_i) als Punkte in einem rechtwinkeligen Koordinatensystem darstellt.

- Der Mittelpunkt des Streudiagramms ist der Punkt $(\overline{x}, \overline{y})$. Physikalisch gesehen handelt es sich um den Schwerpunkt der Punktwolke.

- Projiziert man die Punkte des Streudiagramms auf die x–Achse, so entsteht das Punktediagramm der Datenliste $x_1, x_2, \ldots, x_n$. Die horizontale Ausdehnung der Punktwolke hängt also von der Streuung der Datenliste $x_1, x_2, \ldots, x_n$ ab.
- Projiziert man die Punkte des Streudiagramms auf die y–Achse, so entsteht das Punktediagramm der Datenliste $y_1, y_2, \ldots, y_n$. Die vertikale Ausdehnung der Punktwolke hängt von der Streuung der Datenliste $y_1, y_2, \ldots, y_n$ ab.

(20.2) Definition *Die Merkmale X und Y sind positiv gekoppelt, wenn für ihre Ausprägungen tendenziell die Aussagen gelten:*

Je größer x, desto größer y

Je kleiner x, desto kleiner y.

Die Merkmale X und Y sind negativ gekoppelt, wenn für ihre Ausprägungen tendenziell die Aussagen gelten:

Je größer x, desto kleiner y.

Je kleiner x, desto größer y.

Sind zwei Merkmale positv oder negativ gekoppelt, so liegt eine monotone Koppelung vor.

Das standardisierte Streudiagramm

Unter dem standardisierten Streudiagramm versteht man das Streudiagramm der Standard–Scores der Daten:

$$\text{Datenpunkt: } (x, y) \ldots \text{Standard–Scores: } z_x = \frac{x - \overline{x}}{s_x},\ z_y = \frac{y - \overline{y}}{s_y}.$$

Ein standardisiertes Streudiagramm hat folgende Eigenschaften:

- Der Mittelpunkt des standardisierten Streudiagramms ist der Nullpunkt.
- Projiziert man die Punkte des standardisierten Streudiagramms auf die x–Achse, so entsteht das Punktediagramm der Standard–Scores $z_{x_1}, z_{x_2}, \ldots, z_{x_n}$. Die horizontale Ausdehnung der Punktwolke entspricht also einer Datenliste mit der Varianz 1.
- Projiziert man die Punkte des standardisierten Streudiagramms auf die y–Achse, so entsteht das Punktediagramm der Standard–Scores $z_{y_1}, z_{y_2}, \ldots, z_{y_n}$. Die vertikale Ausdehnung der Punktwolke entspricht auch einer Datenliste mit der Varianz 1.

Korrelation

(20.5) Definition *Unter dem Korrelationskoeffizienten einer Liste von Datenpaaren (x_i, y_i) versteht man den Mittelwert*

$$r = \frac{1}{n}(z_{x_1} z_{y_1} + z_{x_2} z_{y_2} + \cdots + z_{x_n} z_{y_n})$$

der Produkte der Standard–Scores.

„Lehrbuchformel": $r = \dfrac{s_{xy}}{s_x s_y}$

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

(20.11) Definition *Unter der Korrelation einer Liste von Datenpaaren (x_i, y_i) versteht man die Bindung der Punktwolke an eine steigende oder fallende Hauptachse.*

Varianz der Punktwolke: Die Varianz einer Punktwolke hängt davon ab, von welchem Punkt der Zeichenebene aus man sie betrachtet. Legt man orthogonal zur Blickrichtung eine Gerade, und projiziert man die Punkte der Punktwolke auf diese Gerade, so entsteht ein projiziertes Punktediagramm, das dem optischen Bild der Punktwolke entspricht.

Hauptachse: Die Hauptachse geht durch den Mittelpunkt der Punktwolke und ist parallel zu jener Geraden, auf der das projizierte Punktediagramm die größte Varianz besitzt.

Bindung an die Hauptachse: Die Bindung an die Hauptachse (=Korrelation) ist groß, wenn die Punktwolke schwach um die Hauptachse streut. Sie ist klein, wenn die Punktwolke stark um die Hauptachse streut.

Interpretation des Korrelationskoeffizienten:

$\operatorname{sgn} r$: Das Vorzeichen des Korrelationskoeffizienten gibt an, ob die Hauptachse steigt oder fällt.

Im Fall $r > 0$ spricht man von positiver Korrelation. Positive Korrelation ist eine spezielle Form einer positiven Koppelung.

Im Fall $r < 0$ spricht man von negativer Korrelation. Negative Korrelation ist eine spezielle Form einer negativen Koppelung.

$|r|$: Der Betrag des Korrelationskoeffizienten gibt die Stärke der Korrelation (=Stärke der Bindung an die Hauptachse) an.

Der Korrelationskoeffizient liegt immer zwischen -1 und $+1$. Die beiden Extremfälle treten dann ein, wenn alle Datenpunkte auf der Hauptachse liegen.

Ist der Korrelationskoeffizient $r = 0$, dann kann das zwei Ursachen haben. Entweder gibt es keine Hauptachse, weil die Punktwolke aus allen Richtungen betrachtet die gleiche Varianz besitzt, oder die Hauptachse ist parallel zu einer der beiden Koordinatenachsen.

Korrelation bedeutet Bindung an eine lineare (= geradlinig verlaufende) Hauptachse.

Sind die Daten an eine nichtlineare Kurve gebunden, so ist der Korrelationskoeffizient wenig oder gar nicht interpretierbar.

Die Frage, ob es sinnvoll ist, den Korrelationskoeffizienten zu berechnen, entscheidet man anhand des Streudiagramms.

Die graphische Beurteilung der Daten ist also vor der quantitativen Analyse vorzunehmen.

Varianz einer Summe von Datenlisten:

(20.14) *Gegeben sei eine Liste von Datenpaaren (x_i, y_i), $i = 1, 2, \ldots, n$. Die Varianz der Summen $(x_i + y_i)$ beträgt*

$$s^2_{x+y} = s^2_x + s^2_y + 2r s_x s_y.$$

- Addiert man positiv korrelierte ($r > 0$) Datensätze, so ist die Varianz der Summe größer als die Summe der Varianzen.
- Addiert man negativ korrelierte ($r < 0$) Datensätze, so ist die Varianz der Summe kleiner als die Summe der Varianzen.
- Addiert man unkorrelierte ($r = 0$) Datensätze, so ist die Varianz der Summe gleich der Summe der Varianzen.

21 Lineare Regression

Regressionsmodelle

Wir betrachten ein Zufallsexperiment, bei dem eine Zufallsgröße Y beobachtet werden kann. Im Gegensatz zu den bisherigen Überlegungen wiederholen wir das Zufallsexperiment aber nicht unter identischen Bedingungen, sondern wir ändern die Versuchsbedingungen von Versuch zu Versuch systematisch ab. Es interessiert uns, ob und wie die Eigenschaften der Zufallsgröße Y von den Versuchsbedingungen abhängen.

(21.1) Anwendung: **Landwirtschaft**

Es soll die Abhängigkeit des Weizenertrags Y von der eingesetzten Menge X an Stickstoffdünger untersucht werden.

$X =$	50	60	70	80	90	100
$Y = 31-35$				2	6	3
$26-30$			5	12	7	2
$21-25$		4	8	8	6	
$16-20$		2	7		1	
$11-15$	1	3				
$6-10$	3	1				
$1-5$	1					

(21.2) Anwendung: **Technologie**

Eine Untersuchung über die Abhängigkeit der Zugfestigkeit f (in kg/cm^2) von der Trockenzeit T (in Tagen) bei Beton ergab folgende Werte für f:

$T =$	1	2	3	7	28
$f =$	13,0	21,9	29,8	32,4	41,8
	13,3	24,5	28,0	30,4	42,6
	11,8	24,7	24,1	34,5	40,3
			24,2	33,1	35,7
			26,2	35,7	37,3

Kontrolliertes Merkmal: Prädiktormerkmal (die „unabhängige" Variable)

Zufallsgröße Y: Responsevariable (die „abhängige" Variable)

(21.3) Definition *Unter einem Regressionsmodell versteht man eine Prädiktor–Response Modell, bei dem das Prädiktormerkmal nur den Erwartungswert der Responsevariablen beeinflußt:*

$$E_x(Y) = f(x).$$

Die Funktion $y = f(x)$, die die Abhängigkeit des Erwartungswerts vom Prädiktor beschreibt, heißt Regressionsfunktion.

(21.4) DEFINITION *Unter einem linearen Regressionsmodell versteht man ein Regressionsmodell mit einer linearen Regressionsfunktion.*

Ein lineares Regressionsmodell hat die Form

$$E_x(Y) = a + bx \text{ oder } Y = a + bx + U \text{ wobei } E_x(U) = 0.$$

Die Regressionsfunktion $f(x) = a + bx$ heißt Regressionsgerade.

DIE EMPIRISCHE REGRESSIONSGERADE

Die empirische Regressionsgerade ist jene Gerade, die im Sinne des Prinzips der kleinsten Quadrate die Werte y_i am besten vorhersagen kann.

Das Prinzip der kleinsten Quadrate:

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

prognostizierte Werte $\hat{y}_i = \alpha + \beta x_i$

Prognosefehler $y_i - \hat{y}_i$

Ziel:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{ Minimum !}$$

(21.5) *Es sei (x_i, y_i), $i = 1, 2 \ldots, n$ eine Liste von Datenpaaren. Die nach dem Prinzip der kleinsten Quadrate konstruierte Gerade zur optimalen Vorhersage der Werte y_i aus den Werten x_i hat die Form $\hat{f}(x) = \hat{a} + \hat{b}x$, wobei*

$$\hat{b} = r\frac{s_y}{s_x} \text{ und } \hat{a} = \overline{y} - \hat{b}\overline{x}.$$

Die Gerade $y = \hat{a} + \hat{b}x$ heißt die empirische Regressionsgerade des Merkmals Y nach dem Merkmal X.

1. Die „Lehrbuchformel" für $\hat{b}$ lautet $\hat{b} = \dfrac{s_{xy}}{s_x^2}$.
2. Es gibt auch eine Regressionsgerade des Merkmals X nach dem Merkmal Y. Die beiden Regressionsgeraden unterscheiden sich.
3. Jede empirische Regressionsgerade geht durch den Mittelpunkt $(\overline{x}, \overline{y})$ der Punktwolke. Daraus kann man die Formel für $\hat{a}$ gewinnen:
   $\overline{y} = \hat{a} + \hat{b}\overline{x} \Rightarrow \hat{a} = \overline{y} - \hat{b}\overline{x}$.

(21.9) AUFGABE Zeichnen Sie das Streudiagramm und berechnen Sie die empirische Regressionsgerade für das Beispiel (21.2).

Wir haben $n = 21$ Datenpunkte (y_i, t_i). Mit der Datentransformation $x = \log t$ ergibt sich $\overline{x} = 1,6173$ und $s_x = 1,1338$. Außerdem ist $\overline{y} = 28,89$ und $s_y = 8,7018$. Daraus folgt $r = 0,9423$, $\hat{b} = 7,3217$ und $\hat{a} = 16,9825$.

Die Regressionsgerade unterscheidet sich von der Hauptachse der Punktwolke.

Bei der Regressionsgeraden geht es um die vertikalen Abstände der Punkte von der Geraden während es bei der Hauptachse um die orthogonalen Abstände geht.

Die Regressionsgerade von Y nach X verläuft grundsätzlich flacher als die Hauptachse;

die Regressionsgerade von X nach Y verläuft steiler als die Hauptachse.

(21.10) AUFGABE Wie lauten die Gleichungen der Regressionsgeraden im standardisierten Streudiagramm ?

Im standardisierten Streudiagramm werden die Standardscores (z_x, z_y) aufgetragen. Die Standardscores haben jeweils den Mittelwert 0 und die Varianz 1. Ihr Korrelationskoeffizient ist aber der gleiche wie bei den ursprünglichen Daten. Daraus folgt $\hat{b} = r$ und $\hat{a} = 0$. Dies gilt sowohl für die Regressionsgerade von z_y nach z_x als auch für die Regressionsgerade von z_x nach z_y.

Die Regressionsgerade von z_y nach z_x lautet also $z_y = rz_x$. Da $|r| \leq 1$, verläuft sie flacher als die Hauptachse.

Die Regressionsgerade von z_x nach z_y lautet also $z_x = rz_y$, dh. $z_y = \frac{1}{r} z_x$. Da $|\frac{1}{r}| \geq 1$, verläuft sie steiler als die Hauptachse.

ANWENDUNG: **Das Regressionsphänomen**

Der Unterschied zwischen Regressionsgerade und Hauptachse sowie seine Interpretation kann am standardisierten Streudiagramm besonders gut erklärt werden.

Wir nehmen an, daß unsere Datenliste eine positive Korrelation $r > 0$ besitzt. In diesem Fall lautet die Gleichung der Hauptachse $z_y = z_x$ und die Gleichung der Regressionsgeraden $z_y = rz_x$. Auf der Hauptachse liegen also die Datenpaare, deren Komponenten gleiche Standard–Scores besitzen. Das sind die Datenpaare, deren Komponenten gleich über– oder unterdurchschnittlich sind.

Die Erscheinung, daß Paare, deren eine Komponente extrem ist, im Durchschnitt eine weniger extreme zweite Komponente besitzen, heißt Regressionsphänomen.

(21.12) Aufgabe

Eine empirische Untersuchung von Ehepaaren zeigt, daß überdurchschnittlich intelligente Ehefrauen im Durchschnitt intelligenter sind als ihre Ehemänner, während unterdurchschnittlich intelligente Ehefrauen eher weniger intelligent als ihre Ehemänner sind. Folgt daraus eine erfolgreiche Dominanzstrategie intelligenter Frauen bei der Partnerwahl ?

(21.13) Aufgabe

Werden bei Trainingsprogrammen die Teilnehmer vor und nach dem Training getestet, so sind zu Beginn überdurchschnittliche Teilnehmer nach dem Training im Durchschnitt weniger überdurchschnittlich. Wirken Trainingsprogramme nivellierend ?

Varianzanalyse

Qualität der empirischen Regressionsgeraden als Prognoseinstrument

$\hat{y}_i - \overline{y}$ erklärbare Streuung

$y_i - \hat{y}_i$ Reststreuung

$$\begin{aligned}(y_i - \overline{y}) &= (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i) \\ &= \hat{b}(x_i - \overline{x}) + (y_i - \hat{y}_i)\end{aligned}$$

Komponenten der Streuung:

SS^*	$=$	$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \hat{b}^2 \sum_{i=1}^{n}(x_i - \overline{x})^2$:
	$=$	Erklärbare Streuung
SS_R	$=$	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$:
	$=$	Reststreuung
SS_T	$=$	$\sum_{i=1}^{n}(y_i - \overline{y})^2$:
	$=$	Gesamtstreuung der y_i

(21.14) Satz von der Streuungszerlegung

$$SS_T = SS^* + SS_R$$

Die Gesamtstreuung SS_T ist die Summe aus der erklärbaren Streuung SS^* und der Reststreuung SS_R.

$$\begin{aligned}SS^* &= r^2 n s_y^2 \\ SS_T &= n s_y^2 \\ SS_R &= (1 - r^2) n s_y^2\end{aligned}$$

(21.15) *Das Bestimmtheitsmaß, also der Prognosewert der empirischen Regressionsgeraden, bezogen auf den vorliegenden Datensatz, beträgt*

$$\frac{SS^*}{SS_T} = r^2.$$

Das Quadrat des Korrelationskoeffizienten ist ein Maß für die Genauigkeit der Regressionsgeraden.

(21.16) Definition *Unter einem Test einer linearen Regression versteht man ein Prüfverfahren, das zwischen den Aussagen*

Nullhypothese: $b = 0$

Alternative: $b \neq 0$

eine Entscheidung herbeiführt. Die Entscheidung wird wird auf Grund empirischer Daten getroffen.

Ein Test dieser Hypothese kann mit Hilfe der ANOVA–Tabelle leicht konstruiert werden. Die ANOVA–Tabelle für ein lineares Regressionsmodell lautet:

	SS	df	MSS
$*$	SS^*	1	MSS^*
R	SS_R	$n-2$	MSS_R
T	SS_T	$n-1$	

Um das Signifikanzproblem zu beantworten, bildet man die F–Größe

$$F = \frac{MSS^*}{MSS_R} = (n-2)\frac{r^2}{1-r^2}.$$

(21.17) Aufgabe Führen Sie für das Beispiel (**??**) die Varianzanalyse durch. Testen Sie die Nullhypothese $b = 0$.

$$\begin{aligned} SS_T &= ns_y^2 = 82 \cdot 44,57 = 3654,7 \\ SS^* &= r^2 \cdot SS_T = 0,748^2 \cdot 3654,7 = 2044,84 \\ SS_R &= 3654,7 - 2044,84 = 1609,85 \end{aligned}$$

	SS	df	MSS
$*$	$2044,84$	1	$2044,84$
R	$1609,85$	80	$20,12$
T	$3654,69$	81	

22 Bivariate Statistik

Das Korrelationsproblem: Besteht zwischen den beiden Zufallsgrößen ein Zusammenhang, der es erlaubt, die Werte der einen Zufallsgröße auf Grund der Werte der anderen Zufallsgröße vorherzusagen ?

Das Symmetrieproblem: Stimmen die Erwartungswerte $E(X)$ und $E(Y)$ überein oder sind sie verschieden ?

(22.1) Beispiel

Beispiele für Korrelationsprobleme sind folgende Fragen:

(a) Besteht in den DEMO–Daten zwischen den Variablen Alter und Intelligenz ein Zusammenhang ?

(b) Besteht in den DEMO–Daten zwischen den Variablen Alter und Blutdruck ein Zusammenhang ?

Ein Beispiel für ein Symmetrieproblem ist die Frage:

(c) Hat das allgemeine Konditionstraining einen Effekt auf die Leistungsfähigkeit der Personen in den DEMO–Daten ?

(22.2) Anwendung: **Marktforschung**

Ein Marktforschungsinstitut untersucht, ob Ereignisse des Vortages einen Einfluß auf den Absatz von Tageszeitungen haben. Es werden die Verkaufszahlen von 10 Zeitungen an zwei verschiedenen Tagen erhoben.

Tag 1: Politische Umwälzungen im Nachbarland, Europacupfinale
Tag 2: Keine bedeutenden Ereignisse

	Verkaufszahlen in Tausend									
Tag 1	410	350	180	60	50	43	22	22	21	20
Tag 2	350	280	160	61	42	40	18	23	20	17

Wir stellen in diesem Beispiel die Frage: Haben die Ereignisse des Vortags einen Einfluß auf die Verkaufszahlen ?

(22.3) Anwendung: **Verkehrspsychologie**

Es sollte untersucht werden, ob Alkohol die Risikobereitschaft bei Autofahrern erhöht. Als ein Maß der Risikobereitschaft wurde die Höchstgeschwindigkeit verwendet, die ein Autofahrer bei vorgegebenen Verkehrsverhältnissen auf einer Teststrecke erreichte. Als Versuchsperson dienten 20 eineiige Zwillinge, von denen jeweils einer zufällig der Bedingung ,,ohne Alkohol" und der andere der Bedingung ,,mit Alkohol" zugeordnet wurde.

Das Korrelationsproblem

(22.4) Definition *Unter dem Korrelationskoeffizienten von zwei Zufallsgrößen X und Y versteht man*

$$\rho = \rho_{XY} = E\left(\frac{X-\mu_X}{\sigma_X} \cdot \frac{Y-\mu_Y}{\sigma_Y}\right)$$

das ist der Erwartungswert des Produkts der Standardscores der Zufallsgrößen.

Sind X und Y zwei Zufallsgrößen mit endlichen Varianzen, so ist

$$\rho(X,Y) = \lim_{n\to\infty} r_n$$

(22.9) *Es seien X und Y zwei Zufallsgrößen in einem Zufallsexperiment. Dann gibt es eine lineare Prognosefunktion $a+bX$, die die Zufallsgröße Y im Sinn des Prinzips der kleinsten Quadrate*

$$E((Y-a-bX)^2) = min!$$

optimal vorhersagt. Diese optimale lineare Prognosefunktion ist gegeben durch

$$b = \rho\frac{\sigma_y}{\sigma_x} \text{ und } a = \mu_Y - b\mu_X.$$

Der Prognosefehler der optimalen linearen Prognosefunktion beträgt

$$E\big((Y-a-bX)^2\big) = (1-\rho^2)\sigma_Y^2.$$

Das Quadrat des Korrelationskoeffizienten gibt also an, um welchen Prozentsatz die Unsicherheit verringert wird, wenn man eine lineare Funktion von X zur Prognose von Y heranzieht.

(22.10) Definition *Die optimale lineare Prognosefunktion $a+bX$ heißt die lineare Regressionsfunktion der Zufallsgröße Y nach der Zufallsgröße X.*

(22.12) Definition *Ein statistischer Test auf Korrelation zweier Zufallsgrößen X und Y ist ein Prüfverfahren, das eine Entscheidung zwischen den Aussagen*

Nullhypothese: $\rho(X,Y) = 0$

Alternative: $\rho(X,Y) \neq 0$

herbeiführt. Die Entscheidung wird auf Grund empirischer Daten getroffen.

Prüfverfahren:

Um das Signifikanzproblem zu beantworten, bildet man die F–Größe

$$F = \frac{MSS^*}{MS_R} = (n-2)\frac{r^2}{1-r^2}$$

der Varianzanalyse im Regressionsproblem.

(22.13) Aufgabe Prüfen Sie den Zusammenhang der Variablen Alter und Intelligenz in der DEMO–Daten.

Der empirische Korrelationskoeffizient zwischen AG und IN beträgt $r = -0,1327$. Daraus ergibt sich als Testgröße

$$(n-2)\frac{r^2}{1-r^2} = 98\frac{0,1327^2}{1-0,1327^2} = 1,7566.$$

Das Symmetrieproblem

Beim Symmetrieproblem geht es um den Vergleich der Erwartungswerte $E(X)$ und $E(Y)$ zweier gekoppelter Zufallsgrößen X und Y.

Das Problem des Vergleichs von Erwartungswerten $E(X)$ und $E(Y)$ gekoppelter Zufallsgrößen behandelt man, indem man den Erwartungswert der Zufallsgröße $D = X - Y$ untersucht.

(22.14) Aufgabe

Prüfen Sie den Unterschied der Erwartungswerte der Variablen T_0 und T_1 in den DEMO–Daten.

Wir berechnen

$$\overline{d} = \overline{x} - \overline{y} = 12,14 - 11,95 = 0,19$$

und

$$s_d^2 = s_x^2 + s_y^2 - 2rs_xs_y = 0,29 + 0,3 - 2 \cdot 0,9543 \cdot 0,54 \cdot 0,55 = 0,02315.$$

Daraus ergibt sich $s_d = 0,15214$ und $s_{d,n-1} = 0,15367$. Die Testgröße beträgt

$$\frac{\overline{d}}{\frac{s_{d,n-1}}{\sqrt{n}}} = \frac{0,19}{\frac{0,15367}{10}} = 12,36.$$

(22.15) Aufgabe

Beantworten Sie das Symmetrieproblem im Beispiel (22.2).

Wir berechnen $\overline{d} = 16,7$ und $s_{d,n-1} = 26,27$. Daraus ergibt sich als Wert der Testgröße $2,012$. Dieser Wert ist schwach signifikant. Wir schließen daraus mit Vorsicht, daß die Eregnisse des Vortags einen Einfluß auf die Verkaufszahlen haben.

(22.16) Aufgabe

Beantworten Sie das Symmetrieproblem im Beispiel (22.3)

Wir berechnen $\overline{d} = 4,75$ und $s_{d,n-1} = 6,06$. Daraus ergibt sich als Wert der Testgröße $3,505$. Dieser Wert ist signifikant. Wir schließen daraus mit Vorsicht, daß Alkohol die Risikobereitschaft erhöht.