Inhaltsverzeichnis

Teil I

Deskriptive Statistik

1 Motivation

Beispiel 1: Indexzahlen

Stand des Dax am14.10.02 um 14^{10}h: 2883,76

PI für die Lebenshaltung:

Jan 1995: 100

Sep 2001: $110,0$

Sep 2002: $111,1$

Sep 2001 $\rightarrow$ Sep 2002 $(\frac{111,1}{110} - 1) \cdot 100 = 1\%$

Dez 2001: $109,6$; in 7 Jahren: $9,6\%$Wachstum

Durchschnitt $\frac{9,7\%}{7} = 1,37\%$ FALSCH!!!

Dito: Kurs einer Aktie:$100 \rightarrow 160 \rightarrow 80$

$60\% - 50\%$; Durchschnitt $= \frac{60\%-50\%}{2} = 5\%$???

Beispiel 2: Aktienkennziffern

Beispiel 3: Optionsbewertung

Beispiel 4: Armut und Ungleichheit

Land A: $0,0,0,3,3,3,5$

Land B: $1,1,1,1,1,1,8$

Durchschnittseinkommen (im Sinne des arithmetischen Mittel) ist in beiden Ländern identisch $(= 2)$

Beispiel 5: Demographie:

P (Ehepaar 20/25 erlebt goldene Hochzeit)=?

P(noch am Leben)= $0,8 \times 0,47 = 0,376$

Δ (Lebenserwartung) bei Elimination von Krebs = $2,9$ Jahre

Beispiel 6: Linguistik: Gedicht von Shakespeare

Wieviele Wörter hat Shakespeare in seinem Leben geschrieben: 864647; davon erschienen: 31544

Weitere Anwendungen

- Wahlhochrechnungen
- Marketing
- Versicherungstabellen
- Portfolio-Management

2 Mittelwerte

2.1 Das arithmetische Mittel

Beispiel 1: Merkmal: Einkommen (quantitativ alias metrisch)

Merkmalsausprägungen: $0, 0, 1, 3, 16$

gesucht: "durchschnittliches" Einkommen

Antwort: arithmetisches Mittel: $\bar{x}_a = \dfrac{0+0+1+3+16}{5} = \dfrac{20}{5} = 4$

allgemein: $\bar{x}_a = \dfrac{x_1 + \ldots + x_n}{n} = \dfrac{1}{n}x_1 + \dfrac{1}{n}x_2 + \ldots + \dfrac{1}{n}x_n$

Im Beispiel: $n = 5$

$x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 3, x_5 = 16 \Rightarrow \bar{x}_a = \frac{1}{5}\cdot 0 + \frac{1}{5}\cdot 0 + \frac{1}{5}\cdot 1 + \frac{1}{5}\cdot 3 + \frac{1}{5}\cdot 16 = \underline{\underline{4}}$

Problem: ungewogenes arithmetisches Mittel oft irreführend

Definition: gewogenes arithmetisches Mittel

Sei X ein metrisches Merkmal mit Ausprägungen $x_1, \ldots, x_n$. Seien $g_1, g_2, \ldots, g_n$ nicht negative, reelle Zahlen mit $\sum_{i=1}^{n} g_i = 1$. Dann heißt $\bar{x}_{ag} := g_1x_1 + g_2x_2 + \ldots + g_nx_n$ **das gewichtete (= gewogene) arithmetische Mittel** der x_i.

Beispiel 2: Durchschnitts-Kosten des Autofahrens:

Kostenanteil Benzin: $90\% = \frac{9}{10} = 0,9$

Kostenanteil Öl: $10\% = \frac{1}{10} = 0,1$

x = Preisanstieg

Preisanstieg von Benzin x_1 : + 50% ; Preisanstieg von Öl x_2: + 10%

$x_1 = 50\%, x_2 = 10\%, g_1 = 0,9,\ g_2 = 0,1 \Rightarrow \bar{x}_a^g = 0,9 \cdot 50\% + 0,1 \cdot 10\% = 46\%$

Eigenschaften des arithmetischen Mittels:

Satz 2.1:

Seien X,Y,Z metrische Merkmale mit Ausprägungen x_i, y_i, z_i ($i = 1, \ldots, n$) und $z_i = ax_i + by_i$. Dann gilt:

$$\bar{\mathbf{z}}_\mathbf{a} = \mathbf{a}\bar{\mathbf{x}}_\mathbf{a} + \mathbf{b}\bar{\mathbf{y}}_\mathbf{a}$$

Achtung: Das gilt für andere Durchschnitte im allgemeinen nicht!

2.2 Der Median (= Zentralwert)

in Beispiel 1: Median = $\bar{x}_m$= 1

Definition:

Der Median ist diejenige Merkmalsausprägung, die bei Anordnung der Größe nach in der Mitte steht.

Vorteile:

- robust gegen "Ausreißer"
- Wert ist immer eine tatsächlich vorkommende Merkmalsausprägung
- auch bei ordinalen Merkmalen anwendbar

Beispiel 3:

5 Restaurants: miserabel, schlecht, mäßig, gut, hervoragend ⇒ Median: mäßig

Weitere Eigenschaften des arithmetischen Mittels und des Medians:

Satz 2.2:

$$\bar{x}_a = argmin_{z \in \mathbb{R}} \left(\sum_{i=1}^{n} (x_i - z)^2 \right)$$

$$\bar{x}_m = argmin_{z \in \mathbb{R}} \left(\sum_{i=1}^{n} |x_i - z| \right)$$

2.3 Das geometrische Mittel

Begründung für den Namen "geometrisch":

Definition:

Sei X ein metrisches Merkmal mit nicht negativen Ausprägungen $x_1, \ldots, x_n$.

Dann heißt $\bar{x}_g := \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$ das **geometrische Mittel** von $x_1, \ldots, x_n$.

Beispiel: $x_1 = 1,\ x_2 = 2,\ x_3 = 4 \ \Rightarrow \bar{x}_g = \sqrt[3]{1 \cdot 2 \cdot 4} = \sqrt[3]{8} = 2$

Hauptanwendung: Durchschnittliche Wachstumsraten

Periode	1	2	3
Kurs X_t	100	160	80
W-Rate r_t		+ 60 % = +0,6	-50 % = - 0,5

gesucht: Durchschnittliche Wachstumsrate?

Todsünde: arithmetisches Mittel = 5%

Korrekt: geometrisches Mittel = $\sqrt[2]{1,6 \cdot 0,5} - 1 = -0,1056 = -10,56\%$

Betrachte zur Begründung ein allgemeines Beispiel:

Anfangskapital: K_0

nach 1 Periode: $K_0 + r_1 \cdot K_0 = K_0(1 + r_1) = K_1$ ($1 + r_1$= Wachstumsfaktor)

nach 2 Periode: $K_2 = K_1(1 + r_2) = K_0(1 + r_1) \cdot (1 + r_2)$

$\vdots$

nach n Perioden: $K_n = K_1(1 + r_1)(1 + r_2) \dots (1 + r_n)$

gesucht: geeigneter Durchschnitt von $r_1, r_2, \dots, r_n (= \bar{r}\ )$

Anforderungen an $\bar{r}$:

$K_0(1 + \bar{r}) \cdot (1 + \bar{r}) \dots (1 + \bar{r}) = K_0(1 + r_1)(1 + r_2) \dots (1 + r_n) = K_0(1 + \bar{r})^n$

Auflösung nach $\bar{r}$: $(1+\bar{r})^n = (1+r_1)\cdot \dots \cdot(1+r_n) \Rightarrow (1+\bar{r}) = \sqrt[n]{(1 + r_1) \dots (1 + r_n)}$

Durchschnittliche W-Rate: $\bar{r} = \sqrt[n]{(1 + r_1)(1 + r_2) \dots (1 + r_n)} - 1$

im Beispiel: $\bar{r} = \sqrt[2]{1,6 \cdot 0,5} - 1 = \sqrt{0,8} - 1 = 0,8944 - 1 = -0,1056$

2.4 Das harmonische Mittel

Zum Namen: 2 Gitarrensaiten der Länge 1 und 1/2 → "harmonisches Mittel": 2/3

Definition:

Sei X ein metrisches Merkmal mit positiven Ausprägungen $x_1, \dots, x_n$

Dann heißt

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

das **harmonische Mittel** von $x_1, \dots, x_n$.

Anwendung: Autofahrt Do → Duisb. → Do, einfache Strecke: 50 km

hin: 1/2 h, d.h. Geschwindigkeit = 100km/h

zurück: 1 h , d.h. Geschwindigkeit = 50 km/h

Mittlere Geschwindigkeit: $\frac{\text{Gesamtstrecke}}{\text{Gesamte Zeit}} = \frac{100km}{1{,}5h} = 66,67km/h = \frac{1}{\frac{1}{2}(\frac{1}{50}+\frac{1}{100})} = \frac{1}{\frac{3}{200}} = \frac{200}{3} = 66,67 = \bar{x}_h$ km/h im Durchschnitt

2.5 Zusammenhang zwischen arithmetischem, geometrischem und harmonischem Mittel

Beispiel:

$n = 2,\ x_1 = 1,\ x_2 = 3$

$\bar{x}_a = \frac{1+3}{2} = \frac{4}{2} = 2$

$\bar{x}_g = \sqrt{1 \cdot 3} = \sqrt{3} = 1,732$

$\bar{x}_h = \frac{1}{\frac{1}{2}(1+\frac{1}{3})} = \frac{6}{4} = 1,5$

Satz 2.3:

Es gilt immer $\bar{x}_h \leq \bar{x}_g \leq \bar{x}_a$

Beweis, daß $\bar{x}_g \leq \bar{x}_a$ für $n = 2$

zu zeigen: $\sqrt{x_1 \cdot x_2} \leq \frac{1}{2}(x_1 + x_2)$

offenbar gilt: $0 \leq (x_1 - x_2)^2 = x_1^2 - 2x_1x_2 + x_2^2| + 4x_1x_2$

$4x_1x_2 \leq x_1^2 + 2x_1x_2 + x_2^2 \mid : 4$

$x_1x_2 \leq \frac{1}{4}(x_1^2 + 2x_1x_2 + x_2) \mid_{\surd}$

$\sqrt{x_1x_2} \leq \frac{1}{2}(x_1 + x_2)$

3 Streuungsmaße

3.1 Problemstellung

Beispiel :

X = Einkommen im Land A: $0, 0, 1, 3, 16$ $\quad \bar{x}_a = 4$

Y = Einkommen im Land B: $3, 3, 4, 5, 5$ $\quad \bar{y}_a = 4$

gesucht: geeignetes Maß für die Streuung

Definition:

$R_x = x_{max} - x_{min}$ heißt **Spannweite** (= "range") von X.

im Beispiel. $R_x = 16 - 0 = 16$; $R_y = 5 - 3 = 2$

Nachteil: Bei der Spannweite wird die kleinste Zahl von der größten subtrahiert. Alles, was zwischen kleinstem und größtem Wert passiert, geht in die Spannweite nicht mit ein.

Definition:

$d_x = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \bar{x}_m|$ heißt **mittlere absolute Abweichung** (vom Median).

$\Delta_x = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$ heißt **mittlerer absoluter Abstand** .

im Beispiel: $d_x = \frac{1}{5}(1 + 1 + 0 + 2 + 15) = \frac{19}{5} = 3,8$

$\Delta x = \frac{1}{25}(0 + 0 + 1 + 3 + 16 + 0 + 0 + 1 + 3 + 16 + 1 + 1 + 0 + 2 + 15 + 3 + 3 + 2 + 0 + 13 + 16 + 16 + 15 + 13 + 0) = \frac{140}{25} = 5,6$

3.2 Die Standardabweichung

Definition:
$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_a)^2$ heißt **mittlere quadratische Abweichung** (alias empirische Varianz).

im Beispiel: $s_x^2 = \frac{1}{5} \left((-4)^2 + (-4)^2 + (-3)^2 + (-1)^2 + (12)^2\right)$
$= \frac{1}{5}(16 + 16 + 9 + 1 + 144) = \frac{186}{5} = 37,2$

Trick für praktische Berechnung

Satz 3.1:
$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x}_a)^2$

Im Beispiel: $\frac{1}{5} \sum_{i=1}^{5} x_i^2 = \frac{1}{5}(0 + 0 + 1 + 9 + 256) = \frac{266}{5} = 53,2$
$(\bar{x}_a)^2 = 4^2 = 16$, d.h. $\frac{1}{n} \sum x_i^2 - (\bar{x}_a)^2 = 53,2 - 16 = 37,2$

Nachteil: $y_i = ax_i \rightarrow s_y^2 = a^2 s_x^2$

Definition:
$s_x = \sqrt{s_x^2}$ heißt **Standardabweichung** .
$y_i = ay \Rightarrow s_y = |a| s_x$
im Beispiel: $s_x = \sqrt{37,2} = 6,099$

3.3 Eigenschaften von s_x und s_x^2

Satz 3.2:

Es gilt immer:

(i) $y_i = ax_i \Rightarrow s_y^2 = a^2 s_x^2$

$s_y = |a| s_x$

(ii) $y_i = x_i + b \Rightarrow s_x^2 = s_y^2$

Beispiel:

$y_i = 0, 0, 2, 6, 32$ $(y_i = 2x_i$ $(x_i = 0, 0, 1, 3, 16)$ $)$

$\bar{y}_a = 2\bar{x}_a = 2 \cdot 4 = 8$

$s_y^2 = \frac{1}{5}(64 + 64 + 36 + 4 + 576) = 148,8 = 4 \cdot 37,2$

4 Maße für Konzentration und Ungleichheit

Beispiel 1:

X = Einkommen in Land A: $0, 0, 1, 3, 16$

Y = Einkommen in Land B: $16, 16, 17, 19, 32$

$\Big\} s_x^2 = s_y^2 = 37,2 \Rightarrow s_x = s_y = 6,099$

offenbar gilt:

	in A	in B	
die 20% Ärmsten haben	0%	16%	des Gesamteinkommens
die 40% Ärmsten haben	0%	32%	des Gesamteinkommens
die 60% Ärmsten haben	5%	49%	des Gesamteinkommens
die 80% Ärmsten haben	20%	68%	des Gesamteinkommens
alle genannten haben	100%	100%	des Gesamteinkommens

4.1 Die Lorenzkurve

Anteil der i Ärmsten : $\dfrac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_j}$ $\nwarrow$ der Größe nach sortiert, von klein nach groß

Definition:

Der Polygonzug durch die Punkte $(\dfrac{i}{n}, \dfrac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_j})$ mit $(i = 0, \ldots, n)$ heißt

"Lorenzkurve" .

Beispiel 2: Einkommen privater Haushalte in Deutschland 1998 (netto DM/Monat)

	HH (Mio)	ges. Einkommen (Mio)
< 2500	$7,77$	14127
$2500 - 5000$	$14,13$	51666
$5000 - 10000$	$11,99$	82566
≥ 10000	$2,92$	39786
zusammen	$36,81$	188145

← gruppierte Daten

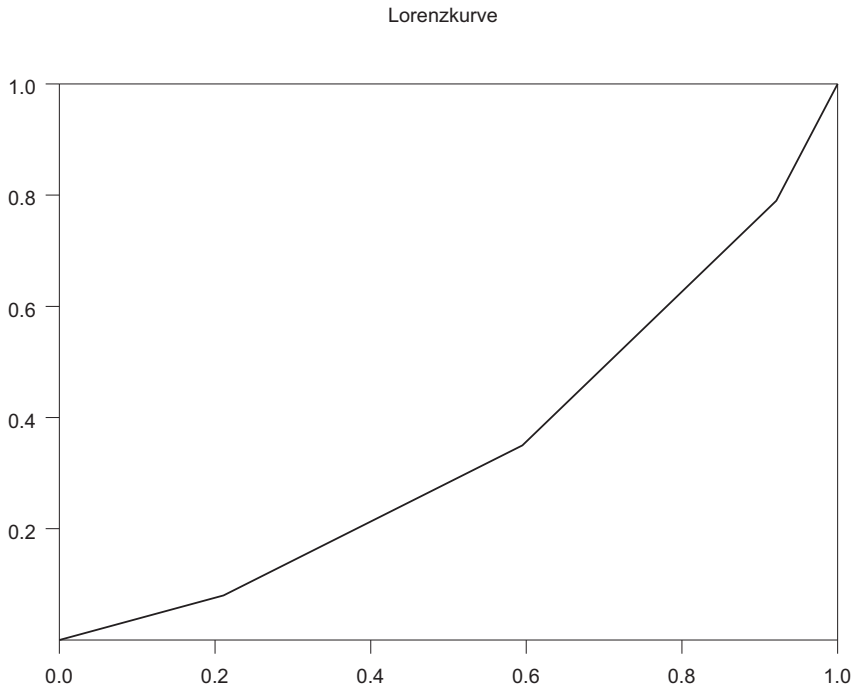
Lorenzkurve der Einkommen von Deutschland 1988:

die $\frac{7,77}{36,81} = 21\%$ Ärmsten haben $\frac{14127}{188145} = 7,5\%$ des Gesamteinkommens,

die $\frac{7,77+14,13}{36,81} = 59\%$ Ärmsten haben $\frac{14127+51666}{188145} = 35\%$ des Gesamteinkommens,

die $\frac{7,77+11,99+14,13}{36,81} = 92\%$ Ärmsten haben 82% des Gesamteinkommens,

die 100% Ärmsten haben 100% des Gesamteinkommens.

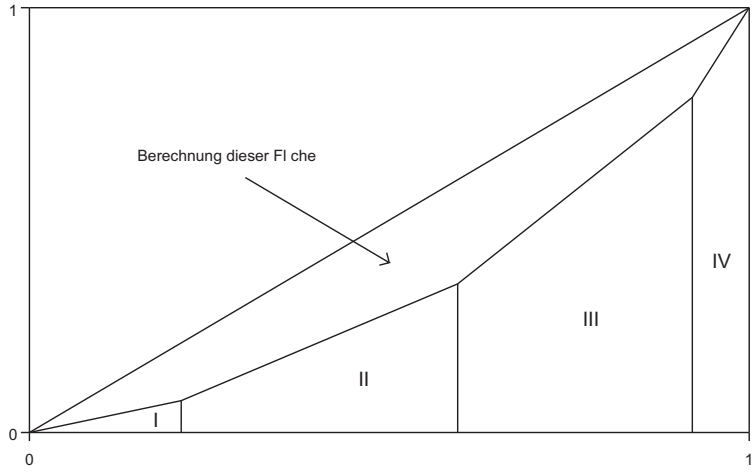


4.2 Eigenschaften der Lorenzkurve

Satz 4.1:

- Die Lorenzkurve geht immer durch die Punkte $(0, 0)$ und $(1, 1)$.
- Sie verläuft nie oberhalb der Winkelhalbierenden.
- Lorenzkurve = Winkelhalbierende $\Leftrightarrow$ Alle Merkmalsausprägungen sind identisch.
- Ungleichheit umso größer, je weiter Lorenzkurve von der Winkelhalbierenden entfernt
- Die Lorenzkurve bleibt gleich, wenn man alle Einkommen mit dem gleichen Faktor multipliziert.

Definition:
Das Doppelte der Fläche zwischen Lorenzkurve und Winkelhalbierender (= "Konzentrationsfläche") heißt **"Gini-Koeffizient"** (nach Corrado Gini, ital. Statistiker).
im Ausgangsbeispiel:



I: $\frac{0,2\times 0,05}{2} = 0,005$

II: $\frac{0,2\times 0,25}{2} = 0,025$

III: $0,2 \cdot \frac{1,2}{2} = 0,12$

d.h. Konzentrationsfläche $= 0,5 - 0,005 - 0,025 - 0,12 = 0,35$

$\Rightarrow$ Gini-Koeffizient: $G_x = 0.35 \cdot 2 = 0,7$

4.3 Eigenschaften des Gini-Koeffizienten

Satz 4.2:

- $0 \leq G_x \leq 1$: Gini-Koeffizient liegt immer zwischen 0 und 1.
- $G_x = 0 \Leftrightarrow$ alle x_i sind gleich.
- $y_i = ax_i \Rightarrow G_y = G_x$
- $G_x = \dfrac{\Delta x}{2\bar{x}_a}$ mit $\Delta x = \dfrac{1}{n^2} \sum_i \sum_j |x_i - x_j|$

Alternative Berechnung des Gini-Koeffizienten im Anfangsbeispiel:
$0, 0, 1, 3, 16 \Rightarrow \bar{x} = 4$, $\Delta x = \dfrac{140}{25} \Rightarrow G_x = \dfrac{140}{25 \times 8} = \dfrac{140}{200} = 0,7$

$\Delta x = |0-0|+|0-0|+|0-1|+|0-3|+|0-16|+|0-0|+|0-0|+|0-1|+|0-3|$
$+|0-16|+|1-0|+|1-0|+|1-3|+\ldots = 140$

Beispiel 3: 3 Firmen, mit Umsätzen $100, 40, 10$
$G_x = \dfrac{\Delta x}{2\bar{x}_a}$ mit $\bar{x}_a = 50$; $2\bar{x}_a = 100$; $\Delta x = \frac{1}{9}(60 + 90 + 60 + 30 + 90 + 30) = \dfrac{360}{9} = 40$; $G_x = \dfrac{40}{100} = 0,4$
Angenommen, weitere Firma mit Umsatz 0 kommt dazu:
$\bar{x} = \frac{150}{4}$, $\Delta x = \frac{360+100+40+10+100+40+10}{16} = \frac{360+300}{16} = \frac{660}{16}$
$\Rightarrow G_x = \frac{\frac{660}{16}}{\frac{300}{4}} = \frac{660 \cdot 4}{300 \cdot 16} = 0,55$ (größer als alter Gini-Koeffizient)

4.4 Der Koeffizient von Herfindahl

Definition:
$H_x = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i} \right)^2$ heißt **Herfindahl-Koeffizient.**

im Beispiel:
$H_x = \left(\frac{100}{150}\right)^2 + \left(\frac{40}{150}\right)^2 + \left(\frac{10}{150}\right)^2 + 0^2 = 0,52 \leftarrow$ ist der gleiche vor - und nach Hinzunahme des Unternehmens mit Umsatz 0

Eigenschaften:

Satz 4.3:

- $\frac{1}{n} \leq H_x \leq 1$
- $H_x = 1 \Leftrightarrow$ alle x_i außer einem sind 0
- $H_x = \frac{1}{n} \Leftrightarrow$ alle x_i sind gleich

Häufiges Problem: Umsätze kleiner Firmen unbekannt

Firma Nr.	Umsatzanteil	
1	40 %	} 95%
2	25 %	
3	20 %	
4	10 %	

gesucht: $H_x : \underbrace{0,4^2 + 0,25^2 + 0,2^2 + 0,1^2}_{=0,2725} + \underbrace{\ldots}_{?}$
es gilt: $0 < Rest \leq 0,05^2 = 0,0025$
$0,2725 + 0,05^2 = 0,2750$, d.h. $0,2725 < H_x \leq 0,2750$

5 Maße für Korrelation und Abhängigkeit

5.1 Problemstellung

bisher: 1 Merkmal (= Variable) pro Merkmalsträger

jetzt: 2 Merkmale (nur für den Fall, daß beide Merkmale metrisch sind)

Beispiel:

Merkmalsträger	Merkmal 1	Merkmal 2
gebrauchter PKW	Alter	Preis
Mietwagen	Größe	Mietpreis
Börsentag	Rendite BMW	Rendite Daimler
Bundesliga	Tabellenpunkte	geschossene Tore
erwachsener Bundesbürger	Schulbildung	Einkommen

usw...

Wie kann man das Ausmaß für den Zusammenhang zwischen 2 Merkmalen sinnvoll festhalten?

Im weiteren: beide Merkmale metrisch!

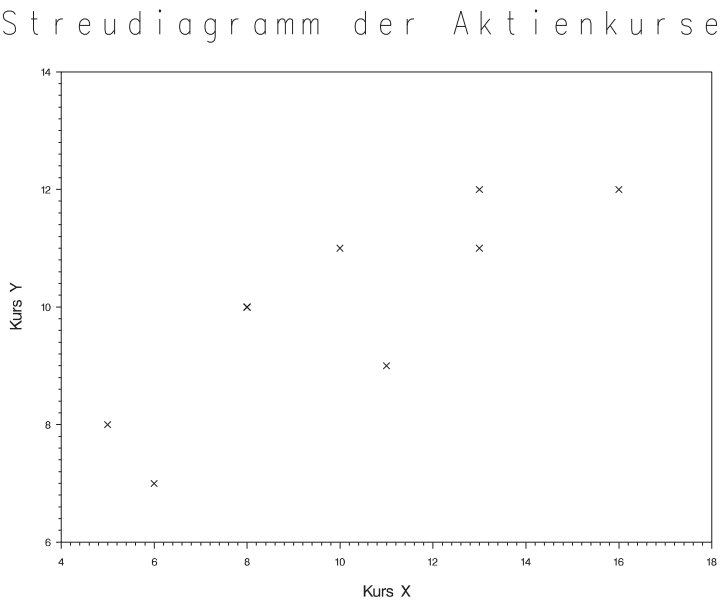
5.2 Die (empirische) Kovarianz

Beispiel: Kurse zweier Aktien X und Y:

t	1	2	3	4	5	6	7	8	9
X	5	6	11	8	13	8	10	16	13
Y	8	7	9	10	11	10	11	12	12

mit $\bar{x} = 10$ und $\bar{y} = 10$

1. Schritt: Streudiagramm :



Unterschiedliche Gewichtung der Datenpunkte:

$$(x_i - \bar{x})(y_i - \bar{y}) \begin{cases} > 0 & \text{, falls} \quad x_i > \bar{x}, y_i > \bar{y} \quad \text{oder} \quad x_i < \bar{x}, y_i < \bar{y} \\ = 0 & \text{, falls} \quad x_i = \bar{x} \quad \text{oder} \quad y_i = \bar{y} \\ < 0 & \text{, falls} \quad x_i > \bar{x}, y_i < \bar{y} \quad \text{oder} \quad x_i < \bar{x}, y_i > \bar{y} \end{cases}$$

Definition:

$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = s_{xy}$ heißt **(empirische) Kovarianz** von X und Y.

im Beispiel:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$x_i y_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
5	8	-5	-2	10	40	25	4
6	7	-4	-3	12	42	16	9
11	9	1	-1	-1	99	1	1
8	10	-2	0	0	80	4	0
13	11	3	1	3	143	9	1
8	10	-2	0	0	80	4	0
10	11	0	1	0	110	0	1
16	12	6	2	12	192	36	4
13	12	3	2	6	156	9	4
$\sum = 90$	$\sum = 90$	$\sum = 0$	$\sum = 0$	$\sum = 42$	$\sum = 942$	$\sum = 104$	$\sum = 24$

Ergebnis: $s_{xy} = \frac{42}{9} = 4,67$

Eigenschaften der (empirischen) Kovarianz:

Satz 5.1:

- $s_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$
  im Beispiel: $\frac{942}{9} - (10 \cdot 10) = 104,67 - 100 = 4,67$

- $z_i = y_i + a \Rightarrow s_{xz} = s_{xy}$ die empirische Kovarianz ändert sich nicht, wenn eine Konstante zu einem Merkmal addiert wird.

- $z_i = a y_i \Rightarrow s_{xz} = a s_{xy}$
  allgemein: $s_{(ax+b)(cy+d)} = ac s_{xy}$

- $|s_{xy}| \leq s_x s_y$

- $|s_{xy}| = s_x s_y \Leftrightarrow y_i = a x_i + b$ mit $a \neq 0$

5.3 Der Bravais-Pearson Korrelationskoeffizient

im Beispiel: $s_{xy} = 4,67$

Wie ist dieser Wert zu interpretieren?

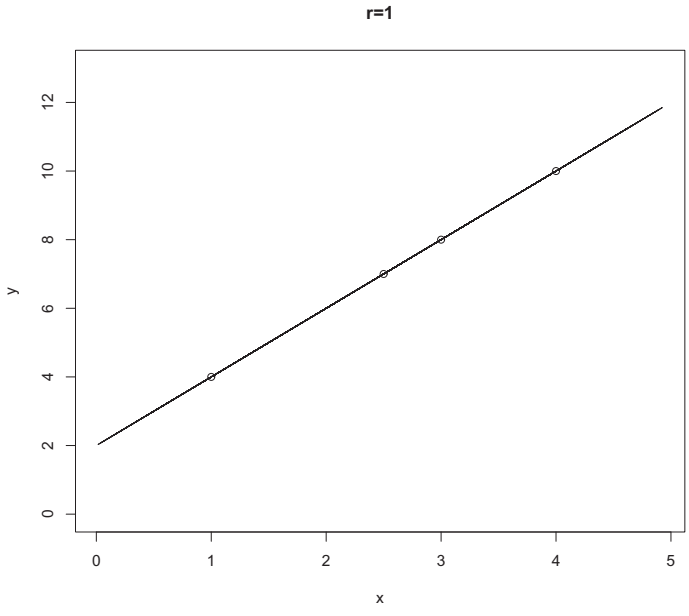
Definition:

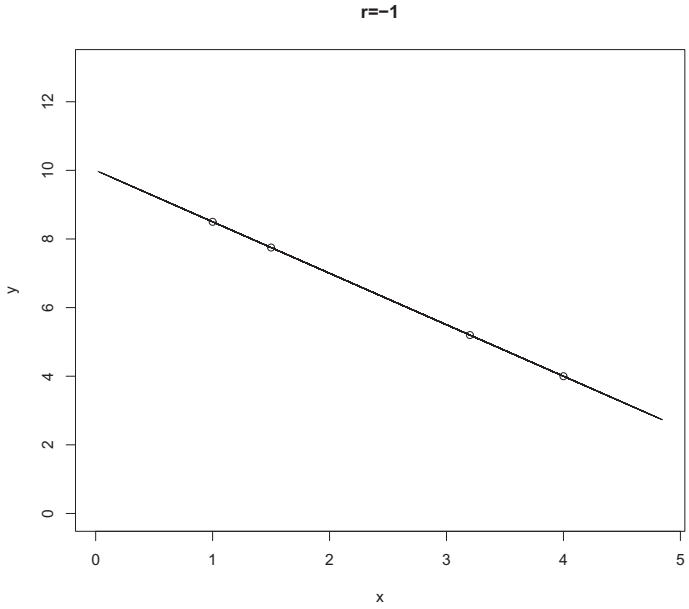
$$r_{xy} := \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

heißt **Bravais-Pearson Korrelationskoeffizient.**

Eigenschaften von r_{xy}:

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} = +1 \Leftrightarrow y_i = ax_i + b$ mit $a > 0$ (größte positive Korrelation)
- $r_{xy} = -1 \Leftrightarrow y_i = ax_i + b$ mit $a < 0$ (größte negative Korrelation)





$r_{xy} = 0$ bedeutet, dass kein linearer Zusammenhang besteht, aber es sind andere Zusammenhänge möglich.

5.4 Korrelation und Kausalität

Ausgangspunkt: 2 metrische Variablen X und Y; n Wertepaare $(x_1, y_1), \ldots, (x_n, y_n)$.
$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ Empirische Kovarianz
$R_{xy} = \frac{s_{xy}}{s_x s_y}$ Korrelationskoeffizient wobei $-1 \leq \frac{s_{xy}}{s_x s_y} \leq 1$

häufiger Trugschluß: x und y sind korreliert $\Rightarrow$ x ist Ursache für y.

alternative Erklärung:

- y ist Ursache für x
- $z \begin{array}{l} \nearrow x \\ \searrow y \end{array}$ beide Variablen hängen von einer dritten Variable ab.
  (wichtigste dritte Variable ist die Zeit)
- $z \rightarrow y \leftarrow x$ x verursacht y. x beeinflusst y negativ, aber dennoch ist der Korrelationskoeffizient positiv.
  (z.B. Länge der Studiendauer und Gehalt)

DAX-Kennzahlen

$Adidas_1, \ldots, Adidas_{30}$: letzte 30 Adidas-Renditen

$Dax_1, \ldots, Dax_{30}$: letzte 30 Dax-Renditen

$r_{Adidas,DAX} = 0,4979 = \frac{s_{Adidas,DAX}}{s_{Adidas} s_{DAX}}$

Beta $= \frac{s_{Adidas,DAX}}{s^2_{DAX}} = 0,4930$

CAPM $= E(Rendite_{Adidas}) = \underbrace{r_f}_{\text{risikofreie Rendite}} + Beta_{Adidas}$

d.h.: Die erwartete Rendite einer Anlage in ein risikoreiches Papier ist gleich der Rendite für ein festverzinsliches Wertpapier plus Beta mal die Differenz der erwarteten Rendite einer Marktanlage minus dem festverzinslichen Wertpapier.

6 Elementare Regressionsrechnung

Ziel: Beschreibung des Zusammenhangs zwischen 2 Merkmalen durch eine lineare Funktion

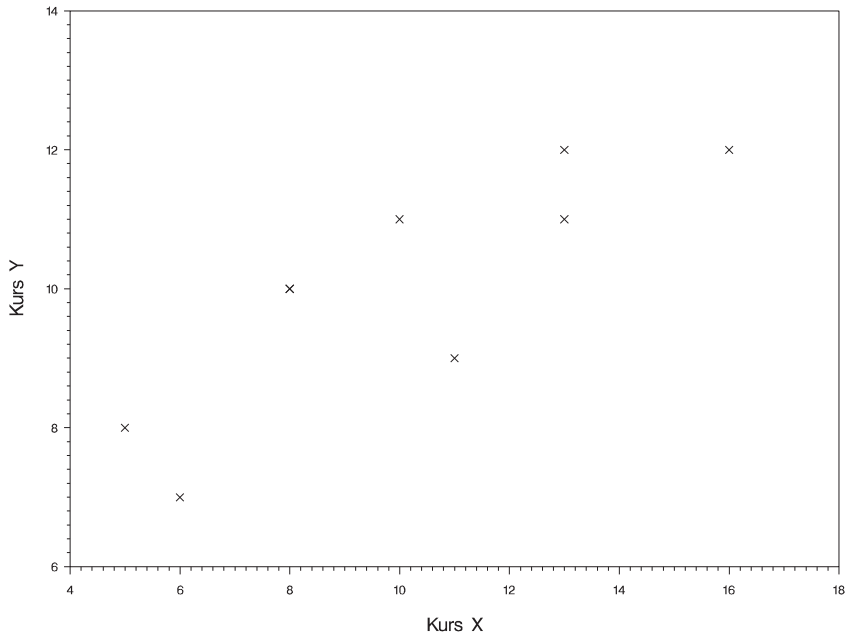
6.1 Die Methode der kleinsten Quadrate

Beispiel:

Aktienkurse $(x_i\ y_i)$ aus Kapitel 5 mit $\bar{x} = 10$ und $\bar{y} = 10$

t	1	2	3	4	5	6	7	8	9
X	5	6	11	8	13	8	10	16	13
Y	8	7	9	10	11	10	11	12	12

Streudiagramm der Aktienkurse



gesucht: Gerade , welche diese Punkte $(x_i; y_i)$ möglichst gut approximiert.

Vorschläge:

- nach Augenmaß
- verbinde Extrempunkte
- minimiere die Summe der absoluten Abweichungen
- minimiere die Summe der quadrierten Abstände (nach Gauß)

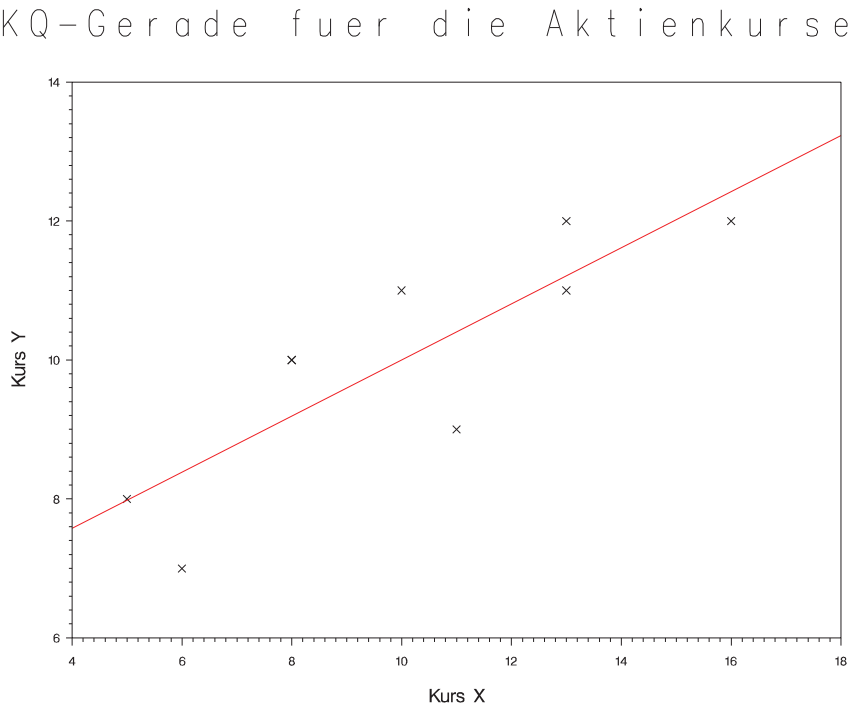
Definition:
Die Gerade $y = a + bx$ durch die Punktewolke $\{(x_i; y_i)\}$, welche die Summe der quadrierten vertikalen Abstände minimiert, heißt **KQ-Ausgleichsgerade.**

Eigenschaften:

- die KQ-Gerade geht immer durch den Punkt $(\bar{x}, \bar{y})$
- die Steigung der KQ-Geraden ist gegeben durch
  $b = \dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{s_{xy}}{s_x^2}$
- der Achsenabschnitt der KQ-Geraden ist gegeben durch $a = \bar{y} - b\bar{x}$

Bestimmung der KQ-Geraden im Beispiel der Aktienkurse:
Steigung der KQ-Geraden:
$b = \frac{s_{xy}}{sx^2} = \frac{4{,}67}{\frac{104}{9}} = \frac{4{,}67}{11{,}56} = 0,40$
Achsenabschnitt der KQ-Geraden:
$a = \bar{y} - b\bar{x} = 10 - 0,4 \cdot 10 = 6$
$\Rightarrow KQ - Gerade: \ y = 6 + 0,4 \cdot x$

Graphische Darstellung:



6.2 Das lineare Regressionsmodell

Ausgangspunkt:

Zwei metrische Variablen

x: Regressor, unabhängige Variable, erklärende Variable, Design-Variable

y: Regressand, abhängige Variable, zu erklärende Variable

Beispiele:

x	y
Einkommen	Konsum
Werbeausgaben	Umsatz
Alter gebrauchter PKW	Preis
Größe einer Wohnung	Miete
Menge Düngemittel	Ernteertrag
⋮	⋮

wichtig:

- a priori bekannt: x verursacht y
- nur eine erklärende Variable
  (sonst: "multiple Regressionsanalyse")
- y hängt linear von x ab: $y = a + bx$

Beispiel für nichtlinearen Zusammenhang:

x = Alter PKW, y = Preis

$y = a \cdot \frac{1}{x} + b$

Lösung: Definiere neuen Regressor $x^* = \frac{1}{x}$

Beispiel: Cobb-Douglas Produktionsfunktion

Output $= \lambda \cdot \text{Arbeit}^{\beta} \cdot \text{Kapital}^{\gamma}$ ist nicht linear

Lösung: logarithmieren

$ln(\text{Output}) = ln(\lambda) + \beta ln(A) + \beta ln(K)$

im weiteren: 1 Regressor x und linearer Zusammenhang: (entweder alle übrigen Einflußgrößen konstant oder nur eine Ursache):

$y \approx a + bx \Rightarrow y = a + bx +$ Störung

Problem: Bestimmung von a und b

Lösung: Approximiere a und b durch die Koeffizienten der KQ-Geraden

wichtig: ceteris-paribus Bedingung (alles andere bleibt gleich)

7 Preisindizes

7.1 Die Indexformel nach Laspeyres

(Etienne Laspeyres, deutscher Statistiker 1834 - 1913)

Beispiel: Konsumausgaben eines ausgewählten Wirtschaftssubjektes

	Periode 0 (=Basisperiode)		Periode 1 (=Berichtsperiode)	
	Preis	Menge	Preis	Menge
	p_0	q_0	p_1	q_1
Zigaretten	2,-	8	4,-	4
Fertigpizza	5,-	4	3,-	9
Kino	6,-	2	11,-	1
Rotwein	3,-	4	2,-	6
	$\bar{p}_0 = \frac{16}{4} = 4$		$\bar{p}_1 = \frac{20}{4} = 5$	

gesucht: durchschnittliche Preisänderung = ?

grober Unfug: Vergleich der Durchschnittspreise

Genauso dumm: Vergleich der Gesamtausgaben

$GA_0 = \sum_{i=1}^{4} p_0(i)q_0(i) = 60$

$GA_1 = \sum_{i=1}^{4} p_1(i)q_1(i) = 66$

Definition:

$P_{0t}^L := \frac{\sum_{i=1}^{n} p_t(i)q_0(i)}{\sum_{i=1}^{n} p_o(i)q_o(i)}$ heißt **Preisindex nach Laspeyres** mit Basisperiode 0 und Berichtsperiode t.

Im Beispiel:
$\sum_{i=1}^{4} p_1(i)q_o(i) = 4 \cdot 8 + 3 \cdot 4 + 11 \cdot 2 + 2 \cdot 4 = 74$
d.h. $P_{01}^L = \frac{74}{60} = 1,233$

d.h.: mittlerer Preisanstieg von $23,3\%$

Satz 7.1**:**

Sei $g_0(i) = \frac{p_0(i)q_0(i)}{\sum_{j=1}^{n} p_0(j)q_0(j)}$. Dann gilt:
$P_{0t}^L = \sum_{i=1}^{n} \frac{p_t(i)}{p_0(i)} g_0(i)$
(gewogenes arithmetisches Mittel der individuellen Preisverhältnisse)

Im Beispiel: $\sum_{i=1}^{4} \frac{p_1(i)}{p_0(i)} g_0(i) = \frac{4}{2} \cdot \frac{16}{60} + \frac{3}{5} \cdot \frac{20}{60} + \frac{11}{6} \cdot \frac{12}{60} + \frac{2}{3} \cdot \frac{12}{60} = 1,233$

7.2 Der Preisindex nach Paasche

(nach Herrmann Paasche, deutscher Statistiker 1851 - 1922)

Definition:

$P_{0t}^P = \frac{\sum_{i=1}^{n} p_t(i) \cdot q_t(i)}{\sum_{i=1}^{n} p_0(i) \cdot q_t(i)}$ heißt **Preisindex nach Paasche** .

Im Beispiel:
$\sum_{i=1}^{4} p_0(i) \cdot q_1(i) = 2 \cdot 4 + 5 \cdot 9 + 6 \cdot 1 + 3 \cdot 6 = 77,-$
d.h.: $P_{01}^P = \frac{66}{77} = 0,857 = 85,7$ gesunken!

$\rightarrow$ Nach Laspeyres gestiegen, nach Paasche gesunken

Satz 7.2:
Sei $g_t(i) = \frac{p_0(i)q_t(i)}{\sum_{j=1}^{n} p_0(j)q_t(j)}$. Dann gilt: $P_{0t}^P = \sum_{i=1}^{n} \frac{p_t(i)}{p_0(i)} g_t(i)$.

Andere Gewichte als bei Laspeyres!

Im Beispiel: $P_{01}^P = \frac{4}{2} \cdot \frac{8}{77} + \frac{3}{5} \cdot \frac{45}{77} + \frac{11}{6} \cdot \frac{6}{77} + \frac{2}{3} \cdot \frac{18}{77} = \frac{6}{7} = 0,857$

Definition:
$P_{0t}^F = \sqrt{P_{0t}^L \cdot P_{0t}^P}$ heißt **"idealer Preisindex nach Fisher"** .

Im Beispiel: $P_{01}^F = \sqrt{1,223 \cdot 0,857} = 1,028$

7.3 Preisindex für die Lebenshaltung

Grundlage: Indexformel von Laspeyres
Vorteil: Verbrauchsdaten müssen nur für Basisperiode erhoben werden!

5 Teilprobleme:

- Bestimmung des Warenkorbs (aktuell: Warenkorb von 2000, n = 750 Güter)
- Auswahl von "Preisrepräsentanten"
- Messung der Preise
- Berücksichtigung von Qualitätsänderungen

7.4 Spezialprobleme von Aktienindices

Eigenheit: Was vorher grober Unfug war, ist jetzt erlaubt!

1. Fall: Dow-Jones (eigentlich D. J. Industrial Average * 26.05.1896)

Problem: Austausch alter Werte gegen Neue

Beispiel:

vorher: $\frac{60+70+110}{3} = \frac{240}{3} = 80$

nachher: $\frac{100+70+110}{3} = \frac{280}{3} = 93\frac{1}{3}$

Lösung: Teile 280 nicht durch 3, sondern durch $3,5 \quad \rightarrow \frac{280}{3,5} = 80$

2. Fall: DAX (* 31.12.1987)

$DAX_t = \frac{\text{Gesamtwert des Portfolios heute}}{\text{Gesamtwert des Portfolios am } 31.12.1987} \cdot 1000$

Ausgangspunkt: Portfolio von 30 Werten

Teil II

Wahrscheinlichkeitsrechnung

8 Zufällige Ereignisse und ihre Wahrscheinlichkeiten

8.1 Ausgewählte Beispiele

P(6 Richtige im Lotto) $= \frac{1}{13.983.816} = 0,000000071$

P(Aktienkurs steigt an 3 von 5 Tagen an) $= \frac{10}{32}$

P(Bei 30 zufällig ausgewählten Personen haben mind. 2 den gleichen Geburtstag) $= 71\%$

usw...

Preisfrage:

Wie rechnen wir solche Wahrscheinlichkeiten aus?

8.2 Zufällige Ereignisse

(Bamberg/Baur, Kapitel 7.1 - 7.3)

Beispiel 1: Einmaliges Würfeln ("Zufallsvorgang")

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

Ereignisse: Teilmengen von Ω

Elementarereignisse =1-elementige Ereignisse: $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$ und $\{6\}$.

Verbal	Mengendarstellung
Gerade Zahl	$A = \{2, 4, 6\}$
Ungerade Zahl	$B = \{1, 3, 5\}$
Primzahl	$C = \{1, 2, 3, 5\}$
Keine Primzahl	$D = \{4, 6\}$
Zahl > 3	$E = \{4, 5, 6\}$

Definition:

$\bar{A}$ = Menge aller Elemente von Ω, die nicht in A liegen. (= Komplementärmenge von A)

Definition:

Vereinigungsmenge:

$A \cup B$ = Menge aller Elemente von Ω, die in A oder B oder in beiden liegen.

Definition:

Schnittmenge:

$A \cap B$ = Menge aller Elemente von Ω, die sowohl in A als auch in B liegen.

Definition:

Zwei Ereignisse A und B heißen **unvereinbar** (=**disjunkt**),

$\Leftrightarrow A \cap B = \emptyset$.

Verbal	Mengendarstellung
Ungerade Zahl <u>oder</u> Zahl > 3	$B \cup E = \{1, 3, 4, 5, 6\}$
Primzahl <u>und</u> Zahl > 3	$C \cap E = \{5\}$
Keine Primzahl	$\bar{C} = \{4, 6\}$
Gerade Zahl <u>und</u> ungerade Zahl	$A \cap B = \emptyset$

Beispiel 2: Zweimaliges Würfeln

$$\Omega = \{(1,1)(1,2)(1,3)(1,4)(1,5)(1,6)$$
$$(2,1)(2,2)(2,3)(2,4)(2,5)(2,6)$$
$$\vdots$$
$$(6,1)(6,2)(6,3)(6,4)(6,5)(6,6)\}$$

$= \{1,2,3,4,5,6\} \otimes \{1,2,3,4,5,6\} \rightarrow$ kartesisches Produkt

$|\Omega| = 6 \cdot 6 = 6^2 = 36$

Beispiel 3: 3-maliger Münzwurf

$\Omega = \{(KKK),(KKZ),(KZK),(KZZ),(ZKK),(ZKZ),(ZZK),(ZZZ)\} =$

$\{K,Z\} \otimes \{K,Z\} \otimes \{K,Z\}$

$|\Omega| = 2 \cdot 2 \cdot 2 = 2^3 = 8$

Satz 8.1:

Allgemein gilt: Wird ein einfacher Zufallsvorgang mit K Elementarereignissen n-mal wiederholt, so hat der zusammengesetzte Zufallsvorgang K^n Elementarereignisse.

8.3 Wahrscheinlichkeiten von zufälligen Ereignissen

im Beispiel 2:

A = beide Zahlen sind gleich = $\{(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)\}$

B = keine 6 = $\{(1,1)\ldots(5,5)\}$

C = nur ungerade Zahlen = $\{(1,3)(1,5)(3,1)(3,5)(5,1)(5,3)(1,1)(3,3)(5,5)\}$

D = Augensumme gleich 7 = $\{(1,6)(2,5)(3,4)(4,3)(5,2)(6,1)\}$

gesucht: zugehörige Wahrscheinlichkeiten

Annahme: Alle Elementarereignisse sind gleichwahrscheinlich (= Laplace-Experiment)

Satz 8.2:

In einem Laplace Experiment gilt:
$P(A) = \frac{\text{Anzahl aller günstigen Ereignisse}}{\text{Anzahl aller möglichen Ereignisse}}$

- $P(A) = \frac{|A|}{|\Omega|} = 6/36 = 1/6$
- $P(B) = 25/36$
- $P(C) = 9/36 = 1/4$
- $P(D) = 6/36 = 1/6$

Rechenregeln für Wahrscheinlichkeiten:

Satz 8.3:

Es gilt immer (auch außerhalb von Laplace-Experimenten)

- $P(\Omega) = 1$
- $P(\emptyset) = 0$
- $P(\bar{A}) = 1 - P(A)$
- falls A und B disjunkt (unvereinbar): $P(A \cup B) = P(A) + P(B)$
- allgemein: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Beispiel: 2-maliges Würfeln

A: nur gerade Zahlen

B: nur ungerade Zahlen

$P(A) = P(B) = \frac{9}{36} = \frac{1}{4}$

$A \cap B = \emptyset \Rightarrow P(A \cup B) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

8.4 Unabhängige Ereignisse und bedingte Wahrscheinlichkeiten

Beispiel: 2-maliges Würfeln

A: erster Wurf eine 6

B: zweiter Wurf eine 6

$$\Omega = \{(1,1)(1,2)(1,3)(1,4)(1,5)(1,6)$$
$$(2,1)(2,2)(2,3)(2,4)(2,5)(2,6)$$
$$\vdots$$
$$(6,1)(6,2)(6,3)(6,4)(6,5)(6,6)\}$$

Definition:

Zwei Ereignisse A und B heißen **unabhängig** $\Leftrightarrow$

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

im Beispiel:

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

$P(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} \rightarrow$ A und B sind unabhängig !

Beispiel für nicht unabhängige Ereignisse:

A: mindestens eine 6 $\quad P(A) = \frac{11}{36}$

B: Augensumme $\geq 6 \quad P(B) = \frac{26}{36}$

$P(A \cap B) = \frac{11}{36} \neq P(A) \cdot P(B) = \frac{11}{36} \cdot \frac{26}{36}$

Preisfrage: P(A), wenn ich weiß, daß B eingetreten ist?

Definition:

Die Wahrscheinlichkeit für A in einem neuen Zufallsexperiment mit Ω = B heißt **bedingte Wahrscheinlichkeit von A gegeben B.** Formal: $P(A|B)$

Satz 8.4:

Falls $P(B) \neq 0$ gilt: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

im Beispiel:

$P(A \cap B) = P(A) = \frac{11}{36}$; $P(B) = \frac{26}{36}$; $P(A|B) = \frac{11/36}{26/36} = \frac{11}{26}$

Satz 8.5:

A und B unabhängig $\Rightarrow P(A|B) = P(A)$

Beweis:

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

Weitere Anwendungen von Satz 8.3:

Satz 8.6:

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

Anwendung:

P(2 Asse bei 2-maligem Ziehen ohne Zurücklegen aus einem 32-er Kartenspiel)

A_1: Ass bei ersten Zug

A_2: Ass beim zweiten Zug

gesucht: $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{4}{32} \cdot \frac{3}{31}$

8.5 Weitere Anwendungen

Beispiel 1: Geburtstagsproblem

N zufällig ausgewählte Personen

A=„mindestens 2 Personen haben gleichen Geburtstag"

gesucht: $P(A)$

Trick: Betrachte stattdessen:

$\bar{A}$ = „alle Geburtstage sind verschieden"

$P(\bar{A}) = 1 - P(A) \Rightarrow P(A) = 1 - P(\bar{A})$

Voraussetzung: 365 Tage, alle als Geburtstage gleich wahrscheinlich

$\Omega = \{1, 2, 3, \ldots, 365\} \otimes \{1, 2, 3, \ldots, 365\} \otimes .... \otimes \{1, 2, 3, \ldots, 365\}$ (n-mal)

$|\Omega| = 365^n$

$|\bar{A}| = 365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - n + 1)$

$\Rightarrow P(\bar{A}) = \frac{365 \cdot 364 \cdot ..... \cdot (365-n+1)}{365^n}$

Ausrechnen ergibt:

N	P($\bar{A}$)	P(A)=1- P($\bar{A}$)
2	364/365 = 0,997	0,003
4	0,98	0,02
6	0,95	0,05
8	0,92	0,08
10	0,88	0,12
15	0,74	0,26
20	0,58	0,42
25	0,43	0,57
30	0,29	0,71

Beispiel 2: Fluktuation von Aktienkursen

Angenommen, P(Kurs steigt) = P(Kurs fällt) $= 0,5$

gesucht: P(Kurs steigt an 3 von 5 Börsentagen einer Woche)

Ω = {Menge aller 5- elementigen Folgen von + und - }

$= \{+-\} \otimes \{+-\} \otimes \{+-\} \otimes \{+-\} \otimes \{+-\}$

$|\Omega| = 2^5 = 32$

A = { Menge aller Folgen mit 3 mal + }

gesucht: $|A|$ = | Menge aller 3- elementigen Teilmengen einer Menge aus 5 Elementen |

Satz 8.7:

Sei $k \leq n$. Dann gibt es $\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$ k - elementige Teilmengen einer Menge vom Umfang n.

im Beispiel:

$|A| = \binom{5}{3} = \frac{5!}{3! \cdot 2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1) \cdot (2 \cdot 1)} = 10$

d.h.: $P(A) = \frac{|A|}{|\Omega|} = \frac{10}{32} = \frac{5}{16}$

9 Zufallsvariablen und Verteilungsfunktionen

9.1 Definitionen und Überblick

Beispiele für Zufallsvariablen:

Z_1=Augensumme bei 2-maligem Würfeln

Z_2=Anzahl "Kopf" bei 3-maligem Münzwurf

Z_3=Lebensdauer einer zufällig ausgewählten Glühbirne

Z_4=Anzahl erfolgloser Tipps bis zum 1. Lotto-Hauptgewinn

Z_5=log(Kursänderung einer Aktie an einem Börsentag)

Z_6=Anzahl positiver Kursänderungen an 10 Börsentagen

etc.

Definition:

Eine Variable Z, deren mögliche Werte ("Realisationen") vom Ausgang eines Zufallsvorgangs abhängen, heißt Zufallsvariable.

Z heißt stetig $\Leftrightarrow$ Z kann (evtl. innerhalb gewisser Grenzen) alle möglichen reellen Zahlen als Wert annehmen.

Z heißt diskret $\Leftrightarrow$ Z kann nur endlich viele (bzw. abzählbar viele) Werte annehmen.

von Interesse:

- welchen Wert nimmt die Variable im Mittel an? ($\rightarrow$ "Erwartungswert")
- wie stark schwankt die Variable um den Erwartungswert? ($\rightarrow$ "Varianz", "Standardabweichung")

Variable	Typ	Wertebereich
Z_1	diskret	$\{2, 3, 4, 5, ..., 12\}$
Z_2	diskret	$\{0, 1, 2, 3\}$
Z_3	stetig	$[0, \infty)$
Z_4	diskret	$\{0, 1, 2, 3, 4, ...\}$
Z_5	stetig	$(-\infty, \infty)$
Z_6	diskret	$\{0, 1, ..., 10\}$

Notation im weiteren:

Variablen selbst: große Buchstaben

mögliche Werte: kleine Buchstaben

später:

Z_2, Z_6:binomialverteilt

Z_3:exponentialverteilt

Z_4:geometrisch verteilt

Z_5:normalverteilt

9.2 Wahrscheinlichkeits- und Verteilungsfunktion bei diskreten Zufallsvariablen

X = Anzahl Kopf bei 3-maligem Münzwurf

Ergebnismenge =

$\{(ZZZ)$	(KZZ)	(KKZ)	(ZKZ)	(ZZK)	(KZK)	(ZKK)	$(KKK)\}$
0	1	2	1	1	2	2	3

offenbar gilt:

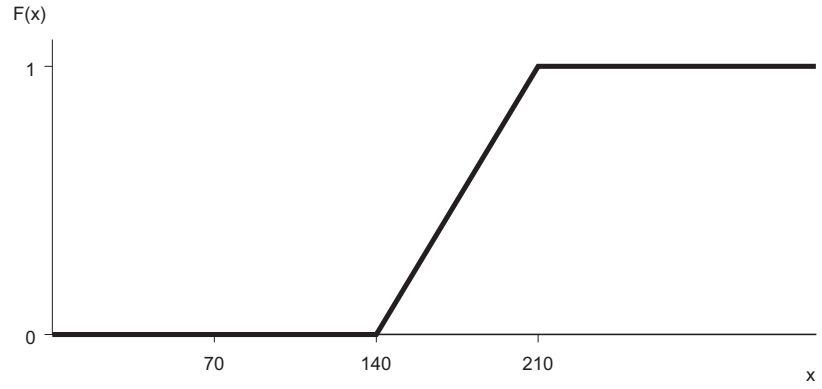
$P(X = 0) = \frac{1}{8}$; $P(X = 1) = \frac{3}{8}$

$P(X = 2) = \frac{3}{8}$; $P(X = 3) = \frac{1}{8}$

Verteilungsfunktion ist eine Sprungfunktion in Form einer Treppe

Wahrscheinlichkeits- und Verteilungsfunktion bei stetigen Zufallsvariablen z.B.

Körpergröße eine Mannes über $18 = X$;$140 \leq X \leq 210$



$P(X = 180,3598) \approx 0$, $P(170 \leq X \leq 180) = ?$

f(x) Dichtefunktion ist die Ableitung der Verteilungsfunktion

Satz 9.1:

Sei X eine beliebige Zufallsvariable mit Verteilungsfunktion F(x):

- $P(a < X \leq b) = F(b) - F(a)$
- F(x) monoton steigend
- $\lim\limits_{x \to -\infty} F(x) = 0$
- $\lim\limits_{x \to \infty} F(x) = 1$
- $f(x) \geq 0$, $F(x) \geq 0$
- $\lim\limits_{x \to -\infty} f(x) = 0$, $\lim\limits_{x \to \infty} f(x) = 0$

Für stetige Zufallsvariablen gilt zusätzlich: $F(b) - F(a) = \int\limits_a^b f(x)dx$

10 Erwartungswert und Varianzen von Zufallsvariablen

10.1 Motivation

Definition:

Sei X eine diskrete ZV mit den Werten $x_1, \ldots, x_n$ und Wahrscheinlichkeitsfunktion $f(x_i)$, dann heißt:

$$E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

Erwartungswert von X .

Beispiel: Augenzahl beim einmaligen Würfeln

x_i	P(X=x_i)	$x_i \cdot f(x_i)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6

$\sum_{i=1}^{n} x_i f(x_i) = \frac{21}{6} = 3,5 = E(X)$

Definition:

Sei X eine stetige Zufallsvariable mit Dichtefunktion $f(x)$. Dann ist der Erwartungswert definiert als $E(X) = \int_{-\infty}^{\infty} x f(x) dx$ definiert.

Achtung: $E(X)$ muss nicht notwendigerweise existieren !!!

Beispiel:

$X \sim$ GV $[0, 5]$

$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \frac{1}{5} \int_{0}^{5} x dx$

$= \frac{1}{5} \left[\frac{1}{2} x^2\right]_0^5 = \frac{1}{5} \left[\frac{25}{2} - 0\right] = \frac{25}{10} = 2,5$

10.2 Eigenschaften von Erwartungswerten

Satz 10.1: (Gesetz der großen Zahlen)
Seien $x_1, x_2, \ldots, x_n$ unabhängige Beobachtungen einer Zufallsvariablen X (genauer: Realisationen von n unabhängigen ZV'en, die alle die gleiche Verteilungsfunktion wie X haben). Dann gilt immer:

$$\lim_{x \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_i = E(X)$$

Problem:
gegeben: ZV X; $E(X)$
gesucht: $E(10X)$, $E(X^2)$, $E(X/2)$, $E(X+Y)$, $E(X \cdot Y)$

Satz 10.2:
Für beliebige ZV'en $X_1, X_2, \ldots, X_n$ gilt immer:

(i) $E(aX + b) = aE(X) + b$

(ii) $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$

(iii) $E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i)$

(iv) Für unabhängige ZV'en X und Y gilt darüber hinaus:
$E(X \cdot Y) = E(X) \cdot E(Y) \leftarrow$ im Allgemeinen falsch!!

Beispiel: X = Augenzahl beim einmaligen Würfeln
bekannt: $E(X) = 3,5$
gesucht: $E(X^2) = ?$
Vermutung: $E(X^2) = [E(X)]^2 = 3,5^2 = 12,25$ FALSCH!!!

Werte x_i	W'keiten $f(x_i)$	$x_i \cdot f(x_i)$
1	1/6	1/6
4	1/6	4/6
9	1/6	9/6
16	1/6	16/6
25	1/6	25/6
36	1/6	36/6

d.h.: $E(X^2) = 1/6(1 + 4 + 9 + 16 + 25 + 36) = 91/6 = 15,1\bar{6} > (E(X))^2$

10.3 Die Varianz von Zufallsvariablen

$E[X - E(X)] = E(X) - E(X) = 0$ (Satz 10.2)

Frage hier: Wie stark schwankt die ZV um ihren Erwartungswert?

Definition:

Sei X eine beliebige ZV. Dann heißt $Var(X) = E[(X - E(X))^2] = \sigma_x^2$ die **Varianz von X,** und $\sigma_X :=\sqrt{\sigma_X^2}$ heißt die **Standardabweichung von X** .

Satz 10.3:

$Var(X) = E(X^2) - [E(X)]^2$

Beispiel: X = Augenzahl beim einmaligen Würfeln

bekannt: $E(X) = 3,5$

gesucht: $Var(X) = E[(X - 3,5)^2]$

Werte	W'keiten
$(1\text{-}3{,}5)^2 = 6,25$	$1/6$
$(2\text{-}3{,}5)^2 = 2,25$	$1/6$
$(3\text{-}3{,}5)^2 = 0,25$	$1/6$
$(4\text{-}3{,}5)^2 = 0,25$	$1/6$
$(5\text{-}3{,}5)^2 = 2,25$	$1/6$
$(6\text{-}3{,}5)^2 = 6,25$	$1/6$

$E[(X-3,5)^2] = \frac{1}{6}(6,25+2,25+0,25+0,25+2,25+6,25) = \frac{1}{6}\cdot 17,5 = 2,91\bar{6}$

alternativ:

$Var(X) = E(X^2) - (E(X))^2 = 15,1\bar{6} - (3,5)^2 = 15,1\bar{6} - 12,25 = 2,91\bar{6}$

Satz 10.4:

Seien X und Y beliebige ZV'en. Dann gilt immer:

(i) $Var(X) \geq 0$, $Var(Y) \geq 0$

(ii) $Var(aX) = a^2 Var(X)$

(iii) $Var(X+a) = Var(X)$

(iv) Falls X, Y unabhängig: $Var(X+Y) = Var(X) + Var(Y)$

(v) Allgemein: für n unabhängige ZV'en $X_1, X_2, \ldots X_n$ gilt:
$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \; Var(X_i)$

Vorsicht: $Var(X-Y) = Var(X) - Var(Y)$ ist FALSCH für X,Y unabhängig

Sondern: $Var(X-Y) = Var(X+(-Y)) = Var(1\cdot X + (-1)\cdot Y)$

$= 1^2 Var(X) + (-1)^2 Var(Y) = Var(X) + Var(Y)$

10.4 Kovarianz und Korrelation von Zufallsvariablen

Definition:

Seien X und Y zwei ZV, mit dem gleichen zugrundeliegenden Zufallsexperiment.

$Cov(X,Y) := E[(X - E(X))(Y - E(Y))]$ heißt **Kovarianz** von X und Y.

$\rho_{X,Y} := \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ heißt **Korrelation** von X und Y.

Nützlich für praktische Berechnung:

Satz 10.5:

Seien X und Y beliebige ZV.

$Cov(X,Y) = E(XY) - ((E(X)) \cdot (E(Y)))$

Beweis:

$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] \rightarrow$ Ausmultiplizieren und Erwartungswert bilden.

Beispiel: 3-maliger Münzwurf

X = Anzahl Kopf

Y = Anzahl Zahl

gesucht: $Cov(X;Y)$, ρ_{XY}

$\Omega = \{$	(ZZZ)	(KZZ)	(KKZ)	(ZKZ)	(ZZK)	(KZK)	(ZKK)	$(KKK)\}$
X	0	1	2	1	1	2	2	3
Y	3	2	1	2	2	1	1	0
$X \cdot Y$	0	2	2	2	2	2	2	0

Werte von $X \cdot Y$	zugehörige W'keit
0	$2/8 = 1/4$
2	$6/8 = 3/4$

X	$P(X=x)$	Y	$P(Y=y)$
0	1/8	0	1/8
1	3/8	1	3/8
2	3/8	2	3/8
3	1/8	3	1/8

$E(XY) = 0 \cdot 1/4 + 2 \cdot 3/4 = 6/4 = 3/2$

$E(X) = 0 \cdot 1/8 + 1 \cdot 3/8 + 2 \cdot 3/8 + 3 \cdot 1/8 = 12/8 = 3/2 = E(Y) \rightarrow Cov(X,Y) = 3/2 - (3/2 \cdot 3/2) = -3/4$

Korrelation: $E(X^2) = 0^2 \cdot 1/8 + 1^2 \cdot 3/8 + 2^2 \cdot 3/8 + 3^2 \cdot 1/8 = 12/8 = 24/8 = 3 = E(Y^2)$

$Var(X) = E(X^2) - [E(X)]^2 = 3 - (3/2)^2 = 3/4 = Var(Y)$

$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{-3/4}{\sqrt{3/4}\sqrt{3/4}} = -1$

Satz 10.6:

Seien X und Y beliebige ZV:

(i) $|Cov(X,Y)| \leq \sqrt{Var(X)} \cdot \sqrt{Var(Y)}$

(ii) X, Y stochastisch unabhängig: $Cov(X,Y) = 0$. Daraus folgt, daß auch die Korrelation $\rho_{XY} = 0$ ist.

(iii) $Var(a \cdot X + b \cdot Y) = a^2 Var(X) + b^2 Var(Y) + 2a \cdot b \cdot Cov(X,Y)$

11 Ausgewählte Typen von Zufallsvariablen im Detail

11.1 Die binomialverteilte Zufallsvariable

Beispiel: Betrachte 5 Börsentage lang den DAX.

Annahme: DAX steigt oder fällt.

P(DAX steigt) = P(DAX fällt) = $0,5$

X:= Anzahl der Tage, an denen DAX steigt.

gesucht: $P(X = i)$ mit $i = 0, 1, 2, 3, 4, 5$

Wahrscheinlichkeit, daß DAX an genau 2 Börsentagen steigt. $P(X = 2) =?$

Definition:

Eine diskrete ZV heißt **binomialverteilt** mit Parametern n und $p \Leftrightarrow$

X zählt die Erfolge bei n unabhängigen Versuchen mit Erfolgswahrscheinlichkeit p.

Satz 11.1:

X $\sim$ Bin(n,p) $\Longrightarrow f(x)$=$P(X$=$x)$= $\binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$

Interpretation der Wahrscheinlichkeitsfunktion:

- x = Anzahl der Erfolge (mit Wahrscheinlichkeit p)
- $n - x$ = Zahl der Mißerfolge (mit Wahrscheinlichkeit $(1 - p)$)
- $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ = Zahl der möglichen Anordnungen von Erfolgen und Mißerfolgen

Im Beispiel: $X \sim \text{Bin}(5, 0.5)$

Somit: P(X= 2)= $\binom{5}{2} \cdot 0,5^2 \cdot (1-0,5)^{5-2} = \frac{5!}{2!\cdot 3!} = \frac{20}{2} \cdot \frac{1}{32} = 0,3125 = 31,25\%$

Beispiel: Februar 2003 hat 20 Börsentage.

gesucht: Wahrscheinlichkeit, daß DAX im Feb. an mehr als 8 Tagen steigt.

X $\sim$ Bin(20,0.5)

gesucht: $P(X > 8)$

Es gilt $P(X > 8) = 1 - P(X \leq 8) = 1 - F(8)$

$= 1 - \sum_{x=0}^{8} \binom{20}{x} \cdot 0,5^x \cdot 0,5^{20-x} \rightarrow$ siehe Verteilungstafel

$= 1 - 0,2517 = 0,7483 = 74,83\%$

wichtiger Spezialfall: n = 1 $\rightarrow$ X $\sim$ Bin(1,p) "Bernoulli Verteilung"

Es gilt: $X_1, \ldots, X_n$, u.i.v. $X_i \sim$ Bin (1,p) $\Rightarrow \sum_{i=1}^{n} X_i \sim$ Bin(n,p)

Satz 11.2:

X $\sim$ Bin(n,p). Dann gilt:

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

(ii) $Var(X) = n \cdot p \cdot (1-p)$

11.2 Normalverteilung

Definition:

Eine stetige ZV mit Dichtefunktion

$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$; $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$ heißt **normalverteilt** mit Parametern μ und σ^2; kurz $X \sim N(\mu, \sigma^2)$.

Satz 11.3:

Eigenschaften der Normalverteilung. Sei $X \sim N(\mu, \sigma^2)$

(i) $E(X) = \mu$

(ii) $Var(X) = \sigma^2$

(iii) $f(\mu - x) = f(\mu + x)$, d.h. die Dichte ist symmetrisch um μ

(iv) argmax f(x) $= \mu$

Problem: Sei $X \sim N(5, 9)$, dann lässt sich $\int \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$ nur sehr schwer bestimmen.

Ausweg:

(i) Führe beliebige Normalverteilung auf N(0,1) zurück. (Standardnormalverteilung)

(ii) Bestimme F(x) $= \phi(x)$ für $X \sim N(0, 1)$ mit Hilfe numerischer Methoden und trage die Ergebnisse in Tabellen ein.
$X \sim N(\mu, \sigma^2) \rightarrow Z := \frac{X-\mu}{\sigma} \sim N(0, 1)$.
Die Verteilungsfunktion der N(0,1) $= \Phi$ ist vertafelt.

(iii) Der Zentrale Grenzwertsatz
gesucht: Verteilungsfunktionen von Summen von Zufallsvariablen.

Teil III

Induktive Statistik

12 Punktschätzungen

(Bamberg/Baur 12.1)

12.1 Problemstellung

bisher: Gegeben eine Zufallsvariable X mit bekannter Verteilungsfunktion F(x).
gesucht: $P(X \leq a)$; $P(a < X \leq b)$; $E(X)$, $Var(X)$ usw. (= "Parameter")
jetzt: Gegeben n unabhängige Realisationen einer Zufallsvariablen X.
(konkret: Realisationen von $X_1, \ldots X_n$, die alle die gleiche Verteilungsfunktion wie X besitzen.)
Aber: Verteilungsfunktion unbekannt!!!

Problem: Rückschluss von $X_1, \ldots, X_n$ (= "Stichprobe") auf $E(X)$, $Var(X)$, $P(X \leq a)$;
bzw. allgemein: Rückschluss auf unbekannten Parameter θ.

Beispiel:
X = Körpergröße eines zufällig ausgewählten Bundesbürgers > 18
Y = Rendite BMW an einem bestimmten Börsentag
Z = Lebensdauer eines VW-Golf Motors

12.2 Schätzung unbekannter Erwartungswerte

gegeben: X_i = Rendite BMW am Börsentag Nr. i ($i = 1, \ldots n$); n Realisationen $x_1, \ldots, x_n$

gesucht: $\mu = E(X_i)$

Lösung: Approximiere μ durch $\hat{\mu} = \frac{1}{n}(X_1 + \ldots + X_n)$

Satz 12.1:

Seien $X_1, \ldots, X_n$ unabhängige ZV'en mit identischer Verteilungsfunktion F(x) und $E(X_i) = \mu$. Dann ist

$\hat{\mu}(X_1, \ldots, X_n) := \frac{1}{n}(X_1 + \ldots + X_n)$ eine erwartungstreue Schätzfunktion für μ.

Beweis:

$E\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] = \frac{1}{n}\left[E(X_1) + \ldots E(X_n)\right] = \frac{1}{n}\underbrace{[\mu + \mu + \ldots + \mu]}_{n-mal}$

$= \frac{1}{n} \cdot n \cdot \mu = \mu$

12.3 Schätzung unbekannter Wahrscheinlichkeiten

gesucht: $P(X_i > 3\%) = ?$

Lösung: Approximiere P durch $\hat{p}$ = Stichprobenanteil

Satz 12.2:

Seien $X_1, \ldots, X_n$ unabhängige ZV'en mit identischer Verteilungsfunktion F(x) und $P(X_i \leq a) = p$ ($\leftarrow$ unbekannt). Dann ist $\hat{p} = \frac{\sum\limits_{X_i \leq a} X_i}{n}$ eine erwartungstreue Schätzfunktion für p.

Spezialfall: X := Merkmal eines zufällig aus einer real existierenden Grundgesamtheit ausgewählten Merkmalsträgers.

Beispiel:

X = IQ eines zufällig ausgewählten BWL-Studenten.

Angenommen es gibt 100.000 BWL-Studenten mit IQ's

$x_1, \ldots x_{100.000}$ ($\leftarrow$ die möglichen Werte von X.)

$E(X) = x_1 \cdot P(X = x_1) + \ldots + x_{100.000} \cdot P(X = x_{100.000}) =$

$x_1 \cdot \frac{1}{100.000} + \ldots + x_{100.000} \cdot \frac{1}{100.000} = \frac{1}{100.000}(x_1 + \ldots + x_{100.000}) =$

arithmetisches Mittel der Grundgesamtheit =: μ

Angenommen: $\mu = 100$

Ziehe Stichproben vom Umfang $n = 3$

Erste Stichprobe: $90, 99, 120 \rightarrow \bar{x}^{(1)} = 103 = \hat{\mu}^{(1)}$ (Schätzung, nicht wahres arithmetisches Mittel). $\hat{\mu}$ ist ZV.

Zweite Stichprobe: $107, 96, 100 \rightarrow \bar{x}^{(2)} = 101 = \hat{\mu}^{(2)}$

Dritte Stichprobe: $110, 105, 118 \rightarrow \bar{x}^{(3)} = 111 = \hat{\mu}^{(3)}$

Diese Zufallsvariablen schwanken um den wahren Mittelwert herum. Im Mittel stimmen die Schätzungen $\hat{\mu}$ mit dem Mittelwert μ überein: $E(\hat{\mu}) = \mu$.

12.4 Schätzung unbekannter Varianzen

$Var(X) = \sigma^2 = E\left[(X - E(X))^2\right]$

gesucht: Schätzung für σ^2 basierend auf Stichprobe $X_1, \ldots, X_n$

bereits in vorherigen Kapiteln gesehen: $\frac{1}{n} \sum_{i=1}^{n} \left(X_i - E(X)\right)^2$ ist erwartungstreue Schätzung für $\sigma^2 = E\left[(X - E(X))^2\right]$.

Problem: E(X) ist auch unbekannt.

Lösung: Ersetze $E(X)$ durch Schätzung $\frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$

Aber: $\frac{1}{n} \sum_{i=1}^{n} \left(X_i - E(X)\right)^2 \geq \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$\rightarrow \sigma^2 = E[\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X))^2] \geq E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2]$
d.h.: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ unterschätzt das wahre σ^2

Satz 12.3:

Unter den Bedingungen von Satz 12.2 ist

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

eine erwartungstreue Schätzung für $\sigma^2 := Var(X_i)$.

13 Intervallschätzungen (=Konfidenzintervalle)

13.1 Motivation

bisher: Versuch, unbekannten Parameter θ punktgenau zu treffen (Punktschätzung).
jetzt: Versuch, θ in einem Intervall "einzufangen".

Linke Intervallgrenze: V_u
Rechte Intervallgrenze: V_o
Konfidenzintervall KI: $[V_u, V_o]$
$P([V_u, V_o] \not\ni \theta)$ = Irrtumswahrscheinlichkeit α
$P([V_u, V_o] \ni \theta)$ = Vertrauenswahrscheinlichkeit bzw. Konfidenzniveau $= 1 - \alpha$

13.2 KI'e für unbekannte Erwartungswerte μ bei normalverteilten Stichproben-Variablen mit bekannter Varianz σ^2

Beispiel:
X = Durchschnittseinkommen (EUR in Tausend/Jahr) eines zufällig ausgewählten WiSo-Absolventen mit 2-jähriger Berufserfahrung.
Zufallsstichprobe: $(x_1, x_2, x_3, x_4, x_5) = (35, 70, 58, 63, 74)$
Aus dem Kapitel 12: optimale Schätzung für $\mu = E(X)$:
arithmetisches Mittel der Stichprobe $= \bar{x} = 60$.
gesucht: KI $= [V_u, V_o]$, so daß Wahrscheinlichkeit $P([V_u, V_o] \ni \mu) = 95\%$, d.h. $\alpha = 5\%$.

Es gilt allgemein:

(i) $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{\sigma^2}} \sim N(0,1)$ (aus Satz 11.4)
$\rightarrow c = 97,5\%$ Fraktil der Standardnormalverteilung

(ii) d.h.: $P\left(-c \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq c\right) = 95\%$.

Aber: $\left(-c \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq c\right) \Rightarrow -c\sigma \leq \sqrt{n}(\bar{X}-\mu) \leq c\sigma$
$\Rightarrow \frac{-c\sigma}{\sqrt{n}} \leq (\bar{X}-\mu) \leq \frac{c\sigma}{\sqrt{n}}$
$\Rightarrow -c\frac{\sigma}{\sqrt{n}} \leq \mu - \bar{X} \leq c\frac{\sigma}{\sqrt{n}}$
$\Rightarrow \underbrace{\bar{X} - c\frac{\sigma}{\sqrt{n}}}_{V_u} \leq \mu \leq \underbrace{\bar{X} + c\frac{\sigma}{\sqrt{n}}}_{V_o}$

Satz 13.1:
Sei c das $\left(1-\frac{\alpha}{2}\right)$- Fraktil der Standardnormalverteilung. Seien die Stichprobenvariablen $X_1, \ldots, X_n$ normalverteilt mit $E(X) = \mu$ und bekannter Varianz σ^2. Dann ist ein KI für μ zum Konfidenzniveau $1-\alpha$ gegeben durch:
$[V_u, V_o]$ mit $V_u = \bar{X} - c\frac{\sigma}{\sqrt{n}}$; $V_o = \bar{X} + c\frac{\sigma}{\sqrt{n}}$.

im Beispiel:
$\alpha = 5\% \rightarrow c = 1,96$ ($\leftarrow$ aus Tabelle)
$\bar{x} = 60$, $\sigma^2 = 100$ ($\sigma = 10$) sei bekannt.
$\Rightarrow V_u = 60 - 1,96\frac{10}{\sqrt{5}} = 51,23$
$\Rightarrow V_o = 60 + 1,96\frac{10}{\sqrt{5}} = 68,77$
d.h.: $P([51,23; 68,77]$ umfasst wahres Durchschnittseinkommen$) = 95\%$.

Länge des KI's: $V_o - V_u = \frac{2\sigma c}{\sqrt{n}}$

Das KI ist umso kürzer:

- je größer n
- je kleiner σ^2
- je größer α (denn je größer α, desto kleiner ist c)

13.3 KI'e für μ bei normalverteilten X_i und unbekanntem σ^2

Bei unbekannten σ^2: Ersetze σ durch $\sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})} =: S$.

Das liefert $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$, was leider nicht mehr standardnormalverteilt ist, sondern $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$ hat eine sogenannte t-Verteilung mit n-1 Freiheitsgraden.

Faustregel: Wenn $n \geq 20$: Nehme Formel wie bei bekanntem σ^2.

Wenn $n \geq 30$: Die Annahme normalverteilter X_i ist nicht mehr nötig.

13.4 KI's für unbekannte Wahrscheinlichkeiten

(Bamberg/Baur 13.3)

Beispiel:

θ = unbekannter wahrer Wähleranteil einer Partei A

gesucht: KI für θ

Zufallsstichprobe: $X_1, \ldots, X_n$ mit

$$X_i = \begin{cases} 1 & \text{i-te Person wählt Partei A} \\ 0 & \text{sonst} \end{cases} \quad \leftarrow \text{Bernoulli-Variable}$$

d.h.: $E(X_i) = P(X_i = 1) = \theta$ = wahrer unbekannter Wähleranteil,

$Var(X_i) = E(X_i^2) - (E(X_i))^2 = \theta - \theta^2 = \theta \cdot (1 - \theta)$.

Schätzung für θ: $\hat{\theta} = \bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ = Stichprobenanteil für Partei A
2 Möglichkeiten σ^2 zu schätzen:

- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- $\hat{\sigma}^2 = \bar{X}(1-\bar{X})$: für unbekanntes θ $\bar{X}$ einsetzen
  d.h. $\hat{\sigma} = \sqrt{\bar{X}(1-\bar{X})} = \sqrt{\hat{\theta}(1-\hat{\theta})}$

Satz 13.2:
Sei c das $\left(1 - \frac{\alpha}{2}\right)$ Fraktil der Standardnormalverteilung, $n \geq 30$; $n\bar{X} \geq 5$,
$n(1-\bar{X}) \geq 5$. Dann ist ein KI für zum Niveau $1-\alpha$ gegeben durch $\left[\bar{X} - c\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + c\frac{\hat{\sigma}}{\sqrt{n}}\right]$.

Problem: Länge des Intervalls $L = V_o - V_u = 2\frac{c\hat{\sigma}}{\sqrt{n}} = 2\frac{c\sqrt{\bar{X}(1-\bar{X})}}{\sqrt{n}}$ hängt von
$\hat{\sigma} = \sqrt{\bar{X}(1-\bar{X})}$ ab!
Aber: $\bar{X}(1-\bar{X}) \leq \frac{1}{4}$, d.h.: $\hat{\sigma} \leq 0,5$, d.h.: $L \leq \frac{c}{\sqrt{n}}$.

14 Statistische Signifikanztests

(Bamber/Baur: 14.1)

14.1 Problemstellung

bisher: keine Vorinformationen, Punkt- und Intervallschätzungen für unbekannte Parameter

jetzt: Es liegt bereits eine Vermutung ("Nullhypothese" H_0) zu einem unbekannten Parameter oder sonstigen Eigenschaften von ZVen vor.

Beispiel 1:

μ_M = durchschnittlicher IQ aller Männer

μ_F = durchschnittlicher IQ aller Frauen

$H_0 : \mu_M = \mu_F$

Beispiel 2:

μ = Erwartungswert der Laufleistung eines zufällig ausgewählten VW-Golf-Motors

$H_0 : \mu \geq 200.000$ km (beispielsweise)

Beispiel 3:

θ = P(DAX fällt an einem zufällig ausgewählten Börsentag um mehr als 10%)

$H_0 : \theta \leq \frac{1}{1000}$ (wichtig für "value at risk")

Beispiel 4:

Der IQ eines zufällig ausgewählten BWL-Studenten ist eine normalverteilte ZV

Beispiel 5:

Die ZVen X=Einkommen und Y=Religion (mit Y=1 für evangelisch und Y=0 sonst) sind unabhängig

u.s.w.

Vorgangsweise immer die gleiche:

- H_0 formulieren
- Stichprobe ziehen
- Entscheiden aufgrund der Stichprobe, ob H_0 ablehnen oder nicht

	Lehne H_0 ab	Lehne H_0 nicht ab
H_0 richtig	**Fehler 1. Art**	Korrekte Entscheidung
H_0 falsch	Korrekte Entscheidung	**Fehler 2. Art**

Definition:
P(Fehler 1. Art) heißt **Signifikanzniveau** eines Tests ($= \alpha$).

Traditionelle Vorgehensweise der Statistik:

- Gebe maximale P(Fehler 1. Art) vor; üblicherweise 5%.
- Suche Entscheidungsregel, die unter dieser Restriktion die Wahrscheinlichkeit für einen Fehler 2. Art minimiert.

14.2 Testen von Hypothesen über Erwartungswerte normalverteilter Zufallsvariablen

Beispiel aus 13.2: X = Einkommen (Euro in Tsd/Jahr) eines zufällig ausgewählten BWL-Absolventen in Deutschland ($\sim N(\mu, \sigma^2)$)
$H_0 : E(X) = \mu \geq 65 =: \mu_0$
Stichprobe:
Annahme: $X \sim N(\mu, \sigma^2)$ mit $\sigma^2 = 100$ bekannt
1. Schritt: Wähle Signifikanzniveau (etwa $\alpha = 5\%$)

2. Schritt: Berechne sog. "Prüfgröße" (="Teststatistik") V, von der wir die Entscheidung abhängen lassen.

Hier: $V = \frac{\bar{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} \sim N(0,1)$ (falls $\mu = \mu_0$)

3. Schritt: Bestimme sogenannten "Ablehnungsbereich".

Hier: Lehne ab für $V \leq 1,645 \leftarrow 5\%$ Fraktil der Standardnormalverteilung, d.h. Ablehnungsbereich $= (-\infty, 1.645)$.

Im Beispiel: $\alpha = 5\%$ $(\Rightarrow c_\alpha = 1,645)$

4. Schritt: Prüfe, ob V $\in$ Ablehnungsbereich.

V $= \sqrt{5} \cdot \frac{(60-65)}{\sqrt{100}} = -1,12$, d.h. H_0 wird nicht abgelehnt.

Probleme:

(i) σ^2 unbekannt. Lösung: ersetze σ^2 durch S^2
Aber: Dann hat V keine Normalverteilung, sondern eine sogenannte t-Verteilung (für n $\geq$ 20 irrelevant).

(ii) Die X_i sind nicht normalverteilt. Lösung: Berufung auf den zentralen Grenzwertsatz. Ab n $\geq$ 30 verfahre wie gehabt.

Verteilung der Stichprobenvariablen X_i	Prüfgröße (unter H_0 exakte od. approx. Standardnormalverteilung)
normal, σ^2 bekannt	$V = \frac{\bar{X}-\mu_0}{\sigma}\sqrt{n}$ "Gausstest"
normal, σ^2 unbekannt	$V = \frac{\bar{X}-\mu_0}{S}\sqrt{n}$ "t-Test"
beliebig, $n \geq 30$	$V = \frac{\bar{X}-\mu_0}{S}\sqrt{n}$ "approximativer Gausstest"
$X_i = 1$ oder 0, $\mu = E(X_i) = p$, $n\bar{x} \geq 5$, $n(1-\bar{x}) \geq 5$, $n \geq 30$	$V = \frac{\bar{X}-p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$ "approximativer Gausstest"

Satz 14.1: Zusammenhang zwischen Nullhypothese und Ablehnungsbereich: Sei c_α das α- Fraktil der Standardnormalverteilung. Dann sind die Ablehnungsbereiche für verschiedene Nullhypothesen zum Niveau α gegeben wie folgt:

H_0	Ablehnungsbereich
$\mu = \mu_0$	$\left(-\infty, c_{\alpha/2})\right) \cup \left(c_{1-\alpha/2}, \infty\right)$
$\mu \geq \mu_0$	$(-\infty, c_\alpha)$
$\mu \leq \mu_0$	$(c_{1-\alpha}, \infty)$

Weitere Signifikanztests betreffen Hypothesen über:

- Varianzen (hier nicht relevant)
- Kovarianzen (hier nicht relevant)
- komplette Verteilungsfunktionen.

14.3 Der χ^2 - Unabhängigkeitstest

gegeben 2 diskrete ZVen X (mit l Ausprägungen) und Y (mit k Ausprägungen).
H_0: X,Y sind unabhängig.

Beispiel:
X = Geschlecht , Y = Kaufverhalten
$n = 1000$ Kunden, $\alpha = 10\%$, "Kreuztabelle":

	kaufen		nicht kaufen		Randhäufigkeiten	
Männer	180	h_{11}	170	h_{12}	350	$h_{1.}$
Frauen	240	h_{21}	410	h_{22}	650	$h_{2.}$
	420	$h_{.1}$	580	$h_{.2}$	1000	

Bei Unabhängigkeit: $h_{ij} = \dfrac{h_{i.} \cdot h_{.j}}{n} =: \tilde{h}_{ij}$

Prüfgröße: $V = \sum_{i=1}^{k} \sum_{j=1}^{l} \dfrac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \dfrac{\left(h_{ij} - \frac{h_{i.} - h_{j.}}{n}\right)^2}{\frac{h_{i.} - h_{j.}}{n}}$

$= \sum\sum \dfrac{\text{beobachtete Zellhäufigkeit - erwartete Zellhäufigkeit}}{\text{erwartete Zellhäufigkeit}}$

Lehne ab, falls V "zu groß". Was heißt "zu groß"?

V hat unter H_0 approximativ eine sogenannte χ^2- Verteilung mit $(l-1)(k-1)$ Freiheitsgraden (falls alle $\tilde{h}_{ij} \geq 5$).

Hier: $h_{11} = 180$, $h_{21} = 240$, $h_{12} = 170$, $h_{22} = 410$

$\tilde{h}_{11} = \frac{350 \cdot 420}{1000} = 147$, $\tilde{h}_{12} = \frac{350 \cdot 580}{1000} = 203$,

$\tilde{h}_{21} = \frac{650 \cdot 420}{1000} = 273$, $\tilde{h}_{22} = \frac{650 \cdot 580}{1000} = 377$

χ^2- Approximation gerechtfertigt, da alle $\tilde{h}_{ij} \geq 5$

Erwartete Kreuztabelle bei Unabhängigkeit:

	Kaufen	Nicht kaufen
Männer	147	203
Frauen	273	377

h_{ij} = tatsächlich beobachtete Häufigkeit in den Zellen

$\tilde{h}_{ij}$ = theoretische Häufigkeit in den Zellen

$\sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\left(h_{ij} - \tilde{h}_{ij}\right)^2}{\tilde{h}_{ij}} = \frac{(180-147)^2}{147} + \frac{(170-203)^2}{203} + \frac{(240-273)^2}{273} + \frac{(410-377)^2}{377} =$
$7,408 + 5,365 + 3,989 + 2,889 = 19,651$

Wenn $19,651$ zu groß ist $\rightarrow$ lehne H_0 ab.

Ablehnungsbereich $= \left(\chi^2_{(k-1)(l-1);1-\alpha}, \infty\right) = \left(\chi^2_{1,0.9}, \infty\right) = (2,706, \infty)$

$\Rightarrow V \in$ Ablehnungsbereich

$\Rightarrow H_0$ ablehnen

Die Hypothese, daß das Kaufverhalten nicht vom Geschlecht abhängt, wird zum Niveau $\alpha = 0,1$ verworfen.