ÜBUNGSSKRIPT STATISTIK

A. Ploner H. Strelec C. Yassouridis M. Melcher

Universität für Bodenkultur

Department für Raum, Landschaft und Infrastruktur
Institut für Statistik

Peter-Jordan-Strasse 82, 1190 Wien

http://statistik.boku.ac.at/

2. März 2020

Inhaltsverzeichnis

Kapitel 1

Beschreibende Statistik

Beispiele

Lage- und Streuungsmaße

Beispiel 1.1 62 Frauen mit Bluthochdruck werden untersucht, wobei u.a. das Gewicht gemessen wurde. Die folgende Tabelle zeigt jeweils das Verhältnis des beobachteten Körpergewichtes zum sogenannten Normgewicht (in %):

117	89	107	102	95	132	163	123
103	144	108	89	118	123	92	140
102	112	115	123	130	102	119	119
114	119	100	106	115	145	115	108
122	113	153	116	81	85	122	120
113	140	117	119	121	130	114	107
125	94	96	118	123	91	120	125
125	103	133	102	131	147		

a) Zeichnen Sie ein Histogramm und das dazugehörende Summenhäufigkeitspolygon.

b) Berechnen Sie den arithmetischen Mittelwert sowie den klassierten Mittelwert.

c) Berechnen Sie Median, Varianz, Standardabweichung und den Variationskoeffizienten.

d) Berechnen Sie das 0.05-, 0.10-, 0.90- und 0.95-Quantil, die Quartile und den Interquartilsabstand.

e) Zeichnen Sie einen Boxplot der Daten.

Beispiel 1.2 Am Brillenschötchen *(Biscutella laevigata)* werden botanisch-morphologische Untersuchungen durchgeführt. Dabei ist u.a. an 40 Exemplaren die Spaltöffnungslänge (in µm) gemessen worden:

27	25	23	27	23	25	28	30	28	32
25	29	28	33	32	28	25	22	25	23
26	28	30	32	31	31	34	29	36	33
30	29	27	27	29	26	23	24	26	27

a) Zeichnen Sie ein Histogramm und das dazugehörende Summenhäufigkeitspolygon.

b) Berechnen Sie den arithmetischen Mittelwert sowie den Median.

c) Berechnen Sie Varianz, Standardabweichung und den Variationskoeffizienten.

d) Berechnen Sie das 0.05-, 0.10-, 0.90- und 0.95-Quantil, die Quartile und den Interquartilsabstand.

e) Zeichnen Sie einen Boxplot der Daten.

Beispiel 1.3 Bei einem Sojabohnenversuch im Forschungsglashaus wurde u.a. der Ertrag verschiedener Sorten untersucht. Für die Sorte 'Labrador' konnten in 55 Töpfen folgende Ertragswerte (in g) festgestellt werden:

95	77	89	96	97	88	88	80	84	106	108
98	110	114	63	83	88	106	91	89	79	108
85	81	72	119	94	91	105	103	98	103	100
77	94	92	115	86	77	77	93	100	98	108
87	105	108	86	74	82	99	95	95	102	101

a) Zeichnen Sie ein Histogramm der Daten.

b) Berechnen Sie den arithmetischen Mittelwert sowie den Median.

c) Berechnen Sie Varianz und Standardabweichung der Stichprobe.

d) Berechnen Sie die Quartile des Datensatzes. Unter welchem Wert liegen die niedrigsten 30% der Erträge? Über welchem Wert liegen die größten 30% der Erträge?

e) Zeichnen Sie einen Boxplot der Daten.

Beispiel 1.4 Im Mai und Juni 1992 wurden Luftmessungen in Roosevelt Island durchgeführt. Unter anderem ergaben sich für die Temperaturen (in Grad Celsius) folgende Messprotokolle:

Mai:

19.5	22.1	23.4	16.7	13.4	18.7	18.5	15.0	16.2	20.4
23.5	20.5	18.8	20.0	14.1	17.9	18.6	13.7	20.2	16.9
15.1	22.5	16.3	16.1	13.9	14.2	13.7	19.2	27.1	26.1
24.5									

Juni:

25.7	23.1	19.5	28.6	29.2	26.4	27.8	30.5	32.4	30.5
33.7	33.4	27.8	26.5	26.1	25.2	22.1	18.5	22.9	24.7
25.0	24.5	24.4	24.5	23.3	25.3	22.8	26.6	25.3	28.2

a) Zeichnen Sie ein Histogramm für den Monat Mai.

b) Berechnen Sie den arithmetischen Mittelwert und den Median für beide Monate.

c) Berechnen Sie Varianz, Standardabweichung und den Variationskoeffizienten für beide Monate.

d) Berechnen Sie das 0.05-, 0.10-, 0.90- und 0.95-Quantil und die Quartile für beide Monate.

e) Zeichnen Sie einen Boxplot für beide Monate getrennt, aber mit einer gemeinsamen Skalierung.

f) Vergleichen Sie die Temperaturverteilung der beiden Monate verbal auf Grund der von Ihnen berechneten Kennzahlen und der angefertigten Grafiken.

Beispiel 1.5 Eine Saatbaugenossenschaft prüft gängige Kartoffelsorten auf ihre Ertragsstabilität. In der nachfolgenden Tabelle ist der Knollenertrag der Sorte Bintje in dt/ha in klassierter Weise dargestellt ($n = 80$).

Ertrag	Anzahl	Ertrag	Anzahl
]340, 350]	2	]390, 400]	15
]350, 360]	9	]400, 410]	3
]360, 370]	16	]410, 420]	2
]370, 380]	19	]420, 430]	1
]380, 390]	12	]430, 440]	1

Berechnen Sie Stichprobenmittel und -streuung.

Beispiel 1.6 Die Zuckerindustrie testet die Zuckerrübensorte Antonia auf 24 verschiedenen Standorten im gesamten Zuckerrübenanbaugebiet. Unter anderem wird der bereinigte Zuckerertrag in dt/ha festgestellt.

10.58	10.89	10.30	10.69	10.98	10.15	12.25	11.31	10.72
10.61	11.45	11.13	11.25	9.87	11.09	10.08	11.89	11.32
11.56	10.18	11.53	10.79	11.75	10.97			

Berechnen Sie Stichprobenmittel und -streuung.

Beispiel 1.7 Bei der Schlachtkörperbewertung von Rindvieh wird auch die Schlachtkörperlänge festgestellt. In der nachfolgenden Tabelle ist die klassierte Schlachtkörperlänge von 44 Bullen der Rasse Fleckvieh in cm dargestellt.

Länge	Anzahl	Länge	Anzahl
]134, 136]	2	]142, 144]	12
]136, 138]	3	]144, 146]	9
]138, 140]	7	]146, 148]	4
]140, 142]	6	]148, 150]	1

Berechnen Sie Stichprobenmittel und -streuung.

Beispiel 1.8 Auf einer Obstplantage wurden 80 Äpfel der Sorte 'Golddelicious' auf ihr Gewicht (in dag) hin untersucht. Dabei konnte man folgende Werte beobachten (sortiert):

13.2	13.4	13.4	13.4	13.5	13.7	13.8	13.8	13.8	13.9
14.2	14.2	14.3	14.4	14.4	14.5	14.6	14.6	14.7	14.7
14.7	14.7	14.8	14.8	14.8	14.9	14.9	15.0	15.0	15.2
15.3	15.3	15.3	15.3	15.3	15.4	15.4	15.4	15.5	15.5
15.6	15.6	15.6	15.6	15.7	15.7	15.7	15.8	15.8	15.8
15.9	16.0	16.0	16.0	16.0	16.0	16.1	16.2	16.2	16.2
16.3	16.3	16.4	16.4	16.5	16.7	17.0	17.3	17.4	17.5
17.7	17.7	17.7	18.1	18.5	18.5	18.7	19.0	19.0	19.4

a) Zeichnen Sie ein Histogramm und das dazugehörige Summenhäufigkeitspolygon.

b) Berechnen Sie den arithmetischen Mittelwert.

c) Berechnen Sie Median, Varianz, Standardabweichung und den Variationskoeffizienten.

d) Berechnen Sie das 0.05-, 0.10-, 0.90- und 0.95-Quantil, die Quartile und den Interquartilsabstand.

e) Zeichnen Sie einen Boxplot der Daten.

Lösungen

Lösung 1.1
Gewichtsverhältnis von Bluthochdruckpatientinnen

a) **Histogramm** und **Summenhäufigkeitspolygon**: Es bietet sich eine Einteilung der Daten in $k = 9$ Klassen an, wodurch sich *schöne* Klassengrenzen 80, 90, ..., 170 ergeben. Man beachte: jede Klasse enthält ihre rechte (obere) Klassengrenze (aber nicht ihre linke). Im Histogramm werden die absoluten Häufigkeiten in Balkenform aufgetragen, im Summenhäufigkeitspolygon die kumulierten relativen Häufigkeiten über der rechten Klassengrenze.

Klasse Nr.	**untere Grenze**	**obere Grenze**	**absolute Häufigkeit**	**kumulierte relative Häufigkeit**
1	80	90	4	$4/62 = 0.06$
2	90	100	6	$10/62 = 0.16$
3	100	110	11	$21/62 = 0.34$
4	110	120	19	$40/62 = 0.65$
5	120	130	12	$52/62 = 0.84$
6	130	140	5	$57/62 = 0.92$
7	140	150	3	$60/62 = 0.97$
8	150	160	1	$61/62 = 0.98$
9	160	170	1	$62/62 = 1.00$

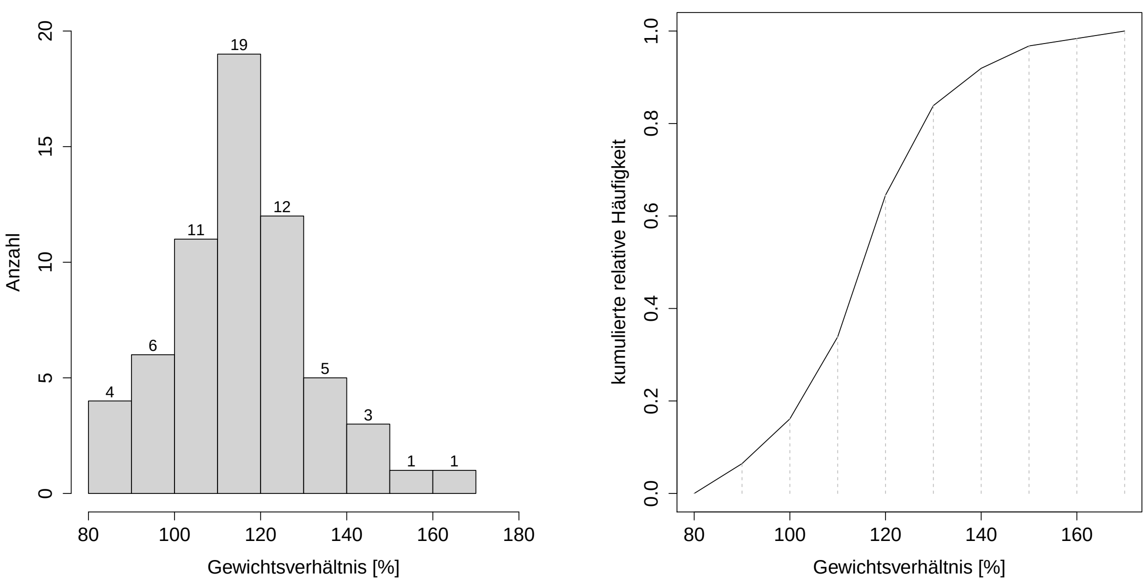


Abbildung 1.1: Histogramm (links) und Summenhäufigkeitspolygon (rechts) für Beispiel 1.1

b, c) **Lage-** und **Streuungsmaße**: Für die händische Berechnung von Mittelwerten etc. ist es günstig, zunächst die Summe der Werte und der quadrierten Werte zu berechnen.

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n = 7195 \qquad \sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \ldots + x_n^2 = 852043$$

Mittelwert:

$$\bar{x} \;=\; \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i = \tfrac{1}{62}(117 + 123 + \ldots + 122 + 118) = \tfrac{7195}{62} = 116.05$$

Mittelwert – klassiert:

$$\bar{x}^K \;=\; \tfrac{1}{n}\textstyle\sum_{j=1}^{k} c_j h_j \quad \text{wobei} \quad \begin{array}{ll} c_j & \text{Klassenmitte} \\ k & \text{Klassenanzahl} \\ h_j & \text{Klassenhäufigkeit} \end{array}$$

Für die vorliegenden Daten ergibt sich unter Verwendung der Klassierung des Histogramms:

$$\bar{x}^K \;=\; \tfrac{1}{62}(85 \cdot 4 + 95 \cdot 6 + \ldots + 155 \cdot 1 + 165 \cdot 1) = \tfrac{7180}{62} = 115.8$$

Median:

$$\tilde{x} \;=\; \tfrac{1}{2}(x_{(31)} + x_{(32)}) = \tfrac{1}{2}(117 + 117) = 117$$

Varianz:

$$s^2 \;=\; \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = \tfrac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = 279.92$$

Standardabweichung:

$$s \;=\; 16.73$$

Variationskoeffizient (sinnvoll wegen $\bar{x} \gg 0$):

$$v \;=\; s/\bar{x} = \tfrac{16.73}{116.05} = 0.144$$

d) **Quantile** und **Quartile**: Um etwa das 5%-Quantil zu bestimmen, wird zunächst die Anzahl n an Beobachtungen mit dem entsprechenden $\alpha = 0.05$ multipliziert. Ist der Wert eine ganze Zahl, so liegt der Quantilswert zwischen dem $(n \cdot \alpha)$-ten und dem $(n \cdot \alpha + 1)$-ten Wert der sortierten Stichprobe. Ist der Wert keine ganze Zahl, so ist der Quantilswert der $(\lfloor n \cdot \alpha \rfloor + 1)$-te Wert der sortierten Stichprobe. Hier ergibt sich $n \cdot \alpha = 3.1$, das Quantil ist also der 4. Wert.

α	$n \cdot \alpha$	Rang	q_α
0.05	3.1	4	89
0.10	6.2	7	94
0.25	15.5	16	103
0.75	46.5	47	123
0.90	55.8	56	140
0.95	58.9	59	145

Interquartilsabstand: $IQR = Q_3 - Q_1 = 123 - 103 = 20$

e) **Boxplot**: Die benötigten Kenngrößen wurden bereits berechnet:

- Median: $\tilde{x} = 117$
- Quartile: $Q_1 = 103$, $Q_3 = 123$
- Interquartilsabstand: $IQR = 20$

Damit kann der Boxplot konstruiert werden:

- Maximale Länge der Whisker: $1.5 \cdot IQR = 30$
- Grenzen: $Q_1 - 1.5 \cdot IQR = 73$, $Q_3 + 1.5 \cdot IQR = 153$
- Kleinster/größter Wert innerhalb der Grenzen: 81, 153
- Ausreißer (Werte außerhalb der Grenzen): 163

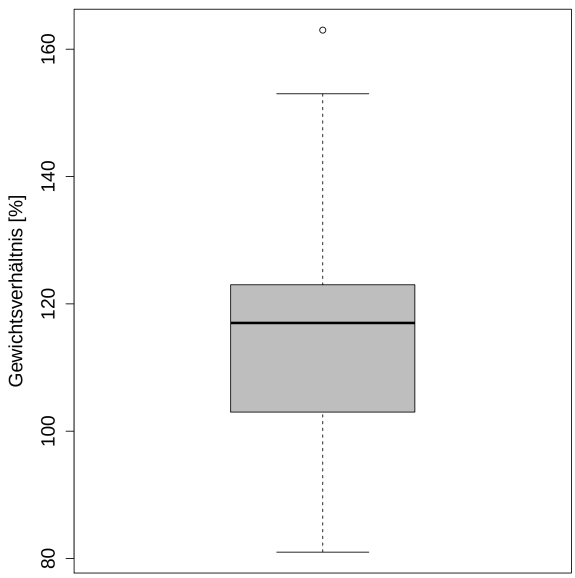


Abbildung 1.2: Boxplot für Beispiel 1.1

Lösung 1.2
Spaltöffnungslänge des Brillenschötchens

a) **Histogramm** und **Summenhäufigkeitspolygon**: siehe Abbildung. Für die Erstellung des Histogramms wurde eine Klassenbreite von $d = 2$ µm auf dem Intervall $[20; 36]$ gewählt (was somit $k = 8$ Klassen ergibt).

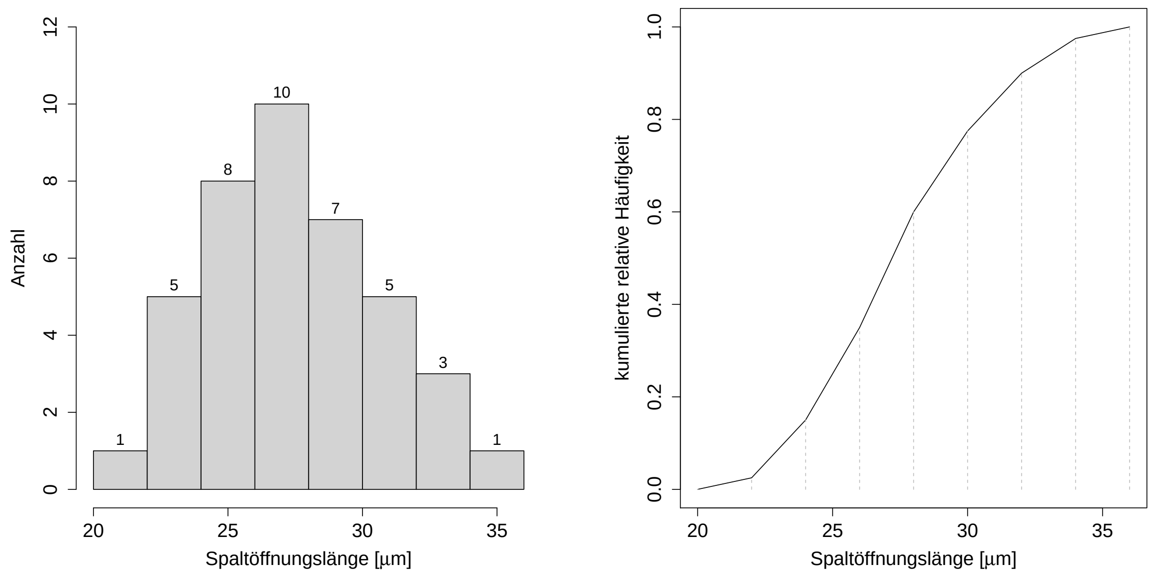


Abbildung 1.3: Histogramm (links) und Summenhäufigkeitspolygon (rechts) für Beispiel 1.2

b, c) **Lage-** und **Streuungsmaße**:

$$n = 40 \qquad \sum_{i=1}^{n} x_i = 1116 \qquad \sum_{i=1}^{n} x_i^2 = 31582$$

Mittelwert: $\bar{x} = 27.90$

Median: $\tilde{x} = 28$

Varianz: $s^2 = 11.43$

Standardabweichung: $s = 3.38$

Variationskoeffizient: $v = 0.12$

d) **Quantile** und **Quartile**:

α	$n \cdot \alpha$	Ränge		$q_\alpha \in$		q_α (Mitte)
0.05	2	2	3	23,	23	23.0
0.10	4	4	5	23,	23	23.0
0.25	10	10	11	25,	25	25.0
0.75	30	31	31	30,	30	30.0
0.90	36	36	37	32,	33	32.5
0.95	38	38	39	33,	34	33.5

Interquartilsabstand: $IQR = Q_3 - Q_1 = 5$

e) **Boxplot**: siehe Abbildung - keine Ausreißer.

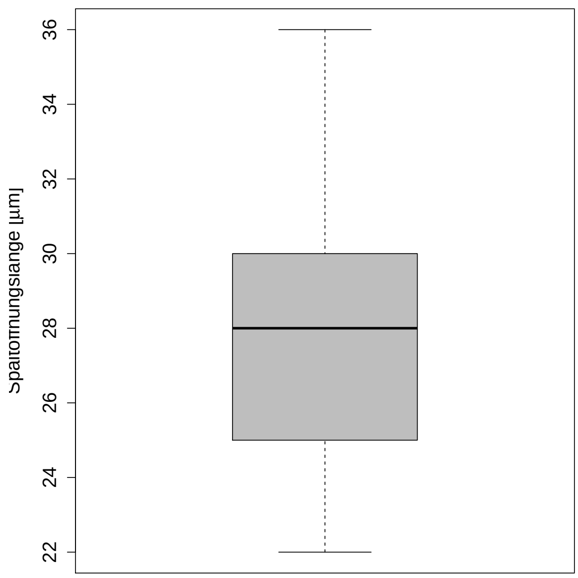


Abbildung 1.4: Boxplot für Beispiel 1.2

Lösung 1.3
Ertrag im Sojabohnenversuch

a) **Histogramm**:

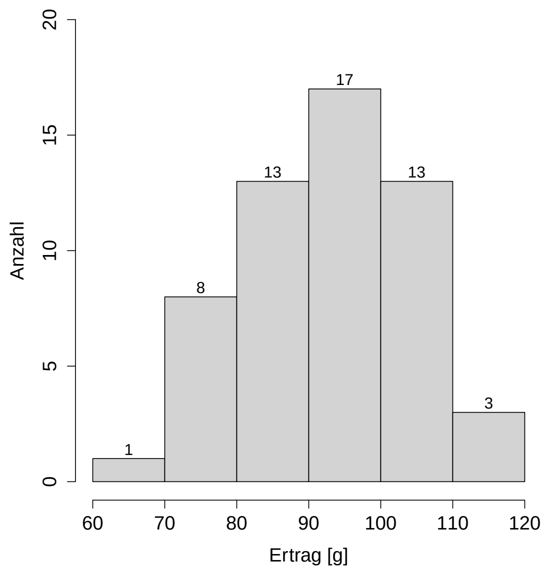


Abbildung 1.5: Histogramm für Beispiel 1.3

b, c) **Lage-** und **Streuungsmaße**:

$$\sum_{i=1}^{n} x_i = 5139 \qquad \sum_{i=1}^{n} x_i^2 = 488019$$

Mittelwert: $\bar{x} = 93.44$

Median: $\tilde{x} = x_{(28)} = 94$

Varianz: $s^2 = (488019 - 5139^2/55)/54 = 145.36$

Standardabweichung: $s = 12.06$

d) **Quantile** und **Quartile**:

α	$n \cdot \alpha$	Rang	q_α
0.25	13.75	14	85
0.30	16.50	17	87
0.70	38.50	39	100
0.75	41.25	42	103

30% der Erträge liegen unter 87 g. 30% der Erträge liegen über 100 g.

e) **Boxplot**: siehe Abbildung 1.6. Berechnet wurden bereits

Median: $\tilde{x} = 94$ erstes Quartil: $Q_1 = 85$ drittes Quartil: $Q_3 = 103$

womit folgt

$IQR = 18$ untere Grenze: 58 obere Grenze: 130

Da kleinster und größter Ertrag (63/119) innerhalb der Grenzen liegen, werden keine Ausreißer eingetragen.

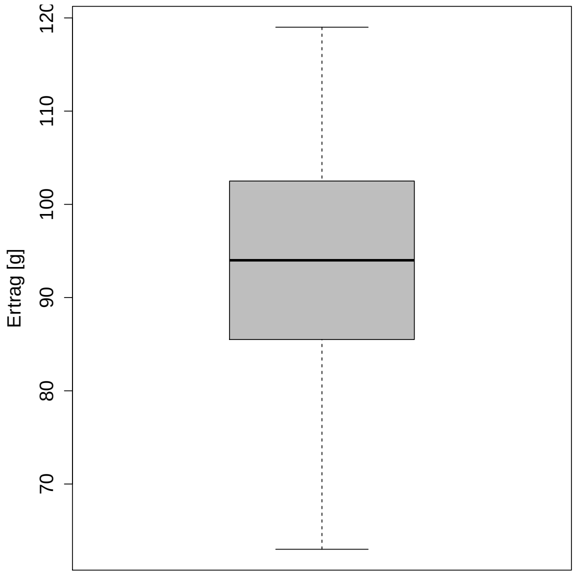


Abbildung 1.6: Boxplot für Beispiel 1.3

Lösung 1.4
Temperaturmessungen im Mai/Juni

a) **Histogramm Mai**: siehe Abbildung 1.7. Beachten Sie, dass das Histogramm auf Basis der ursprünglichen Werte (inklusive Zehntelstelle) gebildet wird!

b, c) **Lage-** und **Streuungsmaße**:

	$\sum x_i$	$\sum x_i^2$	$\bar{x}$	$\tilde{x}$	s^2	s	v
Mai	576.8	11168.88	18.61	18.60	14.56	3.82	0.21
Juni	784.5	20899.33	26.15	25.50	13.26	3.64	0.14

d) **Quantile** (links … Mai, rechts … Juni):

α	$n \cdot \alpha$	Rang	q_α
0.05	1.55	2	13.7
0.10	3.10	4	13.9
0.25	7.75	8	15.1
0.75	23.25	24	20.5
0.90	27.90	28	23.5
0.95	29.45	30	26.1

α	$n \cdot \alpha$	Ränge		$q_\alpha \in$		q_α (Mitte)
0.05	1.5	2				19.50
0.10	3.0	3	4	22.1	22.8	22.45
0.25	7.5	8				24.40
0.75	22.5	23				28.20
0.90	27.0	27	28	30.5	32.4	31.45
0.95	28.5	29				33.40

e) **Boxplots**: siehe Abbildung 1.7

Mai: untere Grenze = 7 obere Grenze = 28.6
Juni: untere Grenze = 18.7 obere Grenze = 33.9

f) **Interpretation**: Wie aus dem Boxplot ersichtlich, unterscheiden sich die Temperaturverteilungen der beiden Monate durch ihre Lage: die mittlere Temperatur im Juni ist um mehr als 7 Grad höher als im Mai. Die Streuung – gemessen als Standardabweichung – ist für beide ähnlich. Beide Verteilungen erscheinen annähernd symmetrisch um ihre Mitte, im Boxplot erscheint der Juni symmetrischer (Median in der Mitte der Box, Whisker annähernd gleich lang), während für den Mai Mittel und Median besser übereinstimmen. Insgesamt erscheinen die Temperaturen im Mai einer flacheren Verteilung zu folgen als im Juni.

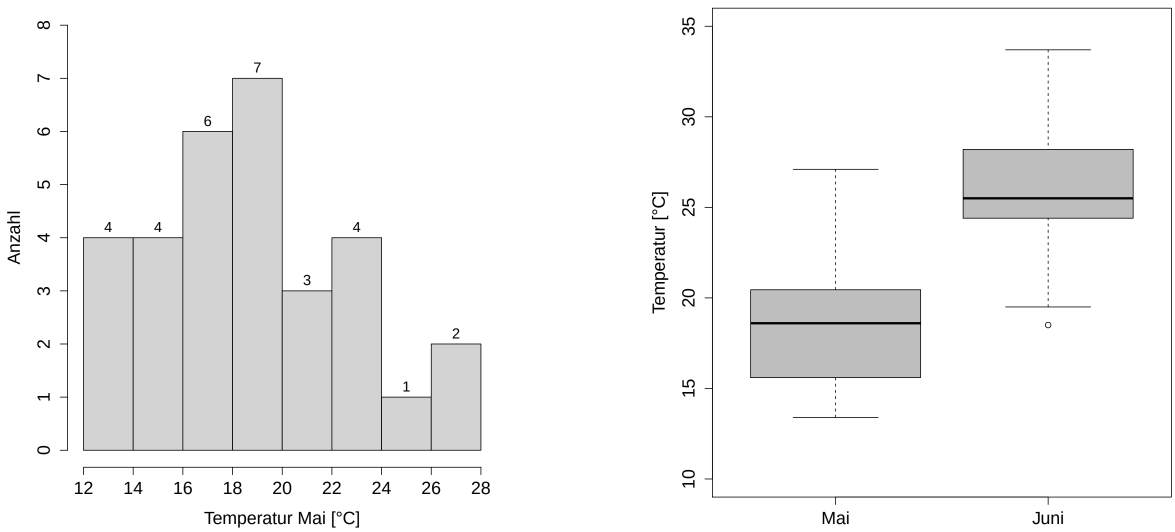


Abbildung 1.7: Histogramm (Mai) und Boxplots für Beispiel 1.4

Lösung 1.5
Knollenertrag von Kartoffeln

Mittels des Stichprobenumfangs n, der Klassenanzahl k, der Anzahl h_j an Beobachtungen in Klasse j und den Klassenmitten c_j ergibt sich für den klassierten Mittelwert $\bar{x}^K$ und die klassierte Varianz $s^{2,K}$ bzw. die klassierte Standardabweichung s^K:

Klassierter Mittelwert:

$$\bar{x}^K = \frac{1}{n}\sum_{j=1}^{k} c_j h_j = \frac{1}{80}(2 \cdot 345 + 9 \cdot 355 + \ldots + 1 \cdot 425 + 1 \cdot 435) = \frac{30300}{80} = 378.75$$

Klassierte Varianz:

$$s^{2,K} = \frac{1}{n-1}\sum_{j=1}^{k} h_j (c_j - \bar{x}^K)^2 = \frac{1}{n-1}\left(\sum_{j=1}^{k} h_j c_j^2 - \frac{1}{n}\left(\sum_{j=1}^{k} h_j c_j\right)^2\right)$$

$$= \frac{1}{79}\left(11501200 - \frac{30300^2}{80}\right) = 317.41$$

Klassierte Standardabweichung: $s^K = 17.82$

Lösung 1.6
Zuckerertrag der Sorte *Antonia*

Mittels der Hilfsgrößen

$$\sum_{i=1}^{n} x_i = 263.34 \qquad \text{und} \qquad \sum_{i=1}^{n} x_i^2 = 2898.049$$

ergibt sich für das Stichprobenmittel

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{24} \cdot 263.34 = 10.94$$

und für die empirische Varianz

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \\ &= \frac{1}{23}\left(2898.049 - 24 \cdot 263.34^2\right) = 0.372 \end{aligned}$$

Lösung 1.7
Schlachtkörperlänge von Bullen

Analog zu Beispiel 1.5 berechnet man

Klassierter Mittelwert:

$$\bar{x}^K = \frac{1}{n}\sum_{j=1}^{k} c_j h_j = \frac{1}{44}(2 \cdot 135 + 3 \cdot 137 + \ldots + 4 \cdot 147 + 1 \cdot 149) = 142.23$$

Klassierte Varianz:

$$\begin{aligned} s^{2,K} &= \frac{1}{n-1}\sum_{j=1}^{k} h_j (c_j - \bar{x}^K)^2 = \\ &= \frac{1}{43}\left[2 \cdot (135 - 142.23)^2 + 3 \cdot (137 - 142.23)^2 + \ldots + 1 \cdot (149 - 142.23)^2\right] = 11.20 \end{aligned}$$

Klassierte Standardabweichung: $s^K = 3.35$

Lösung 1.8
Gewicht von Äpfeln der Sorte 'Golddelicious'

a) **Histogramm** und **Summenhäufigkeitspolygon**: siehe Abbildung. Es wurden ganzzahlige Klassengrenzen von 13 bis 20 dag gewählt, d.h. $k = 7$ Klassen.

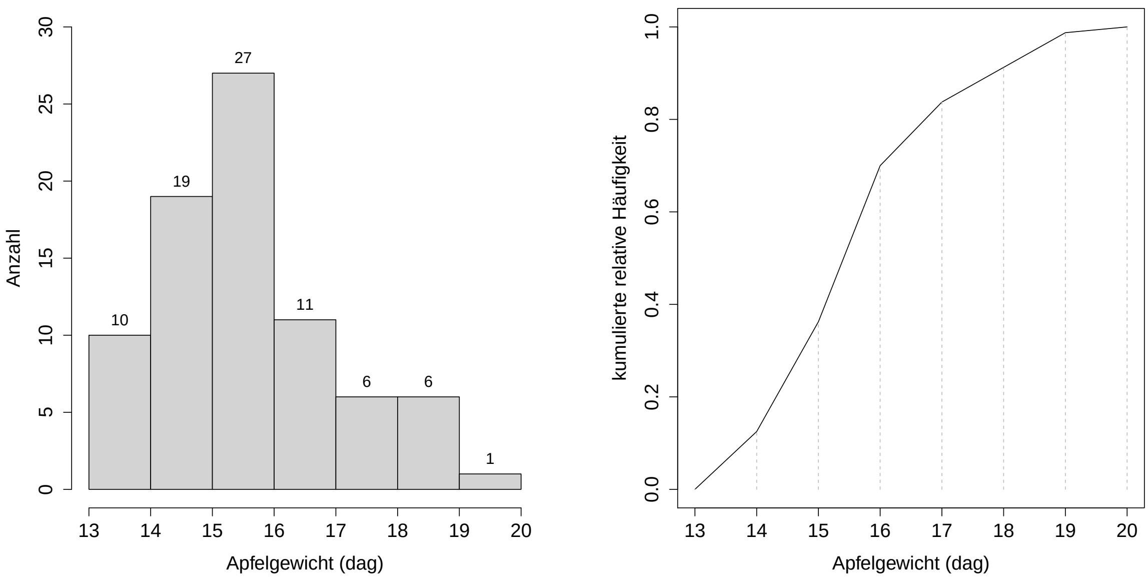


Abbildung 1.8: Histogramm (links) und Summenhäufigkeitspolygon (rechts) für Beispiel 1.8

b, c) **Lage-** und **Streuungsmaße**:

$$n = 80 \qquad \sum_{i=1}^{n} x_i = 1252.6 \qquad \sum_{i=1}^{n} x_i^2 = 19774.36$$

Mittelwert: $\bar{x} = 15.658$

Median: $\tilde{x} = 15.55$

Varianz: $s^2 = 2.048$

Standardabweichung: $s = 1.431$

Variationskoeffizient: $v = 0.0914$

d) **Quantile** und **Quartile**:

α	$n \cdot \alpha$	Ränge		$q_\alpha \in$		q_α (Mitte)
0.05	4	4	5	13.4,	13.5	13.45
0.10	8	8	9	13.8,	13.8	13.80
0.25	20	20	21	14.7,	14.7	14.70
0.75	60	60	61	16.2,	16.3	16.25
0.90	72	72	73	17.7,	17.7	17.70
0.95	76	76	77	18.5,	18.7	18.60

Interquartilsabstand: $IQR = Q_3 - Q_1 = 16.25 - 14.7 = 1.55$

e) **Boxplot**: Die benötigten Kenngrößen wurden bereits berechnet:

- Median: $\tilde{x} = 15.55$
- Quartile: $Q_1 = 14.7$, $Q_3 = 16.25$
- Interquartilsabstand: $IQR = 1.55$

Damit kann der Boxplot konstruiert werden:

- Maximale Länge der Whisker: $1.5 \cdot IQR = 1.5 \cdot 1.55 = 2.325$
- Grenzen: $Q_1 - 1.5 \cdot IQR = 12.375$, $Q_3 + 1.5 \cdot IQR = 18.525$
- Kleinster/größter Wert innerhalb der Grenzen: 13.2, 18.5
- Ausreißer (Werte außerhalb der Grenzen): 18.7, 19.0 (2×), 19.4

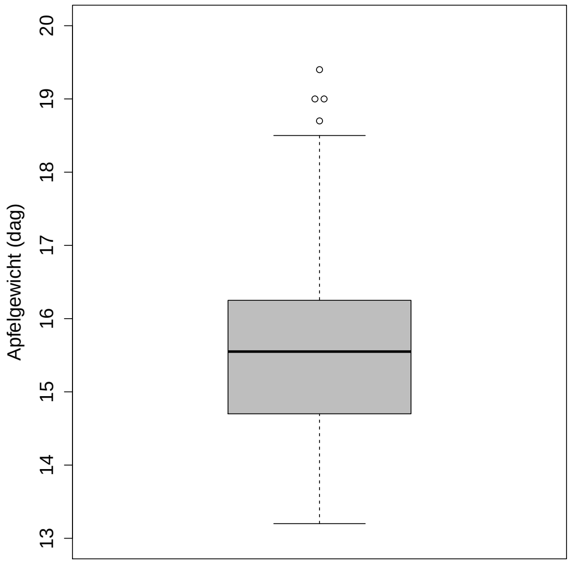


Abbildung 1.9: Boxplot für Beispiel 1.8

Kapitel 2

Wahrscheinlichkeitsrechnung

Beispiele

Beispiel 2.1 Im Lotto 6 aus 45 besteht ein Tipp aus dem Setzen von 6 Zahlen zwischen 1 und 45. Wie groß ist die Wahrscheinlichkeit für den Gewinn in den folgenden Kombinationen:

a) *Sechser*: alle 6 Tipps sind richtig;

b) *Fünfer mit Zusatzzahl*: 5 der 6 Tipps sind richtig und der falsche Tipp ist gerade die bei der Ziehung als siebente Zahl gezogene Zusatzzahl;

c) *Fünfer*: 5 von 6 Tipps sind richtig, die sechste ist *nicht* die Zusatzzahl;

d) *Vierer*: 4 von 6 Tipps sind richtig;

e) *Dreier*: 3 von 6 Tipps sind richtig.

Beispiel 2.2 Beim Kartenpoker mit 52 Karten werden an die Spieler jeweils fünf Karten verteilt. Es gibt dabei folgende Gewinnmöglichkeiten:

One Pair: Unter den 5 Spielkarten befinden sich genau 2 gleiche Kartenwerte, z.B. Herz–Dame und Treff–Dame.

Two Pairs: 2×2 gleiche Kartenwerte, z.B. Herz–Bube und Treff–Bube sowie Herz–Dame und Karo–Dame.

Three of a Kind: Genau 3 gleiche Kartenwerte (z.B. Herz–As, Karo–As und Pik–As).

Straight: 5 Karten in aufsteigender Reihenfolge (z.B. Herz–Acht, Pik–Neun, Pik–Zehn, Karo–Bube und Pik–Dame).

Flush: Alle 5 Karten sind von einer Spielfarbe.

Full House: 2 gleiche und 3 gleiche Kartenwerte (z.B. Karo–As und Herz–As sowie Herz–Neun, Treff–Neun und Pik–Neun).

Four of a Kind: ("Poker") genau 4 gleiche Kartenwerte (z.B. Herz–Acht, Karo–Acht, Treff–Acht und Pik–Acht sowie Karo–Dame).

Straight Flush: ein Straight, bei dem alle Karten von *einer* Farbe sind.

Royal Flush: Straight Flush mit den höchsten Kartenwerten, also z.B. Herz–As, Herz–König, Herz–Dame, Herz–Bube und Herz–Zehn.

Berechnen Sie die einzelnen Gewinnwahrscheinlichkeiten.

Beispiel 2.3 Von den Studenten der Boku fahren 40% der Frauen und 50% der Männer mit dem Fahrrad zur Uni. Die Anzahl der weiblichen und männlichen Studenten stehen an der Boku in einem Verhältnis von 35 zu 65.

a) Wie groß ist die Wahrscheinlichkeit, dass eine Studentin (Student) mit dem Fahrrad zur Uni kommt?

b) Wie groß ist die Wahrscheinlichkeit, dass eine studierende Person, die mit dem Fahrrad zur Uni kommt, weiblich (männlich) ist?

c) Wie groß ist die Wahrscheinlichkeit, ein vierstelliges Fahrradzahlenschloss innerhalb einer halben Stunde aufzubekommen, wenn alle zwei Sekunden eine Kombination eingestellt werden kann?

Beispiel 2.4 Angenommen in einer bestimmten Region kommt es innerhalb eines Jahres mit einer Wahrscheinlichkeit von 10 % zu einem Erdbeben. Wir betrachten das Auftreten innerhalb eines Jahres als ein binäres Zufallsereignis mit den beiden Optionen:

- „Erdbeben tritt in dem Jahr ein“ und
- „Erdbeben tritt in dem Jahr nicht ein“.

Nun betrachten wir einen Zeitraum von zehn Jahren und verzeichnen für jedes Jahr, ob ein Erdbeben eingetreten ist oder nicht. Wir nehmen dabei an, dass das Auftreten in den einzelnen Jahren unabhängig voneinander ist (d.h. das Auftreten/Ausbleiben in einem Jahr beeinflusst nicht die Wahrscheinlichkeit des Auftretens in den folgenden Jahren). Die Zufallsvariable X bezeichnet nun die Gesamtanzahl der Jahre im Beobachtungszeitraum, in denen ein Erdbeben eintritt.

a) Welche Verteilung hat die Zufallsvariable X?

b) Wie groß ist die Wahrscheinlichkeit, dass in genau drei (von zehn) Jahren ein Erdbeben eintritt?

c) Wie groß ist die Wahrscheinlichkeit, dass in höchstens drei (von zehn) Jahren ein Erdbeben eintritt?

d) Wie groß ist die Wahrscheinlichkeit, dass in mindestens einem und höchstens drei (von zehn) Jahren ein Erdbeben eintritt?

Normalverteilung

Beispiel 2.5 Sei X eine normalverteilte Zufallsgröße mit Erwartungswertwert μ und Varianz σ^2. Im folgenden soll eine Reihe von Toleranzintervallen zu vorgegebener Länge bzw. zu einer vorgegebenen Wahrscheinlichkeit angegeben werden.

a) $P(|X-\mu| \leq k\sigma) = ?$ für $k = 1, 1.5, 2, 2.5, 3, 3.5$

b) $P(|X-\mu| \leq k\sigma) = p$ für $p = 0.90, 0.95, 0.99, 0.995, 0.999$ $k = ?$

c) $P(X-\mu \leq t\sigma) = p$ für $p = 0.90, 0.95, 0.99, 0.995, 0.999$ $t = ?$

Beispiel 2.6 In der Firma eines Spaghettifabrikanten wird Qualitätskontrolle durchgeführt. Um die Verpackung bzw. die Gestaltung der Verpackung gut planen zu können stellen sich einige Fragen. Man nimmt an, dass die Spaghettilänge normalverteilt ist und im Mittel 30 cm beträgt, bei einer mittleren Abweichung von 0.6 cm.

a) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Spaghetti kürzer als 28.5 cm ist?

b) Wie groß ist die Wahrscheinlichkeit, dass ein Spaghetti länger als 31.1 cm ist?

c) Wie groß ist die Wahrscheinlichkeit, dass ein Spaghetti zwischen 29.1 und 30.8 cm lang ist?

d) Um nicht allzu viel Ausschuss zu erzeugen muss man sicherstellen dass möglichst viele Spaghetti der Länge nach gut in die Schachtel passen. In welcher Länge muss die Schachtel produziert werden, damit 95% Wahrscheinlichkeit der Produktion hineinpassen?

Beispiel 2.7 Auf einer Apfelplantage werden Äpfel der Sorte 'Golddelicios' angebaut. Wir betrachten das Gewicht (in dag) eines einzelnen Apfels und modellieren dieses als Zufallsvariable X. Aus Erfahrung weiß man, dass das Gewicht X annähernd einer Normalverteilung folgt. Außerdem wiegt ein Apfel im Mittel 15 dag mit einer Standardabweichung von 1.5 dag. Es gilt also, dass $X \sim N(\mu = 15, \sigma^2 = 1.5^2)$. Man will nun folgende Fragen zum Apfelgewicht beantworten:

a) Wie groß ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Apfel weniger als 17 dag wiegt?

b) Wie groß ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Apfel mehr als 16 dag wiegt?

c) Wie groß ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Apfel zwischen 14 und 17 dag wiegt?

Lösungen

Lösung 2.1
Gewinnwahrscheinlichkeiten im Lotto 6 aus 45

a) **Sechser**: Da es auf die Reihenfolge der Ziehung nicht ankommt, gibt es $\binom{45}{6}$ mögliche Tipps (= mögliche Fälle), von denen genau einer richtig ist (= günstige Fälle). Die Wahrscheinlichkeit für einen Sechser ergibt sich daher als

$$\frac{1}{\binom{45}{6}} = 1.23 \cdot 10^{-7}$$

Man kann dieses Ergebnis auch aus der hypergeometrischen Verteilung (Abschnitt 3.5.4 des Vorlesungsskriptums) herleiten: Es gibt $N = 45$ Elemente, von denen $A = 6$ "Richtige" sind. Aus dieser Grundgesamtheit werden $n = 6$ Elemente gezogen, die Wahrscheinlichkeit, dass davon $i = 6$ richtig sind, ergibt sich als

$$h_{45,6,6}(6) = \frac{\binom{6}{6}\binom{39}{0}}{\binom{45}{6}} = 1.23 \cdot 10^{-7}$$

b, c) **Beliebiger Fünfer**: Falls wir die Zusatzzahl (ZZ) zunächst ignorieren, so gibt es immer noch $\binom{45}{6}$ mögliche Tipps. Um einen Fünfer zu erhalten, müssen wir zunächst 5 von 6 Richtigen auswählen, wofür es $\binom{6}{5}$ Möglichkeiten gibt. Diese können wir dann mit irgendeiner der $45 - 6 = 39$ verbliebenen "falschen" Zahlen kombinieren. Wir erhalten damit als Wahrscheinlichkeit

$$\frac{\binom{6}{5} \cdot 39}{\binom{45}{6}} = 2.87 \cdot 10^{-5}$$

Derselbe Wert ergibt sich aus der hypergeometrischen Verteilung:

$$h_{45,6,6}(5) = \frac{\binom{6}{5}\binom{39}{1}}{\binom{45}{6}} = 2.87 \cdot 10^{-5}$$

Fünfer mit ZZ: Hier wählt man wieder 5 von 6 Richtigen ($\binom{6}{5}$ Möglichkeiten), aber die sechste Zahl muss die ZZ sein, die Wahrscheinlichkeit ergibt sich daher zu

$$\frac{\binom{6}{5}}{\binom{45}{6}} = 7.38 \cdot 10^{-7}$$

Fünfer ohne ZZ: Hier hat man für die sechste Zahl 38 Möglichkeiten (39 falsche Tipps, minus die ZZ), die Wahrscheinlichkeit ist daher

$$\frac{\binom{6}{5} \cdot 38}{\binom{45}{6}} = 2.80 \cdot 10^{-5}$$

Man beachte, dass die Wahrscheinlichkeit für einen beliebigen Fünfer gerade die Summe der Teilwahrscheinlichkeiten ist.

d) **Vierer**:

$$h_{45,6,6}(4) = \frac{\binom{6}{4}\binom{39}{2}}{\binom{45}{6}} = 1.37 \cdot 10^{-3}$$

e) **Dreier**:

$$h_{45,6,6}(3) = \frac{\binom{6}{3}\binom{39}{3}}{\binom{45}{6}} = 2.24 \cdot 10^{-2}$$

Lösung 2.2
Gewinnwahrscheinlichkeiten beim Kartenpokern.

Insgesamt vorhandene Zuteilungsmöglichkeiten eines 5-Karten-Blattes:

$$\binom{52}{5} = \frac{52!}{47! \cdot 5!} = 2598960$$

One Pair: Für einen fixen Kartenwert aus $2, 3, \ldots,$ As gibt es zunächst $\binom{4}{2}$ Möglichkeiten für ein Paar, bei 13 verschiedenen Kartenwerten somit $13 \cdot \binom{4}{2}$. Die restlichen 3 Karten werden aus den verbleibenden 48 gezogen – der im Paar vorkommende Kartenwert ist nicht *günstig*, zumal man sonst eine höherwertige Kombination (*Three of a kind*, *Poker*, ...) erhalten würde. Von diesen $\binom{48}{3}$ Möglichkeiten müssen nun die Fälle abgezogen werden, wo 2 der 3 Karten gleich sind (insgesamt hätte man dann zwei Paare) bzw. wo alle 3 Karten gleich sind (dann hätte man ein *Full House*):

$$\mathrm{P(One\ Pair)} = \frac{13 \cdot \binom{4}{2} \cdot \left(\binom{48}{3} - 12 \cdot \binom{4}{2} \cdot 44 - 12 \cdot \binom{4}{3}\right)}{2598960} = \frac{1098240}{2598960} = 0.4226$$

Two Pairs: Zunächst gibt es $\binom{13}{2}$ Möglichkeiten, 2 der 13 Kartenwerte auszusuchen, für die Paare vorliegen. Für jeden Kartenwert existieren $\binom{4}{2}$ Möglichkeiten für ein Paar. Die letzte der 5 Karten kann eine beliebige der $52 - 8 = 44$ restlichen Karten sein:

$$\mathrm{P(Two\ Pairs)} = \frac{\binom{13}{2} \cdot \binom{4}{2} \cdot \binom{4}{2} \cdot 44}{2598960} = \frac{123552}{2598960} = 0.0475$$

Three of a Kind: Es gibt 13 Möglichkeiten, einen der 13 Kartenwerte auszusuchen, für den ein Drilling vorliegen soll. Für einen fixen Kartenwert gibt es dann $\binom{4}{3}$ Möglichkeiten, einen Drilling zusammenzustellen. Die restlichen beiden Karten dürfen weder untereinander den gleichen Kartenwert besitzen (dies wäre ein *Full House*), noch darf eine der beiden den gleichen Kartenwert wie der Drilling besitzen (dies wäre ein Poker). Es verbleiben somit 12 Kartenwerte, aus denen 2 gezogen werden und pro Kartenwert jeweils 4 Möglichkeiten (da 4 Farben):

$$\mathrm{P(Three\ of\ a\ Kind)} = \frac{13 \cdot \binom{4}{3} \cdot \binom{12}{2} \cdot 4 \cdot 4}{2598960} = \frac{54912}{2598960} = 0.0211$$

Straight: Lässt man zunächst die Farbe außer acht, so gibt es 10 mögliche *Straights*: beginnend mit *As*, endend mit 5; beginnend mit 2, endend mit 6, ..., beginnend mit 10, endend mit *As*. Da jede dieser 5 Karten eine beliebige Farbe haben kann, folgt

$$\mathrm{P(Straight)} = \frac{10 \cdot 4^5 - 40}{2598960} = \frac{10200}{2598960} = 3.92 \cdot 10^{-3}$$

40 Möglichkeiten werden abgezogen, da für jede der 4 Farben 9 Möglichkeiten eines *Straight Flush* und eine Möglichkeit eines *Royal Flush* existieren, welche jeweils eigene Kategorien darstellen (siehe unten).

Flush: Es gibt 4 Möglichkeiten für die Auswahl einer Farbe sowie $\binom{13}{5}$ Möglichkeiten, 5 der 13 Werte dieser Farbe zu ziehen. Davon werden die insgesamt $4 \cdot 10 = 40$ Möglichkeiten abgezogen, die einen *Straight Flush* oder *Royal Flush* darstellen.

$$\mathrm{P(Flush)} = \frac{4 \cdot \binom{13}{5} - 4 \cdot 10}{2598960} = \frac{5108}{2598960} = 1.97 \cdot 10^{-3}$$

Full House: Zunächst gibt es 13 Möglichkeiten, einen Kartenwert für den Drilling zu wählen, für den es dann wiederum $\binom{4}{3}$ Möglichkeiten gibt. Von den verbleibenden 12 Kartenwerten wird dann eine für das Paar bestimmt, für das wiederum $\binom{4}{3}$ Möglichkeiten existieren:

$$\mathrm{P(Full\ House)} = \frac{13 \cdot \binom{4}{2} \cdot 12 \cdot \binom{4}{2}}{2598960} = \frac{3744}{2598960} = 1.44 \cdot 10^{-3}$$

Four of a Kind: Für jede der 13 Kartenwerte existiert $\binom{4}{4} = 1$ Möglichkeit, den Poker zusammenzustellen. Diese 4 Karten können mit jeder der verbleibenden $52 - 4 = 48$ Karten kombiniert werden, wodurch sich ergibt

$$\mathrm{P(Four\ of\ a\ Kind)} = \frac{13 \cdot \binom{4}{4} \cdot 48}{2598960} = \frac{624}{2598960} = 2.40 \cdot 10^{-4}$$

Straight Flush: Für jede der 4 Farben existieren 9 Möglichkeiten, *Straights* bestehend lediglich aus dieser Farbe zu erhalten:

$$\mathrm{P(Straight\ Flush)} = \frac{36}{2598960} = 1.39 \cdot 10^{-5}$$

Royal Straight: Es gibt lediglich eine günstige Möglichkeit für jede Farbe, einen *Roal Flush* zu erhalten.

$$\mathrm{P(Royal\ Flush)} = \frac{4}{2598960} = 1.54 \cdot 10^{-6}$$

Lösung 2.3
Radfahrer unter den Boku-Studenten/-innen

a) Wie groß ist die Wahrscheinlichkeit, dass eine Studentin mit dem Fahrrad fährt? Die Grundgesamtheit ist also die Menge der Boku-Studentinnen – zum gegebenen Zeitpunkt gerade m Personen. Von diesen wählen wir zufällig eine aus und befragen sie nach ihrem Radfahrstatus. Als "günstig" in unserem Sinne betrachten wir es, wenn wir eine Radfahrerin erwischt haben; die Anzahl der günstigen Fälle beträgt lt. Angabe gerade 40% der Grundgesamtheit, also $0.4m$, womit sich die gesuchte Wahrscheinlichkeit nach der elementaren Formel aus dem Vorlesungsskriptum (Abschnitt 3.1.1) als

$$\mathrm{P}\left(\text{Fahrrad} \mid \text{weiblich}\right) = \frac{g}{m} = \frac{0.4m}{m} = 0.4$$

ergibt. Es genügt hier also, einfach die Angabe einzusetzen, und es ergibt sich analog für die Studenten

$$\mathrm{P}\left(\text{Fahrrad} \mid \text{männlich}\right) = 0.5$$

Man beachte hier die Schreibweise $\mathrm{P}\left(F \mid w\right)$: dadurch geben wir sehr konzise zu erkennen, dass wir an Fahrrädern (F) nur interessiert sind, wenn das Geschlecht (w) stimmt – es wird angegeben, auf welche Grundgesamtheit (innerhalb aller Bokuangehörigen) sich die Wahrscheinlichkeit bezieht.

b) Wie groß ist die Wahrscheinlichkeit, dass ein Fahrradfahrer weiblich ist? Als Zugriffsmechanismus können wir uns vorstellen, dass wir vor einem zufälligen Fahrradständer der Boku bei einem zufällig ausgewählten Rad warten bis die BesitzerIn auftaucht, woraufhin wir das Geschlecht feststellen. Die Grundgesamtheit innerhalb der Bokuistas ist also nun die Menge der Fahrradfahrer, und was wir suchen ist

$$\mathrm{P}\,(\text{weiblich} \mid \text{Fahrrad}) = \,?$$

oder kurz $\mathrm{P}\,(w \mid F)$. Es geht also im wesentlichen um das umgekehrte Problem wie unter Punkt a), und für diese Art von Inversionen ist der Satz von Bayes (Abschnitt 3.2) zuständig, der hier angewendet besagt, dass

$$\mathrm{P}\,(w \mid F) = \frac{\mathrm{P}\,(F \mid w)\,\mathrm{P}\,(w)}{\mathrm{P}\,(F \mid w)\,\mathrm{P}\,(w) + \mathrm{P}\,(F \mid m)\,\mathrm{P}\,(m)}$$

Hier steht auf der linken Seite, was wir suchen, und auf der rechten nur Dinge, die wir schon kennen: die bedingten Wahrscheinlichkeiten $\mathrm{P}\,(F \mid w)$ und $\mathrm{P}\,(F \mid m)$ aus Punkt a), und die Wahrscheinlichkeiten, dass ein beliebig ausgewählter Bokuista weiblich resp. männlich ist, aus der Angabe ($\mathrm{P}\,(w) = 0.35$ bzw. $\mathrm{P}\,(m) = 0.65$). Eingesetzt ergibt sich damit

$$\mathrm{P}\,(\text{weiblich} \mid \text{Fahrrad}) = \frac{0.4 \cdot 0.35}{0.4 \cdot 0.35 + 0.5 \cdot 0.65} = 0.301$$

und für das andersrum gelagerte Problem

$$\mathrm{P}\,(m \mid F) = \frac{\mathrm{P}\,(F \mid m)\,\mathrm{P}\,(m)}{\mathrm{P}\,(F \mid w)\,\mathrm{P}\,(w) + \mathrm{P}\,(F \mid m)\,\mathrm{P}\,(m)} = 0.699$$

Man beachte: Der Satz von der totalen Wahrscheinlichkeit (ebenfalls Abschnitt 3.2) garantiert, dass

$$\mathrm{P}(F) = \mathrm{P}\,(F \mid w)\,\mathrm{P}\,(w) + \mathrm{P}\,(F \mid m)\,\mathrm{P}\,(m) = 0.465$$

Im Nenner des Satz von Bayes steht also nur die Wahrscheinlichkeit, dass einE zufällig ausgewählteR BokuangehörigeR Fahrrad fährt.

c) Nur zum Drüberstreuen: bei 2 Sekunden pro Kombination kann man 900 Kombinationen in einer halben Stunde ausprobieren. Wenn eine davon die richtige ist, dann ist das günstig (für den Knacker); möglich sind natürlich 10000 Kombinationen von 0000 bis 9999, also gilt

$$\mathrm{P}(\text{erfolreicher Bruch}) = \frac{g}{m} = \frac{900}{10000} = 0.09$$

Lösung 2.4
Auftrittswahrscheinlichkeiten von Erdbeben

a) Zuerst muss die Verteilung der Zufallsvariable X bestimmt werden. Es handelt sich dabei um eine Binomialverteilung. Diese hat die beiden Parameter n (= Gesamtanzahl an betrachteten Zufallsereignissen) und p (= 'Erfolgswahrscheinlichkeit'; hier bedeutet 'Erfolg', dass ein Erdbeben innerhalb des Jahres eintritt). Es gilt nach der Angabe: $n = 10$

(wir betrachten zehn Jahre) und $p = 0.1$ (innerhalb eines Jahres tritt das Erdbeben mit Wahrscheinlichkeit von 10 % ein). Kurz schreiben wir, dass $X \sim B(10, 0.1)$.

Die Wahrscheinlichkeitsfunktion einer Binomialverteilung lautet allgemein für $i = 0, 1, \dots, n$

$$\mathrm{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$$

mit dem Binomialkoeffizienten

$$\binom{n}{i} = \frac{n!}{i! \cdot (n-i)!}$$

Für die hier gefragte Verteilung ergibt sich also

$$\mathrm{P}(X = i) = \binom{10}{i} \cdot 0.1^i \cdot 0.9^{10-i}$$

für $i = 0, 1, \dots, 10$.

b) Hier suchen wir die Wahrscheinlichkeit, dass es in genau drei von zehn Jahren zu einem Erdbeben kommt. Gefragt ist also $\mathrm{P(X = 3)}$. Wir ersetzen in obiger Formel daher i durch den Wert 3 und erhalten

$$\mathrm{P}(X = 3) = \binom{10}{3} \cdot 0.1^3 \cdot 0.9^{10-3} = 120 \cdot 0.1^3 \cdot 0.9^7 = 0.0574$$

mit

$$\binom{10}{3} = \frac{10!}{7! \cdot 3!} = \frac{10 \cdot 9 \cdot 8 \cdot \ldots \cdot 2 \cdot 1}{(7 \cdot 6 \cdot \ldots \cdot 2 \cdot 1) \cdot (3 \cdot 2 \cdot 1)} = \frac{3628800}{5040 \cdot 6} = \frac{3628800}{30240} = 120$$

c) Hier soll die Wahrscheinlichkeit berechnet werden, dass es in höchstens drei von zehn Jahren zu einem Erdbeben kommt, also $\mathrm{P}(X \leq 3)$. Wir teilen die Berechnung in Wahrscheinlichkeiten für disjunkte Ereignisse auf und berechnen dann

$$\begin{aligned}
\mathrm{P}(X \leq 3) &= \mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3) \\
&= \binom{10}{0} \cdot 0.1^0 \cdot 0.9^{10-0} + \binom{10}{1} \cdot 0.1^1 \cdot 0.9^{10-1} + \binom{10}{2} \cdot 0.1^2 \cdot 0.9^{10-2} \\
&\quad + \binom{10}{3} \cdot 0.1^3 \cdot 0.9^{10-3} \\
&= 1 \cdot 0.1^0 \cdot 0.9^{10} + 10 \cdot 0.1^1 \cdot 0.9^9 + 45 \cdot 0.1^2 \cdot 0.9^8 + 120 \cdot 0.1^3 \cdot 0.9^7 \\
&= 0.34868 + 0.3874 + 0.1937 + 0.0574 \\
&= 0.9872
\end{aligned}$$

d) Hier soll die Wahrscheinlichkeit berechnet werden, dass es in mindestens einem aber in höchstens drei von zehn Jahren zu einem Erdbeben kommt, also $\mathrm{P}(1 \leq X \leq 3)$. Wir teilen diese Wahrscheinlichkeit wieder auf und berechnen analog zu Teil c) des Beispiels

$$\begin{aligned}
\mathrm{P}(1 \leq X \leq 3) &= \mathrm{P}(X = 1) + \mathrm{P}(X = 2) + \mathrm{P}(X = 3) \\
&= \binom{10}{1} \cdot 0.1^1 \cdot 0.9^{10-1} + \binom{10}{2} \cdot 0.1^2 \cdot 0.9^{10-2} + \binom{10}{3} \cdot 0.1^3 \cdot 0.9^{10-3} \\
&= 10 \cdot 0.1^1 \cdot 0.9^9 + 45 \cdot 0.1^2 \cdot 0.9^8 + 120 \cdot 0.1^3 \cdot 0.9^7 \\
&= 0.3874 + 0.1937 + 0.0574 \\
&= 0.6385
\end{aligned}$$

Lösung 2.5
Toleranzintervalle für normalverteilte Größen

Wir vereinfachen den Fall der Normalverteilung stets auf den der Standardnormalverteilung:

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Gesucht ist nun zunächst $p = \mathrm{P}(|X - \mu| \leq k\sigma)$, also die Wahrscheinlichkeit, dass ein Wert aus einer Normalverteilung nicht weiter als k-mal die Standardabweichung σ vom Mittelwert μ entfernt liegt (siehe Abbildung 2.1). Umformung liefert hier

$$\begin{aligned}
\mathrm{P}(|X - \mu| \leq k\sigma) &= \mathrm{P}(\mu - k\sigma \leq X \leq \mu + k\sigma) \\
&= \mathrm{P}\left(-k \leq \frac{X - \mu}{\sigma} \leq k\right) \\
&= \Phi(k) - \Phi(-k) \\
&= \Phi(k) - [1 - \Phi(k)] \\
&= 2\Phi(k) - 1 = p
\end{aligned}$$

d.h. die Wahrscheinlichkeit hängt nur von k ab (Φ bezeichne wie immer die Verteilungsfunktion der Standardnormalverteilung).

a) Hier soll p für verschiedene Werte von k bestimmt werden. Für $k = 1$ etwa finden wir in Tabelle A.1 des VL-Skriptums den Wert $\Phi(1)$ in der Zeile 1.0 und der Spalte 0.00; daraus ergibt sich

$$p = 2\Phi(1) - 1 = 2 \cdot 0.8413 - 1 = 0.683$$

Ein beliebiger Wert aus einer Normalverteilung liegt also mit ca. 68%-iger Wahrscheinlichkeit im Intervall $\mu \pm \sigma$. Die weiteren Ergebnisse sind wie folgt:

k	1.0	1.5	2.0	2.5	3.0	3.5
p	0.683	0.866	0.954	0.988	0.997	0.9995

b) Hier soll für festes p ein passendes k gefunden werden. Indem wir obige Gleichung noch einmal umformen, erhalten wir

$$\Phi(k) = (p+1)/2$$

Betrachten wir also $p = 0.9$, so sollen wir den Wert k bestimmen, für den

$$\Phi(k) = 0.95$$

gilt. In Tabelle A.1 finden wir die Einträge 0.9495 und 0.9505 in der Zeile 1.6 und den Spalten 0.04 bzw. 0.05. Das "wahre" k wird also zwischen 1.64 und 1.65 liegen, also ca. bei 1.645.

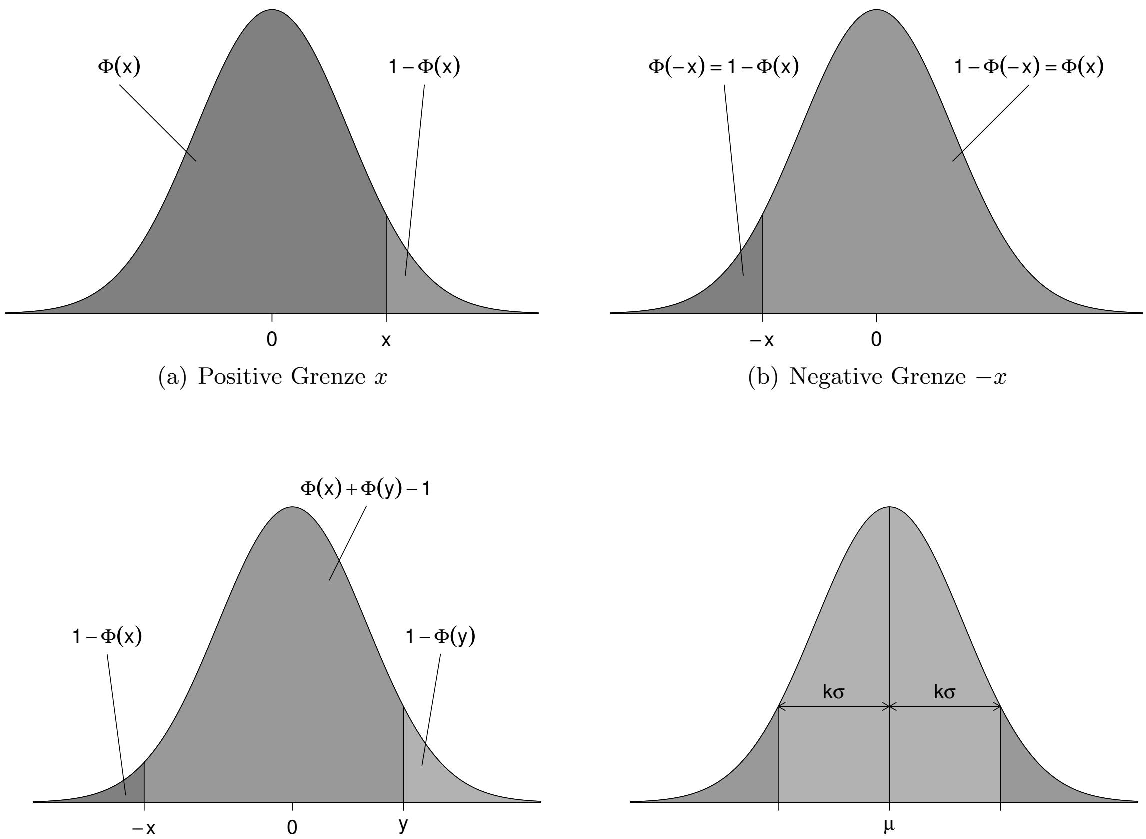
p	0.90	0.95	0.99	0.995	0.999
k	1.645	1.960	2.575	2.810	3.270

c) Hier fehlt der Betrag, also soll die Wahrscheinlichkeit bestimmt werden, dass ein zufälliger Wert aus einer Normalverteilung nicht um mehr als t-mal die Standardabweichung größer als der Mittelwert ist. Der obenstehende Ansatz lässt sich leicht modifizieren:

$$\begin{aligned} \mathrm{P}(X - \mu \leq t\sigma) &= \mathrm{P}\left(\frac{X-\mu}{\sigma} \leq t\right) \\ &= \Phi(t) = p \end{aligned}$$

Mit der selben Überlegung wie unter Punkt b) lassen sich damit die folgenden Ergebnisse erzielen:

p	0.90	0.95	0.99	0.995	0.999
t	1.280	1.645	2.325	2.575	3.080



(a) Positive Grenze x (b) Negative Grenze $-x$

(c) Beidseitige Grenzen (d) Allgemeine symmetrische Grenzen

Abbildung 2.1: Dichte der Standardnormalverteilung für Beispiel 2.5.

Lösung 2.6
Qualitätskontrolle bei der Spaghettiherstellung

Wie schon in Beispiel 2.5 vereinfachen wir wieder auf die Standardnormalverteilung:

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$$

a) Für den 1. Fall ist nun die Wahrscheinlichkeit gesucht, dass ein Spaghetti kürzer als 28.5 cm ist, also $\mathrm{P}(X \leq 28.5)$. Daraus lesen wir ab bei gegebener Standardabweichung, dass offenbar

$$\mathrm{P}(X \leq \mu - 2.5\sigma)$$

und wegen obiger Definition der Standardnormalverteilung ist

$$\mathrm{P}\left(\frac{X-\mu}{\sigma} \leq -2.5\right) = \Phi(-2.5) = 1 - \Phi(2.5) = 0.0062$$

Lediglich 0.62% der Spaghettiproduktion sollten also kürzer als 28.5 cm sein.

b) Hier suchen wir die Wahrscheinlichkeit, dass eine Spaghetti länger als 31.1 cm ist, also $\mathrm{P}(X \geq 31.1) = 1 - \mathrm{P}(X \leq 31.1)$. Analog zum ersten Teil des Beispiels berechnet sich die

Wahrscheinlichkeit:

$$\mathrm{P}(X \leq \mu + 1.83\sigma) = 1 - \Phi(1.83) = 1 - 0.9664 = 0.0336$$

Mit 3.36%iger Wahrscheinlichkeit ist ein Spaghetti daher länger als 31.1 cm.

c) Hier soll die Wahrscheinlichkeit berechnet werden, mit der eine Spaghetti zwischen 2 Werten liegt, also $\mathrm{P}(29.1 \leq X \leq 30.8)$. Daraus folgt wiederum

$$\mathrm{P}(\mu - 1.5\sigma \leq X \leq \mu + 1.33\sigma) = \mathrm{P}\left(-1.5 \leq \frac{X-\mu}{\sigma} \leq 1.33\right)$$

und daher berechnen wir

$$\Phi(1.33) + \Phi(1.5) - 1 = 0.9082 + 0.9332 - 1 = 0.8414$$

84.14% der Spaghetti sind daher zwischen 29.1 und 30.8 cm lang.

d) Hier ist die Wahrscheinlichkeit bereits bekannt und das dazugehörige Quantil der Verteilungsfunktion ist gesucht. Die Frage stellt sich daher in folgender Form dar:

$$\mathrm{P}(X \leq a) = 0.95$$

wobei a natürlich wie oben schon in der Form $a = \mu + z\sigma$ vorliegt.

$$\mathrm{P}(X \leq \mu + z\sigma) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \leq z\right) = \Phi(z) = 0.95$$

In der Normalverteilungstabelle finden sich somit $z = 1.645$. Mit dieser Information kann letztendlich a berechnet werden:

$$a = 30 + 1.645 \cdot 0.6 = 30.987$$

Die Schachtel muss daher 30.987 cm lang sein, damit 95% der Spaghetti hineinpassen.

Lösung 2.7
Apfelgewicht

Wie schon in Beispiel 2.5 und 2.6 vereinfachen wir wieder auf die Standardnormalverteilung:

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1) \Rightarrow F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

a) Es ist nun die Wahrscheinlichkeit gesucht, dass ein Apfel weniger als 17 dag wiegt, also $\mathrm{P}(X < 17)$. Es gilt wegen der Stetigkeit der Normalverteilung, dass

$$\mathrm{P}(X < 17) = \mathrm{P}(X \leq 17) = F_X(17)$$

Um die gesuchte Wahrscheinlichkeit in der Tabelle ablesen zu können, verwenden wir die oben angegebene Vereinfachung auf die Standardnormalverteilung:

$$F_X(17) = \Phi\left(\frac{17-\mu}{\sigma}\right) = \Phi\left(\frac{17-15}{1.5}\right) = \Phi\left(\frac{4}{3}\right) \approx \Phi(1.33)$$

In der Tabelle finden wir $\Phi(1.33) = 0.9082$ was der gesuchten Wahrscheinlichkeit entspricht. Ca. 90.82 % der Äpfel wiegen unter 17 dag.

b) Hier suchen wir die Wahrscheinlichkeit, dass ein Apfel mehr als 16 dag wiegt, also $\mathrm{P}(X > 16) = 1 - \mathrm{P}(X \leq 16)$. Analog zum ersten Teil des Beispiels berechnet sich die Wahrscheinlichkeit:

$$\begin{aligned} 1 - \mathrm{P}(X \leq 16) &= 1 - F_X(16) = 1 - \Phi\left(\frac{16-15}{1.5}\right) = 1 - \Phi\left(\frac{2}{3}\right) \\ &\approx 1 - \Phi(0.67) = 1 - 0.7486 = 0.2514 \end{aligned}$$

Mit ca. 25.14 %iger Wahrscheinlichkeit ist ein Apfel schwerer als 16 dag.

c) Hier soll die Wahrscheinlichkeit berechnet werden, mit der das Gewicht eines Apfels zwischen 14 und 17 dag liegt, also $\mathrm{P}(14 < X < 17)$. Es gilt

$$\mathrm{P}(14 < X < 17) = \mathrm{P}(14 < X \leq 17)$$

wegen der stetigen Verteilung. Wir berechnen dann, dass

$$\begin{aligned} \mathrm{P}(14 < X \leq 17) &= F_X(17) - F_X(14) = \Phi\left(\frac{17-15}{1.5}\right) - \Phi\left(\frac{14-15}{1.5}\right) \\ &\approx \Phi(1.33) - \Phi(-0.67) = \Phi(1.33) - (1 - \Phi(0.67)) \\ &= \Phi(1.33) + \Phi(0.67) - 1 = 0.9082 + 0.7486 - 1 \\ &= 0.6568 \end{aligned}$$

Somit liegt bei ca. 65.68 % der Äpfel das Gewicht zwischen 14 und 17 dag.

Kapitel 3

Normalverteilungsverfahren

Beispiele

Eine Stichprobe

Beispiel 3.1 Bei der Messung des Eiweißgehaltes von Rohmilch werden folgende (normalverteilte) Werte (in %) gemessen:

3.57	3.66	3.48	3.62	3.88	3.80	3.82	3.73	3.23	3.58	3.70
3.52	3.54	3.34	3.62	3.28	3.33	3.46	3.38	3.68	3.68	3.81

a) Geben Sie ein beidseitiges 95%-Konfidenzintervall für den mittleren Eiweißgehalt an.

b) Geben Sie ein beidseitiges 95%-iges Konfidenzintervall für den Mittelwert an, wenn die theoretische Standardabweichung mit $\sigma = 0.2$ als bekannt angenommen werden kann.

c) Geben Sie ein einseitiges, nach unten beschränktes 95%-iges Konfidenzintervall für den mittleren Eiweißgehalt an.

Beispiel 3.2 Bei der Messung der Sprosslänge von Ackerehrenpreis (*Veronica agrestis*) wurden folgende (normalverteilte) Werte (in mm) gemessen:

298	345	183	340	350	380	190	351	443	290
160	298	185	370	245	377	92	380	195	265

a) Geben Sie ein beidseitiges 95%-iges Konfidenzintervall für die mittlere Sprosslänge an.

b) Geben Sie ein beidseitiges 95%-iges Konfidenzintervall für die Standardabweichung der Sprosslängen an.

c) Geben Sie ein einseitiges, nach oben beschränktes 95%-iges Konfidenzintervall für die Standardabweichung der Sprosslängen an.

Beispiel 3.3 Im Zuge der Molekulargewichtsbestimmung vom Genom des Parapoxvirus *SPV 660* wurden mit Hilfe des Elektronenmikroskopes die Länge von 40 Molekülen (in mm) bestimmt. Sie können im folgenden davon ausgehen, dass diese Werte normalverteilt sind.

9.02	8.26	9.19	8.32	9.10	9.25	8.42	8.30	8.74	8.84
8.60	8.45	9.11	9.42	8.45	8.89	9.37	8.66	8.68	9.19
8.26	8.47	8.25	8.98	8.82	8.39	8.95	9.02	8.32	8.22
8.44	8.35	8.53	8.82	8.78	8.87	8.93	8.66	8.57	8.73

a) Geben Sie beidseitige 99%-Konfidenzintervalle für die mittlere Moleküllänge und die Standardabweichung an.

b) Testen Sie mit Sicherheit $1 - \alpha = 0.95$ die Nullhypothese, dass die mittlere Moleküllänge gleich 9.0 mm ist.

c) Testen Sie mit Sicherheit $1 - \alpha = 0.95$ die Nullhypothese, dass die Standardabweichung gleich 0.5 mm ist.

Beispiel 3.4 Bei der Messung des Laktosegehaltes von Rohmilch werden folgende (normalverteilte) Werte (in %) gemessen:

4.57	4.66	4.48	4.62	4.88	4.80	4.82	4.73	5.23
5.02	4.54	4.94	4.92	5.08	5.03	4.76	5.08	5.18

a) Geben Sie ein nach unten beschränktes einseitiges 95%-Konfidenzintervall für den Mittelwert μ und sowie ein nach oben beschränktes 95%-Konfidenzintervall für die Standardabweichung σ an.

b) Testen Sie die Hypothese, dass der mittlere Laktosegehalt gleich 4.8% ist, auf dem Signifikanzniveau $\alpha = 0.01$.

c) Testen Sie mit Sicherheit $1 - \alpha = 0.95$ die Nullypothese, dass der mittlere Laktosegehalt größer gleich 5% ist.

d) Testen Sie die Nullhypothese, dass die Standardabweichung der Laktosegehalte gleich 0.3% (Signifikanzniveau $\alpha = 0.01$).

e) Die Nullhypothese, dass die Standardabweichung σ kleiner gleich 0.15% ist, soll mit Sicherheit $1 - \alpha = 0.95$ getestet werden.

Zwei Stichproben

Beispiel 3.5 Im Rahmen eines Kälbermastversuches wurde der Einsatz eines neuen Eiweißkonzentrates als alternativer Eiweißträger im Milchaustauschfutter (M1) mit einem herkömmlichen Milchaustauschfutter (M2) verglichen. Nachfolgend sind die Tageszunahmen (in kg) im Endmastabschnitt aufgelistet.

M1:	1.36	1.24	1.38	1.40	1.12	1.45	1.56	1.43	1.26
	1.32	1.42	1.24	1.43	1.48	1.35			
M2:	1.31	1.28	1.16	1.29	1.08	1.23	1.45	1.40	1.30
	1.21	1.25	1.19	1.35	1.38				

a) Geben Sie beidseitige 95%-Konfidenzintervalle für Mittel und Standardabweichung von M1 und M2 an.

b) Überprüfen Sie mit Sicherheit $1-\alpha = 0.95$ die Hypothese, dass die mittleren Tageszunahmen für beide Futtermittel gleich sind, wenn die entsprechenden Standardabweichungen σ_{M1} und σ_{M2} als gleich angenommen werden können.

c) Testen Sie die gleiche Hypothese wie unter b), wenn die Standardabweichungen der Futtermittel nicht als gleich angenommen werden können.

Beispiel 3.6 Ein Nahrungsmittelproduzent hat sich auf die Erzeugung von Nudeln spezialisiert. Er verwendet für die Erzeugung seiner als wohlschmeckend bekannten Spaghetti zwei unterschiedliche Produktionsverfahren. Um zu wissen, welches der Produktionsverfahren besser läuft, wird das Streuverhalten der Nudellänge (in mm) untersucht. Die Messwerte beider Verfahren scheinen normalverteilt zu sein.

PV I	370	378	375	371	381	382	376	379	380	369	375	374
	371	378	373									
PV II	375	389	371	372	370	385	381	377	385	381	374	390

a) Geben Sie beidseitige 95%-Konfidenzintervalle für die Streuungen der Produktionsverfahren an.

b) Für beide Produktionsverfahren soll die Hypothese getestet werden, dass die jeweilige Standardabweichung unter 5 mm liegt (Sicherheit $1 - \alpha = 0.95$).

c) Testen Sie die Hypothese, dass die Streuungen beider Produktionsverfahren gleich hoch sind, auf dem Signifikanzniveau $\alpha = 0.05$.

d) Überprüfen Sie mittels eines geeigneten Tests, ob die mittlere Nudellänge für beide Verfahren gleich ist (Signifikanzniveau $\alpha = 0.05$). Wählen Sie das dafür geeignete Verfahren auf Grund des Ergebnisses aus Punkt c).

e) Kommentieren Sie die unter b) bis d) anfallenden Fehlerwahrscheinlichkeiten für Ihre Entscheidungen.

Beispiel 3.7 Bei einem Forschungsprojekt über die Schadstoffbelastung der Luft wird an 21 Messpunkten an zwei Terminen (März/Juli) der Asbestgehalt untersucht. Gemessen wird die Anzahl der Asbestfasern pro Kubikmeter Luft. Die Messpunkte befinden sich im Freien in einem einmetrigen Abstand von asbesthältigen Gebäudeteilen.

450/530	480/510	250/450	600/780	530/500	420/490	580/600
670/780	460/550	320/440	200/380	430/300	520/450	480/540
650/530	580/610	750/550	540/710	430/540	320/490	680/800

a) Geben Sie jeweils ein 95%-Konfidenzintervall für Mittel und Standardabweichung der Messtermine März und Juli an.

b) Testen Sie mit Sicherheit $1 - \alpha = 0.95$ die Nullhypothese, dass der Asbestgehalt der Luft an beiden Messterminen gleich ist, wenn bekannt ist, dass die Differenzen normalverteilt sind.

c) Geben Sie ein 95%-iges Konfidenzintervall für den wahren mittleren Wert der Differenzen an.

d) Testen Sie mit $1 - \alpha = 0.99$ die Nullhypothese, dass das Mittel des Monats März größer gleich dem Mittel des Monats Juli ist.

Beispiel 3.8 Zwei Bäckereien in der näheren Umgebung eines Großbetriebes mit eigener Werksküche beliefern diese mit Semmeln. Da die Küchenleitung, durch entsprechende Beschwerden der Mitarbeiter bestärkt, daran interessiert ist, möglichst gleichmäßige Qualität, d.h. unter anderem annähernd gleich große Semmeln, auszugeben, entschließt sie sich, anhand zweier Stichproben das Streuverhalten im Gewicht des Gebäcks zu vergleichen. Folgende Messwerte (in g) liegen vor, wobei das Gewicht offensichtlich normalverteilt zu sein scheint.

Fabrik A	49.7	51.9	49.0	49.6	51.2	49.2	49.8	49.9	46.5	51.7
Fabrik B	48.4	46.3	51.1	49.2	48.7	49.4	49.9	48.3		

a) Lässt sich die Hypothese, dass beide Produktionsstreuungen übereinstimmen, widerlegen? (Signifikanzniveau $\alpha = 0.05$)

b) Lässt sich die Hypothese, dass das mittlere Semmelgewicht beider Fabriken übereinstimmt, widerlegen? (Signifikanzniveau $\alpha = 0.05$) Wählen Sie das geeignete Verfahren auf Grund des Ergebnisses von Punkt a).

Beispiel 3.9 Im Zuge einer Honigleistungserhebung wurden an einem Standort in zwei aufeinanderfolgenden Jahren (1992/1993) pro Bienenvolk folgende normalverteilte Jahreshonigleistungen (in kg) ermittelt. Insgesamt wurden 35 Bienenvölker untersucht.

21/17	25/26	12/ 9	21/19	15/16	14/10	23/18
22/20	19/17	26/24	13/14	15/13	16/12	18/17
21/18	11/ 9	13/10	18/18	15/12	23/20	14/12
26/18	16/17	23/21	13/15	16/19	15/11	27/17
22/17	20/19	23/20	17/15	24/20	13/15	16/11

a) Testen Sie auf dem Signifikanzniveau $\alpha = 0.05$ die Hypothese, dass der mittlere Ertrag für beide Jahre gleich ist. Sie können davon ausgehen, dass die Differenzen normalverteilt sind.

b) Geben Sie ein 95%iges Konfidenzintervall für den wahren mittleren Wert der Differenzen an.

c) Testen Sie mit Sicherheit $1 - \alpha = 0.99$ die Hypothese, dass das Mittel des Jahres 1993 größer als das Mittel des Jahres 1992 ist.

d) Kommentieren Sie die unter a) und c) anfallenden Fehlerwahrscheinlichkeiten für Ihre Entscheidungen.

Lösungen

Lösung 3.1
Eiweißgehalt von Rohmilch

Wir berechnen zunächst Mittel und Standardabweichung der Stichprobe (ein bis zwei Stellen mehr als die Datengenauigkeit hergibt):

$$\bar{x} = 3.578 \qquad \text{und} \qquad s = 0.185$$

a) Beidseitiges Konfidenzintervall für den Mittelwert:

Wenn die Einzelmessungen wie hier unabhängig und identisch normalverteilt sind, dann gilt wie im Vorlesungsskript (Abschnitt 5.1) dargestellt, dass der Ausdruck

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

einer t-Verteilung mit $n-1$ Freiheitsgraden folgt. Für eine beliebige Wahrscheinlichkeit α gilt dann:

$$\mathrm{P}\left(t_{n-1;\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1;1-\alpha/2}\right) = 1 - \alpha$$

(So sind ja die Quantile, hier die der t-Verteilung, gerade definiert.) Außerdem gilt, weil die t-Verteilung symmetrisch um 0 ist, dass $t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$. Damit kann man die obenstehende Ungleichung umformen und erhält:

$$\mathrm{P}\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Damit erhalten wir für ein beidseitiges Konfidenzintervall für den Mittelwert μ mit Überdeckungswahrscheinlichkeit $1 - \alpha$

$$\bar{x} \pm t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

den gleichen Ausdruck wie in Tabelle 5.1 des Vorlesungsskripts; hier ist alles bekannt:

$$\bar{x} = 3.578 \qquad s = 0.185 \qquad t_{21;0.975} = 2.080$$

und Einsetzen ergibt

$$3.496 \leq \mu \leq 3.660$$

als das gesuchte Konfidenzintervall.

b) Beidseitiges Konfidenzintervall bei bekannter theoretischer Standardabweichung $\sigma = 0.2$ (ein eher theoretischer Fall, weil woher sollte man i.a. das theoretische σ exakt kennen?):

Die Herleitung ist ähnlich wie im allgemeinen Fall unter a), nur dass hier die Tatsache verwendet wird, dass der Ausdruck

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

einer Standardnormalverteilung $N(0,1)$ folgt (siehe z.B. Beispiel 4.7 oder 4.9 im Vorlesungsskript). Damit kann man mit den Quantilen z_α der Normalverteilung wieder schreiben:

$$\mathrm{P}\left(z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

Weil die Standardnormalverteilung (ebenso wie die t-Verteilung) symmetrisch um 0 ist, gilt für die Quantile $z_{\alpha/2} = -z_{1-\alpha/2}$, und mit Umformung ergibt sich

$$\mathrm{P}\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Damit erhalten wir für die Grenzen des gesuchten Konfidenzintervalls den Ausdruck

$$\bar{x} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

gerade wie in Tabelle 5.1; wieder sind alle Größen bekannt:

$$\bar{x} = 3.578 \qquad \sigma = 0.2 \qquad z_{0.975} = 1.96$$

und durch Einsetzen erhält man:

$$3.494 \leq \mu \leq 3.661$$

c) Nach unten beschränktes Konfidenzintervall für den Mittelwert:

Wenn man nur an einer Absicherung nach unten interessiert ist (hier also einem Mindesteiweißgehalt der Rohmilch), dann lässt sich ein einseitiges Konfidenzintervall wie unter a) aus der t-Verteilung (mit $n-1$ Freiheitsgraden) des Ausdrucks

$$\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

konstruieren. Für eine beliebige Wahrscheinlichkeit α gilt ja auch:

$$\mathrm{P}\left(\frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{n-1;1-\alpha}\right) = 1-\alpha$$

Umformen ergibt dann:

$$\mathrm{P}\left(\bar{X} - t_{n-1;1-\alpha}\frac{S}{\sqrt{n}} \leq \mu\right) = 1-\alpha$$

Damit erhalten wir für ein einseitiges, nach unten beschränktes Konfidenzintervall für den Mittelwert μ mit Überdeckungswahrscheinlichkeit $1-\alpha$

$$\bar{x} - t_{n-1;1-\alpha} \cdot \frac{s}{\sqrt{n}}$$

den gleichen Ausdruck wie in Tabelle 5.1 des Vorlesungsskripts; hier ist alles bekannt:

$$\bar{x} = 3.58 \qquad s = 0.1849 \qquad t_{21;0.95} \; = \; 1.721$$

und Einsetzen ergibt

$$3.512 \leq \mu$$

Beachten Sie, dass hier die untere Grenze für den wahren mittleren Eiweißgehalt größer ist als die untere Grenze des beidseitigen Intervalls unter a): Indem wir beliebige Abweichungen nach oben akzeptieren, können wir eine schärfere Schranke für die Abweichungen nach unten ziehen.

Lösung 3.2
Sprosslänge von Ackerehrenpreis

Wie in Beispiel 3.1 berechnen wir zunächst das arithmetische Mittel und die empirische Standardabweichung:

$$\bar{x} = 286.85 \qquad \text{und} \qquad s = 93.64$$

a) Beidseitiges Konfidenzintervall für den Mittelwert:

Wie in Beispiel 3.1 oder aus Tabelle 5.1 des Vorlesungsskriptums erhalten wir als beidseitiges Konfidenzintervall für $\alpha = 0.05$:

$$\bar{x} \pm t_{n-1;0.975} \cdot \frac{s}{\sqrt{n}}$$

Mit dem entsprechenden Quantil $t_{19;0.975} = 2.093$ und den zuvor berechneten Kenngrößen ergibt sich als Konfidenzintervall

$$243.02 \leq \mu \leq 330.68$$

b) Beidseitiges Konfidenzintervall für die Standardabweichung:

Es ist mathematisch bequemer, zuerst ein Konfidenzintervall für die Varianz zu berechnen und aus diesem dann (durch Wurzelziehen) ein Konfidenzintervall für die Standardabweichung zu gewinnen. Dafür verwendet man die Tatsache, dass für normalverteilte und unabhängige Einzelmessungen der Ausdruck

$$\frac{(n-1)S^2}{\sigma^2}$$

einer χ^2-Verteilung mit $n-1$ Freiheitsgraden folgt; für eine beliebige Wahrscheinlichkeit α gilt dann mit den entsprechenden Quantilen dieser Verteilung

$$\mathrm{P}\left(\chi^2_{n-1;\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1;1-\alpha/2}\right) = 1-\alpha$$

Durch Äquivalenzumformungen der Ungleichungen ergibt sich damit:

$$\mathrm{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right) = 1-\alpha$$

Das Konfidenzintervall für die Varianz σ^2 lautet also

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right]$$

wie in Tabelle 5.3 des Vorlesungsskriptums angegeben. Für die vorliegenden Daten gilt:

$$s^2 = 8768.766 \qquad \chi^2_{19;0.975} = 32.852 \qquad \chi^2_{19;0.025} = 8.907$$

(Da die χ^2-Verteilung nicht symmetrisch ist, müssen beide Quantile bestimmt werden.) Das Konfidenzintervall für die Varianz berechnet sich zu

$$5071.24 \leq \sigma^2 \leq 18704.45$$

und entsprechend ergibt sich für die Standardabweichung

$$71.21 \leq \sigma \leq 136.76$$

c) Einseitiges, nach oben beschränktes Konfidenzintervall für die Standardabweichung:

In vielen Anwendungen (z.B. Produktionsprozessen) ist man oft nur an einer oberen Schranke für die Variabilität der Daten (gemessen durch die Varianz bzw. Standardabweichung) interessiert, d.h. es ist egal (oder sogar gut), wenn die Varianz möglichst klein wird. Für ein entsprechendes einseitiges Konfidenzintervall verwendet man wie unter b) die χ^2-Verteilung mit $n-1$ Freiheitsgraden; aufgrund der Definition des Quantils gilt für beliebiges α

$$\mathrm{P}\left(\chi^2_{n-1;\alpha} \leq \frac{(n-1)S^2}{\sigma^2}\right) = 1-\alpha$$

Umformen der Ungleichung ergibt das Konfidenzintervall

$$\mathrm{P}\left(\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha}}\right) = 1-\alpha$$

und damit wie in Tabelle 5.3 des Vorlesungsskriptums

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1;\alpha}}$$

Mit

$$s^2 = 8768.766 \qquad \chi^2_{19;0.05} = 10.117$$

ergibt sich

$$\sigma^2 \leq 16467.98 \qquad \text{bzw.} \qquad \sigma \leq 128.33$$

Lösung 3.3
Moleküllängen

Zunächst berechnen wir Mittel, Varianz und Standardabweichung der Stichprobe:

$$\bar{x} = 8.7155 \qquad s = 0.3382 \qquad s^2 = 0.11436$$

a) Beidseitige Konfidenzintervalle für Mittelwert und Standardabweichung zu $\alpha = 0.01$: Da das benötigte Quantil der t-Verteilung $t_{39;0.995}$ in Tabelle A.2 nicht tabelliert ist, verwenden wir das nächstgelegene tabellierte Quantil mit **weniger** Freiheitsgraden (d.h. 35 Freiheitsgrade)

$$t_{39;0.995} \simeq t_{35;0.995} = 2.724$$

und erhalten damit für das 99% Konfidenzintervall für die mittlere Moleküllänge gemäß Tabelle 5.1 des Vorlesungsskriptums

$$8.570 \leq \mu \leq 8.861$$

Analog gehen wir bei den Quantilen zu 39 Freiheitsgraden der χ^2-Verteilung in Tabelle A.3 vor:

$$\chi^2_{39;0.005} \simeq \chi^2_{35;0.005} = 17.192 \qquad \text{bzw.} \ \chi^2_{39;0.995} \simeq \chi^2_{35;0.995} = 60.275$$

Einsetzen in die entsprechende Formel in Tabelle 5.3 des Vorlesungsskripts ergibt sodann

$$0.074 \leq \sigma^2 \leq 0.2594 \qquad \text{bzw.} \qquad 0.272 \leq \sigma \leq 0.509$$

b) Test der Hypothese $\mu = \mu_0 = 9$ für $\alpha = 0.05$:

Für die Nullhypothese $\mu = \mu_0$ wählen wir als Teststatistik t gemäß Tabelle 5.2 den Ausdruck

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Stammen die Daten aus einer Normalverteilung und gilt die Nullhypothese, dass also $\mu = 9$, so folgt die Teststatistik t einer t-Verteilung mit $n-1$ Freiheitsgraden. Mit $\bar{x} = 8.7155$ und $s = 0.3382$ ergibt sich die Teststatistik zu

$$t = \frac{8.7155 - 9}{0.3382/\sqrt{40}} = -5.32$$

Die kritischen Werte c_u und c_o ergeben sich nun als Quantile der t-Verteilung mit $n-1$ Freiheitsgraden:

$$c_u = t_{n-1;\alpha/2} \qquad c_o = t_{n-1;1-\alpha/2}$$

Wegen der Symmetrie der t-Verteilung gilt (siehe Beispiel 3.1 des Übungsskripts)

$$c_u/c_o = \pm t_{n-1;1-\alpha/2} = \pm t_{39;0.975}$$

Wie in Teil a) des Beispiels wählen wir wieder das nächstliegende Quantil mit weniger Freiheitsgraden:

$$c_u/c_o = \pm t_{39;0.975} \simeq \pm t_{35;0.975} = \pm 2.030$$

Es gilt also:

$$t = -5.32 < -2.030 = c_u$$

Die Teststatistik liegt unterhalb der unteren kritischen Grenze, also wird die Nullhypothese verworfen: die mittlere Moleküllänge ist signifikant (auf dem Niveau $\alpha = 0.05$) verschieden von 9 mm.

c) Test der Hypothese $\sigma = 0.5$ für $\alpha = 0.05$:

Wie beim Konfidenzintervall ist es einfacher, die Hypothese für die Varianz statt für die Standardabweichung zu testen. Wir verwenden also die äquivalente Hypothese

$$H_0 : \sigma^2 = 0.5^2 = 0.25$$

Die passende Teststatistik ist hier (siehe Tabelle 5.4 des Vorlesungsskriptums)

$$t = \frac{(n-1)s^2}{\sigma_0^2}$$

Stammen die Beobachtungen aus einer Normalverteilung und gilt die Nullhypothese, dass $\sigma^2 = 0.25$, ist dieser Ausdruck χ^2-verteilt mit $n-1$ Freiheitsgraden. Mit $s^2 = 0.11436$ ergibt sich damit

$$t = \frac{39 \cdot 0.11436}{0.25} = 17.84$$

Die kritischen Werte ergeben sich nun gerade als die Quantile der χ^2-Verteilung mit $n-1$ Freiheitsgraden:

$$c_u = \chi^2_{n-1;\alpha/2} = \chi^2_{39;0.025} \qquad c_o = \chi^2_{n-1;1-\alpha/2} = \chi^2_{39;0.975}$$

Wählt man wiederum die nächstgelegenen tabellierten Quantile mit weniger Freiheitsgraden in Tabelle A.3, also

$$c_u = \chi^2_{39;0.025} \simeq \chi^2_{35;0.025} = 20.569 \qquad c_o = \chi^2_{39;0.975} \simeq \chi^2_{35;0.975} = 53.203$$

so erhält man

$$t = 17.84 < 20.569 = c_u$$

Die Teststatistik liegt unterhalb der unteren kritischen Grenze, also wird die Nullhypothese verworfen: die Standardabweichung der Moleküllängen ist signifikant (auf dem Niveau $\alpha = 0.05$) verschieden von 0.5 mm.

Lösung 3.4
Laktosegehalt von Rohmilch

Zunächst erhalten wir für Mittel, Standardabweichung und Varianz der Stichprobe

$$\bar{x} = 4.85 \qquad s = 0.225 \qquad s^2 = 0.0506$$

a) Einseitige Konfidenzintervalle für Mittelwert und Standardabweichung mit $\alpha = 0.05$:

Für ein nach unten beschränktes Konfidenzintervall für den Mittelwert erhalten wir gemäß Tabelle 5.1 des Vorlesungsskripts

$$\bar{x} - t_{n-1;1-\alpha} \cdot \frac{s}{\sqrt{n}} \leq \mu$$

Mit Hilfe des entsprechenden t-Quantils $t_{17;0.95} = 1.74$ ergibt sich

$$4.76 \leq \mu$$

Für ein nach oben beschränktes Konfidenzintervall für die Standardabweichung entnehmen wir Tabelle 5.3 des Vorlesungsskriptums die Formel

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1;\alpha}}$$

und erhalten mit dem entsprechenden χ^2-Quantil $\chi^2_{17;0.05} = 8.672$

$$\sigma^2 \leq 0.0992 \qquad \text{bzw.} \qquad \sigma \leq 0.3150$$

b) Hypothese: $\mu = \mu_0 = 4.8$ für $\alpha = 0.01$:

Gemäß Tabelle 5.2 des Vorlesungsskripts berechnet sich die Teststatistik

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

berechnet sich mit $\bar{x} = 4.85$ und $s = 0.225$ zu $t = 0.943$. Die kritischen Werte sind die Quantile der t-Verteilung mit 17 Freiheitsgraden:

$$c_u/c_o = \pm t_{n-1;1-\alpha/2} = \pm t_{17;0.995} = \pm 2.898$$

Die Teststatistik liegt zwischen den kritischen Grenzen, also wird H_0 beibehalten. Es besteht kein signifikanter Einwand gegen die Hypothese, dass der mittlere Laktosegehalt μ gleich 4.8% ist.

c) Hypothese: $\mu \geq 5.0$ für $\alpha = 0.05$:

Die Teststatistik berechnet sich für diesen einseitigen Test (auf kleiner/größer) genauso wie für einen zweiseitigen Fall (auf Gleichheit) unter b), also

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4.85 - 5}{0.225/\sqrt{18}} = -2.83$$

Die kritischen Werte, oder genauer *der* kritische Wert, stammen wieder aus einer t-Verteilung mit $n-1 = 17$ Freiheitsgraden

$$c_u = t_{n-1;\alpha} = -t_{n-1;1-\alpha} = -t_{17;0.95} = -1.74$$

Die obere Grenze $c_o = \infty$ ist rein formal – große Werte der Teststatistik kommen ja gerade zustande, wenn der Mittelwert der Beobachtungen über dem hypothetischen μ_0 liegt, was hier gerade für die Nullhypothese spricht. Nur kleine Werte der Teststatistik können gegen die H_0 sprechen, deshalb ist auch nur die untere Grenze interessant.

Hier ist

$$t = -2.83 < -1.74 = c_u$$

Die Teststatistik liegt also außerhalb der kritischen Grenzen, weshalb H_0 abgelehnt wird. Der mittlere Laktosegehalt der Rohmilch ist signifikant kleiner als 5% (auf dem Niveau $\alpha = 0.05$).

d) Hypothese: $\sigma = 0.3$ für $\alpha = 0.01$:

Wir testen wieder die äquivalente Hypothese $\sigma^2 = 0.09$. Die Teststatistik berechnet sich gemäß Tabelle 5.4 des Vorlesungsskripts mit $s^2 = 0.0506$ zu

$$t = \frac{(n-1)s^2}{\sigma_0^2} = 9.56$$

Die kritischen Werte sind die Quantile der χ^2-Verteilung mit 17 Freiheitsgraden:

$$c_u = \chi^2_{n-1;\alpha/2} = \chi^2_{17;0.005} = 5.697 \qquad \text{und} \qquad c_o = \chi^2_{n-1;1-\alpha/2} = \chi^2_{17;0.995} = 35.718$$

Die Teststatistik liegt zwischen den kritischen Grenzen, also wird H_0 beibehalten. Es besteht also kein signifikanter Einwand gegen die Hypothese, dass die theoretische Standardabweichung σ der Laktosegehalte gleich 0.3% ist.

e) Hypothese: $\sigma \leq 0.15$ für $\alpha = 0.05$:

Wieder betrachten wir die äquivalente Hypothese $\sigma^2 \leq 0.15^2 = 0.0225$. Die Teststatistik hat für diesen einseitigen Test die gleiche Form wie im zweiseitigen Fall:

$$t = \frac{(n-1)s^2}{\sigma_0^2} = \frac{17 \cdot 0.0506}{0.0225} = 38.23$$

Der obere kritische Wert ergibt sich wieder aus einer χ^2-Verteilung mit $n-1 = 17$ Freiheitsgraden:

$$c_o = \chi^2_{n-1;1-\alpha} = \chi^2_{17;0.95} = 27.587$$

Die untere Grenze $c_u = 0$ ist für die Nullhypothese $\sigma \leq 0.15$ unerheblich, da kleine Werte der Teststatistik ja nicht gegen diese Hypothese sprechen, siehe Tabelle 5.4 des Vorlesungsskriptums. Hier liegt die Teststatistik über der oberen kritischen Grenze:

$$t = 38.23 > 27.587 = c_o$$

Die Nullhypothese wird also verworfen: Die Standardabweichung der Laktosegehalte ist signifikant größer als 0.15% (für $\alpha = 0.05$).

Lösung 3.5
Gewichtszunahme im Kälbermastversuch

Wir berechnen zunächst für die beiden Futtermittel M1 und M2 Stichprobenmittel und -standardabweichungen

$$\bar{x}_{\mathrm{M1}} = 1.363 \qquad s_{\mathrm{M1}} = 0.112 \qquad \bar{x}_{\mathrm{M2}} = 1.277 \qquad s_{\mathrm{M2}} = 0.100$$

a) Beidseitige 95%-Konfidenzintervalle für Mittelwert und Standardabweichung beider Futtermittel: Analog zu Beispiel 3.2 ergeben sich die Konfidenzintervalle zu

$$\begin{array}{ll} 1.301 \leq \mu_{\mathrm{M1}} \leq 1.425 & 0.082 \leq \sigma_{\mathrm{M1}} \leq 0.177 \\ 1.219 \leq \mu_{\mathrm{M2}} \leq 1.335 & 0.073 \leq \sigma_{\mathrm{M2}} \leq 0.161 \end{array}$$

b) Test der Hypothese $\mu_{\mathrm{M1}} = \mu_{\mathrm{M2}}$ für $\alpha = 0.05$ unter der Annahme gleicher Varianzen (Annahme: $\sigma_{\mathrm{M1}} = \sigma_{\mathrm{M2}} = \sigma$ unbekannt)

Es handelt sich hier um zwei **unabhängige Stichproben**, da ein Kalb jeweils nur einer Fütterungsgruppe zugeteilt sein kann. Der erste Schritt ist daher die Berechnung eines gemeinsamen (gepoolten) Schätzwertes s_P für σ aus den Standardabweichungen der beiden Stichproben; unter Annahme gleicher Varianzen ist der Ausdruck dafür gemäß Tabelle 5.6 des Vorlesungsskripts

$$\begin{aligned} s_P &= \sqrt{\frac{(n_{\mathrm{M1}}-1)\cdot s_{\mathrm{M1}}^2 + (n_{\mathrm{M2}}-1)\cdot s_{\mathrm{M2}}^2}{n_{\mathrm{M1}}+n_{\mathrm{M2}}-2}} \\ &= \sqrt{\frac{(15-1)\cdot 0.112^2 + (14-1)\cdot 0.100^2}{15+14-2}} = 0.106 \end{aligned}$$

Mit dieser gepoolten Standardabweichung lautet die Teststatistik nun

$$t = \frac{\bar{x}_{\mathrm{M1}} - \bar{x}_{\mathrm{M2}}}{s_P \cdot \sqrt{\frac{1}{n_{\mathrm{M1}}} + \frac{1}{n_{\mathrm{M2}}}}} = \frac{1.363 - 1.277}{0.106 \cdot \sqrt{1/15 + 1/14}} = 2.175$$

Unter Annahme der Nullhypothese (gleiche Mittel) ist diese Teststatistik t-verteilt mit $n_{M1} + n_{M2} - 2 = 27$ Freiheitsgraden. Die kritischen Werte werden wie üblich für einen beidseitigen t-Test als die entsprechenden Quantile der t-Verteilung gewählt (siehe Tabelle 5.6 im Vorlesungsskript):

$$c_u/c_o = \pm t_{n_1+n_2-2,1-\alpha/2} = \pm t_{27;0.975} = \pm 2.052$$

Die Teststatistik liegt also außerhalb des Annahmebereichs; die H_0 wird also abgelehnt, es gibt einen auf dem Niveau $\alpha = 0.05$ signifikanten Unterschied zwischen den Zunahmen der beiden Fütterungsgruppen.

c) Test der Hypothese $\mu_{\mathrm{M1}} = \mu_{\mathrm{M2}}$ für $\alpha = 0.05$ unter der Annahme verschiedener Varianzen (Annahme: $\sigma_{\mathrm{M1}} \neq \sigma_{\mathrm{M2}}$):

Falls wir die Annahme gleicher Varianzen nicht mehr aufrecht erhalten können, können wir immer noch mit einem approximativen t-Test arbeiten. Die Teststatistik vereinfacht sich sogar (Tabelle 5.6 Vorlesungsskript):

$$t = \frac{\bar{x}_{\mathrm{M1}} - \bar{x}_{\mathrm{M2}}}{\sqrt{\frac{s_{\mathrm{M1}}^2}{n_{\mathrm{M1}}} + \frac{s_{\mathrm{M2}}^2}{n_{\mathrm{M2}}}}} = \frac{1.363 - 1.277}{\sqrt{0.112^2/15 + 0.1^2/14}} = 2.18$$

Dieser Ausdruck ist approximativ (also mit wachsender Anzahl von Beobachtungen immer besser) t-verteilt mit folgenden Freiheitsgraden:

$$f = \frac{\left(\frac{s_{\mathrm{M1}}^2}{n_{\mathrm{M1}}} + \frac{s_{\mathrm{M2}}^2}{n_{\mathrm{M2}}}\right)^2}{\left(\frac{s_{\mathrm{M1}}^2}{n_{\mathrm{M1}}}\right)^2/(n_{\mathrm{M1}}-1) + \left(\frac{s_{\mathrm{M2}}^2}{n_{\mathrm{M2}}}\right)^2/(n_{\mathrm{M2}}-1)}$$

Am einfachsten berechnet man dazu zunächst

$$\frac{s_{\mathrm{M1}}^2}{n_{\mathrm{M1}}} = 0.112^2/15 = 0.00084 \qquad \frac{s_{\mathrm{M2}}^2}{n_{\mathrm{M2}}} = 0.1^2/14 = 0.00071$$

Damit erhält man die Freiheitsgrade als

$$f = \frac{(0.00084 + 0.00071)^2}{0.00084^2/14 + 0.00071^2/13} = 26.9$$

Damit folgt die Teststatistik also einer t-Verteilung mit 26.9 Freiheitsgraden. Da diese Verteilung nicht tabelliert ist, rundet man zur Bestimmung der kritischen Werte die Freiheitsgrade auf die nächstkleinere ganze Zahl $\tilde{f} = 26$ ab. Damit erhalten wir für die untere und obere Grenze des Annahmebereichs

$$c_u/c_o = \pm t_{\tilde{f},1-\alpha/2} = \pm t_{26;0.975} = \pm 2.056$$

Da die Teststatistik oberhalb der oberen kritischen Grenze liegt, wird die Nullhypothese verworfen. Im vorliegenden Fall kommen wir zur selben Entscheidung, unabhängig davon, ob wir gleiche Varianzen für beide Futtergruppen annehmen oder nicht. Sogar die Teststatistiken und die Freiheitsgrade sind für b) und c) trotz der unterschiedlichen Berechnungswege sehr ähnlich.

Lösung 3.6
Spaghettilängenvergleich zweier Produktionsverfahren

Für die beiden Stichproben (I und II) ermitteln wir zunächst folgende Kenngrößen:

$$\begin{array}{llll} \bar{x}_{\mathrm{I}} = 375.47 & s_{\mathrm{I}}^2 = 17.12 & s_{\mathrm{I}} = 4.14 & n_{\mathrm{I}} = 15 \\ \bar{x}_{\mathrm{II}} = 379.17 & s_{\mathrm{II}}^2 = 49.06 & s_{\mathrm{II}} = 7.00 & n_{\mathrm{II}} = 12 \end{array}$$

a) 95%-Konfidenzintervalle für die Standardabweichungen:

Das Konfidenzintervall für die Varianz σ^2 lautet gemäß Tabelle 5.3 des Vorlesungsskripts

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right]$$

Für das Produktionsverfahren I ergibt sich mit $\chi^2_{14;0.975} = 26.119$ und $\chi^2_{14;0.025} = 5.629$

$$9.18 \leq \sigma^2 \leq 42.63 \qquad \text{bzw.} \qquad 3.03 \leq \sigma \leq 6.53$$

Analog ergibt sich für das Produktionsverfahren II mit $\chi^2_{11;0.975} = 21.920$ und $\chi^2_{11;0.025} = 3.816$

$$24.59 \leq \sigma^2 \leq 141.25 \qquad \text{bzw.} \qquad 4.96 \leq \sigma \leq 11.88$$

b) Test der Hypothese $\sigma \leq 5$ auf dem Niveau $\alpha = 0.05$ für beide Produktionsverfahren:

Gemäß Tabelle 5.4 des Vorlesungsskripts lauten Teststatistik und kritische Werte

$$t = \frac{(n-1)s^2}{\sigma_0^2} \qquad c_u = 0 \qquad c_o = \chi^2_{n-1;1-\alpha}$$

Damit ergibt sich für Produktionsverfahren I:

$$t = 9.6 \qquad c_u = 0 \qquad c_o = 23.685$$

Die Nullhypothese wird also für das Produktionsverfahren I beibehalten, die Standardabweichung der Nudellängen sind für dieses Verfahren auf dem Niveau $\alpha = 0.05$ nicht signifikant größer als 5 mm.

Für das Produktionsverfahren II erhalten wir:

$$t = 21.6 \qquad c_u = 0 \qquad c_o = 19.675$$

Hier liegt die Teststatistik oberhalb der oberen Grenze, wodurch die Nullhypothese abgelehnt wird. Die Standardabweichung der Spaghettilängen ist für das Produktionsverfahren II auf dem Niveau $\alpha = 0.05$ signifikant größer als 5 mm.

c) Test der Hypothese $\sigma_{\mathrm{I}} = \sigma_{\mathrm{II}}$ auf dem Niveau $\alpha = 0.05$:

Wir testen wieder die äquivalente Hypothese $\sigma_{\mathrm{I}}^2 = \sigma_{\mathrm{II}}^2$. Die zugehörige Teststatistik (siehe Tabelle 5.11 des Vorlesungsskriptums) ist dabei sehr einfach und folgt unter der Nullhypothese einer F-Verteilung:

$$t = \frac{s_{\mathrm{II}}^2}{s_{\mathrm{I}}^2} = \frac{7^2}{4.14^2} = 2.86$$

Bemerkung: Der einzige Punkt auf den man hier achten muss, damit man mit den im Vorlesungsskriptum enthaltenen Quantiltabellen der F-Verteilung etwas anfangen kann, ist die Anordnung der beiden Stichprobenvarianzen – diese ist so zu wählen, dass die Teststatistik *größer als 1* ist; deswegen steht im obigen Ausdruck s_{II}^2 im Zähler und s_{I}^2 im Nenner und nicht umgekehrt!

Die Teststatistik ist also F-verteilt mit $f_1 = n_{\mathrm{II}} - 1 = 11$ und $f_2 = n_{\mathrm{I}} - 1 = 14$ Freiheitsgraden (Zählerfreiheitsgrade vor Nennerfreiheitsgraden), folgt also einer $F_{11,14}$-Verteilung. Die entsprechenden kritischen Werte sind daher

$$c_u = F_{11,14;\alpha/2} = F_{11,14;0.025} < 1 \qquad c_o = F_{11,14;1-\alpha/2} = F_{11,14;0.975}$$

Die untere Grenze ist also kleiner als 1 und damit auch kleiner als die Teststatistik; von dieser Seite ist also aufgrund unserer Wahl der Teststatistik nichts zu befürchten, weswegen die unteren Quantile der F-Verteilung auch nicht im Vorlesungsskriptum enthalten sind. Für die obere Grenze konsultieren wir Tabelle A.4 des Vorlesungsskriptums und sehen, dass die exakten Freiheitsgrade 11 und 14 nicht verfügbar sind; der gesuchte Wert liegt also zwischen $F_{10,12;0.975}$ und $F_{12,15;0.975}$, oder

$$3.374 > F_{11,14;0.975} > 2.963$$

Man kann jetzt entweder mühsam interpolieren, oder man vergleicht einfach die Teststatistik mit den beiden Zahlen:

$$t = 2.86 < 2.963 < F_{11,14;0.975} = c_o$$

Damit liegt die Teststatistik in jedem Fall im Annahmebereich, und wir behalten die Hypothese gleicher Streuungen bei. Es gibt also keinen signifikanten Unterschied (auf dem Niveau $\alpha = 0.05$) zwischen den Standardabweichungen der beiden Produktionsverfahren.

d) Test der Hypothese $\mu_{\mathrm{I}} = \mu_{\mathrm{II}}$ auf dem Niveau $\alpha = 0.05$:

Es handelt sich um **unabhängige Stichproben**, da jeder Spaghetto nur mittels eines Produktionsverfahrens hergestellt wurde; wir können also die Mittelwerte über einen entsprechenden t-Test vergleichen. In Punkt c) haben wir keinen signifikanten Unterschied zwischen den Streuungen der beiden Produktionsverfahren gefunden, in Abwesenheit von andere Informationen wählen wir also einen t-Test unter Annahme gleicher Varianzen wie in Tabelle 5.6 des Vorlesungsskriptums beschrieben.

Berechnung der gepoolten Standardabweichung:

$$\begin{aligned} s_P &= \sqrt{\frac{(n_\mathrm{I}-1)\cdot s_\mathrm{I}^2+(n_\mathrm{II}-1)\cdot s_\mathrm{II}^2}{n_\mathrm{I}+n_\mathrm{II}-2}} \\ &= \sqrt{\frac{(15-1)\cdot 4.14^2+(12-1)\cdot 7.00^2}{15+12-2}} = 5.582 \end{aligned}$$

Berechnung der Teststatistik:

$$t = \frac{\bar{x}_\mathrm{I}-\bar{x}_\mathrm{II}}{s_P\cdot\sqrt{\frac{1}{n_\mathrm{I}}+\frac{1}{n_\mathrm{II}}}} = \frac{375.47-379.17}{5.582\cdot\sqrt{1/15+1/12}} = -1.711$$

Kritische Werte

$$c_u/c_o = \pm t_{n_\mathrm{I}+n_\mathrm{II}-2,1-\alpha/2} = \pm t_{25;0.975} = \pm 2.06$$

Die Hypothese gleicher Mittelwerte für beide Produktionsverfahren kann somit beibehalten werden.

e) Kommentiere die Fehlerwahrscheinlichkeiten der Entscheidungen für die statistischen Tests:

ad b) Für das Produktionsverfahren I wurde die Nullhypothese $\sigma \leq 5$ beibehalten. Damit können wir hier nur einen Fehler 2. Art begehen (Beibehalten der Nullhypothese, obwohl die Alternative korrekt wäre), dessen Wahrscheinlichkeit wir aber nicht explizit kennen. Für das Produktionsverfahren II wurde dieselbe Hypothese abgelehnt, d.h. wenn diese Entscheidung falsch ist, handelt es sich um einen Fehler 1. Art (Ablehnen einer richtigen Alternativhypothese); diese Fehlerwahrscheinlichkeit ist aber gerade durch das Signifikanzniveau $\alpha = 0.05$ beschränkt, deswegen ist für dieses Produktionsverfahren die Irrtumswahrscheinlichkeit kleiner gleich 0.05.

ad c) Die Hypothese gleicher Streuungen wurde beibehalten, die Irrtumswahrscheinlichkeit für einen Fehler 2. Art ist aber unbekannt.

ad d) Die Hypothese gleicher Mittelwerte wurde beibehalten, die Irrtumswahrscheinlichkeit für einen Fehler 2. Art ist aber unbekannt.

Lösung 3.7
Asbestgehalt der Luft

Für die arithmetischen Mittel und empirischen Standardabweichungen der Asbestgehalte der Monate März und Juli erhalten wir zunächst

$$\begin{aligned} \bar{x}_\text{März} &= 492.38 & s_\text{März} &= 143.04 \\ \bar{x}_\text{Juli} &= 549.05 & s_\text{Juli} &= 129.57 \end{aligned}$$

a) Die 95%-Konfidenzintervalle für Mittelwert und Standardabweichung der Asbestwerte für beide Monate lauten mit den Formeln der Tabellen 5.1 und 5.3 im Vorlesungsskriptum

$$427.27 \leq \mu_{\text{März}} \leq 557.49 \qquad 109.43 \leq \sigma_{\text{März}} \leq 206.57$$
$$490.07 \leq \mu_{\text{Juli}} \leq 608.03 \qquad 99.13 \leq \sigma_{\text{Juli}} \leq 187.12$$

b) Test der Nullhypothese $\mu_{\text{März}} = \mu_{\text{Juli}}$ auf dem Niveau $\alpha = 0.05$:

Hier handelt es sich offensichtlich um eine **verbundene Stichprobe**: 21 Standorte wurden zweimal untersucht, für jeden der Standorte wurde ein März- und ein zugehöriger Juli-Asbestwert gemessen. Für verbundene Stichproben ist nur die Differenz der Werte interessant, d.h. statt dem Wertepaar $(\text{Asbest}_{\text{März}}, \text{Asbest}_{\text{Juli}})$ betrachten wir für jede Station nur einen Wert, nämlich $\text{Asbest}_{\text{März}} - \text{Asbest}_{\text{Juli}}$. Damit haben wir statt eines Zweistichproben- nur noch ein Einstichprobenproblem.

Die äquivalente Hypothese für die Differenzen lautet $\mu_D = 0$, also die Differenz zwischen März- und Juli-Asbestwerten soll im Mittel Null sein. Falls die Differenzen normalverteilt sind, kann man für diese Hypothese einen t-Test verwenden. Dazu benötigen wir aber den Mittelwert und die Standardabweichungen der Differenzen. Der Mittelwert $\bar{d}$ ergibt sich leicht aus den Mittelwerten der Einzelmonate:

$$\bar{d} = \bar{x}_{\text{März}} - \bar{x}_{\text{Juli}} = 492.38 - 549.05 = -56.67$$

Die Standardabweichung lässt sich nicht so leicht ableiten; im Zweifelsfall muss man ganz altmodisch zunächst die Differenzen berechnen

-80	-30	-200	-180	30	-70	-20
-110	-90	-120	-180	130	70	-60
120	-30	200	-170	-110	-170	-120

und aus diesen dann die Standardabweichung:

$$s_D = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{250466.67}{20}} = 111.91$$

Die **Teststatistik** lautet (siehe Tabelle 5.8 Vorlesungsskriptum) sodann

$$t = \frac{\bar{x}_{\text{März}} - \bar{x}_{\text{Juli}}}{s_D/\sqrt{n}} = \frac{\bar{d} - 0}{s_D/\sqrt{n}} = \frac{-56.67}{111.91/\sqrt{21}} = -2.32$$

Die **kritischen Werte** sind gerade die Quantile der t-Verteilung mit $n-1$ Freiheitsgraden:

$$c_u/c_o = \pm t_{n-1,1-\alpha/2} = \pm t_{20;0.975} = \pm 2.086$$

Die Teststatistik liegt also unterhalb der unteren Grenze, und die Hypothese wird abgelehnt: Es gibt einen signifikanten (auf dem Niveau $\alpha = 0.05$) Unterschied zwischen der Asbestbelastung im Monat März und im Monat Juli.

c) 95%-Konfidenzintervall für den mittleren Unterschied zwischen den Asbestgehalten im März und Juli:

Falls die Differenzen normalverteilt sind, ist für verbundene Stichproben das Konfidenzintervall für die mittlere Differenz (Tabelle 5.7 Vorlesungsskriptum) ein ganz normales Konfidenzintervall für den Mittelwert μ_D, also mit Teststatistik

$$\bar{d} \pm t_{n-1;1-\alpha/2} \cdot \frac{s_D}{\sqrt{n}}$$

Alle Terme in diesem Ausdruck wurden schon unter b) berechnet, wir müssen nur noch einsetzen und erhalten:

$$-107.61 \leq \mu_D \leq -5.73$$

d) Test der Nullhypothese $\mu_{\text{März}} \geq \mu_{\text{Juli}}$ auf dem Niveau $\alpha = 0.01$.

Der einseitige Test für eine verbundene Stichprobe hat dieselbe Teststatistik wie der beidseitige Test (Tabelle 5.8 Vorlesungsskriptum), der unter b) gerechnet wurde, also

$$t = \frac{\bar{d}}{s_D/\sqrt{n}} = -2.32$$

Der untere kritische Wert ergibt sich wieder aus einer t-Verteilung mit $n-1$ Freiheitsgraden:

$$c_u = t_{n-1;\alpha} = -t_{n-1;1-\alpha} = -t_{20;0.99} = -2.528$$

Die obere Grenze ist wieder nur formal, $c_o = \infty$, da große Werte der Teststatistik nicht gegen die Nullhypothese sprechen. Die Nullhypothese wird also angenommen, es gibt keinen signifikanten Einwand (auf dem Niveau $\alpha = 0.01$) dagegen, dass die Asbestbelastung im Mittel im März höher ist als im Juli.

Lösung 3.8
Vergleich des Semmelgewichts zweier Fabriken

Für die beiden Stichproben (Semmelgewichte zweier Bäckereien A und B) ergeben sich folgende Kenngrößen:

$$\begin{array}{llll} \bar{x}_{\text{A}} = 49.85 & s_{\text{A}}^2 = 2.4339 & s_{\text{A}} = 1.56 & n_{\text{A}} = 10 \\ \bar{x}_{\text{B}} = 48.91 & s_{\text{B}}^2 = 1.9413 & s_{\text{B}} = 1.39 & n_{\text{B}} = 8 \end{array}$$

a) Test der Hypothese $\sigma_A = \sigma_B$ auf dem Niveau $\alpha = 0.05$:

Die **Teststatistik** lautet gemäß Tabelle 5.11 des Vorlesungsskripts

$$t = \frac{s_{\text{A}}^2}{s_{\text{B}}^2} = \frac{2.4339}{1.9413} = 1.254$$

Bemerkung: Wieder steht die größere Varianz im Zähler, sodass die Teststatistik größer als eins ist.

Kritische Werte aus der F-Verteilung mit $f_1 = n_{\text{A}} - 1 = 9$ und $f_2 = n_{\text{B}} - 1 = 7$ Freiheitsgraden:

$$c_u = F_{9,7;0.025} < 1 \qquad c_o = F_{9,7;0.975} = 4.823$$

Die Teststatistik liegt im Annahmebereich, die Hypothese gleicher Streuungen wird also auf dem Niveau $\alpha = 0.05$ beibehalten.

b) Test der Hypothese $\mu_\text{A} = \mu_\text{B}$ für $\alpha = 0.05$:

Es handelt sich um **unabhängige Stichproben** (eine Semmel kann nur aus der einen oder anderen Bäckerei stammen). Aufgrund unserer Entscheidung in a) wählen wir daher einen t-Test unter Annahme gleicher Standardabweichungen $\sigma_\text{A} = \sigma_\text{B}$ der Semmelgewichte für beide Bäckereien.

Für die **gepoolte Standardabweichung** erhalten wir zunächst

$$s_P = \sqrt{\frac{(n_\text{A}-1)\cdot s_\text{A}^2 + (n_\text{B}-1)\cdot s_\text{B}^2}{n_\text{B}+n_\text{B}-2}} = 1.489$$

und damit für die **Teststatistik**:

$$t = \frac{\bar{x}_\text{A} - \bar{x}_\text{B}}{s_P \cdot \sqrt{\frac{1}{n_\text{A}} + \frac{1}{n_\text{B}}}} = 1.33$$

Die **kritische Werte** lauten

$$c_u/c_o = \pm t_{n_\text{A}+n_\text{B}-2,1-\alpha/2} = \pm t_{16;0.975} = \pm 2.12$$

weshalb die Nullhypothese im Mittel gleicher Semmelgewichte für die beiden Bäckereien A und B auf dem Niveau $\alpha = 0.05$ beibehalten wird.

Lösung 3.9
Honigleistungserhebung

Die Kenngrößen der Honigleistung für die beiden Jahre lauten zunächst

$$\bar{x}_{92} = 18.457 \qquad s_{92} = 4.604$$
$$\bar{x}_{93} = 16.171 \qquad s_{93} = 4.148$$

a) Test der Hypothese $\mu_{92} = \mu_{93}$ auf dem Niveau $\alpha = 0.05$:

Hier handelt es sich um eine **verbundene Stichprobe**, da die Messwerte beider Jahres an den selben 35 Bienenvölkern erhoben wurden. Wir betrachten also im folgenden nur noch die **Differenzen** $d_i = x_i^{(92)} - x_i^{(93)}$:

4	-1	3	2	-1	4	5
2	2	2	-1	2	4	1
3	2	3	0	3	3	2
8	-1	2	-2	-3	4	10
5	1	3	2	4	-2	5

Die zugehörigen Kennzahlen der Differenzen sind Mittelwert

$$\bar{d} = \bar{x}_{92} - \bar{x}_{93} = 2.286$$

und Standardabweichung

$$s_D = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = 2.707$$

Da wir laut Angabe von normalverteilten Differenzen ansehen können, verwenden wir einen t-Test für die Hypothese $\mu_D = 0$ (siehe Tabelle 5.8 im Vorlesungsskriptum).

Die **Teststatistik** lautet

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_D/\sqrt{n}} = 4.991$$

Für die **kritischen Werte** ergibt sich

$$c_u/c_o = \pm t_{n-1;1-\alpha/2} = \pm t_{34;0.975} \simeq \pm t_{30;0.975} = \pm 2.042$$

Wieder wird das nächste tabellierte Quantil mit weniger Freiheitsgraden verwendet. Die Teststatistik liegt außerhalb des Intervalls $[-2.042; +2.042]$, wir lehnen daher H_0 ab: Es gibt einen auf dem Niveau $\alpha = 0.05$ signifikanten Unterschied zwischen dem mittleren Ertrag der Jahres 1992 und 1993.

b) Konfidenzintervall für die mittlere Differenz μ_D mit Überdeckungswahrscheinlichkeit 0.95:

Ein ganz normales Konfidenzintervall für den Mittelwert einer Normalverteilung, hier für die Differenzen der beiden Jahre; Die Tabellen 5.1 und 5.7 des Vorlesungsskriptums liefern Formel

$$\bar{d} \pm t_{n-1;1-\alpha/2} \cdot \frac{s_D}{\sqrt{n}}$$

Mit dem passenden Quantil $t_{34;0.975} \simeq t_{30;0.975} = 2.042$ (wir nehmen also das nächstgelegene tabellierte mit weniger df) ergibt sich daher das Konfidenzintervall für $\mu_D = \mu_1 - \mu_2$:

$$1.35 \leq \mu_D \leq 3.22$$

c) Test der Hypothese $\mu_{92} \leq \mu_{93}$ auf dem Niveau $\alpha = 0.01$:

Die **Teststatistik** ist identisch zu der für die beidseitige Hypothese in a), also

$$t = 4.991$$

Kleine Werte der Teststatistik sprechen für die Nullhypothese, große Werte dagegen. Entsprechend sind wir hier an einer oberen Grenze für die Teststatistik interessiert:

$$c_o = t_{n-1;1-\alpha} = t_{34,0.99} \simeq t_{30,0.99} = 2.457$$

Die untere Grenze $c_u = -\infty$ ist hier nur formal (Tabelle 5.8 Vorlesungsskript). Für die obere Grenze wird wieder das nächstgelegene tabellierte Quantil mit weniger df verwendet. Entsprechend wird die Hypothese, dass der Ertrag im Jahr 92 kleiner als im Jahr 93 war, verworfen: Der Ertrag des Jahres 1992 war somit auf dem Niveau $\alpha = 0.01$ signifikant größer als der des Jahres 1993.

d) Kommentiere Fehlerwahrscheinlichkeiten der Entscheidungen in a) und c):

In beiden Punkten wurde die Nullhypothese verworfen. Der einzig mögliche Fehler ist daher ein Fehler 1. Art (Verwerfen einer richtigen Nullhypothese); die Wahrscheinlichkeit für einen solchen Fehler ist aber gerade durch das jeweils gewählte α beschränkt, d.h. die Fehlerwahrscheinlichkeit liegt unter 0.05 in a) und unter 0.01 in c).

Kapitel 4

Varianzanalyse

Beispiele

Einfache Varianzanalyse

Beispiel 4.1 Ein und dieselbe Käsesorte wird von vier verschiedenen Molkereien A bis D angeboten. Aus jeder Produktion werden einige Käseziegel als Stichprobe entnommen, und das Gewicht von Normziegeln (in g) ermittelt.

Molkerei A	997.1	991.1	998.5	1007.1	1001.7	1000.9
	997.8	985.5	1007.6	998.3		
Molkerei B	976.1	992.9	964.6	980.3	988.1	980.3
	967.2	992.2	973.9	985.3	984.5	962.4
Molkerei C	1000.5	997.9	998.3	996.7	1000.9	1005.3
	994.7	997.8	1000.2			
Molkerei D	990.3	987.6	989.2	993.5	990.9	992.1
	991.4	990.7	988.6	994.1	989.3	

Prüfen Sie die Frage, ob die Molkereien hinsichtlich des Durchschnittsgewichtes gleich produzieren (Signifikanzniveau $\alpha = 0.05$). Formulieren Sie Hypothese und Entscheidung als ganzen Satz.

Beispiel 4.2 Bei der Rinderschlachtleistungsprüfung wird unter anderem auch die Nettozunahme pro Tag (Schlachtkörpergewicht/Mastdauer) festgestellt. Nachfolgend ist die Nettozunahme in g von fünf verschiedenen Rinderrassen in achtfacher Wiederholung dargestellt.

Fleckvieh	784	747	831	780	810	824	844	790
Grauvieh	705	735	768	750	781	715	721	733
Blondvieh	734	756	706	780	768	731	741	763
Pinzgauer	809	780	830	773	798	805	802	815
Schwarzbunt	641	728	654	678	708	730	692	675

Überprüfen Sie, ob es zwischen den Rinderrassen Unterschiede bezüglich der Nettozunahme gibt (Signifikanzniveau $\alpha = 0.05$).

Zweifache Varianzanalyse ohne Wechselwirkungen

Beispiel 4.3 Bei der Silageeignungsprüfung von Grünschnittroggen wird unter anderem auch der Sorteneinfluss auf den Rohproteinertrag untersucht. In der nachfolgenden Tabelle ist der Rohproteinertrag von 5 Sorten in 7 Prüffeldern (Pf) in dt/ha angegeben.

	Pf 1	Pf 2	Pf 3	Pf 4	Pf 5	Pf 6	Pf 7
Sorte 1	6.3	6.7	6.1	6.9	6.6	6.0	7.1
Sorte 2	7.3	7.5	6.9	8.4	7.3	7.5	7.9
Sorte 3	6.3	6.4	5.8	6.7	6.4	6.1	6.3
Sorte 4	7.2	7.4	7.8	7.2	7.0	6.8	7.1
Sorte 5	6.2	6.4	6.8	6.2	6.3	6.0	6.2

Überprüfen Sie, ob sich der Rohproteinertrag der 5 Sorten unterscheidet, wenn der Rohproteinertrag als annähernd normalverteilt angenommen werden kann, und auch die Varianzen der einzelnen Sorten übereinstimmen ($\alpha = 0.05$).

Zweifache Varianzanalyse mit Wechselwirkungen

Beispiel 4.4 Ein Biobauer will die Wirkung von Steinmehl in der Gülle überprüfen. Er untersucht dabei in vierfacher Wiederholung die Wirkung von Gülle mit und ohne Steinmehl auf den Ertrag von drei Winterroggensorten. Der Ertrag ist in dt/ha dargestellt.

	Gülle mit Steinmehl				Gülle ohne Steinmehl			
Eho Kurz	42.0	45.2	46.3	44.7	41.6	43.7	41.5	40.8
Motto	39.9	42.4	41.8	40.4	42.6	43.7	41.8	42.4
Kustro	39.5	41.3	39.6	41.9	37.4	39.0	36.2	37.4

Testen Sie auf dem 5%-Signifikanzniveau, ob es Unterschiede zwischen den Winterroggensorten bzw. den Güllevarianten gibt, und ob Wechselwirkungen bestehen. Nehmen Sie an, dass die Daten normalverteilt und die Varianzen homogen sind.

Beispiel 4.5 Im Rahmen eines Forschungsprojektes über nachwachsende Rohstoffe wird die Ertragsleistung von Chinaschilf (*Miscanthus sinensis Giganteus*) bei drei unterschiedlichen Stickstoff- (0 – 60 – 120 kg N/ha) und zwei verschiedenen Kalivarianten (70 – 140 kg K_2O/ha) untersucht. Der Versuch wird mit vier Wiederholungen vollständig randomisiert angelegt. Der Trockensubstanzertrag ist in t/ha angegeben.

	70 kg K_2O				140 kg K_2O			
0 kg N/ha	18.3	17.8	16.8	17.1	15.4	16.9	16.4	17.8
60 kg N/ha	19.7	21.2	21.6	21.1	21.9	21.6	22.8	21.3
120 kg N/ha	20.9	19.8	21.9	21.3	21.6	21.9	22.9	22.5

Auf dem 5%-Signifikanzniveau soll getestet werden, ob die Unterschiede zwischen den Stickstoffdüngerstufen und zwischen den Kalidüngerstufen signifikant sind, und ob Wechselwirkungen zwischen den Stickstoff- und Kalidüngerstufen bestehen.

Gemischte Beispiele

Beispiel 4.6 Bei Untersuchungen über das Wanderverhalten von zwei verschiedenen Blattlausarten wird u. a. die Aufenthaltsdauer (in min) vor der Wanderphase auf jeweils 15 Blättern (Bl) der Kartoffelpflanze *Solanum tuberosum* untersucht. Da man weiß, dass die Aufenthaltsdauer der Blattläuse vor allem vom Alter der Blätter abhängt, werden die zwei Blattlausarten jeweils gleichzeitig und zufällig auf ein Blatt (jede auf eine Blatthälfte) gesetzt.

	Bl 1	Bl 2	Bl 3	Bl 4	Bl 5	Bl 6	Bl 7	Bl 8
Blattlausart A	280	289	287	246	268	308	338	285
Blattlausart B	260	209	196	213	193	210	240	166
	Bl 9	Bl 10	Bl 11	Bl 12	Bl 13	Bl 14	Bl 15	
Blattlausart A	269	246	317	276	329	265	356	
Blattlausart B	309	189	316	172	242	153	327	

Überprüfen Sie, ob sich die Aufenthaltsdauer der beiden Blattlausarten auf den Blättern unterscheidet, wenn die Aufenthaltsdauer als annähernd normalverteilt angenommen werden kann und auch die Varianzen der Blattlausarten übereinstimmen ($\alpha = 0.05$).

Beispiel 4.7 Bei Untersuchungen über die Krankheitsanfälligkeit der Ackerbohne (*Vicia faba*) wird bei drei neuen Zuchtlinien die Virulenz von zwei Pseudomonasrassen (*Pseudomonas phaseolicola*) untersucht. Dabei werden im 3- bis 4-Blattstadium acht Pflanzen pro Zuchtlinie zufällig ausgewählt und jeweils vier Pflanzen mit einer Pseudomonasrasse geimpft. Nach zehn Tagen wird das Ausmaß der Chlorose gemessen (in cm^2).

	Rasse A				Rasse B			
Zuchtlinie I	2.12	1.67	1.23	0.91	3.54	3.76	4.09	2.82
Zuchtlinie II	2.11	2.62	1.84	2.78	2.48	2.31	2.24	2.51
Zuchtlinie III	1.36	1.92	1.75	2.43	3.65	4.11	3.89	2.59

Testen Sie auf dem 5%-Signifikanzniveau, ob die Unterschiede zwischen den Zuchtlinien und zwischen den Rassen signifikant sind, und ob Wechselwirkungen zwischen den Zuchtlinien und Rassen bestehen.

Lösungen

Lösung 4.1
Gewicht von Normziegeln bei der Käseherstellung

Es handelt sich um eine **einfache Varianzanalyse**, zumal *eine* Gruppierungsvariable (Molkerei) mit $I = 4$ Stufen (A bis D) vorliegt. Die abhängige quantitative Variable ist das Gewicht der Käseziegel.

a) **Hypothese**: Die Nullhypothese der einfachen Varianzanalyse lautet, dass die Gruppenerwartungswerte gleich sind, d.h. also, dass die Gruppierungsvariable keinen Einfluss auf die abhängige Größe hat. Für das vorliegende Beispiel lautet die Nullhypothese also:

> Es besteht kein Unterschied zwischen dem theoretischen mittleren Ziegelgewicht der Molkereien.

b) **ANOVA-Tabelle**:

Zunächst berechnet man die $I = 4$ Gruppenmittel sowie das Gesamtmittel:

$$\begin{aligned}
\bar{y}_{\mathrm{A}} &= \bar{y}_{1.} = 9985.6/10 = 998.56 \\
\bar{y}_{\mathrm{B}} &= \bar{y}_{2.} = 11747.8/12 \simeq 978.98 \\
\bar{y}_{\mathrm{C}} &= \bar{y}_{3.} = 8992.3/9 \simeq 999.14 \\
\bar{y}_{\mathrm{D}} &= \bar{y}_{4.} = 10897.7/11 = 990.70 \\
\bar{y}_{\mathrm{Gesamt}} &= \bar{y}_{..} = (9985.6 + 11747.8 + 8992.3 + 10897.7)/42 \simeq 991.03
\end{aligned}$$

Bemerkung: Man beachte, dass das Gesamtmittel $\bar{y}_{..}$ im allgemeinen **nicht** mit dem Mittel der einzelnen Gruppenmittel ident ist. Dies gilt nur im Falle gleich großer Stichproben in den einzelnen Gruppen.

Dann berechnet man die Summe $\mathrm{SS_T}$ der quadrierten Abweichungen jeder Beobachtung y_{ij} vom Gesamtmittelwert $\bar{y}_{..}$:

$$\begin{aligned}
\mathrm{SS_T} &= \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{..})^2 = \\
&= (997.1 - 991.03)^2 + (991.1 - 991.03)^2 + \ldots + (994.1 - 991.03)^2 + (989.3 - 991.03)^2 = \\
&= 4595.41
\end{aligned}$$

Man beachte, dass diese Quadratsumme bis auf den Faktor $1/(n-1)$ der Stichprobenvarianz entspricht; SS_T kann also leicht aus der Varianz bzw. Standardabweichung der Gewichte berechnet werden.

Der Anteil an $\mathrm{SS_T}$, der durch den Faktor *Molkerei* erklärt wird, ist durch die gewichtete Summe der quadrierten Abweichungen der Gruppenmittelwerte vom Gesamtmittelwert gegeben:

$$\begin{aligned}
\mathrm{SS_{Molkerei}} &= \sum_{i=1}^{I} J_i \cdot (\bar{y}_{i.} - \bar{y}_{..})^2 = \\
&= 10 \cdot (998.56 - 991.03)^2 + 12 \cdot (978.98 - 991.03)^2 + \\
&+ 9 \cdot (999.14 - 991.03)^2 + 11 \cdot (990.7 - 991.03)^2 = \\
&= 2902.27
\end{aligned}$$

Die Fehlerquadratsumme $\mathrm{SS_E}$ (Fehler bzw. Error) ist die Restvariation der Daten, die nicht durch die Molkereien erklärt werden kann. Sie kann praktischerweise als Differenz von $\mathrm{SS_T}$ und $\mathrm{SS_{Molkerei}}$ berechnet werden:

$$\mathrm{SS_E} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_{i.})^2 = \mathrm{SS_T} - \mathrm{SS_{Molkerei}} = 1693.14$$

Damit sind alle Zutaten für die ANOVA-Tabelle vorbereitet:

	SS	d.f.	MS	F	p
Molkerei	2902.27	3	967.42	21.71	0.0000
Fehler	1693.14	38	44.56		
Total	4595.41	41			

Die Spalte SS (Sum of Squares) enthält gerade die soeben berechneten Quadratsummen, die Spalte d.f. (degrees of freedom) die zugehörigen **Freiheitsgrade**: für die Gesamtquadratsumme SS_T gerade $\sum_{i=1}^{I} J_i - 1 = n - 1$ (wie für die Stichprobenvarianz), für den Gruppierungsfaktor *Molkerei* die Anzahl an Faktorstufen minus 1, $I - 1$, und für den Fehler wieder die Differenz aus beiden, d.h. $n - 1 - (I - 1) = n - I$. Die MS (**mean squares**) ergeben sich dann als SS/d.f., und die F-Statistik als Quotient der MS:

$$F = \frac{\mathrm{MS}_{\mathrm{Molkerei}}}{\mathrm{MS}_{\mathrm{Fehler}}}$$

Der in dieser Tabelle enthaltene p-Wert kann mit den in der Vorlesung beschriebenen Methoden nicht händisch berechnet werden und ist hier nur ergänzend angeführt, um im weiteren die Interpretation von Computerausdrucken zu üben!

c) **Entscheidung**: Für “zu Fuß” gerechnete Beispiele wird die in der ANOVA-Tabelle berechnete F-Statistik mit dem **kritischen Wert** der F-Verteilung verglichen; die Freiheitsgrade der F-Verteilung sind gerade die Freiheitsgrade der beteiligten MS, hier also ist der kritische Wert

$$F_{I-1,n-I;1-\alpha} = F_{3,38;0.95}$$

Dieser Wert kann aus Tabelle A.4 des VL-Skriptums durch Interpolation näherungsweise berechnet werden und stellt eine obere kritische Grenze dar: liegt die F-Statistik darüber, so wird die Nullhypothese verworfen, andernfalls beibehalten. Einfacher als Interpolieren ist allerdings Abschätzen aus der Tabelle; dazu muss man nur wissen, dass die Quantile der F-Verteilung mit wachsendem Freiheitsgrad immer kleiner werden, es gilt also hier:

$$F_{3,38;0.95} < F_{3,30;0.95} = 2.922 < F_{\mathrm{Molkerei}} = 21.71$$

Damit wird H_0 abgelehnt: Es besteht also ein signifikanter Unterschied bezüglich des mittleren Ziegelgewichts zwischen den Molkereien, oder anders gesagt, der Faktor *Molkerei* hat einen signifikanten Einfluss auf das Ziegelgewicht.

Falls wie oben in der Tabelle der p-Wert angegeben ist, kann man sich das Ablesen von kritischen Werten überhaupt sparen. Generell gilt für alle Testverfahren, dass die Nullhypothese (hier: die Gleichheit aller theoretischen Gruppenmittelwerte) verworfen wird, wenn der p-Wert kleiner als das gewählte Signifikanzniveau ist. Hier wird der p-Wert als $0.0000 < 0.05 = \alpha$ angegeben, die Nullhypothese wird also wie oben abgelehnt.

Lösung 4.2
Nettozunahme von 5 Rinderrassen

In diesem Beispiel haben wir als qualitative Variable die *Rinderrasse* mit $I = 5$ Stufen (Ausprägungen) und $J_i = J = 8$ Beobachtungen der quantitativen Variable *Gewichtszunahme* in jeder Faktorstufe (und daher insgesamt $n = \sum_{i=1}^{I} J_i = I \cdot J = 5 \cdot 8 = 40$ Beobachtungen).

a) **Hypothese**: Die Rinderrasse hat keinen Einfluss auf die Tageszunahmen bzw. die Rinderrassen unterscheiden sich nur zufällig bezüglich ihrer Tageszunahmen.

b) **ANOVA-Tabelle**

Für die **Gruppenmittelwerte** der Rinderrassen (Abkürzungen **F**leckvieh, **G**rauvieh, **B**londvieh, **P**inzgauer, **S**chwarzbunt):

$$\begin{aligned}\bar{y}_\mathrm{F} &= \bar{y}_{1.} = 801.250\\ \bar{y}_\mathrm{G} &= \bar{y}_{2.} = 738.500\\ \bar{y}_\mathrm{B} &= \bar{y}_{3.} = 747.375\\ \bar{y}_\mathrm{P} &= \bar{y}_{4.} = 801.500\\ \bar{y}_\mathrm{S} &= \bar{y}_{5.} = 688.250\end{aligned}$$

Aufgrund der gleichen Beobachtungszahlen in den 5 Gruppen kann das **Gesamtmittel** einfach als Mittelwert der obigen Gruppenmittel berechnet werden:

$$\bar{y}_{..} = (801.25 + 738.5 + 747.375 + 801.5 + 688.25)/5 = 755.375$$

Für die Gesamtquadratsumme $\mathrm{SS_T}$ erhält man damit

$$\begin{aligned}\mathrm{SS_T} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{..})^2 =\\ &= (n-1)s^2 =\\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}^2 - n(\bar{y}_{..})^2\\ &= 22922077 - 40 \cdot (755.375)^2 = 98421\end{aligned}$$

Der durch die Rinderrasse erklärte Anteil der Variation lautet:

$$\begin{aligned}\mathrm{SS_{Rinderrasse}} &= \sum_{i=1}^{I} J_i(\bar{y}_{i.} - \bar{y}_{..})^2 = J \cdot \sum_{i=1}^{I}(\bar{y}_{i.} - \bar{y}_{..})^2 =\\ &= 8 \cdot \big[(801.25 - 755.375)^2 + (738.5 - 755.375)^2 + (747.375 - 755.375)^2 +\\ &+ (801.5 - 755.375)^2 + (688.25 - 755.375)^2\big] = 72692\end{aligned}$$

Die **Fehlerquadratsumme** ergibt sich wiederum als Differenz:

$$\mathrm{SS_E} = \mathrm{SS_T} - \mathrm{SS_{Rinderrasse}} = 25729$$

Damit können wir die **ANOVA-Tabelle** konstruieren:

	SS	d.f.	MS	F	p
Rinderrasse	72692	4	18173.1	24.7	$8.9 \cdot 10^{-10}$
Fehler	25729	35	735.1		
Total	98421	39			

c) **Entscheidung**: die kritische (obere) Grenze c_o ist durch das Quantil $F_{4,35;0.95}$ gegeben, das allerdings nicht in der Tabelle A.4 des Vorlesungsskripts abgelesen werden kann. Eine Abschätzung (F-Quantile werden mit wachsenden Freiheitsgraden kleiner) ergibt

$$c_o = F_{4,35;0.95} < F_{4,30;0.95} = 2.69 < 24.7 = F_{\text{Rinderrasse}}$$

Die F-Statistik ist also deutlich größer als die kritische Grenze, daher wird die **Nullhypothese verworfen**: Es gibt also einen signifikanten (auf dem Niveau $\alpha = 0.05$) Einfluss der Rinderrasse auf die Tageszunahmen. Zum selben Ergebnis kommt man durch Betrachtung des p-Werts $p = 8.9 \cdot 10^{-10} < 0.05 = \alpha$.

Lösung 4.3
Sorteneinfluß auf Rohproteinertrag

Es handelt sich um eine zweifaktorielle Varianzanalyse – Rohproteinertrag von Grünschnittroggen (quantitative Variable) für $I = 5$ verschiedene Roggensorten (erste qualitative Variable mit 5 Faktorstufen) und auf $J = 7$ verschiedenen Prüffeldern (zwete qualitative Variable). Es können keine Wechselwirkungen der beiden Faktoren betrachtet werden, da jede Kombination von Sorte und Prüffeld nur einmal beobachtet wurde.

a) Wir haben für jeden der beiden Faktoren eine Null- bzw. Alternativhypothese:

- H_{Sorte}: Die Roggensorten unterscheiden sich nicht bezüglich ihres mittleren Rohproteingehalts.
- H_{Feld}: Die Prüffelder unterscheiden sich nicht bezüglich des mittleren auf ihnen zu erzielenden Rohproteingehalts.

Diese Hypothesen können in einem Aufwaschen (in einer ANOVA-Tabelle) getestet werden.

b) **ANOVA-Tabelle**: Zuerst werden wieder die Gruppenmittelwerte berechnet, und zwar sowohl für die Roggensorten als auch die Prüffelder; beginnend mit Sorte 1 also etwa

$$\bar{y}_{\text{Sorte }1} = \bar{y}_{1.} = (6.3 + 6.7 + 6.1 + 6.9 + 6.6 + 6.0 + 7.1)/7 = 6.529,$$

ebenso für Sorte 2, 3 etc., und genau so für die verschiedenen Prüffelder:

$$\bar{x}_{\text{Feld }1} = \bar{y}_{.1} = (6.3 + 7.3 + 6.3 + 7.2 + 6.2)/5 = 6.66$$

und so fort für alle sieben Felder. Tatsächlich sind dies, wenn wir die Daten wie in der Angabe rechteckig anordnen, gerade die Zeilen- und Spaltenmittelwerte:

	Pf 1	Pf 2	Pf 3	Pf 4	Pf 5	Pf 6	Pf 7	$\bar{y}_{i.}$
Sorte 1	6.3	6.7	6.1	6.9	6.6	6.0	7.1	6.529
Sorte 2	7.3	7.5	6.9	8.4	7.3	7.5	7.9	7.543
Sorte 3	6.3	6.4	5.8	6.7	6.4	6.1	6.3	6.286
Sorte 4	7.2	7.4	7.8	7.2	7.0	6.8	7.1	7.214
Sorte 5	6.2	6.4	6.8	6.2	6.3	6.0	6.2	6.300
$\bar{y}_{.j}$	6.66	6.88	6.68	7.08	6.72	6.48	6.92	

Der Gesamtmittelwert $\overline{y}_{..}$ lässt sich also auf zweierlei Weise berechnen:

$$\begin{aligned}\bar{y}_{..} &= (6.529 + 7.543 + 6.286 + 7.214 + 6.3)/5 \\ &= (6.66 + 6.88 + 6.68 + 7.08 + 6.72 + 6.48 + 6.92)/7 \\ &= 6.774\end{aligned}$$

wobei die erste Variante arbeitssparender ist, der Vergleich mit der zweiten aber erlaubt zu kontrollieren, wie weit man sich bisher verrechnet hat (oder auch nicht).

Die Gesamtquadratsumme $\mathrm{SS_T}$ hat zwar eine kompliziertere Formel (weil wir über zwei Faktorindices summieren müssen), ist aber im Wesen immer noch die Stichprobenvarianz ohne den Faktor $1/(n-1)$:

$$\mathrm{SS_T} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{..})^2 = (n-1)s^2 = 34 \cdot 0.3831 = 13.03$$

Für Roggensorte und Prüffeld wird nun je eine Quadratsumme berechnet, die den entsprechenden Anteil des jeweiligen Faktors an der Gesamtquadratsumme wiedergibt. Für den ersten Faktor (Sorte) also etwa

$$\begin{aligned}\mathrm{SS_{Sorte}} &= \sum_{i=1}^{I} J(\bar{y}_{i.} - \bar{y}_{..})^2 = \\ &= 7 \cdot (6.529 - 6.774)^2 + \ldots + 7 \cdot (6.3 - 6.774)^2 = \\ &= 9.15\end{aligned}$$

Analog für den zweiten Faktor (Felder):

$$\begin{aligned}\mathrm{SS_{Felder}} &= \sum_{j=1}^{J} I(\bar{y}_{.j} - \bar{y}_{..})^2 = \\ &= 5 \cdot (6.66 - 6.774)^2 + \ldots + 5 \cdot (6.92 - 6.774)^2 = \\ &= 1.18\end{aligned}$$

Bemerkung: Sowohl die Gesamtquadratsumme als auch die beiden Faktorenquadratsummen werden ganz genau gleich berechnet wie für zwei einfache Varianzanalysen mit Roggensorten und Prüffeldern als Faktoren.

Die **Fehlerquadratsumme** $\mathrm{SS_E}$ ergibt sich wieder als die Gesamtquadratsumme minus den anderen Quadratsummen, also

$$\begin{aligned}\mathrm{SS_{Error}} &= \mathrm{SS_T} - \mathrm{SS_{Sorte}} - \mathrm{SS_{Felder}} = \\ &= 13.03 - 9.15 - 1.18 = 2.70\end{aligned}$$

Die mühsam berechneten Quadratsummen lassen sich nun wie bei der einfachen Varianzanalyse leicht und gefällig zu einer **ANOVA-Tabelle** anordnen, mit dem einzigen Unterschied, dass wir für jeden der beiden Faktoren eine Zeile haben:

	SS	d.f.	MS	F	p
Sorte	9.15	4	2.2875	20.33	$1.9 \cdot 10^{-7}$
Felder	1.18	6	0.1967	1.75	0.15
Fehler	2.70	24	0.1125		
Total	13.03	34			

Spalte 1 (SS) enthält die obigen Quadratsummen, Spalte 2 die Freiheitsgrade d.f., Spalte 3 die mittleren Quadratsummen (MS), bei denen die Quadratsummen durch die Freiheitsgrade dividiert werden, und die Spalten 4 und 5 enthalten die entsprechenden F-Statistiken und p-Werte:

$$F_{\text{Sorte}} = \frac{\text{MS}_{\text{Sorte}}}{\text{MS}_{\text{Error}}} = \frac{2.2875}{0.1125} = 20.33$$

$$F_{\text{Felder}} = \frac{\text{MS}_{\text{Felder}}}{\text{MS}_{\text{Error}}} = \frac{0.1967}{0.1125} = 1.75$$

c) **Entscheidung**: Die **kritischen Werte** für beide F-Statistiken ergeben sich wieder als Quantile der F-Verteilung, wobei die Freiheitsgrade direkt aus der ANOVA-Tabelle abgelesen werden können. Die oberen kritischen Grenzen sind also

$$c_{\text{Sorte}} = F_{4,24;0.95} \qquad \text{und} \qquad c_{\text{Felder}} = F_{6,24;0.95}$$

Beide Quantile finden sich nicht in der entsprechenden Tabelle A.4 des Vorlesungsskriptums, können aber auf bewährte Weise abgeschätzt werden (siehe Beispiel 4.2):

$$c_{\text{Sorte}} = F_{4,24;0.95} < F_{4,20;0.95} = 2.866 < 20.33 = F_{\text{Sorte}} \quad \Rightarrow \quad H_{\text{Sorte}} \text{ ablehnen}$$
$$c_{\text{Felder}} = F_{6,24;0.95} > F_{6,30;0.95} = 2.421 > 1.75 = F_{\text{Felder}} \quad \Rightarrow \quad H_{\text{Felder}} \text{ annehmen}$$

Die Roggensorte hat also auf dem gewählten Niveau $\alpha = 0.05$ einen signifikanten Einfluss auf den Rohproteinertrag, die Prüffelder jedoch nicht.

Das gleiche Ergebnis lässt sich an Hand der p-Werte ablesen:

$$p_{\text{sorte}} = 1.9 \cdot 10^{-7} < \alpha = 0.05 \qquad \text{bzw.} \qquad p_{\text{Felder}} = 0.15 > \alpha = 0.05$$

Lösung 4.4
Wirkung von Steinmehl auf Roggenertrag

Zweifaktorielle Varianzanalyse mit dem Ertrag als quantitative Variable in Abhängigkeit von Getreidesorte ($I = 3$ Faktorstufen – *Eho Kurz*, *Motto* und *Kustro*) und Düngungsvariante ($J = 2$ Faktorstufen – *mit/ohne Steinmehl*).

a) **Hypothesen**: Wir haben hier drei Hypothesen, entsprechend den möglichen Haupt- und Wechselwirkungen:

- H_{Sorte}: Die Sorte hat keinen Einfluss auf den (theoretischen) mittleren Ertrag.
- $H_{\text{Steinmehl}}$: Die Beigabe von Steinmehl zur Gülle hat keinen Einfluss auf den theoretischen mittleren Ertrag.

- $H_{\text{Sorte:Steinmehl}}$: Es gibt keine Wechselwirkungen zwischen Sorte und Steinmehl bezüglich des Ertrags. (Oder anschaulicher: Die Beigabe von Steinmehl hat für alle Sorten den gleichen Effekt auf den theoretischen mittleren Ertrag.)

b) **ANOVA-Tabelle**:

Zunächst werden die **Gruppenmittelwerte** für alle Faktorkombinationen berechnet:

$$\begin{array}{lllll lllll}
\bar{y}_{\text{Eho, mit SM}} & = & \bar{y}_{11.} & = & 44.550 & \bar{y}_{\text{Eho, ohne SM}} & = & \bar{y}_{12.} & = & 41.900 \\
\bar{y}_{\text{Motto, mit SM}} & = & \bar{y}_{21.} & = & 41.125 & \bar{y}_{\text{Motto, ohne SM}} & = & \bar{y}_{22.} & = & 42.625 \\
\bar{y}_{\text{Kustro, mit SM}} & = & \bar{y}_{31.} & = & 40.575 & \bar{y}_{\text{Kustro, ohne SM}} & = & \bar{y}_{32.} & = & 37.500
\end{array}$$

Aus diesen können dann die **Mittelwerte** für die **Sorten** und die **Güllevarianten** sowie der **Gesamtmittelwert** berechnet werden:

$$\begin{array}{lllll lllll}
\bar{y}_{\text{Eho}} & = & \bar{y}_{1..} & = & 43.225 & \bar{y}_{\text{Motto}} & = & \bar{y}_{2..} & = & 41.875 \\
\bar{y}_{\text{Kustro}} & = & \bar{y}_{3..} & = & 39.038 & \bar{y}_{\text{mit SM}} & = & \bar{y}_{.1.} & = & 42.083 \\
\bar{y}_{\text{ohne SM}} & = & \bar{y}_{.2.} & = & 40.675 & \bar{y}_{\text{Gesamt}} & = & \bar{y}_{...} & = & 41.379
\end{array}$$

Die **Gesamtquadratsumme** SS_T ergibt sich wieder bis auf den Faktor $1/(n-1)$ als Stichprobenvarianz aller Erträge:

$$\text{SS}_\text{T} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{...})^2 = 139.60$$

Die Quadratsummen für die **Hauptwirkungen von Sorte und Steinmehl** sind wieder die Summen der quadrierten Abweichungen von den jeweiligen Mittelwerten, gewichtet mit der Anzahl von Beobachtungen:

$$\begin{aligned}
\text{SS}_\text{Sorte} &= \sum_{i=1}^{I} J\cdot K(\bar{y}_{i..} - \bar{y}_{...})^2 \\
&= 2\cdot 4\cdot 1.846^2 + 2\cdot 4\cdot 0.496^2 + 2\cdot 4\cdot(-2.341)^2 \\
&= 73.09 \\
\text{SS}_\text{Steinmehl} &= \sum_{j=1}^{J} I\cdot K(\bar{y}_{.j.} - \bar{y}_{...})^2 \\
&= 3\cdot 4\cdot 0.704^2 + 3\cdot 4\cdot(-0.704)^2 \\
&= 11.90
\end{aligned}$$

Die **Quadratsumme** für die **Wechselwirkungen** ist etwas aufwändiger zu berechnen, folgt aber dem gleichen Schema wie die Quadratsummen für die Haupteffekte:

$$\begin{aligned}
\text{SS}_\text{Sorte:Steinmehl} &= \sum_{i=1}^{I}\sum_{j=1}^{J} K(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 = \\
&= 4\cdot(44.550 - 43.225 - 42.083 + 41.379)^2 + \ldots + \\
&+ 4\cdot(37.500 - 39.083 - 40.675 + 41.379)^2 = \\
&= 4\cdot(0.621)^2 + \ldots + 4\cdot(-0.879)^2 = 25.56
\end{aligned}$$

Die **Fehlerquadratsumme** $\mathrm{SS_E}$ kann wieder wie bei der einfachen Varianzanalyse durch Subtraktion berechnet werden:

$$\mathrm{SS_E} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{ij.})^2 =$$
$$= \mathrm{SS_T} - \mathrm{SS_{Sorte}} - \mathrm{SS_{Steinmehl}} - \mathrm{SS_{Sorte:Steinmehl}} = 29.05$$

Damit haben wir alle Quadratsummen und können die **ANOVA-Tabelle** erstellen:

	SS	d.f.	MS	F	p
Sorte	73.09	2	36.55	22.64	0.000
Steinmehl	11.90	1	11.90	7.37	0.014
Sorte:Steinmehl	25.56	2	12.78	7.92	0.003
Fehler	29.05	18	1.61		
Total	139.60	23			

Die **Freiheitsgrade** ergeben sich als Anzahl der Faktorstufen minus 1 für die Hauptwirkungen, und als deren Produkt für die Wechselwirkungen; die Gesamtfreiheitsgrade sind Anzahl aller Beobachtungen minus 1, und die Fehlerfreiheitsgrade ergeben sich wieder als Differenz. Die **MS** ergeben sich als SS durch die jeweiligen Freiheitsgrade, und die F-Statistiken als die jeweilige MS durch die Fehler-MS. Die p-Werte sind wiederum nur der Vollständigkeit halber angeführt und nicht händisch zu berechnen!

c) **Entscheidungen**: Wir können hier wieder die **kritischen Grenzen** für die jeweiligen F-Statistiken aus den Tabellen interpolieren, oder auch hier einfacher, abschätzen:

$$\begin{aligned}
&F_{2,18;0.95} < F_{2,15;0.95} = 3.682 < 22.64 = F_{\text{Sorte}} &&\Rightarrow\ H_{\text{Sorte}} \text{ ablehnen}\\
&F_{1,18;0.95} < F_{1,15;0.95} = 4.543 < 7.37 = F_{\text{Steinmehl}} &&\Rightarrow\ H_{\text{Steinmehl}} \text{ ablehnen}\\
&F_{2,18;0.95} < F_{2,15;0.95} = 3.682 < 7.92 = F_{\text{Sorte:Steinmehl}} &&\Rightarrow\ H_{\text{Sorte:Steinmehl}} \text{ ablehnen}
\end{aligned}$$

Im vorliegen Fall haben also sowohl Sorte als auch die Zugabe von Steinmehl signifikanten Einfluss auf den Ertrag, und es gibt signifikante Wechselwirkungen zwischen den Faktoren. Dasselbe Resultat erhält man, wenn man statt der kritischen Werte die p-Werte mit dem gegebenen $\alpha = 0.05$ vergleicht: alle p-Werte liegen darunter, also werden die entsprechenden Nullhypothesen verworfen.

Lösung 4.5
Ertragsleistung von Chinaschilf

Zweifaktorielle Varianzanalyse mit dem Trockensubstanzertrag von Chinaschilf (quantitative Variable) in Abhängigkeit von der Stickstoff- und Kalidüngerstufe (mit $I = 3$ bzw. $J = 2$ Faktorstufen), in vierfacher Wiederholung ($K = 4$).

a) **Nullhypothesen**:

- H_{N}: Kein Einfluss der Stickstoffmenge auf den Ertrag.
- H_{K}: Kein Einfluss der Kailmenge auf den Ertrag.

- $H_{\mathrm{N:K}}$: Keine Wechselwirkung zwischen der ausgebrachten Stickstoff- und Kalimenge auf den Ertrag

b) **Gruppenmittelwerte** $\bar{y}_{ij.}$:

	70 kg K2O	140 kg K2O
0 kg N/ha	17.500	16.625
60 kg N/ha	20.900	21.900
120 kg N/ha	20.975	22.225

Damit können die entsprechenden Mittelwerte für die einzelnen Faktorstufen $\bar{y}_{i..}$ bzw. $\bar{y}_{.j.}$ berechnet werden:

0 kg N/ha:	17.0625	70 kg K2O:	19.79167
60 kg N/ha:	21.4000	140 kg K2O:	20.25000
120 kg N/ha:	21.6000		

Damit können wie in Beispiel 4.3 die Quadratsummen berechnet und die **ANOVA-Tabelle** aufgestellt werden:

	SS	d.f.	MS	F	p
Stickstoff	105.181	2	52.5904	85.42	$6.5 \cdot 10^{-10}$
Kali	1.260	1	1.2604	2.05	$1.7 \cdot 10^{-1}$
Stickstoff:Kali	5.396	2	2.6979	4.38	$2.8 \cdot 10^{-2}$
Fehler	11.083	18	0.6157		
Total	122.9	23			

c) **Kritische Werte**: siehe Beispiel 4.3:

- H_{N}: abgelehnt
- H_{K}: angenommen
- $H_{\mathrm{N:K}}$: abgelehnt

Bemerkung: Man beachte, dass die Kalimenge keinen signifikanten Einfluss hat (die Nullhypothese H_{K} beibehalten), aber signifikante Wechselwirkungen mit der Stickstoffmenge bestehen ($H_{\mathrm{N:K}}$ wird abgelehnt). Das lässt sich so interpretieren, dass die Kalimenge *allein* betrachtet keinen Einfluss hat, sehr wohl aber in Verbindung mit der ausgebrachten Stickstoffmenge; ein Blick auf die Gruppenmittelwerte zeigt, dass wenn überhaupt kein Stickstoff ausgebracht wird, der Ertrag für die 140 kg Kali gegenüber den 70 kg Kali zurückgeht; wird hingegen mit Stickstoff gedüngt, dann hat die Vergrößerung der Kalimenge einen leicht positiven Effekt.

Lösung 4.6
Wanderverhalten von Blattläusen

Zweifaktorielle Varianzanalyse ohne Wechselwirkungen – Aufenthaltsdauer von Blattläusen (quantitative Variable) in Abhängigkeit von Blattlausart ($I = 2$ Faktorstufen) und Blatt ($J = 15$ Faktorstufen). Wechselwirkungen können nicht berücksichtigt werden, da nur eine Wiederholung vorliegt ($K = 1$), vgl. Beispiel 4.3. Für die **ANOVA-Tabelle** ergibt sich

	SS	d.f.	MS	F	p
Blattlausart	30977	1	30977	29.08	$9.5 \cdot 10^{-5}$
Blatt	42779	14	3056	2.87	$2.9 \cdot 10^{-2}$
Fehler	14915	14	1065		
Total	88671	29			

Lösung 4.7
Krankheitsanfälligkeit der Ackerbohne

Zweifaktorielle Varianzanalyse mit Wechselwirkungen – Chlorose in Ackerbohnenpflanzen (quantitative Variable) in Abhängigkeit von der Zuchtlinie der Ackerbohnen ($I = 3$ Faktorstufen) und aufgebrachter Pseudomonasrasse ($J = 2$ Faktorstufen) in vierfacher Wiederholung, vergleiche Beispiel 4.4. Für die **ANOVA-Tabelle** ergibt sich

	SS	d.f.	MS	F	p
Zuchtlinie	0.4955	2	0.2478	1.04	$3.7 \cdot 10^{-1}$
Pseudomonasrasse	9.6901	1	9.6901	40.76	$5.2 \cdot 10^{-6}$
Zuchtlinie:Pseudomonasrasse	4.6303	2	2.3151	9.74	$1.4 \cdot 10^{-3}$
Fehler	4.2788	18	0.2377		
Total	19.09	23			

Kapitel 5

Regressions– und Korrelationsrechnung

Beispiele

Korrelationsrechnung

Beispiel 5.1 Bei 12 Fichten werden im Rahmen eines Projektes übers Waldsterben die Schadstoffe in den Nadeln untersucht. Unter anderem werden auch die Blei- und Eisengehalte ermittelt (in 10^{-5} mg/g).

Blei	2.8	2.6	2.4	1.7	2.5	1.4	2.1	2.2	1.8	2.5	2.4	1.9
Eisen	36	30	39	31	38	21	41	47	40	37	42	31

a) Wie stark sind Blei und Eisen korreliert?

b) Sind Blei und Eisen in Fichtennadeln voneinander abhängig? (Signifikanzniveau $\alpha = 0.05$)

Beispiel 5.2 Bei 10 Jungstieren wird im Rahmen einer Versteigerung sowohl der Brustumfang (in cm) als auch das Gewicht (in kg) ermittelt.

Brustumfang	176	183	177	180	184	178	179	181	179	182
Gewicht	471	510	463	491	541	481	470	518	496	511

a) Wie stark sind Brustumfang und Gewicht korreliert?

b) Sind die beiden unabhängig? (Signifikanzniveau $\alpha = 0.05$)

Regressionsrechnung

Beispiel 5.3 Bei 15 zufällig ausgewählten Frauen wird das Alter in Jahren und der Blutdruck in mm Hg ermittelt.

Alter	47	52	30	35	59	44	63	38
Blutdruck	129	139	112	119	145	133	152	117
Alter	49	41	32	55	46	51	63	
Blutdruck	145	136	115	137	134	141	157	

a) Stellen Sie die Daten grafisch dar.

b) Wählen Sie einen linearen Regressionsansatz und schätzen Sie die Regressionskoeffizienten und die Modellvarianz σ^2. Tragen Sie die geschätzte Regressionsgerade in die grafische Darstellung aus a) ein.

c) Geben Sie 95%–Konfidenzintervalle für die unter b) berechneten Größen an.

d) Welcher Blutdruck kann im Mittel für eine fünfzigjährige Frau erwartet werden? Geben Sie ein 95%–Konfidenzintervall für diesen Wert an.

e) Berechnen Sie ein 95%–Konfidenzintervall für den beobachteten Blutdruck einer 45jährigen Frau.

Beispiel 5.4 Die Luzerne (*Medicago sativa*) findet in der Landwirtschaft Verwendung als Viehfutter. Im Zuge der Sortenprüfung von Luzernen wird bei der Sorte „Europe“ der Ernteertrag in Abhängigkeit des ausgesääten Grasanteils geprüft. Der Grasanteil in der Saatmischung ist in Gewichtsprozent und der Jahrestrockensubstanzertrag einer Jahresernte in dt/ha angegeben.

Grasanteil	0	5	10	15	20	25	30	35
Ertrag	116.8	117.1	115.4	118.3	118.8	124.6	120.5	122.7

a) Stellen Sie die Daten grafisch dar.

b) Wählen Sie einen linearen Regressionsansatz und schätzen Sie die Parameter a, b und σ^2.

c) Geben Sie 95%–Konfidenzintervalle für diese Parameter an.

d) Beschreiben Sie die Brauchbarkeit des (linearen) Modells anhand einer geeigneten Kenngröße.

e) Ermitteln Sie ein 95%–Konfidenzintervall für den gemessenen Ertrag einer Luzernegrasmischung mit einem Grasanteil von 25%.

Vermischte Aufgaben

Beispiel 5.5 Sieben Rinder der Rasse Fleckvieh erhalten in der Mastperiode I (Masttage 245 bis 305) Futter mit unterschiedlichem Anteil an Rohprotein. Im folgenden sind die Rohproteingehalte (in %) und die durchschnittlichen Tageszunahmen (in g) angegeben.

Rohproteinanteil	11	12	13	14	15	16	17
Tageszunahmen	1212	1167	1204	1258	1389	1276	1445

a) Stellen Sie die Daten grafisch dar.

b) Wählen Sie einen linearen Regressionsansatz und schätzen Sie die Parameter a, b und σ^2.

c) Geben Sie 95%–Konfidenzintervalle für diese Parameter an.

d) Mit welchen Tageszunahmen ist bei einem Rohproteinanteil in der Futterration von 15.5% zu rechnen? Geben Sie ein 95%–Konfidenzintervall für diesen Wert an.

e) Ermitteln Sie ein 95%–Konfidenzintervall für die beobachtete Tageszunahme bei einem Rohproteinanteil in der Futterration von 15%.

Beispiel 5.6 Bei Untersuchungen über die Nährstoffaufnahme von Pflanzen wird unter anderem auch die Aufnahme von Selen aus dem Boden ermittelt. Dazu wird an neun Standorten der Selengehalt (in ppm) der Pflanzen und der pH-Wert des Bodens untersucht.

pH-Wert Boden	7.4	7.3	7.1	7.1	6.2	7.3	6.4	5.7	6.1
Selengehalt Pflanze	131	258	85	250	50	128	40	92	21

a) Wie stark sind der pH-Wert des Bodens und der Selengehalt der Pflanzen korreliert?

b) Sind der pH-Wert des Bodens und der Selengehalt der Pflanzen voneinander unabhängig? (Signifikanzniveau $\alpha = 0.05$)

Beispiel 5.7 Bei einem Forschungsvorhaben über die Auswirkungen eines steigenden Ozongehaltes werden Fichtensämlinge unter kontrollierten Umweltbedingungen unterschiedlichen Ozonkonzentrationen ausgesetzt. In der nachfolgenden Tabelle ist das Gewicht der oberirdischen Sprossmasse von Fichtensämlingen (in g) bei steigender Ozonkonzentration (in ppm) dargestellt.

Ozonkonzentration	0	100	200	300	400	500	600	700	800
Gewicht	650	510	630	840	690	750	930	870	790

a) Stellen Sie die Daten grafisch dar.

b) Wählen Sie einen linearen Regressionsansatz und schätzen Sie die Parameter a, b und σ^2.

c) Geben Sie 95%–Konfidenzintervalle für diese Parameter an.

d) Beurteilen Sie die Brauchbarkeit des Modells.

e) Ermitteln Sie das erwartete Gewicht für eine Ozonkonzentration von 500 ppm und geben Sie dafür ein 95%–Konfidenzintervall an.

f) Ermitteln Sie ein 95%–Konfidenzintervall für die gemessene Ozonkonzentration von 200 ppm.

Lösungen

Lösung 5.1
Korrelation von Blei und Eisen in Fichtennadeln

a) Berechnung des **Korrelationskoeffizienten**: Ist händisch eine reine Einsetzübung, wobei man sich abhängig davon, welche Hilfsgrößen man sich verschafft hat, verschiedene Formeln zu Nutze machen kann. Die im Vorlesungsskriptum (Kapitel 7) angegebene Formel lässt sich als

$$\begin{aligned} r_{X,Y} &= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{\sqrt{(\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n)(\sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2/n)}} \\ &= \frac{963.2 - 26.3 \cdot 433/12}{\sqrt{(59.57 - 26.3^2/12)(16147 - 433^2/12)}} = 0.45 \end{aligned}$$

schreiben; diese Form ist nützlich, wenn man die Summen der Werte und der quadrierten Werte zur Verfügung hat. Man kann aber auch, wenn man die Mittelwerte und Standardabweichungen der beiden Variablen kennt, die Formel

$$\begin{aligned} r_{X,Y} &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} \\ &= \frac{963.2 - 12 \cdot 2.1917 \cdot 36.0833}{11 \cdot 0.4188 \cdot 6.8948} = 0.45 \end{aligned}$$

verwenden.

Vorsicht: Die Formel ist sehr empfindlich gegen Rundungsfehler und funktioniert nur für möglichst genaue Mittelwerte und Standardabweichungen!

Bemerkung: Was man immer zur Berechnung braucht ist ein Ausdruck der Form $\sum_{i=1}^n x_i y_i$; das ist es, was im Prinzip den Korrelationskoeffizienten ausmacht, der Rest ist nur Kosmetik, um einen Wert zwischen -1 und +1 zu erhalten.

Anregung: Überraschen Sie Ihren Vortragenden, indem Sie ihm beweisen, warum diese Formeln algebraisch (wenn schon nicht numerisch) äquivalent zu denen im Vorlesungsskriptum sind.

b) Test, ob die Korrelation signifikant ist ($\alpha = 0.05$):

Die **Nullhypothese** für diesen Test lautet $\rho_{X,Y} = 0$, d.h. der wahre (theoretische) Korrelationskoeffizient ist Null, die Variablen (hier: Blei- und Eisengehalt) sind also unkorreliert. Die **Teststatistik** berechnet sich sehr einfach aus dem empirischen Korrelationskoeffizienten:

$$t = \frac{r_{X,Y}\sqrt{n-2}}{\sqrt{1 - r_{X,Y}^2}} = \frac{0.45 \cdot \sqrt{10}}{\sqrt{1 - 0.45^2}} = 1.59$$

Die Teststatistik folgt einer t-Verteilung mit $n - 2 = 10$ Freiheitsgraden und wird groß bzw. klein, wenn der Korrelationskoeffizient nahe bei +1 bzw. -1 liegt; deswegen sprechen zu große oder zu kleine Werte der Teststatistik gegen die Nullhypothese. Die **kritischen Werte** sind also wieder die Quantile der t-Verteilung:

$$c_u = t_{n-2;\alpha/2} = -t_{n-2;1-\alpha/2} = -t_{10;0.975} = -2.228 \qquad c_o = t_{n-2;1-\alpha/2} = 2.228$$

Die Teststatistik liegt klar zwischen den kritischen Grenzen, die Nullhypothese wird also beibehalten. Wir finden auf dem Niveau $\alpha = 0.05$ keinen signifikanten Einwand gegen die Hypothese, dass der Blei- und Eisengehalt in den Fichtennadeln unkorreliert sind.

Bemerkung: Eine Voraussetzung für die Anwendung dieses Tests ist die (annähernde) Normalverteilung der beiden Merkmale (ganz wie z.B. für die Verfahren des Kapitels 5 des Vorlesungsskriptums). Für normalverteilte Daten ist die Unkorreliertheit (d.h. kein *linearer* Zusammenhang zwischen den Daten) äquivalent zu Unabhängigkeit (d.h. *überhaupt* kein statistischer Zusammenhang zwischen den Daten). Jedesmal wenn wir also diesen Test sicher anwenden dürfen, ist er nicht nur ein Test auf Unkorreliertheit, sondern auch auf Unabhängigkeit. Im Gegensatz dazu können wir den Korrelationskoeffizienten für nicht normalverteilte Daten zwar berechnen, aber i.a. nicht testen.

Lösung 5.2
Korrelation von Brustumfang und Gewicht bei Jungstieren

a) Berechnung der **Korrelation**: gesetzt X=Brustumfang und Y=Gewicht, und gegeben die Hilfsgrößen

$$\begin{array}{lll} \sum_{i=1}^n x_i = 1799 & \sum_{i=1}^n x_i^2 = 323701 & \\ & & \sum_{i=1}^n x_i y_i = 891393 \\ \sum_{i=1}^n y_i = 4952 & \sum_{i=1}^n y_i^2 = 2457794 & \end{array}$$

ergibt sich der Korrelationskoeffizient zu

$$r_{X,Y} = r_{\text{Brustumfang, Gewicht}} = 0.91$$

b) Test auf **Signifikanz der Korrelation** für $\alpha = 0.05$:

Die **Nullhypothese** lautet *Brustumfang und Gewicht der Jungstiere sind unkorreliert*, d.h. $\rho_{X,Y} = 0$. Für die **Teststatistik** ermittelt man

$$t = \frac{r_{X,Y}\sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}} = 6.11$$

Die **kritischen Werte** lauten

$$c_u/c_o = \pm t_{n-2;1-\alpha/2} = \pm 2.306$$

Somit wird die **Nullhypothese abgelehnt**, es besteht also ein signifikanter (auf dem Niveau $\alpha = 0.05$) Zusammenhang zwischen Brustumfang und Gewicht.

Lösung 5.3
Abhängigkeit Blutdruck/Alter

Zunächst ermittelt man die Größen

$$\sum_{i=1}^n x_i = 705 \quad \sum_{i=1}^n y_i = 2011 \quad \sum_{i=1}^n x_i^2 = 34685 \quad \sum_{i=1}^n y_i^2 = 272175 \quad \sum_{i=1}^n x_i y_i = 96387$$

und daraus

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = 47 \quad \text{und} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = 134.067$$

a) **Streudiagramm** der Daten: siehe Abbildung 5.1

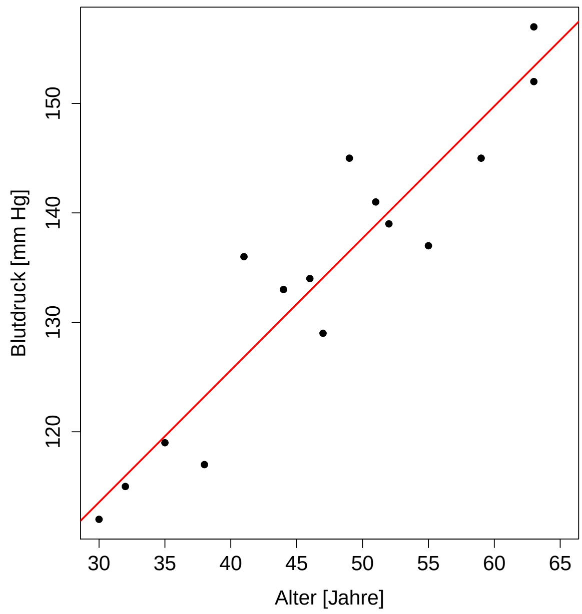


Abbildung 5.1: Streudiagramm und Regressionsgerade für Blutdruck vs. Alter

b) Das **Regressionsmodell** lautet hier sinnvollerweise

$$\text{Blutdruck} = a + b \cdot \text{Alter} + E.$$

Für die **Schätzung der Regressionsparameter** beginnt man günstigerweise mit der Steigung b:

$$\hat{b} = \frac{\frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y})}{\frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)} = \frac{96387 - 15 \cdot 47 \cdot 134.067}{34685 - 15 \cdot 47^2} = 1.206$$

Da das Wertepaar $(\bar{x}, \bar{y})$ (der Schwerpunkt der Punktewolke in Abbildung 5.1) immer auf der Regressionsgeraden liegt, können wir daraus den Schätzer für a berechnen:

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = 134.067 - 1.206 \cdot 47 = 77.385$$

Zur Berechnung der **Modellvarianz** benötigen wir die Stichprobenvarianzen der abhängigen (y bzw. Blutdruck) und unabhängigen (x bzw. Alter) Größe. Man beachte, dass es sich für die unabhängige Größe x nur um eine formale Stichprobenvarianz handelt. Diese verschaffen wir uns wie gehabt aus den angegebenen Summen:

$$s_y^2 = 183.3524 \qquad \text{und} \qquad s_x^2 = 110.7143$$

Damit können wir direkt in die vereinfachte Formel für die Modellvarianz einsetzen:

$$\begin{aligned} s^2 &= \frac{n-1}{n-2} \cdot (s_y^2 - \hat{b}^2 s_x^2) \\ &= \frac{14}{13} \cdot (183.3524 - 1.206^2 \cdot 110.7143) = 24.042 \\ s &= 4.9033 \end{aligned}$$

c) **Konfidenzintervalle** für a, b und σ^2: Die Schätzer $\hat{a}$ und $\hat{b}$ sind wie im Vorlesungsskriptum beschrieben normalverteilt, die Konfidenzintervalle ergeben sich daher aus der t-Verteilung, hier mit $n-2$ Freiheitsgraden. Damit erhalten wir :

$$\begin{aligned}
\hat{a} \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} &= 77.385 \pm 2.16 \cdot 4.9033 \cdot \sqrt{\frac{1}{15} + \frac{47^2}{14 \cdot 110.7143}} \\
&= 77.385 \pm 12.936 = (64.449, 90.321) \\
\hat{b} \pm t_{n-2,1-\alpha/2} \cdot \frac{s}{s_x\sqrt{n-1}} &= 1.206 \pm 2.16 \cdot \frac{4.9033}{10.5221\sqrt{14}} \\
&= 1.206 \pm 0.269 = (0.936, 1.475)
\end{aligned}$$

Die Modellvarianz hingegen folgt wieder einer χ^2-Verteilung, ebenfalls mit $n-2$ Freiheitsgraden:

$$\frac{(n-2)s^2}{\chi^2_{n-2;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-2)s^2}{\chi^2_{n-2;\alpha/2}}, \quad \text{oder} \quad \frac{13 \cdot 24.042}{24.736} \leq \sigma^2 \leq \frac{13 \cdot 24.042}{5.009},$$

$$\text{also} \quad 12.635 \leq \sigma^2 \leq 62.397$$

d) **Konfidenzintervall** für den **erwarteten Blutdruck** mit 50 Jahren: Zunächst berechnen wir den erwarteten Blutdruck, also gerade den entsprechenden Wert auf der Regressionsgeraden:

$$\hat{y}_{50} = 77.385 + 1.206 \cdot 50 = 137.685$$

Das zugehörige Konfidenzintervall gibt an, mit welcher Genauigkeit $\hat{y}_{50}$ auf Grund der vorliegenden Daten geschätzt werden kann:

$$\begin{aligned}
\hat{y} \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} &= 137.685 \pm 2.16 \cdot 4.9033 \cdot \sqrt{\frac{1}{15} + \frac{(50-47)^2}{14 \cdot 110.7143}} \\
&= 137.685 \pm 2.851 = (134.8, 140.5)
\end{aligned}$$

e) **Vorhersageintervall** für den Blutdruck einer 45jährigen: Wieder wird zunächst der erwartete Wert berechnet:

$$\hat{y}_{45} = 77.385 + 1.206 \cdot 45 = 131.655$$

Das entsprechende Vorhersageintervall hat gegenüber dem für den erwarteten Wert einen zusätzlichen 1er unter der Quadratwurzel, die die zusätzliche Variation eines Individuums um die Regressionsgerade beschreibt:

$$\begin{aligned}
&\hat{y} \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} = \\
&= 131.655 \pm 2.16 \cdot 4.9033 \cdot \sqrt{1 + \frac{1}{15} + \frac{(45-47)^2}{14 \cdot 110.7143}} \\
&= 131.655 \pm 10.952 = (120.7, 142.6)
\end{aligned}$$

Lösung 5.4
Ertrag in Abhängigkeit vom Grasanteil

Zunächst berechnen wir Mittel und empirische Varianzen der unabhängigen (*Grasanteil*, x) sowie der abhängigen Größe (*Ertrag*, y):

$$\bar{x} = 17.5 \qquad \bar{y} = 119.275 \qquad s_x^2 = 150 \qquad s_y^2 = 9.805$$

und zusammen mit $\sum_{i=1}^{n} x_i y_i = 16914.5$ den Korrelationskoeffizienten

$$r_{X,Y} = r_{\text{Ertrag,Grasanteil}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} =$$
$$= \frac{16914.5 - 8 \cdot 17.5 \cdot 119.275}{7 \cdot \sqrt{150} \cdot \sqrt{9.805}} = \frac{216}{268.4525} = 0.8046$$

Man beachte: s_x^2 und $r_{X,Y}$ sind nur formal Varianz und Korrelationskoeffizient, da die unabhängige Variable (*Grasanteil*) keine zufällige Größe ist.

a) **Streudiagramm** der Daten: siehe Abbildung 5.2

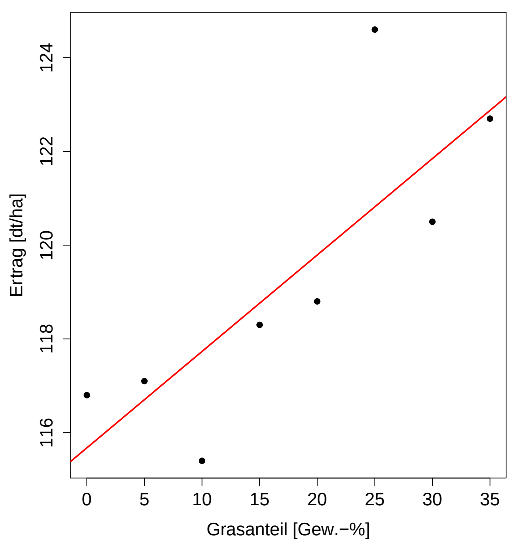


Abbildung 5.2: Streudiagramm und Regressionsgerade für Ertrag/Grasanteil

b) Als **Regressionsmodell** bietet sich hier

$$\text{Ertrag} = a + b \cdot \text{Grasanteil} + E$$

an. Für die Berechnung ist zu beachten, dass als Hilfsgrößen nicht wie üblich die Summen, sondern direkt Mittelwerte, Varianzen und Korrelationskoeffizient gegeben sind. Um nun b zu schätzen kann man entweder die Summen aus den Mittelwerten etc. rückrechnen und die gleiche Formel wie im Vorlesungsskriptum verwenden (mühsam), oder man verwendet folgende nützliche Alternative:

$$\begin{aligned}\hat{b} &= \frac{s_y}{s_x} \cdot r_{x,y} \\ &= \frac{3.1313}{12.2475} \cdot 0.8046 = 0.2057\end{aligned}$$

Sobald man $\hat{b}$ hat, kann man wieder wie in Beispiel 5.3 vorgehen, um Schätzwerte für a und die Modellvarianz σ^2 (bzw. die entsprechende Standardabweichung σ) zu erhalten:

$$\begin{aligned}\hat{a} &= \bar{y} - \hat{b} \cdot \bar{x} = 115.675 \\ s^2 &= \frac{n-1}{n-2} \cdot (s_y^2 - \hat{b}^2 s_x^2) = 4.033 \\ s &= 2.008\end{aligned}$$

c) Mit denselben Formeln wie in Beispiel 5.3 erhält man die **95% Konfidenzintervalle** für die Parameter a und b und für die Modellvarianz σ^2:

$$\begin{aligned}\hat{a} &: (112.50, 118.85) \\ \hat{b} &: (0.054, 0.357) \\ s^2 &: (1.675, 19.56)\end{aligned}$$

d) Das **Bestimmtheitsmaß** R^2 gibt an, welcher Anteil bzw. wieviel Prozent der Variabilität der Erträge durch den Grasanteil erklärt wird und ergibt sich als Quadrat des (formalen) Stichprobenkorrelationskoeffizienten $r_{X,Y}$:

$$R^2 = r_{X,Y}^2 = 0.8046^2 = 0.647$$

Rund 65% der Variabilität der Luzernerträge ist auf die unterschiedlichen Grasanteile zurückzuführen.

e) Ein Konfidenzintervall für den *gemessenen* Ertrag entspricht einem **Prognoseintervall** für einen einzelnen Wert. Der vorhergesagte Wert lautet zunächst für einen Grasanteil von 25 %

$$\hat{y}_{25} = 120.818.$$

Die entsprechende Formel (siehe auch Beispiel 5.3) liefert dann

$$\begin{aligned}&\hat{y} \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} = \\ &= 120.818 \pm 2.447 \cdot 2.008 \cdot \sqrt{1 + \frac{1}{8} + \frac{(25-17.5)^2}{7 \cdot 150}} \\ &= 120.818 \pm 5.335 = (115.48, 126.15)\end{aligned}$$

Lösung 5.5
Tageszunahmen von Rindern in Abhängigkeit des Rohproteinanteils im Futter

a) **Streudiagramm** der Daten: siehe Abbildung 5.3:

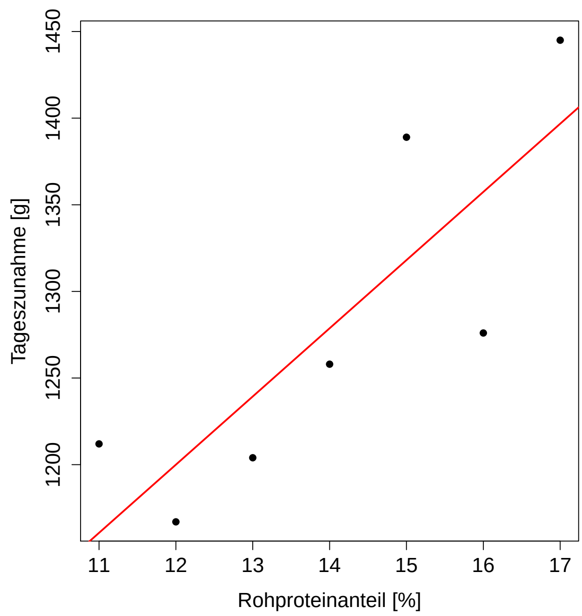


Abbildung 5.3: Streudiagramm und Regressionsgerade für Tageszunahme vs. Rohproteinanteil

b) **Regressionsansatz**:

$$\text{Tageszunahme} = a + b \cdot \text{Rohproteinanteil} + E$$

Mit den üblichen Hilfsgrößen

$$\sum_{i=1}^{n} x_i = 98 \quad \sum_{i=1}^{n} y_i = 8951 \quad \sum_{i=1}^{n} x_i^2 = 1400 \quad \sum_{i=1}^{n} y_i^2 = 11508535 \quad \sum_{i=1}^{n} x_i\, y_i = 126416$$

berechnen wir zunächst die Kenngrößen der beiden Variablen

$$\bar{x} = 14 \qquad s_x^2 = 4.6667 \qquad \bar{y} = 1278.714 \qquad s_y^2 = 10460.57$$

und dann die Schätzwerte für die beiden **Regressionskoeffizienten**:

$$\begin{aligned} \hat{b} &= \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{7 \cdot 126416 - 98 \cdot 8951}{7 \cdot 1400 - 98^2} = 39.357 \end{aligned}$$

Der Achsenabschnitt wird dann wieder rückgerechnet aus dem Punktepaar $(\bar{x}, \bar{y})$, von dem wir wissen, das es immer auf der Regressionsgeraden liegt:

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = 727.714$$

Um die Varianz σ^2 bzw. die Standardabweichung σ des Fehlerterms zu schätzen, können wir ebenfalls die von $\hat{b}$ abhängige Abkürzungsformel wählen:

$$\begin{aligned} s^2 &= \frac{n-1}{n-2} \cdot (s_y^2 - \hat{b}^2 s_x^2) = \\ &= \frac{6}{5} \cdot (10460.57 - 39.357^2 \cdot 4.6667) = 3878.371 \\ s &= 62.277 \end{aligned}$$

c) 95% **Konfidenzintervalle** für die Parameter a, b und σ^2 (Formeln siehe Beispiel 5.3):

$$\begin{aligned} 727.714 \pm 2.571 \cdot 62.277 \cdot \sqrt{\frac{1}{7} + \frac{14^2}{6 \cdot 4.6667}} &\quad \Rightarrow \quad 299.86 \le a \le 1155.6 \\ 39.357 \pm 2.571 \cdot \frac{62.276}{\sqrt{6 \cdot 4.6667}} &\quad \Rightarrow \quad 9.10 \le b \le 69.6 \\ \frac{5 \cdot 3878.371}{12.832}, \frac{5 \cdot 3878.371}{0.831} &\quad \Rightarrow \quad 1511.2 \le \sigma^2 \le 23335.6 \end{aligned}$$

d) Die Fragestellung ist ein bisschen ungenau, aber wenn danach gefragt ist, mit welchen Tageszunahmen für einen Rohproteinanteil von 15.5% zu *rechnen* ist, dann interpretieren wir das als Frage danach, welche Tageszunahmen zu *erwarten* sind, und berechnen daher ein **Konfidenzintervall** für den bei 15.5% zu *erwartenden* Wert.

Zunächst berechnen wir also den erwarteten Wert:

$$y_{15.5} = 727.716 + 39.357 \cdot 15.5 = 1337.749$$

Das zugehörige Konfidenzintervall gibt an, mit welcher Genauigkeit dieser Erwartungswert aus den Daten geschätzt werden kann:

$$\begin{aligned} \hat{y} \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} &= 1337.749 \pm 2.571 \cdot 62.276 \cdot \sqrt{\frac{1}{7} + \frac{(15.5-14)^2}{6 \cdot 4.6667}} \\ &= 1337.749 \pm 75.645 = (1262.1, 1413.4) \end{aligned}$$

e) **Konfidenzintervall** für einen beobachteten Wert (also ein Rindvieh) bei 15% Rohproteinzusatz:

$$y_{15} = 727.716 + 39.357 \times 15 = 1318.071$$

$$\begin{aligned} \hat{y} \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} &= 1318.071 \pm 2.571 \cdot 62.276 \sqrt{1 + \frac{1}{7} + \frac{(15-14)^2}{6 \cdot 4.6667}} \\ &= 1318.071 \pm 173.820 = (1144.3, 1491.9) \end{aligned}$$

Anmerkung: Die Berechnung der Konfidenzintervalle in c)-e) ist offensichtlich von Hand eher mühsam. Allerdings wiederholen sich, wie man hier gut sehen kann, einige Ausdrücke immer wieder, man kann die Berechnungen also durchaus ein bisschen ökonomischer gestalten.

Lösung 5.6
Boden-pH Wert und Selengehalt von Pflanzen

a) **Korrelationskoeffizient**: Mittels

$$\sum_{i=1}^{n} x_i = 60.6 \quad \sum_{i=1}^{n} y_i = 1055 \quad \sum_{i=1}^{n} x_i^2 = 411.26 \quad \sum_{i=1}^{n} y_i^2 = 182839 \quad \sum_{i=1}^{n} x_i y_i = 7384.2$$

ergibt sich für den Korrelationskoeffizienten zwischen dem pH-Wert und dem Selengehalt

$$\begin{aligned} r_{X,Y} &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{(\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n)(\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n)}} \\ &= \frac{7384.2 - \frac{1}{9} \cdot 60.6 \cdot 1055}{\sqrt{(411.26 - 60.6^2/9)(182839 - 1055^2/9)}} = 0.6427 \end{aligned}$$

b) **Test** der Hypothese $\rho_{X,Y} = 0$ für $\alpha = 0.05$

Die **Teststatistik** lautet

$$t = \frac{r_{X,Y}\sqrt{n-2}}{\sqrt{1 - r_{X,Y}^2}} = \frac{0.6427 \cdot \sqrt{7}}{\sqrt{1 - 0.6427^2}} = 2.2195$$

und liegt damit innerhalb der kritischen Grenzen:

$$c_u/c_o = \pm t_{n-2;1-\alpha/2} = \pm t_{7;0.975} = \pm 2.365$$

Die Nullhypothese (*Der pH-Wert und Selengehalt sind unkorreliert bzw. unabhängig*) wird somit beibehalten.

Lösung 5.7
Ozonkonzentration in der Luft vs. Gewicht von Fichtensämlingen.

Mittels der Hilfsgrößen

$$\sum_{i=1}^{n} x_i = 3600 \quad \sum_{i=1}^{n} y_i = 6660 \quad \sum_{i=1}^{n} x_i^2 = 2040000 \quad \sum_{i=1}^{n} y_i^2 = 5069600 \quad \sum_{i=1}^{n} x_i y_i = 2879000$$

ergeben sich unmittelbar die Kennzahlen

$$\bar{x} = 400 \qquad s_x^2 = 75000 \qquad \bar{y} = 740 \qquad s_y^2 = 17650$$

a) **Streudiagramm** der Daten: siehe Abbildung 5.4.

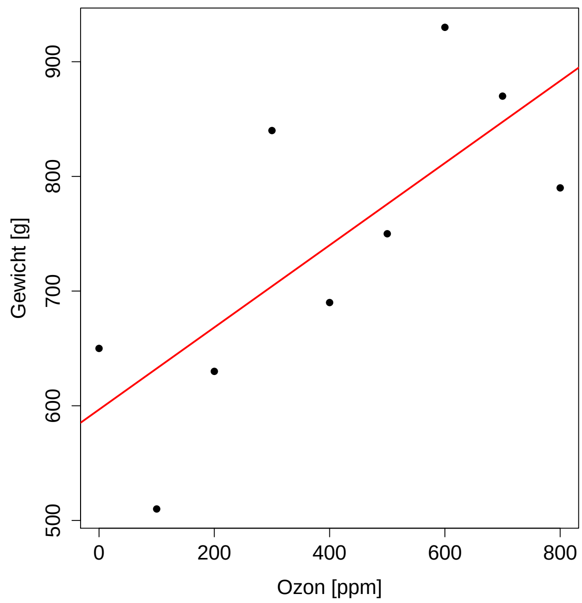


Abbildung 5.4: Streudiagramm und Regressionsgerade für Gewicht vs. Ozonkonzentration

b) **Regressionsmodell**: Sinnvollerweise modellieren wir das Gewicht (Y) als Funktion der Ozonkonzentration (X):

$$\text{Gewicht} = a + b \cdot \text{Ozonkonzentration} + E$$

Mittels der zuvor berechnet Größen ergeben sich die **Schätzwerte** für a, b und σ^2:

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{2879000 - 9 \cdot 400 \cdot 740}{2040000 - 9 \cdot 400^2} = 0.3583$$

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = 740 - 0.3583 \cdot 400 = 596.667$$

$$s^2 = \hat{\sigma}^2 = \frac{n-1}{n-2} \cdot (s_y^2 - \hat{b}^2 s_x^2) = \frac{8}{7} \cdot (17650 - 0.3583 \cdot 75000) = 9165.476$$

$$s = \hat{\sigma} = 95.736$$

c) **Konfidenzintervalle** für Parameter a, b und σ^2: Mittels der entsprechenden Formeln (siehe z.B. Beispiel 5.3) ergeben sich

$$457.50 \leq a \leq 735.83 \qquad 0.066 \leq b \leq 0.651 \qquad 4006.64 \leq \sigma^2 \leq 37963.53$$

d) Das **Bestimmtheitsmaß** R^2 stimmt formal mit dem Quadrat des Korrelationskoeffizienten $r_{X,Y}$ überein. Einfacher ist im vorliegenden Fall allerdings folgende Berechnung (alle einfließenden Größen sind bereits bekannt):

$$R^2 = 1 - \frac{(n-2)\hat{\sigma}^2}{(n-1)s_y^2} = 1 - \frac{7 \cdot 9165.476}{8 \cdot 17650} = 0.5456$$

Rund 55% der Varianz des Sämlingsgewichts geht auf die Ozonbelastung zurück.

e) **Konfidenzintervall** für das *erwartete* Gewicht bei einer Ozonbelastung von 500ppm: Wir berechnen zunächst den geschätzten Wert an der Stelle $x = 500$:

$$\hat{y}_{500} = \hat{a} + \hat{b} \cdot x = 596.667 + 0.3583 \cdot 500 = 775.833$$

Damit ergibt sich für das 95% Konfidenzintervall für den erwarteten Wert:

$$\begin{aligned}
\hat{y} \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} &= \\
= \quad 775.833 \pm 2.365 \cdot 95.736 \cdot \sqrt{\frac{1}{9} + \frac{(500-400)^2}{8 \cdot 75000}} &= \\
= \quad 775.833 \pm 80.935 = (694.9, 856.8) &
\end{aligned}$$

f) **Konfidenzintervall** für ein *gemessenes* Gewicht (i.e. für einen Sämling) bei einer Ozonbelastung von 200ppm:

$$\hat{y}_{200} = \hat{a} + \hat{b} \cdot x = 596.667 + 0.3583 \cdot 200 = 668.333$$

Damit ergibt sich für das 95% Prognoseintervall an der Stelle $x = 200$:

$$\begin{aligned}
\hat{y} \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} &= \\
= \quad 668.333 \pm 2.365 \cdot 95.736 \cdot \sqrt{1 + \frac{1}{9} + \frac{(200-400)^2}{8 \cdot 75000}} &= \\
= \quad 668.333 \pm 245.72 = (422.6, 914.1) &
\end{aligned}$$

Kapitel 6

Nichtparametrische Verfahren

Beispiele

Ersatz für Normalverteilungsmethoden

Beispiel 6.1 An 10 Versuchspersonen wird am Morgen und am Abend die Körpergröße (in cm) gemessen.

Versuchsperson	1	2	3	4	5
Morgen	167.6	170.5	171.5	174.8	166.7
Abend	165.7	168.9	169.8	173.3	166.1
Versuchsperson	6	7	8	8	10
Morgen	172.1	175.2	168.9	167.0	174.1
Abend	173.5	174.5	169.4	165.8	173.7

Testen Sie zum Signifikanzniveau $\alpha = 0.05$, ob sich die morgendlichen und abendlichen Körpergrößen unterscheiden, wenn *nicht* angenommen werden kann, dass die Differenzen normalverteilt sind.

Beispiel 6.2 Für die Herdenbucheintragung wird bei 13 Milchkühen der Rinderrasse *Braunvieh* und bei 11 Kühen der Rasse *Fleckvieh* eine Melkbarkeitsprüfung durchgeführt. Dabei wird unter anderem die Melkdauer (in min) der Kühe pro Tag ermittelt.

Braunvieh	4.9	4.2	4.1	7.9	5.7	5.1	5.6	4.0	7.1	4.8	4.9	7.5	6.7
Fleckvieh	6.1	6.8	5.2	7.1	5.5	6.2	5.3	7.4	8.5	6.3	8.8		

a) Stellen Sie die Melkzeiten je Rinderrasse durch Box-Plots graphisch dar.

b) Gibt es signifikante Unterschiede zwischen den Melkzeiten (Signifikanzniveau $\alpha = 0.1$)? Gehen Sie bei der Wahl des Testverfahrens davon aus, dass die Melkzeiten *nicht* normalverteilt sind.

Verteilungstests

Beispiel 6.3 Bei Untersuchungen über die Saaterbse *Pisum sativum* wird u.a. ein Spaltungsversuch durchgeführt, bei dem die drei Phänotypen (Blütenfarben weiß, rosa und rot) im Verhältnis 1:2:1 erwartet werden. Entspricht das im Versuch beobachtete Spaltungsverhältnis von 14:50:16 dem theoretisch erwarteten? Testen Sie die Hypothese, dass dem so ist, auf dem Signifikanzniveau ($\alpha = 0.05$).

Beispiel 6.4 Ein Saatguterzeuger überprüft, ob bei dem von ihm in den Handel gebrachten Primelsaatgut die Richtlinien bezüglich der Keimung eines anerkannten Saatgutes auch eingehalten werden. Dazu legt er 120 mal jeweils 9 Primelsamen in ein geeignetes Kultursubstrat ein, und beobachtet, wieviele von den Samen tatsächlich keimen.

Gekeimte Primelsamen	0	1	2	3	4	5	6	7	8	9
Anzahl	0	0	1	6	22	29	32	21	6	3

a) Berechnen Sie Mittelwert und Standardabweichung.

b) Durch welche Verteilung lässt sich die Anzahl der gekeimten Primelsamen beschreiben? Testen Sie, ob es sich tatsächlich um eine solche Verteilung handelt ($\alpha = 0.05$).

Beispiel 6.5 Im Rahmen von Bodenuntersuchungsergebnissen wird auf 50 verschiedenen Untersuchungsflächen die Anzahl der Regenwürmer gezählt. In der nachfolgenden Tabelle ist dargestellt, wie oft eine bestimmte Anzahl von Tieren pro Bodenfläche vorgekommen ist.

Anzahl der Regenwürmer	0	1	2	3	4	5	6	7	8	≥ 9
Zahl der Untersuchungsflächen	3	9	11	8	5	9	3	0	1	1

a) Berechnen Sie Mittelwert und Standardabweichung.

b) Durch welche Verteilung lässt sich die Anzahl der Regenwürmer beschreiben? Testen Sie, ob es sich tatsächlich um eine solche Verteilung handelt ($\alpha = 0.05$).

Beispiel 6.6 Eine Obstverwertungsfirma führt bei der Herstellung ihrer Edelbrände laufend Qualitätsuntersuchungen durch. Unter anderem wird die Dichte (in kg/l) eines Apfelvorlaufs untersucht. Die folgende Tabelle gibt die Verteilung von 100 Proben auf sechs Güteklassen an.

K_l	$(-\infty, 0.86]$	$(0.86, 0.88]$	$(0.88, 0.90]$	$(0.90, 0.92]$	$(0.92, 0.94]$	$(0.94, \infty)$
$y_{n;l}$	5	19	30	25	15	6

a) Berechnen Sie Mittelwert und Standardabweichung.

b) Durch welche Verteilung lässt sich die Dichte des Apfelvorlaufs beschreiben? Testen Sie, ob es sich tatsächlich um eine solche Verteilung handelt ($\alpha = 0.05$).

Beispiel 6.7 Eine Saatzuchtgenossenschaft meldet neu gezüchtete Maissorten zur Sortenzulassungsprüfung an und prüft deswegen auch betriebsintern den Ertrag dieser neu gezüchteten Sorten. In der nachfolgenden Tabelle ist der Ertrag (in dt/ha) der Maissorte Figaro in 10-facher Wiederholung dargestellt.

Figaro	121	110	108	99	122	87	98	105	95	107

a) Berechnen Sie Mittelwert und Standardabweichung.

b) Runden Sie Mittelwert und Standardabweichung auf ganze Zahlen μ_0 und σ_0. Durch welche Verteilung lässt sich der Ertrag der Sorte Figaro beschreiben? Testen Sie, ob es sich um eine derartige Verteilung mit dem Mittelwert $\mu = \mu_0$ und der Standardabweichung $\sigma = \sigma_0$ handelt ($\alpha = 0.05$).

Nichtparametrische Varianzanalyse

Beispiel 6.8 Für die Beimischung zu neuen Grünlandsaatgutmischungen werden drei Rotschwingelsorten auf ihre Anbaueignung geprüft. In der nachfolgenden Tabelle ist der Grünmasseertrag in fünffacher Wiederholung für ein Jahr in dt/ha dargestellt.

Futuro	260	265	256	260	271
Condor	280	270	284	279	295
Echo	270	260	255	265	269

Testen Sie, ob es Unterschiede im Grünmasseertrag dieser Sorten gibt, wenn bekannt ist, dass die Ertragswerte nicht normalverteilt sind ($\alpha = 0.05$).

Beispiel 6.9 Ein Lebensmittelverarbeitungsbetrieb erzeugt 3 Müslisorten, die sich vor allem in der mengenmäßigen Zusammensetzung der Bestandteile unterscheiden. Unter anderem wird auch der Trockensubstanzgehalt in % untersucht.

Müsli 1	91.3	92.5	90.1	90.8	
Müsli 2	94.1	94.3	93.1	95.5	94.9
Müsli 3	93.1	96.1	95.5	94.8	95.1

Testen Sie, ob es Unterschiede im Trockensubstanzgehalt der Müslisorten gibt, wenn bekannt ist, dass die Gehaltswerte nicht normalverteilt sind ($\alpha = 0.05$).

Lösungen

Lösung 6.1
Körpergröße Morgen/Abend

Hier handelt es sich offensichtlich um eine **abhängige Stichprobe**, da die Körpergröße zu zwei verschiedenen Zeitpunkten an denselben Versuchspersonen gemessen wird. Die Hypothese *Kein Unterschied zwischen Morgen- und Abendgröße* bedeutet also gerade, dass die

Differenz zwischen Morgen- und Abendgröße im Mittel 0 ist, was in Kapitel 5 des Vorlesungsskriptums als $H_0 : \mu_D = 0$ geschrieben wird. Wären die Differenzen normalverteilt, könnten wir wie dort beschrieben mit einem t-Test für diese fortfahren.

Hier können wir nicht von einer Normalverteilung ausgehen. Damit kommen als Alternative zum t-Test für die Differenzen der **Vorzeichentest** und der **Wilcoxon-Vorzeichenrangtest** in Frage. Diese Verfahren setzen nur voraus, dass die Verteilung der Differenzen wie hier annähernd symmetrisch ist. (Nebenfrage: Wie können Sie die Symmetrie überprüfen?)

Der Vorzeichentest (Kapitel 8.2 im Vorelsungsskriptum) achtet nicht auf die Größe der Differenzen, sondern nur darauf, ob sie positiv oder negativ sind:

Versuchsperson	1	2	3	4	5	6	7	8	9	10
Differenz	1.9	1.6	1.7	1.5	0.6	-1.4	0.7	-0.5	1.2	0.4
Vorzeichen	$+$	$+$	$+$	$+$	$+$	$-$	$+$	$-$	$+$	$+$

Falls die Morgengröße im Mittel gleich der Abendgröße ist, würden wir ungefähr gleich viele negative wie positive Vorzeichen erwarten. Im vorliegenden Fall gibt es einen deutlichen Überhang der positiven Vorzeichen (acht von zehn). Die Frage ist nun, ob dieser Überhang mit Zufall erklärt werden kann, oder ob dies einen signifikanten Einwand gegen die Nullhypothese darstellt.
Die Anzahl positiver Vorzeichen folgt einer Binomialverteilung mit Ereigniswahrscheinlichkeit p und Anzahl von Wiederholungen $n = 10$; die Nullhypothese lautet hier gerade $H_0 : p = 0.5$, d.h. positives und negatives Vorzeichen sind gleich wahrscheinlich. Die Teststatistik ist gerade die Anzahl der positiven Vorzeichen.

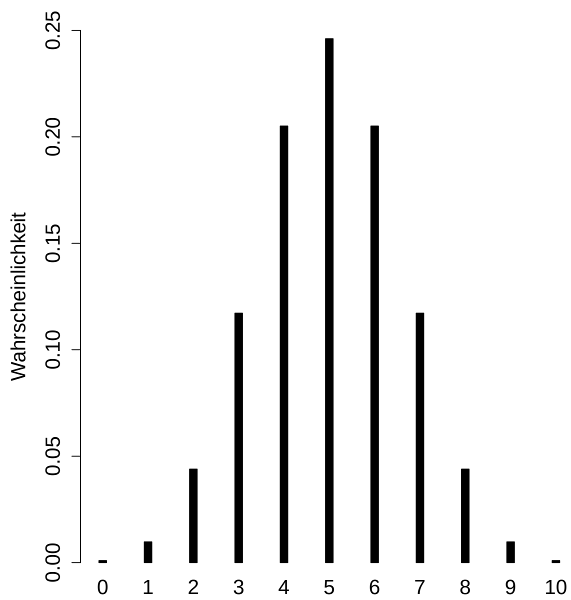


Abbildung 6.1: Wahrscheinlichkeiten für die Anzahl positiver Vorzeichen unter Annahme von $p = 0.5$.

Der Vorzeichentest ist das einzige der in der Vorlesung besprochenen Verfahren, für das wir direkt das empirische Signifikanzniveau (den p-Wert) berechnen können; zur Erinnerung: das ist die Wahrscheinlichkeit, unter Annahme der Nullhypothese eine Teststatistik zu erhalten, die ebenso stark oder noch stärker gegen die Nullhypothese spricht. Im vorliegenden Fall, weil es sich um einen symmetrischen Test handelt, müssen wir Abweichungen von der Nullhypothese in *beide* Richtungen berücksichtigen; mögliche Werte der Teststatistik, die ebenso stark oder noch stärker gegen die Nullhypothese sprechen, sind 8, 9 oder 10 positive Vorzeichen, bzw. am anderen Ende der Skala, 2, 1 oder 0 positive Vorzeichen. Entsprechend ist das empirische Signifikanzniveau gegeben durch

$$p - \text{Wert} = \mathrm{P}(T \geq 8 \text{ oder } T \leq 2 \,|\, p = 0.5)$$

Erfreulicherweise sind die Wahrscheinlichkeiten symmetrisch, wie man z.B. in Abbildung 6.1 sieht oder sich leicht überlegen kann wegen

$$\begin{aligned} \mathrm{P}(T = k \,|\, p = 0.5) &= \binom{n}{k} 0.5^k \cdot 0.5^{n-k} = \binom{n}{k} 0.5^n \\ &= \binom{n}{n-k} 0.5^k \cdot 0.5^{n-k} = \mathrm{P}(T = n-k \,|\, p = 0.5) \end{aligned}$$

Also gilt:

$$\begin{aligned} p - \text{Wert} &= \mathrm{P}(T \geq 8 \text{ oder } T \leq 2 \,|\, p = 0.5) \\ &= 2 \cdot \mathrm{P}(T \geq 8 \,|\, p = 0.5) \\ &= 2 \cdot [\mathrm{P}(T = 8 \,|\, p = 0.5) + \mathrm{P}(T = 9 \,|\, p = 0.5) + \mathrm{P}(T = 10 \,|\, p = 0.5)] \\ &= 2 \cdot \left[\binom{10}{8} \left(\frac{1}{2}\right)^{10} + \binom{10}{9} \left(\frac{1}{2}\right)^{10} + \binom{10}{10} \left(\frac{1}{2}\right)^{10} \right] \\ &= \frac{1}{2^9} \cdot \left[\binom{10}{8} + \binom{10}{9} + \binom{10}{10} \right] \\ &= \frac{1}{512} \cdot \left[\frac{10 \cdot 9}{2 \cdot 1} + \frac{10}{1} + 1 \right] \\ &= \frac{56}{512} = 0.109 > 0.05 = \alpha \end{aligned}$$

Der p-Wert liegt also über dem gewählten Signifikanzniveau α, daher können wir die Nullhypothese nicht mit der gewünschten Sicherheit verwerfen.

Bemerkung: Aus Symmetriegründen kann der Test auf die gleiche Weise und mit dem selben Ergebnis für die Anzahl der negativen Vorzeichen durchgeführt werden.

Der **Wilcoxon-Vorzeichenrangtest** beruht auf den Rängen und Vorzeichen der Differenzen. Dazu werden zunächst die Differenzen nach ihrer absoluten Größe (ohne Berücksichtigung des Vorzeichens) geordnet und mit ihrem Rang versehen (also der Größe nach durchnummeriert):

Differenz	0.4	-0.5	0.6	0.7	1.2	-1.4	1.5	1.6	1.7	1.9
Vorzeichen	+	−	+	+	+	−	+	+	+	+
Rang	1	2	3	4	5	6	7	8	9	10

Die Teststatistik ist gerade die Summe der Ränge für die positiven Vorzeichen:

$$t = t^+ = 1 + 3 + 4 + 5 + 7 + 8 + 9 + 10 = 47$$

Da wir hier die beidseitige Hypothese $H_0 : \mu_D = 0$ testen, ergeben sich nach Tabelle A.9 für $\alpha = 0.05$ die folgenden **kritischen Grenzen**:

$$c_u = w_{10;0.025} = 9 \qquad c_o = w_{10;0.975} = 45$$

Die Teststatistik liegt also über der kritischen Grenze, die Hypothese gleicher mittlerer Morgen- und Abendgröße kann also unter Verwendung des Wilcoxon-Vorzeichenrangtests auf dem Niveau $\alpha = 0.05$ abgelehnt werden.

Bemerkung: Aus Symmetriegründen kann der Test auf die gleiche Weise und mit dem selben Ergebnis für die Summe der negativen Ränge $t^- = 8 < c_u$ durchgeführt werden.

Bemerkung: Dies ist ein schönes Beispiel dafür, dass ein mächtigerer Test (hier: der Wilcoxon-Vorzeichenrangtest) ein signifikantes Ergebnis finden kann, wo dies mit einem schwächeren Test (hier: Vorzeichentest) nicht möglich ist. Es kommt also im Ernstfall darauf an, dass man nicht nur *irgendein* geeignetes Testverfahren, sondern das mächtigste auswählt.

Lösung 6.2
Melkdauer zweier Rinderrassen

a) **Boxplots**: wie in Kapitel 2 des Übungsskriptums beschrieben (siehe Abbildung 6.2).

	Ausr.	Whisker	Quantil	Median	Quantil	Whisker	Ausr.
Braunvieh	–	4	4.5	5.1	6.9	7.9	–
Fleckvieh	–	5.2	5.5	6.3	7.4	8.8	–

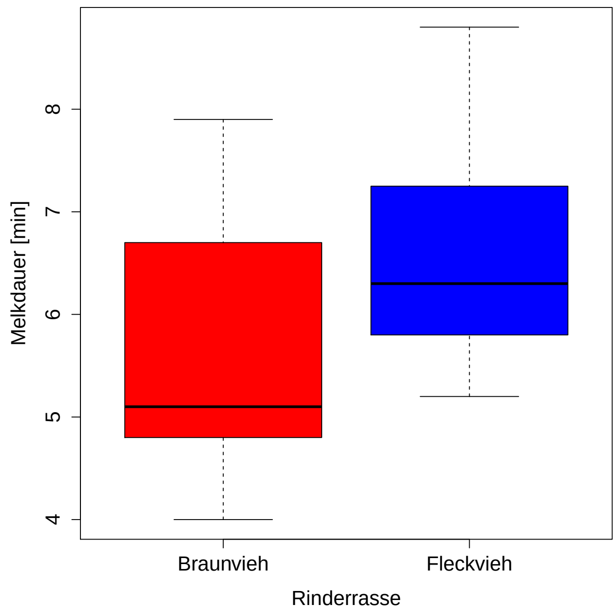


Abbildung 6.2: Boxplots für die Melkzeiten von Braun- und Fleckvieh.

i	Braunv.	Fleckv.	r_{i1}	r_{i2}	i	Braunv.	Fleckv.	r_{i1}	r_{i2}
1	4.0		1		13		6.1		13
2	4.1		2		14		6.2		14
3	4.2		3		15		6.3		15
4	4.8		4		16	6.7		16	
5	4.9		5.5		17		6.8		17
6	4.9		5.5		18	7.1		18.5	
7	5.1		7		19		7.1		18.5
8		5.2		8	20		7.4		20
9		5.3		9	21	7.5		21	
10		5.5		10	22	7.9		22	
11	5.6		11		23		8.5		23
12	5.7		12		24		8.8		24
								128.5	171.5

Tabelle 6.1: Rangsummenberechnung für Melkzeiten zweier Rinderrassen

b) Hier handelt es sich um **unabhängige Stichproben** (Messungen an zwei verschiedenen Gruppen von Rindern). Im Falle einer Normalverteilung der Melkzeiten für beide Rassen könnte die Frage nach signifikanten Unterschieden wieder über einen t-Test beantwortet werden. Im vorliegenden Fall kommen der **Wilcoxon-Rangsummentest** und der **Kolmogorov-Smirnov Zweistichprobentest** in Frage.

Der **Wilcoxon-Rangsummentest** (8.1.2 im Vorlesungsskript) vergleicht die Rangsummen der beiden betrachteten Gruppen (hier: Rinderrassen). Dazu werden die Beobachtungen zunächst gemeinsam der Größe nach sortiert, dann werden die Rangzahlen vergeben und gruppenweise aufsummiert. Unter der Nullhypothese gleicher Verteilungen in beiden Gruppen sollten diese Summen vergleichbar groß sein, insbesonders also nicht zu groß (Überhang an großen Werten für die betreffende Gruppe) und nicht zu klein (detto für kleine Werte).

Tabelle 6.1 zeigt die Berechnung tabellarisch: Die Spalte i enthält die Rangzahl in der gemeinsamen Stichprobe, die Spalten Braunvieh und Fleckvieh die tatsächlichen Melkzeiten, und die Spalten r_{i1} und r_{i2} die Rangzahlen für Braunvieh bzw. Fleckvieh. Zu beachten ist, wie dabei mit mehrfach vorkommenden Werten (sogenannte *ties*, z.B. 4.9 oder 7.1) umgegangen wird: als Rangzahl wird gerade der durchschnittliche Rang mehrfach vergeben; so etwa nimmt der Wert 4.9 die Ränge 5 und 6 in der gemeinsamen Stichprobe ein; der mittlere Rang ist daher 5.5 und wird zweimal vergeben; der nächste Wert (5.1) hat dann natürlich den Rang 7, weil ja 5 und 6 beide verbraucht sind.

Die Rangsummen ergeben sich zu 128.5 und 171.5; zur Kontrolle, ob richtig gerechnet wurde, muss für ihre Summe gelten:

$$\text{Rangsumme Gruppe 1} + \text{Rangsumme Gruppe 2} = \frac{(n_1 + n_2) \cdot (n_1 + n_2 + 1)}{2}$$

$$128.5 + 171.5 \stackrel{!!}{=} \frac{(13 + 11) \cdot (13 + 11 + 1)}{2} = 300, \quad \text{ok}$$

Wir wählen nun als **Teststatistik** die Rangsumme für das Braunvieh, in der Schreibweise des Vorlesungsskriptums

$$t = w_{13,11} = 128.5$$

Den **unteren kritischen Wert** können wir direkt aus Tabelle A.5 des Vorlesungsskriptums ablesen:

$$c_u = w_{13,11}^{2,u} = 134$$

Bemerkung: Tabelle A.5 enthält die *einseitigen* γ-Quantile; für einen beidseitigen Test zum Niveau $\alpha = 0.1$ wie hier wählen wir also $\gamma = 0.05$, d.h. die jeweils zweite Zeile jeden Tabelleneintrags.

Den **oberen kritischen Wert** gewinnen wir durch Subtraktion:

$$\begin{aligned} c_o &= w_{13,11}^{2,o} \\ &= (13+11)(13+11+1)/2 - w_{11,13}^{2,u} = 300 - 109 = 191 \end{aligned}$$

Da die Teststatistik unterhalb der unteren Grenze liegt, können wir also die **Hypothese gleicher Verteilung** der Melkzeiten zwischen den beiden Rassen auf dem Niveau $\alpha = 0.1$ **verwerfen**.

Bemerkung: Aus Symmetriegründen kann auch mit der Rangsumme der 2. Stichprobe (*Fleckvieh*) als Teststatistik der Test durchgeführt werden. In diesem Fall lautet die Teststatistik $t = w_{11,13} = 171.5$, die untere kritische Grenze $c_u = w_{11,13}^{2,u} = 109$ und die obere kritische Grenze ergibt sich aus

$$w_{11,13}^{2,o} = (13+11)(13+11+1)/2 - w_{13,11}^{2,u} = 300 - 134 = 166.$$

H_0 wird also auch in diesem Fall ablehnt, da $t = 171.5 = w_{n1,n2} > w_{n1,n2}^{2,o} = 166$.

Der **Kolmogorov-Smirnov-Test** (Abschnitt 8.3.2 im Vorlesungsskript) für zwei Stichproben beruht auf dem maximalen Absolutabstand zwischen den empirischen Verteilungsfunktionen der beiden Gruppen. Unter der Nullhypothese gleicher (theoretischer) Verteilungen sollte dieser Abstand nicht zu groß werden.

Tabelle 6.2 fasst die notwendigen Berechnungen zusammen: Wieder werden die Beobachtungen beider Gruppen gemeinsam sortiert (erste Spalte); die beiden folgenden Spalten enthalten den Wert der empirischen Verteilungsfunktion für beide Gruppen, und die letzte Spalte deren Differenz. Zu beachten ist dabei, dass ein in einer Gruppe mehrfach vorkommender Wert (hier: 4.9 für Braunvieh) auch entsprechend bei der Berechnung der Verteilungsfunktion berücksichtigt wird: hier also ein doppelt so großer Sprung von 0.3077 auf 0.4615 als bei Werten, die nur einmal vorkommen. Außerdem sei daran erinnert, dass der Wert der empirischen Verteilungsfunktion unterhalb des kleinsten beobachteten Wertes 0 und oberhalb des größten beobachteten Wertes 1 ist: dadurch kommen etwa die Differenzen in Zeile 1 bis 6 zustande. Veranschaulicht ist das Ganze in Abbildung 6.3.

x	$\hat{F}_A(x)$	$\hat{F}_B(x)$	Δ	x	$\hat{F}_A(x)$	$\hat{F}_B(x)$	Δ
4.0	0.0769		0.0769	6.1		0.3636	0.3287
4.1	0.1538		0.1538	6.2		0.4545	0.2378
4.2	0.2308		0.2308	6.3		0.5455	0.1468
4.8	0.3077		0.3077	6.7	0.7692		0.2237
4.9	0.4615		0.4615	6.8		0.6364	0.1328
5.1	0.5385		0.5385	7.1	0.8462	0.7273	0.1189
5.2		0.0909	0.4476	7.4		0.8182	0.0280
5.3		0.1818	0.3567	7.5	0.9231		0.1049
5.5		0.2727	0.2658	7.9	1.0000		0.1818
5.6	0.6154		0.3427	8.5		0.9091	0.0909
5.7	0.6923		0.4196	8.8		1.0000	0.0000

Tabelle 6.2: Verteilungsfunktionen der Melkzeiten für Braunvieh ($\hat{F}_A$) und Fleckvieh ($\hat{F}_B$) und ihre Differenz (Δ)

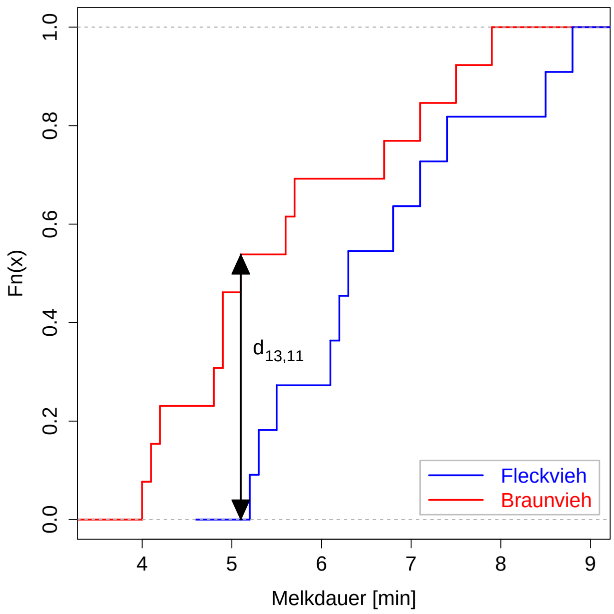


Abbildung 6.3: Empirische Verteilungsfunktionen der Melkzeiten für Braunvieh (rot) und Fleckvieh (blau). Der maximale Abstand zwischen den beiden Verteilungsfunktionen ist durch einen Pfeil gekennzeichnet.

Der maximale Abstand (in Tabelle 6.2 blau gekennzeichnet, in Abbildung 6.3 durch einen vertikalen Pfeil) ergibt sich zu 0.538, und damit erhalten wir als Teststatistik

$$t = \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \cdot d_{n_1,n_2} = \sqrt{\frac{13 \cdot 11}{13 + 11}} \cdot 0.538 = 1.31$$

Der (obere) **kritische Wert** lässt sich aus Tabelle A.8 des Vorlesungsskriptums ablesen: Dabei bedeuten $n_{(1)}$ den kleineren und $n_{(2)}$ den größeren der beiden Stichprobenumfängen, in unserem Fall also $n_{(1)} = n_2 = 11$ und $n_{(2)} = n_1 = 11$. Man liest den kritischen Wert somit bei $\alpha = 0.1$, $n_{(1)} = 4 - 16$ und $n_{(2)} = 10 - 20$ ab, somit

$$c_o = d^{(2)}_{11,13,0.90} = 1.16.$$

Die **Nullhypothese** gleicher Verteilungen wird also auch hier wieder **abgelehnt**.

Lösung 6.3
Spaltungsverhältnis dreier Phänotypen

Ein Spaltungsverhältnis von 1:2:1 in einer Population entspricht einer Multinomialverteilung für die drei möglichen Ergebnisse (den drei Blütenfarben) für ein Individuum (eine Pflanze), mit Ereigniswahrscheinlichkeiten p_1=0.25, p_2=0.5 und p_3=0.25. Dass das Merkmal Blütenfarbe dieser Verteilung folgt, ist also unsere **Nullhypothese**. Die "Zielverteilung" ist hier exakt bekannt, ohne dass zusätzliche Parameter geschätzt werden müssen. Damit ist der **einfache χ^2-Test** ein geeignetes Verfahren.
Aus der Angabe geht hervor, dass insgesamt $n = 80$ Pflanzen bis zur Ausbildung des interessierenden Phänotyps beobachtet wurden; unter der Nullhypothese (also bei Vorliegen der oben angeführten Wahrscheinlichkeiten p_i) ergibt sich die für einen Phänotyp i zu erwartende Häufigkeit als

$$np_i = 80p_i,$$

vgl. z.B. Kapitel 3.5.3 des Vorlesungsskriptums. Abbildung 6.4 stellt tatsächlich beobachtete und erwartete Häufigkeiten einander gegenüber. Tabelle 6.3 zeigt in den Spalten 2 und 4 ebenfalls die Häufigkeiten; da hier für jede Klasse l mehr als fünf Ereignisse erwartet werden ($\hat{e}_l > 5$), ist es nicht notwendig, Klassen zusammenzulegen. Spalte 5 enthält nun die Differenzen zwischen beobachteten und erwarteten Häufigkeiten, und Spalte 6 die quadrierten Differenzen aus Spalte 5, dividiert durch die erwarteten Häufigkeiten aus Spalte 4; addiert man nun die Werte in Spalte 6, erhält man gerade die χ^2-Teststatistik

$$t = \sum_l \frac{(y_{n;l} - \hat{e}_l)^2}{\hat{e}_l} = 1.8 + 2.5 + 0.8 = 5.1$$

l	$y_{n;l}$	p_l	$\hat{e}_l = np_l$	$y_{n;l} - \hat{e}_l$	$(y_{n;l} - \hat{e}_l)^2/\hat{e}_l$
1	14	0.25	20	-6	1.80
2	50	0.50	40	10	2.50
3	16	0.25	20	-4	0.80
$\sum$	80	1.00	80	0	5.10

Tabelle 6.3: Schrittweise Berechnung der χ^2-Teststatistik für den Spaltungsversuch

Unter der Nullhypothese folgt diese Teststatistik einer χ^2-Verteilung; da es sich um einen einfachen χ^2-Test handelt (keine Parameter geschätzt), hat diese $k - 1$ Freiheitsgrade, wobei $k = 3$ die Anzahl von Klassen ist. Der **kritische Wert** des ergibt sich hier zu

$$c_o = \chi^2_{k-1;1-\alpha} = \chi^2_{2;0.95} = 5.99 > 5.1 = t,$$

d.h. H_0 wird beibehalten. Die Daten liefern keinen signifikanten Einwand gegen ein Spaltungsverhältnis von 1:2:1.

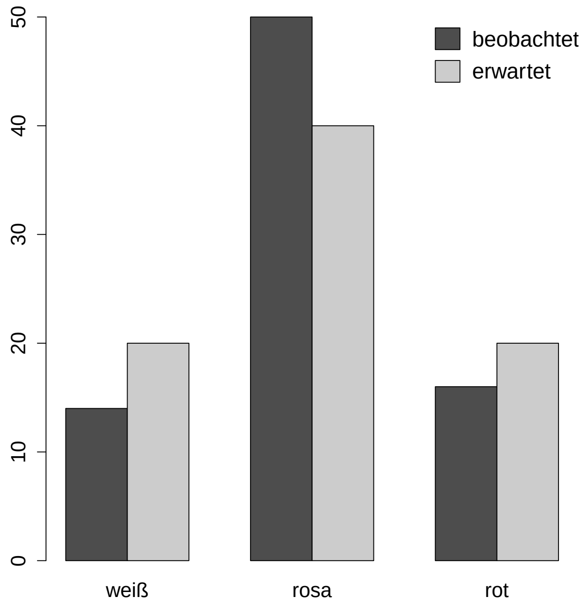


Abbildung 6.4: Häufigkeitsplot für den Spaltungsversuch

Lösung 6.4
Keimung von Primelsamen

a) Mittelwert und Standardabweichung berechnen sich ganz ähnlich wie für klassierte Daten:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{120}(0 \cdot 0 + 1 \cdot 0 + \ldots + 9 \cdot 3) = \frac{667}{120} = 5.56$$

Im Durchschnitt keimen also 5.56 von 9 ausgebrachten Samen (oder in der Realität 5 bis 6). Ähnlich für die Stichprobenvarianz:

$$\begin{aligned} s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i x_i^2 - n \cdot \bar{x}^2\right) \\ &= \frac{1}{119}\left(3943 - 120 \cdot 5.56^2\right) = 1.96 \end{aligned}$$

b) Die naheliegende Vermutung für diese Art von Versuch ist, dass jeder Primelsamen die gleiche Wahrscheinlichkeit p hat erfolgreich zu keimen. Wiederholt man dieses Experiment neunmal, ergibt sich eine Binomialverteilung mit bekanntem Parameter $k = 9$ und unbekanntem Parameter p. Ein guter Schätzer für p (sogar der Maximum-Likelihood-Schätzer, siehe Abschnitt 4.1.2 im Vorlesungsskriptum) ist

$$\hat{p} = \frac{\bar{x}}{n} = \frac{5.56}{9} = 0.618$$

Die **Nullhypothese** lautet also: die beobachteten Häufigkeiten stammen aus einer Binomialverteilung mit festem Parameter $k = 9$ und geschätztem Parameter $\hat{p} = 0.618$. Das geeignete Verfahren ist hier der zusammengesetzte χ^2-Test.

Die Berechnung der **Teststatistik** erfolgt schrittweise wie in Beispiel 6.3, der einzige Unterschied liegt in der Berechnung der unter der Nullhypothese erwarteten Klassenwahrscheinlichkeiten $\hat{p}_l$. Der Verlauf ist in Tabelle 6.4 wieder schematisch dargestellt: Spalte 1 enthält die möglichen Ereignisklassen (zwischen 0 und 9 gekeimten Samen), Spalte 2 die beobachteten Häufigkeiten; die Wahrscheinlichkeiten $\hat{p}_l$ in Spalte 3 berechnen sich nach der Formel für die Binomialverteilung (Abschnitt 3.5.3 im Vorlesungsskriptum), also etwa

$$\begin{aligned} \hat{p}_3 = \binom{k}{3} p^3 q^{k-3} &= \binom{9}{3} 0.618^3 \cdot (1 - 0.618)^{(9-3)} \\ &= \frac{9!}{3!6!} \cdot 0.618^3 \cdot 0.382^6 = 0.0616 \end{aligned}$$

Spalte 4 enthält wieder die unter der Nullhypothese erwarteten Häufigkeiten, die sich stets durch Multiplizieren der Wahrscheinlichkeiten mit der Anzahl von Wiederholungen (hier: $n = 120$) ergeben.

l	$y_{n;l}$	$\hat{p}_l = \binom{k}{l} p^l q^{k-l}$	$\hat{e}_l = np_l$		$(y_{n;l} - \hat{e}_l)^2/\hat{e}_l$
0	0	0.0002	0.024	$\uparrow$	
1	0	0.0025	0.300	2.280	0.72
2	1	0.0163	1.956	$\downarrow$	
3	6	0.0616	7.392		0.26
4	22	0.1495	17.940		0.92
5	29	0.2419	29.028		0.00
6	32	0.2609	31.308		0.02
7	21	0.1809	21.708		0.02
8	6	0.0731	8.772		0.88
9	3	0.0131	1.572		1.30
$\sum$	120	1.0000	120.000		4.12

Tabelle 6.4: Schrittweise Berechnung der χ^2-Teststatistik für den Primelversuch

Hier ist darauf zu achten, dass die erwartete Häufigkeit für jede Klasse ausreichend groß ist, um die annähernde χ^2-Verteilung der Teststatistik zu garantieren; dies ist gewährleistet, wenn $\hat{e}_l > 1$ für jede Randklasse (hier: 0 und 9) und $\hat{e}_l > 5$ für alle anderen Klassen. Offensichtlich funktioniert dies nicht für $l = 0$, wir müssen also solange Klassen zusammenlegen, bis dies erreicht ist; im vorliegenden Fall vereinigen wir die Klassen $l = 0, 1, 2$ und erhalten eine neue Randklasse mit beobachteter Häufigkeit 1 und erwarteter Häufigkeit 2.28, angedeutet durch die Pfeile in Spalte 4. Spalte 5 enthält wieder die Beiträge der einzelnen Klassen zur Teststatistik, die sich durch Aufsummieren zu $t = 4.12$ ergibt.

Unter der Nullhypothese (und unter Beachtung der Mindestanzahlen) folgt diese Teststatistik einer χ^2-Verteilung mit Freiheitsgraden $df = g - r - 1$, wobei g die Anzahl der Klassen ist (hier: g=8 nach Zusammenfassen) und r die Anzahl der geschätzten Parameter (hier: $r = 1$). Damit ist der **kritische Wert**

$$c_o = \chi^2_{g-r-1;1-\alpha} = \chi^2_{6;0.95} = 12.592 > 4.12 = t$$

und wir können die **Nullhypothese beibehalten**.

Lösung 6.5
Anzahl von Regenwürmern pro Bodenfläche

a) **Mittelwert** und **Standardabweichung**: Wieder wie für klassierte Daten, wobei wir als repräsentativen Wert für die letzte Klasse (≥ 9) gerade den Randwert 9 wählen.

$$\begin{aligned} \bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i &= \frac{1}{50}(3 \cdot 0 + 9 \cdot 1 + \ldots + 1 \cdot 9) = \frac{155}{50} = 3.10 \\ s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i x_i^2 - n\bar{x}^2\right) \\ &= \frac{1}{49}(683 - 50 \cdot 3.1^2) = 4.13 \\ s = \sqrt{s^2} &= 2.03 \end{aligned}$$

b) Zum Unterschied vom Primelbeispiel gibt es hier keine natürliche obere Grenze für die *mögliche* Anzahl von Regenwürmern pro Untersuchungsfläche – damit kommt eine Binomialverteilung nicht in Frage. Ein häufige Annahme ist in dieser Situation die der **Poissonverteilung** (siehe Abschnitt 3.5.6 im Vorlesungsskriptum) mit Wahrscheinlichkeitsfunktion

$$P_\mu(l) = e^{-\mu}\frac{\mu^l}{l!}$$

für die Wahrscheinlichkeit von l Ereignissen (hier: Regenwürmern). Der Maximum-Likelihood-Schätzer für den unbekannten Parameter μ ist gerade der Mittelwert:

$$\hat{\mu} = \bar{x} = 3.1$$

Die Nullhypothese besagt also, dass die Anzahl der Regenwürmer einer Poissonverteilung mit geschätztem Parameter $\hat{\mu} = 3.1$ folgt.

Das geeignete Verfahren für diese Art der Hypothese ist wieder ein **zusammengesetzter χ^2-Test**, wie er in Tabelle 6.5 dargestellt ist. Der einzige Unterschied zu den vorhergehenden Beispielen ist wieder die Berechnung der Klassenwahrscheinlichkeiten $\hat{p}_l$, hier z.B.:

$$\hat{p}_2 = e^{-3.1}\frac{3.1^2}{2!} = 0.2165$$

Man beachte, dass man die Wahrscheinlichkeit für die letzte Klasse (≥ 9) leicht durch Subtraktion erhält:

$$\hat{p}_9 = 1 - \hat{p}_0 - \hat{p}_1 - \ldots - \hat{p}_8.$$

Die erwarteten Häufigkeiten $\hat{e}_l$ in Spalte 4 ergeben sich wieder durch Multiplikation mit der Anzahl von Wiederholungen (hier: Untersuchungsflächen, also 50). Nach Berechnung

l	$y_{n;l}$	$\hat{p}_l = e^{-\hat{\mu}}\hat{\mu}^l/l!$	$\hat{e}_l = n\hat{p}_l$		$(y_{n;l} - \hat{e}_l)^2/\hat{e}_l$
0	3	0.0450	2.25		0.250
1	9	0.1397	6.98		0.585
2	11	0.2165	10.82		0.003
3	8	0.2237	11.18		0.905
4	5	0.1733	8.67		1.551
5	9	0.1075	5.37		2.447
6	3	0.0555	2.78	$\uparrow$	
7	0	0.0245	1.23	4.72	0.017
8	1	0.0095	0.48	$\vert$	
$\geq$9	1	0.0047	0.23	$\downarrow$	
$\sum$	50	1	50		5.754

Tabelle 6.5: Schrittweise Berechnung der χ^2-Teststatistik für den Regenwurmversuch.

der $\hat{e}_l$ wird wieder überprüft, ob diese ausreichend groß sind; im vorliegenden Fall wird zweimal zusammengefasst: zunächst die Klassen ≥ 9 und 8, damit die neue Randklasse eine erwartete Häufigkeit größer als 1 aufweist; diese neue Randklasse wird dann mit Klasse 6 und 7 zusammengelegt, weil ansonsten zwei Innenklassen mit $\hat{e}_l < 5$ existieren; die neue Randklasse ist mit einer erwarteten Häufigkeit von 4.72 dann mehr als ausreichend besetzt. Die **Teststatistik** berechnet sich wieder als Summe der einzelnen Beiträge in Spalte 5. Die Freiheitsgrade der χ^2-Verteilung ergeben sich wieder als $df = g-r-1 = 7-1-1 = 5$:

$$t = 5.75 < 11.070 = \chi^2_{5;0.95} = \chi^2_{g-r-1;1-\alpha} = c_o$$

Die H_0 wird also beibehalten.

Lösung 6.6
Dichte des Apfelvorlaufs

a) Berechnung von **Mittelwert** und **Standardabweichung**: Diesmal handelt es sich effektiv um klassierte Daten ($k = 6$ Klassen), die Kenngrößen werden berechnet wie im Abschnitt über beschreibende Statistik beschrieben; als repräsentativen Wert für die beidseitig beschränkten Klassen wählen wir sinnvollerweise die Klassenmitte, für die Randklassen die obere bzw. untere Grenze.

$$\begin{aligned}
\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{x}_i &= \frac{1}{100}(5 \cdot 0.86 + 19 \cdot 0.87 + \ldots + 6 \cdot 0.94) \\
&= \frac{89 \cdot 87}{100} = 0.8987 \\
s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 &= \frac{1}{n-1}\sum_{i=1}^{k} n_i\bar{x}_i^2 - n\bar{x}^2 \\
&= \frac{1}{99}\left(80.8197 - 100 \cdot 0.8987^2\right) = 5.4 \cdot 10^{-4} \\
s = \sqrt{s^2} &= 0.02325
\end{aligned}$$

b) Hier handelt es sich um **klassierte intervallskalierte Daten**, eine diskrete Verteilung wie die Binomial- oder Poissonverteilung kommt also nicht direkt in Frage. Die Darstellung der Klassenhäufigkeiten in Abbildung 6.5 zeigt ein ungefähr symmetrisches Verhalten, also scheint eine Normalverteilung eine plausible Idee zu sein. Als Schätzer für die Parameter μ und σ^2 der Normalverteilung wählen wir wie üblich $\bar{x}$ und s^2.

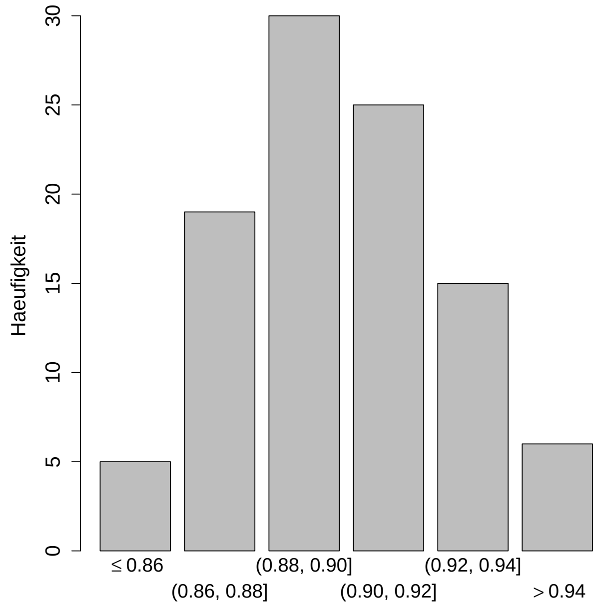


Abbildung 6.5: Beobachtete Häufigkeiten für den Apfelvorlauf

Es handelt sich also wieder um einen **zusammengesetzten χ^2-Test**, diesmal mit **zwei unbekannten Parametern**. Die Berechnung der Teststatistik ist in Tabelle 6.6 dargestellt und folgt dem üblichen Schema, nur dass sich die Klassenwahrscheinlichkeiten diesmal aus der Normalverteilung ergeben.

l	K_l	$y_{n;l}$	$\hat{p}_l$	$\hat{e}_l = n\hat{p}_l$	$(y_{n;l} - \hat{e}_l)^2/\hat{e}_l$
1	$(-\infty, 0.86]$	5	0.0485	4.85	0.0046
2	$(0.86, 0.88]$	19	0.1634	16.34	0.4330
3	$(0.88, 0.90]$	30	0.3120	31.20	0.0462
4	$(0.90, 0.92]$	25	0.2973	29.73	0.7525
5	$(0.92, 0.94]$	15	0.1413	14.13	0.0536
6	$(0.94, \infty)$	6	0.0375	3.75	1.3500
$\sum$		100	1	100	2.64

Tabelle 6.6: Schrittweise Berechnung der χ^2-Teststatistik für den Apfelvorlauf

Beispielsweise berechnet sich die Klassenwahrscheinlichkeit für die zweite Klasse als

$$\begin{aligned}\hat{p}_2 = \mathrm{P}(0.86 < X \leq 0.88 \,|\, \hat{\mu}, \hat{\sigma}) = \\ = \quad & \Phi\left(\frac{0.88 - 0.8987}{0.02325}\right) - \Phi\left(\frac{0.86 - 0.8987}{0.02325}\right) = \\ = \quad & \Phi(-0.80) - \Phi(-1.66) = 0.2119 - 0.0485 = 0.1634\end{aligned}$$

Φ bedeutet dabei wie üblich die Verteilungsfunktion der Standardnormalverteilung. Klassenzusammenlegungen sind diesmal nicht notwendig, und es ergibt sich mit $g = 6$ Klassen und $r = 2$ geschätzten Parametern:

$$t = 2.64 < 7.815 = \chi^2_{g-r-1;1-\alpha} = \chi^2_{3;0.95}$$

Die Nullhypothese wird daher beibehalten.

Lösung 6.7
Ertrag der Maissorte Figaro

a) **Mittelwert** und **Standardabweichung**:

$$\bar{x} = 105.2 \qquad s = 11.01$$

b) **Kolmogorov-Smirnov Einstichprobentest**: Die Nullhypothese lautet: *Die Daten stammen aus einer Normalverteilung mit Mittel* $\mu = 105$ *und Standardabweichung* $\sigma = 11$.

Die **Teststatistik** t beruht auf dem maximalen Absolutabstand d zwischen der empirischen Verteilungsfunktion $\hat{F}_n$ und der theoretischen Verteilung F_0 aus der Nullhypothese. Die Werte der $\hat{F}_n$ ergeben sich wie gehabt, in dem man die Beobachtungen der Größe nach sortiert und der k-ten Beobachtung den Wert k/n zuweist:

$$\hat{F}_n\left(x_{(k)}\right) = \frac{k}{n}$$

Die Werte der F_0 ergeben sich gerade aus der Normalverteilungstabelle:

$$F_0\left(x_i\right) = \Phi\left(\frac{x_i - \mu}{\sigma}\right) = \Phi\left(\frac{x_i - 105}{11}\right)$$

In Tabelle 6.7 finden sich die sortierten (Spalte 2) und standardisierten (Spalte 3) Beobachtungen sowie die zugehörigen Werte der beiden Verteilungsfunktionen (Spalte 4: $F_0\left(x_{(i)}\right)$, Spalte 5: $\hat{F}_n\left(x_{(i)}\right)$). Die letzte Spalte enthält gerade die Differenz zwischen Spalten 4 und 5, die vorletzte Spalte die Differenz zwischen Spalte 4 und Spalte 5 der vorhergehenden Zeile, wobei man sich im Stillen für die erste Zeile noch eine zusätzliche 0 in Spalte 5 dazudenkt. Warum es notwendig ist, die Differenz zwischen der theoretischen und der empirischen Verteilungsfunktion an zwei Stellen zu bilden, kann man in Abbildung 6.6 sehen: da die empirische VF eine Stufenfunktion ist und "springt", muss man den Abstand zur stetigen Kurve der theoretischen VF an beiden Enden der Stufe betrachten, was durch die doppelte Differenzbildung in der Tabelle geschieht.

i	$x_{(i)}$	$\frac{x_{(i)}-105}{11}$	$F_0\left(x_{(i)}\right)$	$\hat{F}_n\left(x_{(i)}\right)$	d_u	d_o
1	87	-1.64	0.0505	0.10	0.0505	-0.0495
2	95	-0.91	0.1814	0.20	0.0814	-0.0186
3	98	-0.64	0.2611	0.30	0.0611	-0.0389
4	99	-0.55	0.2912	0.40	-0.0088	-0.1088
5	105	0.00	0.5000	0.50	0.1000	0.0000
6	107	0.18	0.5714	0.60	0.0714	-0.0286
7	108	0.27	0.6064	0.70	0.0064	-0.0936
8	110	0.45	0.6736	0.80	-0.0264	-0.1264
9	121	1.45	0.9265	0.90	0.1265	0.0265
10	122	1.55	0.9394	1.00	0.0394	-0.0606

Tabelle 6.7: Kolmogorov-Smirnov Einstichprobentest für Maiserträge

Die maximale (absolute) Differenz ergibt sich in Zeile 9, und damit berechnet sich die **Teststatistik** als

$$t = \sqrt{n} \cdot d = \sqrt{10} \cdot 0.1265 = 0.40$$

Der **kritische Wert** $c_o = d^{(1)}_{10;0.95} = 1.29$ findet sich in Tabelle A.6 des Vorlesungsskriptums. Wegen $t = 0.40 < 1.29 = c_o$ wird also die **Nullhypothese beibehalten**.

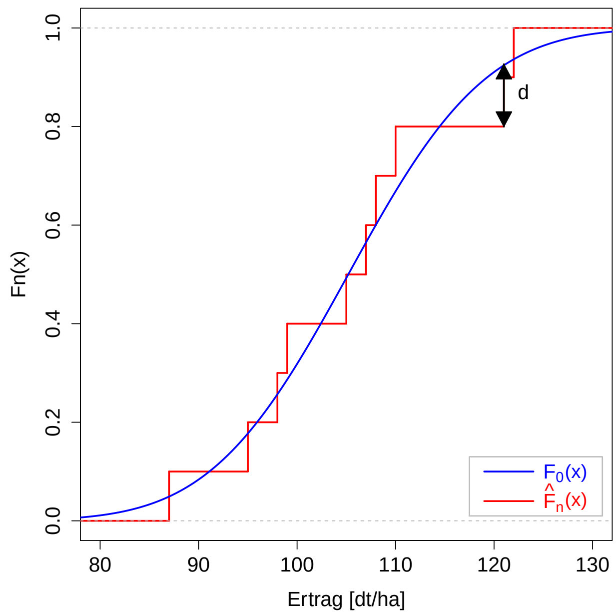


Abbildung 6.6: Verteilungsfunktionen für den Maisertrag.

Bemerkung: Ganz lupenrein ist die Verwendung des Kolmogorov-Smirnov Tests in dieser Weise nicht, da die Parameter aus den Daten geschätzt werden. Tatsächlich müsste je nach angenommener Verteilung eine Korrektur vorgenommen werden, damit der Test Gültigkeit hat. Im Fall der Normalverteilung ist dies die *Lilliefors-Korrektur*.

Lösung 6.8
Rotschwingelsorten, Ertragswerte nicht normalverteilt

Der **Kruskal-Wallis (KW) Test** ist eine Verallgemeinerung des Wilcoxon-Tests für mehr als zwei Stichproben. Mit dem KW-Test kann sogar eine etwas allgemeinere Hypothese als für die einfache Varianzanalyse getestet werden, nämlich, dass die Verteilung des betrachteten Merkmals (hier: die Ertragswerte) für alle Gruppen (hier gegeben durch die drei Rotschwingelsorten) gleich ist, oder anders gesagt, dass die Rotschwingelsorte keinen Einfluss auf den Ertrag hat.

Wie beim Wilcoxon-Test werden zunächst alle Messwerte in einer gemeinsamen Stichprobe angeordnet, die Ränge vergeben, und für jede Gruppe die Summe dieser Ränge gebildet. Tabelle 6.8 erhält die Ergebnisse dieses Schritts. Für Werte, die mehrmals vorkommen (hier: 260 dreimal, 265 und 270 jeweils zweimal), werden wieder die durchschnittlichen Rangzahlen vergeben, also $(3+4+5)/3 = 4$ für 260 und $(6+7)/2 = 6.5$ für 265.

	Futuro	Condor	Echo	Futuro.r	Condor.r	Echo.r
1			255			1.0
2	256			2.0		
3	260			4.0		
4	260			4.0		
5			260			4.0
6	265			6.5		
7			265			6.5
8			269			8.0
9		270			9.5	
10			270			9.5
11	271			11.0		
12		279			12.0	
13		280			13.0	
14		284			14.0	
15		295			15.0	
$\sum$				27.5	63.5	29

Tabelle 6.8: Berechnung der Rangsummen für den Kruskal-Wallis Test für drei verschiedene Rotschwingelsorten

Die **Teststatistik** beruht nun im wesentlichen auf der mittleren quadrierten Rangsumme der beteiligten Gruppen:

$$t = \frac{1}{b}\left(\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{r_{i.}^2}{n_i} - 3(n+1)\right)$$

Dabei ist $k = 3$ die Anzahl von Gruppen, die n_i die Anzahl von Beobachtungen in der i-ten Gruppe und die $r_{i.}$ die Rangsumme für die i-te Gruppe; $n = 15$ ist die Anzahl aller Beobachtungen, und b ist ein Korrekturfaktor, der wegen der mehrfach auftretenden Werte notwendig ist. Gilt nun die Nullhypothese gleicher Verteilungen in allen Gruppen, dann sollten die Rangsummen $r_{i.}$ der Gruppen annähernd gleich groß sein, was bewirkt, dass die Teststatistik nicht zu groß wird; liegen hingegen etwa in einer Gruppe bevorzugt große Werte, so wird sich dies in einer großen Rangsumme für diese Gruppe und damit einem höheren Wert für die Teststatistik niederschlagen.

Zur Berechnung benötigen wir zunächst den **Korrekturfaktor** b:

$$b = 1 - \frac{1}{n^3 - n} \sum_{l=1}^{g} \left(t_l^3 - t_l\right)$$

Dabei ist g die Anzahl von mehrfach auftretenden Werten und t_l die jeweilige 'Vielfachheit' des Wertes; im vorliegenden Fall etwa ist $g = 3$, und die Vielfachheiten sind $t_1 = 3$ (für 260), $t_2 = 2$ (für 265) und $t_3 = 2$ (für 270). Damit ergibt sich b zu

$$\begin{aligned} b &= 1 - \frac{1}{15^3 - 15} \left[\left(3^3 - 3\right) + \left(2^3 - 2\right) + \left(2^3 - 2\right)\right] = \\ &= 1 - \frac{1}{3360} [24 + 6 + 6] = 1 - \frac{36}{3360} = 0.9893 \end{aligned}$$

Diesen Wert können wir nun in der Teststatistik einsetzen:

$$\begin{aligned} t &= \frac{1}{0.9893} \cdot \left[\frac{12}{15 \cdot (15+1)} \cdot \left(\frac{27.5^2}{5} + \frac{63.5^2}{5} + \frac{29^2}{5}\right) - 3 \cdot (15+1)\right] = \\ &= \frac{1}{0.9893} \cdot \left[\frac{12}{1200} \cdot (756.25 + 4032.25 + 841) - 48\right] = 8.38 \end{aligned}$$

Wie bereits angedeutet, sprechen nur große Werte gegen die Nullhypothese, wir benötigen also nur eine obere kritische Schranke c_o. Diese kann für drei Gruppen aus Tabelle A.10 des Vorlesungsskriptums abgelesen werden:

$$c_o = h_{k;(n_1,n_2,n_3);1-\alpha} = h_{3;(5,5,5);0.95} = 5.66$$

Da die Teststatistik mit $t = 8.38$ über diesem Wert liegt, wird die **Nullhypothese verworfen**: Es gibt also einen signifikanten (auf dem Niveau $\alpha = 0.05$) Einfluss der Rotschwingelsorte auf den Ertrag.

Lösung 6.9
Müslitrockensubstanzgehalt, Gehaltswerte nicht normalverteilt

Hier ist die Hypothese zu testen, dass sich die Müslisorten bezüglich ihres Trockensubstanzgehalts nicht unterscheiden ($\alpha = 0.05$). Für den **Kruskal-Wallis-Test** berechnet man zunächst wie in Beispiel 6.8 die Ränge der Beobachtungen in der kombinierten Stichprobe und die Rangsummen für jede Gruppe (Tabelle 6.9).
Es gibt in diesem Datensatz zwei Werte (93.1 und 95.5), die jeweils zweimal auftreten, allen anderen Werte kommen unrepliziert vor. Wir können also sofort zur **Berechnung des Korrekturterms** für die Teststatistik schreiten:

$$b = 1 - \frac{1}{14^3 - 14} \cdot \left((2^3 - 2) + (2^3 - 2)\right) = 1 - 12/2730 = 0.9956$$

	M1	M2	M3	M1.r	M2.r	M3.r
1	90.1			1.0		
2	90.8			2.0		
3	91.3			3.0		
4	92.5			4.0		
5		93.1			5.5	
6			93.1			5.5
7		94.1			7.0	
8		94.3			8.0	
9			94.8			9.0
10		94.9			10.0	
11			95.1			11.0
12		95.5			12.5	
13			95.5			12.5
14			96.1			14.0
$\sum$				10	43	52

Tabelle 6.9: Berechnung der Rangsummen für den Kruskal-Wallis Test für drei verschiedene Müslisorten

Einsetzen des Korrekturterms und der Rangsummen in die **Teststatistik** ergibt

$$t = \frac{1}{0.9956} \cdot \left[\frac{12}{14 \cdot (14+1)} \cdot \left(\frac{10^2}{4} + \frac{43^2}{5} + \frac{52^2}{5} \right) - 3 \cdot (14+1) \right] = 8.5$$

Der **kritische Wert** c_o stammt wieder aus Tabelle A.10 des Vorlesungsskriptums:

$$c_o = h_{14;(4,5,5);0.95} = 5.64$$

Da die Teststatistik mit $t = 8.5$ darüber liegt, wird die **Nullhypothese verworfen**: Die Müslisorten unterscheiden sich signifikant(auf dem Niveau $\alpha = 0.05$) bezüglich ihres Trockensubstanzgehalts.

Kapitel 7

Kontingenztafeln

Beispiele

Beispiel 7.1 Ein Forstbetrieb untersucht an vier verschiedenen Standorten die Verjüngung seines Waldbestandes. Dabei werden als Standort Beschattungsflächen unterschiedlicher Intensität (ausgedrückt in Prozent) gewählt. Bei der Untersuchung werden alle Jungexemplare der Baumarten Fichte, Buche und Bergahorn, die größer als 15 cm sind, erfasst. Überprüfen Sie mittels eines geeigneten Tests, ob sich die Standorte bezüglich ihres verjüngten Baumbestandes unterscheiden ($\alpha = 0.05$).

	Beschattung in %			
Baumarten	20-40	41-60	61-80	81-100
Fichte	35	22	65	90
Buche	90	106	98	155
Bergahorn	32	27	32	18

Beispiel 7.2 Bei einer Umfrage bezüglich Urlaub am Bauernhof werden Informationen betreffend Anbieter, Nachfrager, Angebot, Preise, Investitionen usw. erhoben. Unter anderem werden auch die Anzahl der Betten und die Bettenauslastung (in %) pro Betrieb festgestellt.

	Bettenauslastung				
Bettenanzahl	bis 20%	21-40%	41-60%	61-80%	81-100%
bis 5	8	19	30	20	5
6 bis 10	6	11	27	38	20
über 10	7	8	14	25	12

Testen Sie, ob zwischen Bettenanzahl und Bettenauslastung ein Zusammenhang besteht ($\alpha = 0.05$).

Beispiel 7.3 Eine Handelskette will für die Planung ihrer Logistik wissen, ob die Wohnlage (ländlich/städtisch) mit der Verwendung von Butter oder Margarine als bevorzugtem Brotaufstrich in Zusammenhang steht. Zur Klärung dieser Frage werden 181 zufällig ausgewählte Personen befragt.

Wohnort	bevorzugter Brotaufstrich	
	Margarine	Butter
ländlich	23	45
städtisch	83	30

Testen Sie, ob ein solcher Zusammenhang besteht ($\alpha = 0.05$).

Beispiel 7.4 Eine Konsumentenschutzorganisation will durch eine Marktuntersuchung klären, ob ein Zusammenhang zwischen der Verkaufsform (Wochenmarkt, Einzelhändler und Lebensmittelkette) und der Qualität (in drei Abstufungen) von Frischobst besteht, und führt zu diesem Zweck 434 Testkäufe durch.

Qualität	Verkaufsform		
	Wochenmarkt	Lebensmittelhändler	Supermarkt
gut	65	69	30
mittel	27	82	73
schlecht	33	13	42

Testen Sie über eine geeignete Hypothese, ob die drei Verkaufsformen gleiche Qualität anbieten ($\alpha = 0.05$).

Beispiel 7.5 In zwei österreichischen Bundesländern wurden Schweinefleischproben auf Antibiotika untersucht. Während im ersten Bundesland 6 von 34 Stichproben Antibiotikaspuren aufwiesen, waren es im zweiten 3 von 46. Testen Sie auf dem Niveau $\alpha = 0.05$, ob der Anteil beanstandeter Proben in beiden Bundesländern gleich ist.

Lösungen

Lösung 7.1
Zusammenhang Standort/verjüngter Baumbestand

Das naheliegendste Verfahren für Häufigkeitstabellen wie die vorliegende und die Fragestellung nach dem Zusammenhang zwischen Spalten- und Zeilenmerkmal ist der χ^2-Unabhängigkeitstest, der im Prinzip gerade der χ^2-Anpassungstest aus Kapitel 6 in anderem Gewande ist. Die **Nullhypothese** lautet hier, dass es **keinen Zusammenhang** zwischen Beschattung und der Zusammensetzung des verjüngten Baumbestandes gibt, d.h., dass der Split zwischen den Arten Fichte, Buche und Bergahorn unabhängig vom Beschattungsgrad ist. Zur Berechnung der Teststatistik verschaffen wir uns zunächst die Zeilen- und Spaltenhäufigkeiten:

35	22	65	90	212
90	106	98	155	449
32	27	32	18	109
157	155	195	263	**770**

Damit können wir mit folgender Formel wie gewohnt die unter Annahme der Nullhypothese zu erwartenden Zellenhäufigkeiten schätzen:

$$\hat{e}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

Für die Zelle (1,1) (Fichte, 20-40% Beschattung) ergibt sich also:

$$\hat{e}_{11} = \frac{n_{1.} \cdot n_{.1}}{n} = \frac{212 \cdot 157}{770} = 43.2$$

Für alle Zellen ergeben sich damit folgende – unter der Nullhypothese zu erwartende – geschätzte Häufigkeiten:

43.2	42.7	53.7	72.4
91.5	90.4	113.7	153.4
22.2	21.9	27.6	37.2

Ein Kontrollblick zeigt, dass hier alle **erwarteten Häufigkeiten größer als fünf** sind, wir uns also vom approximativen Vorliegen einer χ^2-Verteilung der Teststatistik unter der Nullhypothese ausgehen können. Die **Teststatistik** berechnet sich wieder wie für den χ^2-Anpassungstest:

$$\begin{aligned} t = \sum_{i,j} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = \\ &= \frac{(35-43.2)^2}{43.2} + \frac{(22-42.7)^2}{42.7} + \ldots + \frac{(32-27.6)^2}{27.6} + \frac{(18-37.2)^2}{37.2} \\ &= 1.6 + 10.0 + 2.4 + 4.3 + 0.0 + 2.7 + 2.2 + 0.0 + 4.3 + 1.2 + 0.7 + 9.9 \\ &= 39.3 \end{aligned}$$

Der **kritische Wert** für das gegebene $\alpha = 0.05$ ergibt sich aus einer χ^2-Verteilung mit $\mathrm{df} = (r-1)(c-1) = 2 \cdot 3 = 6$ Freiheitsgraden:

$$c_o = \chi^2_{(r-1)(c-1);1-\alpha} = \chi^2_{6;0.95} = 12.592 < t = 39.3$$

Die **Nullhypothese** wird also **abgelehnt**, es gibt einen signifikanten Zusammenhang zwischen Beschattungsgrad und Zusammensetzung des nachwachsenden Jungwaldes.

Lösung 7.2
Zusammenhang Bettenanzahl und -auslastung

Als Verfahren ist wieder ein χ^2-Unabhängigkeitstest angebracht. Die **Nullhypothese** lautet hier, dass **kein Zusammenhang** zwischen **Bettenauslastung** und **Bettenanzahl** der Betriebe besteht. Die Zwischenschritte zur Berechnung der Teststatistik sind als Tabelle zusammengefasst: der großgeschriebene Eintrag ist die beobachtete Häufigkeit, der links darunter stehende Wert ist die geschätzte Häufigkeit; der rechte untere Wert in jeder Zelle ist der entsprechende Beitrag zur Teststatistik, also gerade der Ausdruck $(n_{ij} - \hat{e}_{ij})^2/\hat{e}_{ij}$

<table>
<tr><td colspan="2">8</td><td colspan="2">19</td><td colspan="2">30</td><td colspan="2">20</td><td colspan="2">5</td><td rowspan="2">82</td></tr>
<tr><td>6.9</td><td>0.18</td><td>12.5</td><td>3.427</td><td>23.3</td><td>1.935</td><td>27.2</td><td>1.917</td><td>12.1</td><td>4.196</td></tr>
<tr><td colspan="2">6</td><td colspan="2">11</td><td colspan="2">27</td><td colspan="2">38</td><td colspan="2">20</td><td rowspan="2">102</td></tr>
<tr><td>8.6</td><td>0.77</td><td>15.5</td><td>1.308</td><td>29</td><td>0.134</td><td>33.9</td><td>0.505</td><td>15.1</td><td>1.593</td></tr>
<tr><td colspan="2">7</td><td colspan="2">8</td><td colspan="2">14</td><td colspan="2">25</td><td colspan="2">12</td><td rowspan="2">66</td></tr>
<tr><td>5.5</td><td>0.382</td><td>10</td><td>0.412</td><td>18.7</td><td>1.201</td><td>21.9</td><td>0.435</td><td>9.8</td><td>0.51</td></tr>
<tr><td colspan="2">21</td><td colspan="2">38</td><td colspan="2">71</td><td colspan="2">83</td><td colspan="2">37</td><td>250</td></tr>
</table>

Als Summe ergibt sich der Wert $t = 18.9$. Da alle erwarteten Häufigkeiten größer als fünf sind, können wir unter der Nullhypothese annehmen, dass t einer χ^2-Verteilung mit df $= (c-1)(r-1) = 4 \cdot 2 = 8$ Freiheitsgraden folgt:

$$c_o = \chi^2_{(r-1)(c-1);1-\alpha} = \chi^2_{8;0.95} = 15.507 < t = 18.904$$

Die **Nullhypothese** wird also **abgelehnt**, es gibt einen signifikanten Zusammenhang zwischen Auslastung und Bettenanzahl.

Lösung 7.3
Zusammenhang Wohnlage und Vewendung Butter/Margarine

Die **Nullhypothese** lautet hier, dass das **Wohngebiet keinen Einfluss** auf die Verwendung von Butter vs. Margarine hat. Die Zwischenschritte zur Berechnung der Teststatistik sind wieder tabellarisch zusammengefasst:

<table>
<tr><td colspan="2">23</td><td colspan="2">83</td><td rowspan="2">106</td></tr>
<tr><td>39.8</td><td>7.107</td><td>66.2</td><td>4.277</td></tr>
<tr><td colspan="2">45</td><td colspan="2">30</td><td rowspan="2">75</td></tr>
<tr><td>28.2</td><td>10.044</td><td>46.8</td><td>6.044</td></tr>
<tr><td colspan="2">68</td><td colspan="2">113</td><td>181</td></tr>
</table>

Teststatistik und **kritischer Wert** ergeben sich wie in Beispiel 7.1 und 7.2 beschrieben:

$$c_o = \chi^2_{(r-1)(c-1);1-\alpha} = \chi^2_{1;0.95} = 3.841 < t = 27.5$$

Die **Nullhypothese** kann also **abgelehnt** werden, es besteht ein signifikanter Einfluss des Wohngebiets auf die Verwendung von Butter bzw. Margarine.
Alternativ kann man für eine 2×2 Kreuztabelle die Teststatistik auch als

$$\begin{aligned} t &= \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{.1}n_{2.}n_{.2}} \\ &= \frac{181 \cdot (23 \cdot 30 - 45 \cdot 83)^2}{106 \cdot 68 \cdot 75 \cdot 113} \\ &= \frac{1678236525}{61087800} = 27.5 \end{aligned}$$

berechnen.

Lösung 7.4
Zusammenhang zwischen Verkaufsform und Qualität

Die **Nullhypothese** lautet, dass die **Verkaufsform keinen Einfluss** auf die **Qualität** hat. Die Zwischenergebnisse zur Berechnung der Teststatistik (siehe Beispiele 7.1 und 7.2):

65		69		30		164
47.2	6.681	62	0.797	54.8	11.218	
27		82		73		182
52.4	12.326	68.8	2.543	60.8	2.445	
33		13		42		88
25.3	2.312	33.3	12.336	29.4	5.399	
125		164		145		434

Der Vergleich von **kritischem Wert** und **Teststatistik** ergibt

$$c_o = \chi^2_{(r-1)(c-1);1-\alpha} = \chi^2_{4;0.95} = 9.488 < t = 56.1$$

Die Nullhypothese kann also abgelehnt werden, die Verkaufsform hat signifikanten Einfluss auf die beobachtete Qualität des Frischobsts.

Lösung 7.5
Zusammenhang Bundesland/Anteil belasteter Proben

Die **Nullhypothese** lautet, dass es **keinen Unterschied** in der Rate von positiven Antibiotikabefunden zwischen den beiden Bundesländern gibt. Die Daten werden zunächst als Kreuztabelle angeschrieben:

	Antibiotikabefund	
Bundesland	positiv	negativ
A	6	28
B	3	43

Die Zwischenergebnisse auf dem Weg zur χ^2-Teststatistik sind (vgl. Beispiele 7.1 und 7.2):

6		28		34
3.8	1.237	30.2	0.157	
3		43		46
5.2	0.914	40.8	0.116	
9		71		80

Die erwartete Häufigkeit für die linke obere Zelle ist hier kleiner als fünf, nach der im Vorlesungsskriptum erwähnten (strengen) Regel ist also die Verwendung eines χ^2-Tests nicht ratsam.
Eine Alternative für eine solche 2×2-Tabelle ist der **exakte Test nach Fisher**. Dazu wählen wir als Teststatistik die beobachtete Häufigkeit für die linke obere Zelle der Tabelle (man könnte auch jede andere Zelle wählen, aber links oben ist traditionell), also hier $t = 6$.

Unter Annahme der (unveränderten) Nullhypothese folgt die Teststatistik einer **hypergeometrischen Verteilung** (Abschnitt 3.5.4 im Vorlesungsskriptum) mit den Parametern

$$N = n = 80 \qquad A = n_{1.} = 34 \qquad n = n_{.1} = 9$$

Die **kritischen Werte** sind also das $\alpha/2$- und $1 - \alpha/2$-Quantil dieser Verteilung:

$$c_u = h_{80,34,9;0.025} = 1 \qquad c_o = h_{80,34,9;0.975} = 7$$

Damit liegt die Teststatistik zwischen den kritischen Werten, und die **Nullhypothese** wird **beibehalten**: Es gibt keinen signifikanten Unterschied zwischen den Bundesländern bezüglich der Rate von Antibiotikafällen.

Bemerkung: Eine Tabelle der Quantile der hypergeometrischen Verteilung ist *nicht* in Ihrem Vorlesungsskript enthalten, dieses Beispiel ist daher zwar praxis-, aber nicht prüfungsrelevant.