Lösungen zu den Aufgaben im Skript

Vertiefung der Statistik

(Kapitel 16)

17. September 2014

16.1. Aufgaben KE 1

16.1.1. Punkt und Intervallschätzung

Aufgabe 1

Zur Modellierung der Daten dient eine Multinomialverteilung $M(N, (\pi_1, \ldots, \pi_k))$ mit $k = 5$ Parteien.

Kategorien:

A_1 = CDU/CSU, A_2 = SPD, A_3 = FDP, A_4 = Linkspartei, A_5 = Grüne

PARTEI	Prozent der Stimmen im Mai 2009
CDU/CSU	39%
SPD	30%
FDP	10%
Linke	8%
Grüne	9%

a) Welche Grundgesamtheit (GG) hat man mit dieser Frage im Auge?

GG: wahlberechtigte Bevölkerung in ganz Deutschland

b) Welche der Grundgesamtheitsparameter entsprechen den obigen Prozentsätzen? Schätzen Sie diese GG-Parameter.

Stichproben-Vektor (Realisation):

$x_{nj} = (0, 0, 0, 0, 1)$: Person n wählt die Grünepartei ($x_{n5} = 1$)

$x_5 = \sum_n x_{n5}$ (Zahl der Personen, die Grünepartei wählen)

$N = 1343$

$(x_1, \ldots, x_5) = (39\% \cdot 1343, 30\% \cdot 1343, 10\% \cdot 1343, 8\% \cdot 1343, 9\% \cdot 1343) = (523.77, 402.9, 134.3, 107.44, 120.87)$

$\approx (524, 403, 134, 107, 121)$ (Prozentwerte sind gerundet)

c) Betrachten Sie nur die Dichotomie

A: „Jetzige Regierungskoalition CDU/CSU/SPD“
$\bar{A}$: „Jetzige Opposition“

und geben Sie einen Punktschätzer und ein näherungsweises 95%-Vertrauensintervall für den Anteil der jetzigen Regierungskoalition an.

Kategorien:

A: „Jetzige Regierungskoalition CDU/CSU/SPD“
$\bar{A}$: „Jetzige Opposition“

$p_A = \hat{\pi}_A = 39\% + 30\% = 69\% = 0.69$

Standardfehler

$\text{Var}[\hat{\pi}_A] = \frac{1}{N}\pi_A(1 - \pi_A)$

Schätzung:

$\widehat{\text{Var}}[\hat{\pi}_A] = \frac{1}{N}p_A(1 - p_A) = \frac{1}{1343}0.69(1 - 0.69) = 0.00015927$

$\sqrt{\widehat{\text{Var}}[\hat{\pi}_A]} = 0.0126$

95%-Vertrauensintervall:

$p_A \pm 1.96\sqrt{\widehat{\text{Var}}[\hat{\pi}_A]} = 0.69 \pm 1.96 \cdot 0.0126 = 0.69 \pm 0.0247 = (0.6653; 0.7147)$

d) Welche Näherungen muss man bei der Berechnung des Vertrauensintervalls in c) machen?

z-Quantil für das 95%-Vertrauensintervall nähert sich 2.

$z(1 - 0.05/2) = z(0.975) = 1.96 \approx 2$

e) Ist die Aussage:

Der Fehlerbereich beträgt bei einem Parteianteil von 40 Prozent rund +/- drei Prozentpunkte und bei einem Parteianteil von 10 Prozent rund +/- zwei Prozentpunkte.

tatsächlich korrekt?

Für $\pi = 40\%$

$\widehat{\text{Var}}[\hat{\pi}] = \frac{1}{1343}0.4(1 - 0.4) = 0.0001787$

2 Standardabweichungen (approximatives 95%-Konfidenz-Intervall):

$2 \cdot \text{std} = 2\sqrt{\widehat{\text{Var}}[\hat{\pi}]} = 2 \cdot \sqrt{0.0001787} = 0.0267 = 2.67\% \approx 3\%$

Für $\pi = 10\%$

$\widehat{\text{Var}}[\hat{\pi}] = \frac{1}{1343}0.1(1 - 0.1) = 0.000067$

2 Standardabweichungen (approximatives 95%-Konfidenz-Intervall):

$2 \cdot \text{std} = 2\sqrt{\widehat{\text{Var}}[\hat{\pi}]} = 2 \cdot \sqrt{0.000067} = 0.0163 = 1.63\% \approx 2\%$

Dies entspricht etwa den Angaben aus dem Text: *Der Fehlerbereich beträgt bei einem Parteianteil von 40 Prozent rund +/- drei Prozentpunkte und bei einem Parteianteil von 10 Prozent rund +/- zwei Prozentpunkte.*

Aufgabe 2

Gegeben sei eine GG, in der jedes Element die Eigenschaft A oder $\bar{A}$ besitzt, mit den Wahrscheinlichkeiten $P(A) = P(\bar{A}) = 1/2$. Die relative Häufigkeit vom Umfang N ist ein Schätzer für $P(A)$. In Abhängigkeit vom Stichprobenumfang bezeichnen wir sie mit $f_N(A)$.

a) Berechnen Sie im Modell mit Zurücklegen die Wahrscheinlichkeit dafür, dass die relative Häufigkeit $f_N(A)$ die zu schätzende Wahrscheinlichkeit $P(A)$ genau trifft, und zwar für die Stichprobenumfänge $N = 2$, $N = 10$, $N = 15$.

Mit $N = 2$ liegt eine Binomialverteilung $B(2, 0.5)$ vor. Die relative Häufigkeit $f_2(A)$ trifft die Wahrscheinlichkeit $P(A) = 1/2$ genau dann, wenn in der Stichprobe 1 Element mit der Eigenschaft A sind, d.h. $P(X = 1) = 0.5$ ist die gesuchte Wahrscheinlichkeit.

Mit $N = 10$ liegt eine Binomialverteilung $B(10, 0.5)$ vor. Die relative Häufigkeit $f_2(A)$ trifft die Wahrscheinlichkeit $P(A) = 1/2$ genau dann, wenn in der Stichprobe 5 Elemente mit der Eigenschaft A sind, d.h. $P(X = 5) = 0.2461$ ist die gesuchte Wahrscheinlichkeit.

Für $N = 15$ nimmt die gesuchte Wahrscheinlichkeit den Wert 0 an, da für eine ungerade Anzahl die relative Häufigkeit die Werte 7/15 oder 8/15 annehmen kann aber nie genau 1/2.

b) In einer Stichprobe vom Umfang $N = 30$ hatten 12 Elemente die Eigenschaft A. Geben Sie ein 95%-Konfidenzintervall für $P(A)$ an.

$\hat{\pi} = \frac{12}{30} = 0.4$, $N\pi = 15 \geq 5$, $N(1-\pi) = 15 \geq 5$, $\alpha = 0.05$.

Als Schätzer für die Varianz ergibt sich $\frac{1}{N}\hat{\pi}(1-\hat{\pi}) = \frac{1}{30}\frac{12}{30}(1-\frac{12}{30}) = 0.008$.

Als Schätzer für die Standardabweichung ergibt sich $\sqrt{\text{Var}} = \sqrt{0.008} = 0.089$

$$\begin{aligned} KI_{zweiseitig} &= \left[\, \hat{\pi} - z\sqrt{\frac{1}{N}\hat{\pi}(1-\hat{\pi})}\,,\ \hat{\pi} + z\sqrt{\frac{1}{N}\hat{\pi}(1-\hat{\pi})}\, \right] \\ &= [0.4 - 1.96 \cdot 0.089\,,\ 0.45 + 1.96 \cdot 0.089] \\ &= [0.4 - 0.174\,,\ 0.45 + 0.174] = [0.226\,,\ 0.574] \\ &\quad (z = z(1-\alpha/2) = z(0.975)) \end{aligned}$$

$$\begin{aligned} KI_{einseitig} &= \left[\, \hat{\pi} - z\sqrt{\frac{1}{N}\hat{\pi}(1-\hat{\pi})}\,,\ \infty \right] \\ &= [0.4 - 1.65 \cdot 0.089\,,\ \infty] \\ &= [0.253\,,\ \infty] \\ &\quad (z = z(1-\alpha) = z(0.95)) \end{aligned}$$

$$\begin{aligned} KI_{einseitig} &= \left[\, -\infty\,,\ \hat{\pi} + z\sqrt{\frac{1}{N}\hat{\pi}(1-\hat{\pi})}\, \right] \\ &= [-\infty\,,\ 0.4 + 1.65 \cdot 0.089] \\ &= [-\infty\,,\ 0.547] \\ &\quad (z = z(1-\alpha) = z(0.95)) \end{aligned}$$

Aufgabe 3

Eine neue Werbestrategie wird an 300 Probanden erprobt. Sie bringt 210 Erfolge, 90 Mißerfolge. Geben Sie einen Punkt- und einen 95%-Intervallschätzer für die Erfolgswahrscheinlichkeit der Werbestrategie an.

Lösung:

Als Schätzer für Mittelwert ergibt sich $\hat{\pi} = \frac{210}{(210+90)} = \frac{210}{300} = 0.7$

Als Schätzer für die Varianz ergibt sich $\frac{1}{N}\hat{\pi}(1-\hat{\pi}) = \frac{1}{300}0.7(1-0.7) = 0.0007$.

Als Schätzer für die Standardabweichung ergibt sich $\sqrt{\text{Var}} = \sqrt{0.0007} = 0.0264$

$$\begin{aligned} KI_{zweiseitig} &= \left[\hat{\pi} - z\sqrt{\frac{1}{N}\hat{\pi}(1-\hat{\pi})}\ ,\ \hat{\pi} + z\sqrt{\frac{1}{N}\hat{\pi}(1-\hat{\pi})} \right] \\ &= [0.7 - 1.96 \cdot 0.0264\ ,\ 0.7 + 1.96 \cdot 0.0264] \\ &= [0.7 - 0.05\ ,\ 0.7 + 0.05] = [0.65\ ,\ 0.75] \end{aligned}$$

Aufgabe 4

Ein stetiges, metrisch skaliertes Merkmal X (Kundenzufriedenheit) wurde an N Kunden erhoben. Es ergab sich $\bar{x} = 105$ ($\alpha = 0.05$).

a) Bestimmen Sie ein Konfidenzintervall für den Erwartungswert, wenn $N = 10$ und

   - X normalverteilt ist und $\sigma^2 = 400$
   - X normalverteilt ist mit unbekanntem $\sigma^2 (s^2 = 400)$.

   Vergleichen Sie die beiden Konfidenzintervalle.

   Lösung:

   Das Konfidenzintervall für μ , wenn die Varianz bekannt ist:

$$P\left\{ \bar{X} - z\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z\frac{\sigma}{\sqrt{N}} \right\} = 1 - \alpha,$$

   mit Quantil der Normalverteilung $z = z(1 - \alpha/2)$.

$z = z(1 - 0.05/2) = z(0.975) = 1.96.$

$$P\left\{105 - 1.96\frac{\sqrt{400}}{\sqrt{10}} \leq \mu \leq 105 + 1.96\frac{\sqrt{400}}{\sqrt{10}}\right\} = 1 - 0.05,$$
$$P\left\{92.6 \leq \mu \leq 117.4\right\} = 1 - 0.05,$$

Das Konfidenzintervall für μ, wenn die Varianz unbekannt ist:

$$P\left\{\bar{X} - t\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + t\frac{\sigma}{\sqrt{N}}\right\} = 1 - \alpha,$$

mit dem t-Quantil $t = t(1 - \alpha/2, N - 1)$.

$t = t(1 - 0.05/2, 10 - 1) = t(0.975, 9) = 2.262.$

$$P\left\{105 - 2.262\frac{\sqrt{400}}{\sqrt{10}} \leq \mu \leq 105 + 2.262\frac{\sqrt{400}}{\sqrt{10}}\right\} = 1 - 0.05,$$
$$P\left\{90.7 \leq \mu \leq 119.3\right\} = 1 - 0.05,$$

Der Unterschied in der Berechnung der Intervalle in beiden Fällen liegt nur in den für die Berechnung verwendeten Quantilen. Bei der bekannten Varianz wird das Konfidenzintervall mit dem z-Quantil berechnet, bei der unbekannten Varianz mit dem t-Quantil.

b) Bestimmen Sie Konfidenzintervalle für μ, wenn $N = 50$ und $\sigma^2 = 300$ bekannt ist, bzw. nur $s^2 = 300$ gegeben ist.

Lösung:

In großen Stichproben ($N > 30$) sind die Summen von beliebig verteilten Zufallsvariablen approximativ normalverteilt (zentraler Grenzwertsatz).

Es wird das Konfidenzintervall für μ in der großen Stichprobenumfang für die beliebig verteilte Zufallsvariable nach der folgenden Formel berechnet:

$$P\left\{\bar{X} - z\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z\frac{\sigma}{\sqrt{N}}\right\} \approx 1 - \alpha,$$

mit Quantil der Normalverteilung $z = z(1 - \alpha/2)$.

Da $N = 50 > 30$ ist, wird in den beiden Fällen (bei bekannter Varianz und unbekannter Varianz) das Konfidenzintervall für $/mu$ so berechnet:

$$P\left\{105 - 1.96\frac{\sqrt{300}}{\sqrt{50}} \leq \mu \leq 105 + 1.96\frac{\sqrt{300}}{\sqrt{50}}\right\} = 1 - 0.05,$$
$$P\left\{100.2 \leq \mu \leq 109.8\right\} = 1 - 0.05,$$

mit $z = z(1 - 0.05/2) = z(0.975) = 1.96$.

Aufgabe 5

Der AStA einer Universität will schätzen, wie viel Geld die 10000 Studenten der Universität durchschnittlich im Monat zur Verfügung haben. Dazu werden 250 zufällig ausgewählte Studenten am Haupteingang befragt. Diese Stichprobe ergibt einen Schätzwert für das mittlere Einkommen von $\bar{x} = 590$ Euro und eine Stichprobenvarianz von $\sigma^2 = 2500$ Euro^2.

a) Man berechne ein Konfidenzintervall für das durchschnittliche Einkommen der Studenten der Universität ($\alpha = 0.05$).

Lösung:
Da $N = 250 > 30$ ist, gilt der zentrale Grenzwertsatz.

$$P\left\{\bar{X} - z\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z\frac{\sigma}{\sqrt{N}}\right\} \approx 1 - \alpha,$$

mit Quantil der Normalverteilung $z = z(1 - \alpha/2)$.

$$\begin{aligned} KI_{zweiseitig} &= \left[\, \bar{X} - z\frac{\sigma}{\sqrt{N}}\,,\ \bar{X} + z\frac{\sigma}{\sqrt{N}}\,\right] \\ &= \left[590 - 1.96 \cdot \sqrt{\frac{2500}{250}}\,,\ 590 + 1.96 \cdot \sqrt{\frac{2500}{250}}\right] \\ &= [590 - 6.2\,,\ 590 + 6.2] = [583.8\,,\ 596.2] \end{aligned}$$

b) Welcher Stichprobenumfang N muss gewählt werden, um für das in a) berechnete Konfidenzintervall einen Sicherheitsgrad von 0.99 zu erreichen?

Lösung:
Bei $\alpha = 0.99$ ist das Quantil der Normalverteilung $z = z(1 - \alpha/2) = z(0.995) = 2.58$

$$\begin{aligned} KI_{zweiseitig} &= \left[\bar{X} - z\frac{\sigma}{\sqrt{N}} \, , \; \bar{X} + z\frac{\sigma}{\sqrt{N}} \right] \\ &= \left[590 - 2.58 \cdot \sqrt{\frac{2500}{N}} \, , \; 590 + 2.58 \cdot \sqrt{\frac{2500}{N}} \right] \\ &= [590 - 6.2 \, , \; 590 + 6.2] \end{aligned}$$

Daraus folgt, dass $2.58 \cdot \sqrt{\frac{2500}{N}} = 6.2$ und $N = (2.58/1.96)^2 \cdot 250 = 433$

Mit $N = 433$ und $\alpha = 0.99$ ist $\mathrm{KI}_{zweiseitig} = [583.8, 596.2]$

c) Kann man davon ausgehen, dass die für das Konfidenzintervall notwendigen Annahmen erfüllt sind?

Die Annahmen sind erfüllt. Die Berechnung der Konfidenzintervalle basiert auf der Annahme, dass die Variable in der Grundgesamtheit normalverteilt ist. Auch wenn diese Voraussetzung nicht erfüllt ist, kann die Schätzung bei hinreichend großem Stichprobenumfang (z. B. N = 100 oder größer) als gültig angesehen werden.

Aufgabe 6

Im Rahmen einer Marktforschungs-Studie wird jedem Teilnehmer ein Fragebogen vorgelegt. Bei $N = 6$ Personen ergeben sich folgende Testwerte auf den Skalen Kundenzufriedenheit und Loyalität:

Kundenzufriedenheit	46	48	31	32	43	68
Loyalität	102	106	88	90	100	122

a) Geben Sie Punkt- und Intervallschätzer für die unbekannten Populationsparameter μ und σ^2 an (jeweils für beide Variablen) ($\alpha = 0.05$).

Lösung:

$\hat{\mu} = \bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$
$S^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X})^2$

$$P\left\{\bar{X} - t\frac{S}{\sqrt{N}} \leq \mu \leq \bar{X} + t\frac{S}{\sqrt{N}}\right\} \;=\; 1-\alpha$$

$$P\left\{\frac{(N-1)S^2}{\chi^2(1-\alpha/2, N-1)} \leq \sigma^2 \leq \frac{(N-1)S^2}{\chi^2(\alpha/2, N-1)}\right\} = 1-\alpha$$

Schätzungen für die Variable Kundenzufriedenheit:

$\hat{\mu}_X = \bar{X} = \frac{1}{6}(46+48+31+32+43+68) = 44.67$

$S_X^2 = \frac{1}{6-1}((46-44.67)^2 + (48-44.67)^2 + (31-44.67)^2 + (32-44.67)^2 + (43-44.67)^2 + (68-44.67)^2) = 181.47$

Intervallschätzer für die unbekannten Parameter μ:

$t = t(1-0.05/2, 6-1) = t(0.975, 5) = 2.571.$

$$P\left\{44.67 - 2.571\frac{\sqrt{181.47}}{\sqrt{6}} \leq \mu_X \leq 44.67 + 2.571\frac{\sqrt{181.47}}{\sqrt{6}}\right\} = 1-0.05,$$
$$P\left\{30.52 \leq \mu_X \leq 58.81\right\} = 1-0.05,$$

Intervallschätzer für die unbekannten Parameter σ^2:

$\chi^2(1-\alpha/2, N-1) = \chi^2(0.975, 5) = 12.832$

$\chi^2(\alpha/2, N-1) = \chi^2(0.025, 5) = 0.831$

$$P\left\{\frac{(6-1)181.47}{12.832} \leq \sigma_X^2 \leq \frac{(6-1)181.47}{0.831}\right\} = 1 - 0.05$$

$$P\left\{70.7 \leq \sigma_X^2 \leq 1091.87\right\} = 1 - 0.05$$

Schätzungen für die Variable Loyalität:

$\hat{\mu}_Y = \bar{Y} = \frac{1}{6}(102 + 106 + 88 + 90 + 100 + 122) = 101.33$

$S_Y^2 = \frac{1}{6-1}((102 - 101.33)^2 + (106 - 101.33)^2 + (88 - 101.33)^2 + (90 - 101.33)^2 + (100 - 101.33)^2 + (122 - 101.33)^2) = 151.47$

Intervallschätzer für die unbekannten Parameter μ:

$t = t(1 - 0.05/2, 6 - 1) = t(0.975, 5) = 2.571.$

$$P\left\{101.33 - 2.571\frac{\sqrt{151.47}}{\sqrt{6}} \leq \mu_Y \leq 101.33 + 2.571\frac{\sqrt{151.47}}{\sqrt{6}}\right\} = 1 - 0.05,$$

$$P\left\{138.55 \leq \mu_Y \leq 164.39\right\} = 1 - 0.05,$$

Intervallschätzer für die unbekannten Parameter σ^2:

$\chi^2(1 - \alpha/2, N - 1) = \chi^2(0.975, 5) = 12.832$

$\chi^2(\alpha/2, N - 1) = \chi^2(0.025, 5) = 0.831$

$$P\left\{\frac{(6-1)151.47}{12.832} \leq \sigma_Y^2 \leq \frac{(6-1)151.47}{0.831}\right\} = 1 - 0.05$$

$$P\left\{59.02 \leq \sigma_Y^2 \leq 911.37\right\} = 1 - 0.05$$

b) Schätzen Sie den unbekannten Korrelationskoeffizienten ρ und berechnen Sie ein Konfidenzintervall (diskutieren Sie, ob die Annahmen erfüllt sind).

Ein Punktschätzer für den unbekannten Korrelationskoeffizienten ρ:

$$R = \frac{S(X,Y)}{S(X)S(Y)}$$

mit

$$\begin{aligned} S(X,Y) &= \frac{1}{N-1}\sum_{n=1}^{N}(X_n-\bar{X})(Y_n-\bar{Y}) \\ S(X)^2 &= \frac{1}{N-1}\sum_{n=1}^{N}(X_n-\bar{X})^2 = S(X,X) \\ S(Y)^2 &= \frac{1}{N-1}\sum_{n=1}^{N}(Y_n-\bar{Y})^2 = S(Y,Y) \end{aligned}$$

$S(X,Y) = \frac{1}{6-1}((46-44.67)(102-101.33)+(48-44.67)(106-101.33)+(31-44.67)(88-101.33)+(32-44.67)(90-101.33)+(43-44.67)(100-101.33)+(68-44.67)(122-101.33)) = 165.33$

$$R = \frac{S(X,Y)}{S(X)S(Y)} = 165.33/\sqrt{181.47\cdot 151.47} = 0.9972$$

Konfidenzintervall für ρ:

$$P\left\{\frac{e^A-1}{e^A+1} \le \rho \le \frac{e^B-1}{e^B+1}\right\} = 1-\alpha,$$

$$\begin{aligned} A &= \ln\frac{1+R}{1-R} - \frac{2z}{\sqrt{N-3}} \\ B &= \ln\frac{1+R}{1-R} + \frac{2z}{\sqrt{N-3}}. \end{aligned}$$

97.5%-Quantil: $z(0.975) = 1.96$

$$\begin{aligned} A &= \ln\frac{1+0.9972}{1-0.9972} - \frac{2\times 1.96}{\sqrt{6-3}} = \ln(713.285) - 2.2632 \\ &= 6.5699 - 2.2632 = 4.3067 \\ B &= 6.5699 + 2.2632 = 8.8331. \end{aligned}$$

$$\begin{aligned} \frac{e^A-1}{e^A+1} &= \frac{74.195-1}{74.195+1} = 0.9734 \\ \frac{e^B-1}{e^B+1} &= \frac{6857.5-1}{6857.5+1} = 0.9997 \end{aligned}$$

Daraus findet man das 95%-Konfidenzintervall (KI)

$[0.9734, 0.9997]$.

Die Annahme ist nicht erfüllt, da Fischers Z-Transformation nur für große Stichproben gilt. In unserem Beispiel ist $N = 6 < 30$.

c) Was ist der Unterschied zwischen den Populationsparametern (μ, σ^2, ρ) und den Punktschätzern?

Die Populationsparameter (μ, σ^2, ρ) sind die wahren Parameter, die fix, aber unbekannt sind. Die können nur bei einer Vollerhebung exakt bestimmt werden. Die Punktschätzer sind Zufallsvariablen, die in jeder Stichprobe einen anderen Wert annehmen können.

Aufgabe 7

An einer Stichprobe von $N = 5$ Personen wurde ein Merkmal X erhoben, von dem bekannt ist, dass es Erwartungswert $\mu = 100$ und Varianz $\sigma^2 = 100$ aufweist.

a) Welchen Erwartungswert hat die Zufallsvariable $\bar{X} = \frac{1}{N}\sum X_n$ (Stichprobenmittelwert) und welche Varianz weist die auf?
(Annahme: X_i sind voneinander unabhängig)

Erwartungswert und Varianz des Mittelwerts sind

$$\begin{aligned} E[\bar{X}] &= \frac{1}{N}\sum_{n=1}^{N} E[X_n] = \frac{1}{N}\sum_{n=1}^{N}\mu = \mu = 100 \\ \mathrm{Var}[\bar{X}] &= \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{Var}[X_n] = \frac{1}{N^2}\sum_{n=1}^{N}\sigma^2 = \frac{\sigma^2}{N} = 100/5 = 20 \end{aligned}$$

b) Skizzieren Sie die Dichtefunktion der Zufallsvariablen X_n und $\bar{X}$, wenn X_i als normalverteilt $X_n \sim N(\mu = 100, \sigma^2 = 100)$ angenommen wird.

Die Dichtefunktion von X_n ist eine Normalverteilungsdichte $\phi N(\mu = 100, \sigma^2 = 100)$

Die Dichtefunktion von $\bar{X}$ ist eine Normalverteilungsdichte $\phi(x; \mu, \frac{\sigma^2}{N})$, die ein impulsartiges Aussehen hat.

Die Dichtefunktion von X_n ist breiter als die Dichtefunktion von $\bar{X}$.

c) Wie sieht die Dichtefunktion von $\bar{X}$ aus, wenn sehr große Stichproben $N \to \infty$ gezogen werden?

In sehr großen Stichproben ist die Dichtefunktion von $\bar{X}$ eine Normalverteilungsdichte $\phi(x; \mu, \frac{\sigma^2}{N})$, die immer schmaler und höher wird. Da der Schätzer bei μ lokalisiert ist und die Varianz (Breite der Verteilung) gegen 0 geht, strebt $\bar{X}$ gegen die Konstante μ.

Aufgabe 8

a) Definieren Sie die Begriffe Grundgesamtheit, Stichprobe, Parameter, Schätzer, Schätzung, Population, Populationsparameter und grenzen Sie diese voneinander ab.

Die Grundgesamtheit (Population) ist die Menge aller relevanten Merkmalsträger. Eine Stichprobe ist eine Untermenge (Auswahl) aus der Population. Die Populationsparameter (zum Beispiel (μ, σ^2, ρ)) sind die wahren Parameter, die fix, aber unbekannt sind. Die können nur bei einer Vollerhebung exakt bestimmt werden.
Die Punktschätzer sind die Schätzer der wahren Parameter und sind Zufallsvariablen, die in jeder Stichprobe einen anderen Wert annehmen können.

b) Warum ist ein Schätzer eine Zufallsvariable? Welche Resultate erhält man für einen Schätzer (z.B. S^2), wenn die Stichprobe wiederholt gezogen wird? Ein Schätzer ist eine Zufallsvariable, weil in den verschiedenen Stichproben die einen anderen Wert annehmen kann. (Vgl. S. 43 im Skript)

Aufgabe 9

Wie lauten die kritischen Werte, oberhalb derer 5% der möglichen Werte liegen (95%-Quantil):

a) bei der Normalverteilung $N(0,1)$: $z(0.95) = 1.65$

b) bei der t-Verteilung mit $df = 27$: $t(0.95, 27) = 1.703$

c) bei der χ^2-Verteilung mit $df = 3$: $\chi^2(0.95, 3) = 7.815$

d) bei der F-Verteilung mit df (Zähler)$= 2$, df (Nenner)$= 16$: $F(0.95, 2, 16) = 3.63$

16.1.2. Tests für Anteilswerte

Aufgabe 10

Eine Brauerei produziert ein neues alkoholfreies Bier. In einem Geschmackstest erhalten 150 Personen je ein Glas alkoholfreies bzw. gewöhnliches Bier, und sie sollen versuchen, das alkoholfreie Bier zu identifizieren.

a) Das gelingt 98 Personen. Testen Sie anhand dieser Daten die Hypothese, alkoholfreies und gewöhnliches Bier seien geschmacklich nicht zu unterscheiden ($\alpha = 0.1$).

$$\begin{aligned} H_0 &: \pi \leq 0.5 = \pi_0 \\ H_1 &: \pi > 0.5 = \pi_0 \end{aligned}$$

$\hat{\pi} = \bar{X} = \dfrac{9}{15} = 0.6533$

Da $N = 150$ und $N\pi_0 = 75 > 5$ und $N(1-\pi_0) = 75 > 5$ kann die Binomialverteilung durch die Normalverteilung approximiert werden.

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1-\pi_0)}} = \frac{0.6533 - 0.5}{\sqrt{\frac{1}{150}0.5 \cdot (1-0.5)}} = 3.755$$

z-Quantil $z(0,9) = 1.29$

Da Prüfgröße $Z = 3.755$ größer als das Quantil $z(0.9) = 1.29$ ist, wird die Nullhypothese (Alkoholfreies und gewöhnliches Bier seien geschmacklich nicht zu unterscheiden) auf 10%-Niveau abgelehnt.

b) Unter den befragten Personen waren 15 Beschäftigte der Brauerei. Von diesen gelingt 9 die richtige Identifizierung. Überprüfen Sie die Hypothese aus a) für diese Subpopulation.

$$\begin{aligned} H_0 &: \pi \leq 0.5 = \pi_0 \\ H_1 &: \pi > 0.5 = \pi_0 \end{aligned}$$

$$\hat{\pi} = \bar{X} = \frac{9}{15} = 0.6$$

Da $N = 15$ und $N\pi_0 = 7.5 > 5$ und $N(1 - \pi_0) = 7.5 > 5$ kann die Binomialverteilung durch die Normalverteilung approximiert werden.

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1 - \pi_0)}} = \frac{0.6 - 0.5}{\sqrt{\frac{1}{15}0.5 \cdot (1 - 0.5)}} = 0.7746$$

z-Quantil $z(0, 9) = 1.29$

Da Prüfgröße $Z = 0.7746$ kleiner als das Quantil $z(0.9) = 1.29$ ist, kann die Nullhypothese (Alkoholfreies und gewöhnliches Bier seien geschmacklich nicht zu unterscheiden) auf 10%-Niveau nicht abgelehnt werden.

Aufgabe 11

a) In einem Land regnet es auf lange Sicht an 100 Tagen im Jahr. 2005 regnete es nur an 80 Tagen. Hat sich das Klima signifikant verändert ($(\alpha = 0.05)$, Binominaltest)?

$$\pi_0 = \frac{100}{360} = 0.2778$$

$$\begin{aligned} H_0 &: \pi \geq 0.2778 = \pi_0 \\ H_1 &: \pi < 0.2778 = \pi_0 \end{aligned}$$

$$\hat{\pi} = \frac{80}{360} = 0.2222$$

Da $N = 360$ und $N\pi_0 = 100 > 5$ und $N(1 - \pi_0) = 260 > 5$ kann die Binomialverteilung durch die Normalverteilung approximiert werden.

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1 - \pi_0)}} = \frac{0.2222 - 0.2778}{\sqrt{\frac{1}{360}0.2778(1 - 0.2778)}} = -2.3552$$

z-Quantil $z(0,05) = -z(1 - 0,05) = -1.65$

Da Betrag von der Prüfgröße $|Z| = |-2.3552| = 2.3552$ größer als $|z(0.05)| = |-1.65| = 1.65$ ist, wird die Nullhypothese (Das Klima hat sich signifikant nicht verändert) auf 5%-Niveau abgelehnt. Das Klima hat sich signifikant verändert.

b) Diskutieren Sie die Annahmen des Binominaltests für dieses Beispiel.

Ein Binomialtest basiert auf einfachen Annahmen:

- Es liegt eine Stichprobe mit n Beobachtungseinheiten vor. In unserem Beispiel liegen 360 Tage vor, an den es geregnet oder nicht geregnet hat.

- Die Stichprobenwerte sind Ausprägungen eines Alternativmerkmals. Im Beispiel lautet eine Ausprägung „Es hat an dem Tag geregnet“, und Alternativausprägung ist „Es hat an dem Tag nicht geregnet“.

Aufgabe 12

Der Hersteller eines Medikaments behauptet, dass sie in 90% der Fälle eine Allergie wirksam erleichtere. In einer Stichprobe von 200 Personen, die unter Allergien litten, brachte die Medizin 160 Personen Erleichterung. Ist die Behauptung des Herstellers berechtigt? - Man wähle $\alpha = 0.01$.

$$
\begin{aligned}
H_0 &: \pi \geq 0.9 = \pi_0 \\
H_1 &: \pi < 0.9 = \pi_0
\end{aligned}
$$

$$\hat{\pi} = \frac{160}{200} = 0.8$$

Da $N = 200$ und $N\pi_0 = 180 > 5$ und $N(1 - \pi_0) = 20 > 5$ kann die Binomialverteilung durch die Normalverteilung approximiert werden.

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1 - \pi_0)}} = \frac{0.8 - 0.9}{\sqrt{\frac{1}{200}0.9(1 - 0.9)}} = -4.7114$$

z-Quantil $z(0,01) = -z(1 - 0,01) = -2.33$

Die Prüfgröße $Z = -4.7114$ ist kleiner als $z(0.01) = -2.33$, deshalb kann die Nullhypothese (Behauptung des Herstellers Medikament erleichtert in 90 % der Fälle eine Allergie ist berechtigt) auf 1%-Niveau abgelehnt werden.

Aufgabe 13

Eine Münze wird 12 Mal geworfen. Dabei erscheint 8 Mal Kopf. Testen Sie, ob die Münze symmetrisch ist ($\alpha = 0.05$).

Wenn die Münze symmetrisch ist, musste 6 Mal Kopf und 6 Mal Zahl erscheinen und deshalb ist $\pi_0 = 0.5$.

$$
\begin{aligned}
H_0 &: \pi = 0.5 = \pi_0 \\
H_1 &: \pi \neq 0.5 = \pi_0
\end{aligned}
$$

$$\hat{\pi} = \frac{8}{12} = 0.6667$$

Da $N = 12$ und $N\pi_0 = 6 > 5$ und $N(1 - \pi_0) = 6 > 5$ kann die Binomialverteilung durch die Normalverteilung approximiert werden.

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{1}{N}\pi_0(1 - \pi_0)}} = \frac{0.6667 - 0.5}{\sqrt{\frac{1}{12}0.5(1 - 0.5)}} = 1.1547$$

z-Quantil $z(1 - \alpha/2) = z(1 - 0,05/2) = z(0.975) = 1.96$

Die Prüfgröße $Z = 1.1547$ ist kleiner als $z(0.975) = 1.96$, deshalb kann die Nullhypothese (Die Münze ist symmetrisch) auf 5%-Niveau nicht abgelehnt werden.

16.1.3. Anpassungstests/Unabhängigkeitstests

Aufgabe 14

Mendel erhielt bei einem seiner Kreuzungsversuche an Erbsenpflanzen folgende Werte:

315	runde gelbe Erbsen,
108	runde grüne Erbsen,
101	kantige gelbe Erbsen,
32	kantige grüne Erbsen.

Spricht dies für oder gegen die Theorie, dass das Verhältnis der 4 Zahlen 9:3:3:1 sein müsste ($\alpha = 0.05$)?

H_0 : $n_i = N\pi_i$für alle $i = 1, 2, 3, 4$(das Verhältnis der 4 Zahlen ist 9:3:3:1)
H_1 : $n_i \neq N\pi_i$für mindestens ein $i = 1, 2, 3, 4$

	beobachtete Häufigkeit n_i	erwartete Häufigkeit $N\pi_i$	$\frac{(n_i - N\pi_i)^2}{N\pi_i}$
runde gelbe Erbsen	315	$\frac{9}{16} \cdot 556 = 312.75$	0.016187
runde grüne Erbsen	108	$\frac{3}{16} \cdot 556 = 104.25$	0.134892
kantige gelbe Erbsen	101	$\frac{3}{16} \cdot 556 = 104.25$	0.10131894
kantige grüne Erbsen	32	$\frac{1}{16} \cdot 556 = 34.75$	0.2176259
Summe	556	556	$\chi^2 = 0.47$

Nach dem χ^2-Anpassungstest mit der Prüfgröße $\chi^2 = \sum_{i=1}^{l} \frac{(n_i - N\pi_i)^2}{N\pi_i}$ ergibt sich zum Signifikanzniveau $\alpha = 0.05$ das Quantil $\chi^2(l - 1, 1 - \alpha) = \chi^2(3, 0.95) = 7.815$. Mit $7.815 \geq 0.47 = \chi^2$ kann die Hypothese „Das Verhältnis der Zahlen ist 9:3:3:1“ nicht abgelehnt werden.

Aufgabe 15

In einer Statistikklausur erhielt man folgende Kontingenztabelle

	Note				
	1	2	3	4	5
weiblich	14	20	20	4	2
männlich	7	20	11	2	5

a) Geben Sie Schätzer an für die bedingte Wahrscheinlichkeit der Noten in den Subpopulationen und vergleichen Sie diese deskriptiv.

Es sei A=Note und B=Geschlecht . Zu Berechnen sind die bedingten Wahrscheinlichkeiten der Noten für Frauen $P(A|B = Frau)$ und für Männer $P(A|B = Mann)$.

$P(A = 1|B = Frau) = \frac{14}{14+20+20+4+2} = \frac{14}{60} = 0.2333 = 23,33\%$
$P(A = 1|B = Mann) = \frac{7}{7+20+11+2+5} = \frac{7}{45} = 0.1556 = 15,56\%$

Ähnlich kann man die bedingten Wahrscheinlichkeiten für die Noten 2, 3, 4 und 5 berechnen.

	Note				
	1	2	3	4	5
weiblich	23.33%	33.33%	33.33%	6.67%	3.33%
männlich	15.56%	44.44%	24.44%	4.44%	11.11%

b) Überprüfen Sie mit einem geeigneten Testverfahren, ob zwischen den Merkmalen Geschlecht und Klausurergebnis ein Zusammenhang besteht ($\alpha = 0.1$).

$H_0:$ $\pi_{ij} = \pi_{i.} \cdot \pi_{.j}$für alle $i = 1, 2$, $j = 1, 2, 3, 4, 5$
$H_1:$ $\pi_{ij} \neq \pi_{i.} \cdot \pi_{.j}$für mindestens ein i, j

Für den χ^2-Unabhängigkeitstest wird $\chi^2 = \sum_{i,j=1}^{I,J} \frac{(n_{ij}-N\hat{\pi}_{ij})^2}{N\hat{\pi}_{ij}}$ als Testgröße verwendet mit $N\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{N}$ ($N = 60 + 45 = 105$).

	Note					
	1	2	3	4	5	$n_{i.}$
weiblich	14	20	20	4	2	60
männlich	7	20	11	2	5	45
$n_{.j}$	21	40	31	6	7	105

	beobachtete Häufigkeit n_{ij}	erwartete Häufigkeit $N\pi_{ij}$	$\frac{(n_{ij}-N\pi_{ij})^2}{N\pi_{ij}}$
n_{11}	14	$\frac{60\cdot 21}{105} = 12$	0.3333
n_{12}	20	$\frac{60\cdot 40}{105} = 22.86$	0.3571
n_{13}	20	$\frac{60\cdot 31}{105} = 17.71$	0.2949
n_{14}	4	$\frac{60\cdot 6}{105} = 3.43$	0.0952
n_{15}	2	$\frac{60\cdot 7}{105} = 4$	1
n_{21}	7	$\frac{45\cdot 21}{105} = 9$	0.4444
n_{22}	20	$\frac{45\cdot 40}{105} = 17.14$	0.4762
n_{23}	11	$\frac{45\cdot 31}{105} = 13.29$	0.3932
n_{24}	2	$\frac{45\cdot 6}{105} = 2.57$	0.1270
n_{25}	5	$\frac{45\cdot 7}{105} = 3$	1.3333
Summe	105	105	$\chi^2 = 4.8548$

Zum Signifikanzniveau $\alpha = 0.1$ ergibt sich der obere kritische Wert zu $c_o = \chi^2(1-\alpha, (I-1)(J-1)) = \chi^2(0.9, 4) = 7.779$. Mit $\chi^2 = 4.8548 < 7.779$ kann die Nullhypothese nicht abgelehnt werden. Somit kann auf eine Abhängigkeit zwischen Merkmalen Geschlecht und Klausurergebnis nicht geschlossen werden.

Aufgabe 16

In einer Umfrage werden 50 Studenten und 50 Nichtstudenten befragt, wie viele Faschingsbälle sie besuchten. Man erhielt

Ballbesuch	viele	wenig	gar keinen	
Studenten	25	15	10	50
Nicht-Studenten	15	20	15	50
				100

Testen Sie die Hypothese, dass sich die Häufigkeit des Ballbesuchs für die beiden Gruppen nicht unterscheiden ($\alpha = 0.1$).

H_0 : $\pi_{ij} = \pi_{i.} \cdot \pi_{.j}$für alle $i = 1, 2$, $j = 1, 2, 3, 4, 5$

H_1 : $\pi_{ij} \neq \pi_{i.} \cdot \pi_{.j}$für mindestens ein i, j

Für den χ^2-Unabhängigkeitstest wird $\chi^2 = \sum_{i,j=1}^{I,J} \frac{(n_{ij} - N\hat{\pi}_{ij})^2}{N\hat{\pi}_{ij}}$ als Testgröße verwendet mit $N\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{N}$ ($N = 100$).

Ballbesuch	viele	wenig	gar keinen	$n_{i.}$
Studenten	25	15	10	50
Nicht-Studenten	15	20	15	50
$n_{.j}$	40	35	25	100

	beobachtete Häufigkeit n_{ij}	erwartete Häufigkeit $N\pi_{ij}$	$\frac{(n_{ij} - N\pi_{ij})^2}{N\pi_{ij}}$
n_{11}	25	$\frac{50 \cdot 40}{100} = 20$	1.25
n_{12}	15	$\frac{50 \cdot 35}{100} = 17.5$	0.3571
n_{13}	10	$\frac{50 \cdot 25}{100} = 12.5$	0.5
n_{21}	15	$\frac{50 \cdot 40}{100} = 20$	1.25
n_{22}	20	$\frac{50 \cdot 35}{100} = 17.5$	0.3571
n_{23}	15	$\frac{50 \cdot 25}{100} = 12.5$	0.5
Summe	105	105	$\chi^2 = 4.2142$

Zum Signifikanzniveau $\alpha = 0.1$ ergibt sich der obere kritische Wert zu $c_o = \chi^2(1-\alpha, (I-1)(J-1)) = \chi^2(0.9, 2) = 4.605$. Mit $\chi^2 = 4.2142 < 4.605$ kann die Nullhypothese (die Häufigkeit des Ballbesuchs unterscheidet sich nicht für die beiden Gruppen) nicht abgelehnt werden.

Aufgabe 17

Eine Variable X ist normalverteilt mit unbekanntem Erwartungswert μ und der bekannten Varianz $\sigma^2 = 25$. Zu prüfen ist die Nullhypothese $H_0 : \mu \geq \mu_0 = 1000$ beim Signifikanzniveau $\alpha = 0.01$ und bei einem Stichprobenumfang von $N = 64$.

a) Man ermittle die kritische Region des hier zu verwendenden Standardtests.

$$H_0 : \mu \geq \mu_0 \qquad H_1 : \mu < \mu_0$$

Der kritische Wert c_u wird mit $\alpha = 0.01$ und $z = z(0.99) = 2.33$ zu

$$c_u = \mu_0 - z \cdot \sigma_{\bar{X}} = 1000 - 2.33 \cdot \frac{5}{8} = 998.54$$

berechnet. Die Nullhypothese wird abgelehnt, falls < 998.54 gilt.

b) Man ermittle die Wahrscheinlichkeit eines Fehlers zweiter Art, wenn $\mu = \mu_1 = 998$ ist.

Es muss die Wahrscheinlichkeit berechnet werden, die Nullhypothese nicht abzulehnen, obwohl sie falsch ist unter der Voraussetzung $\mu = 998$, d.h. gesucht ist

$$\beta(\mu = 998) = P(\bar{X} \geq 998.54 | \mu = 998).$$

Dies entspricht der Wahrscheinlichkeit, dass $\bar{X}$ größer als der kritische Wert ist unter der Annahme, dass $\bar{X}$ normalverteilt ist mit Erwartungswert 998 und Standardabweichung 5/8.

$$\beta(\mu = 998) = P(\bar{X} \geq 998.54 | \mu = 998) = P(Z \geq 0.864) = 1 - P(Z \leq 0.864)$$

Laut Tabelle gilt $1 - P(Z \leq 0.86) = 0.1949$.

c) Man ermittle den Stichprobenumfang, der im vorliegenden Fall nötig ist, damit die in b) genannte Wahrscheinlichkeit höchstens 0.05 beträgt.

Zu berechnen ist N mit

$$\begin{aligned} & P(\bar{X} \geq 998.54 | \mu = 998) \leq 0.05 \\ \iff \quad & P(Z \geq \frac{0.54}{5/\sqrt{N}}) \leq 0.05 \\ \iff \quad & P(Z \leq \frac{0.54}{5/\sqrt{N}}) \geq 0.95 \end{aligned}$$

Für die Gleichheit $P(Z \leq \frac{0.54}{5/\sqrt{N}}) = 0.95$ ergibt sich somit $\frac{0.54}{5/\sqrt{N}} = 1.65$. Dies nach N aufgelöst ergibt 233.41. Somit muss der Stichprobenumfang mindestens 234 lauten.

Aufgabe 18

Bei einer Umfrage zur Kompetenzeinschätzung der Politiker A und B werden folgende Zufallsvariablen betrachtet

$$X = \begin{cases} 1 & A \text{ ist kompetent} \\ 0 & A \text{ ist nicht kompetent} \end{cases}$$

$$Y = \begin{cases} 1 & B \text{ ist kompetent} \\ 0 & B \text{ ist nicht kompetent} \end{cases}$$

Es wird eine Stichprobe von $N = 100$ befragt. 60 Personen halten A für kompetent, 40 Personen halten B für kompetent, 35 Personen halten beide für kompetent.

a) Geben Sie in einer Kontingenztafel (Kreuztabelle) die gemeinsame (absolute) Häufigkeitsverteilung der Zufallsvariablen X und Y an.

	A ist kompetent X=1	A ist nicht kompetent X=0	
B ist kompetent (Y=1)	35	5	40
B ist nicht kompetent (Y=0)	25	35	60
	60	40	100

b) Testen Sie die Hypothese der Unabhängigkeit von X und Y ($\alpha = 0.05$)

$$H_0: \quad \pi_{ij} = \pi_{i.} \cdot \pi_{.j} \text{für alle } i = 1, 2,\ j = 1, 2, 3, 4, 5$$
$$H_1: \quad \pi_{ij} \neq \pi_{i.} \cdot \pi_{.j} \text{für mindestens ein } i, j$$

Für den χ^2-Unabhängigkeitstest wird $\chi^2 = \sum_{i,j=1}^{I,J} \frac{(n_{ij} - N\hat{\pi}_{ij})^2}{N\hat{\pi}_{ij}}$ als Testgröße verwendet mit $N\hat{\pi}_{ij} = \frac{n_{i.} n_{.j}}{N}$ ($N = 100$).

	beobachtete Häufigkeit n_{ij}	erwartete Häufigkeit $N\pi_{ij}$	$\frac{(n_{ij} - N\pi_{ij})^2}{N\pi_{ij}}$
n_{11}	35	$\frac{40 \cdot 60}{100} = 24$	5.04
n_{12}	5	$\frac{40 \cdot 40}{100} = 16$	7.56
n_{21}	25	$\frac{60 \cdot 60}{100} = 36$	3.36
n_{22}	35	$\frac{60 \cdot 40}{100} = 24$	5.04
Summe	60	40	$\chi^2 = 21$

Zum Signifikanzniveau $\alpha = 0.05$ ergibt sich der obere kritische Wert zu $c_o = \chi^2(1-\alpha, (I-1)(J-1)) = \chi^2(0.95, 1) = 3.841$. Mit $\chi^2 = 21 > 3.841$ kann die Nullhypothese abgelehnt und somit auf eine Abhängigkeit von X und Y geschlossen werden.

c) Bestimmen Sie die bedingten Häufigkeiten, dass B für kompetent bzw. inkompetent gehalten wird, wenn bekannt ist, dass A für kompetent gehalten wurde.

$$P(Y = 1|X = 1) = \frac{35}{60} = 0.5833 = 58.33\%$$
$$P(Y = 0|X = 1) = \frac{25}{60} = 1 - P(Y = 1|X = 1) = 0.4167 = 41.67\%$$

16.1.4. Tests auf zentrale Tendenz

Aufgabe 19

Die Wirksamkeit einer neuen Unterrichtsmethode soll getestet werden. Kriterium sind die Punktwerte X in einem Leistungstest. Für 10 Personen erhielt man nach Unterricht mit der neuen Methode die Werte

$$50, 82, 73, 65, 64, 59, 72, 84, 69, 75$$

a) Schätzen Sie $E(X)$, $\mathrm{Var}(X)$

$$\begin{aligned}
\bar{X} &= \frac{1}{N}\sum_{n=1}^{N} X_n \\
\bar{X} &= \frac{50+82+73+65+64+59+72+84+69+75}{10} = 69.3 \\
S^2 &= \frac{1}{N-1}\sum_{n=1}^{N}(X_n-\bar{X})^2 \\
S^2 &= \frac{1}{10-1}((50-69.3)^2+(82-69.3)^2+ \\
&\quad +(73-69.3)^2+...+(75-69.3)^2) = 106.23
\end{aligned}$$

b) Bei Unterrichtung mit der alten Methode war $\mu_0 = 50$ der durchschnittliche Punktewert in der GG. Testen Sie unter Annahme, dass die Werte normalverteilt sind mit bekannter Varianz $\sigma = 9$ die Hypothese, dass die neue Methode eine Verbesserung bringt ($\alpha = 0.05$).

$H_0 : \mu < \mu_0 = 50$
$H_1 : \mu \geq \mu_0 = 50$

Die Standardabweichung σ ist bekannt und die Werte sind normalverteilt. Als Prüfgröße wird $Z = \frac{\bar{X}-\mu_0}{\sigma_{\bar{X}}}$ verwendet, welche approximativ standardnormalverteilt ist ($N = 10, \alpha = 0.05$). Mit $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{9}{\sqrt{10}} = 2.846$ ergibt sich:

$$Z = \frac{69.3-50}{2.846} = 6.781.$$

Wegen $6.781 > z(0.95) = 1.65$ kann die Nullhypothese abgelehnt werden. Die Behauptung, dass die neue Methode eine Verbesserung bringt,

wurde statistisch bewiesen.

c) Bearbeiten Sie dasselbe Problem wie unter b), ohne die Annahme, dass die Varianz bekannt ist.

$H_0 : \mu < \mu_0 = 50$
$H_1 : \mu \geq \mu_0 = 50$

Die Standardabweichung σ ist unbekannt und muss geschätzt werden. $S^2 = 106.23$ und somit $S = \sqrt{106.23} = 10.31$

Als Prüfgröße wird $T = \frac{\bar{X}-\mu_0}{S_{\bar{X}}}$ verwendet, welche approximativ standardnormalverteilt ist ($N = 10, \alpha = 0.05$). Mit $S_{\bar{X}} = \frac{S}{\sqrt{N}} = \frac{10.31}{\sqrt{10}} = 3.26$ ergibt sich:

$$T = \frac{69.3 - 50}{3.26} = 5.92.$$

Wegen $5.92 > 1.833 = t(0.95, 9) = t(1 - \alpha, N - 1)$ kann die Nullhypothese abgelehnt werden. Die Behauptung, dass die neue Methode eine Verbesserung bringt, wurde statistisch bewiesen.

d) Geben Sie für $E(X)$ ein 90%-Konfidenzintervall an

d_1) unter den Annahmen von b)
Das Konfidenzintervall für μ , wenn die Varianz bekannt ist:

$$P\left\{\bar{X} - z\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z\frac{\sigma}{\sqrt{N}}\right\} = 1 - \alpha,$$

mit Quantil der Normalverteilung $z = z(1 - \alpha/2)$.
$z = z(1 - 0.1/2) = z(0.95) = 1.65$.

$$P\left\{69.3 - 1.65\frac{9}{\sqrt{10}} \leq \mu \leq 69.3 + 1.65\frac{9}{\sqrt{10}}\right\} = 1 - 0.1,$$
$$P\left\{64.6 \leq \mu \leq 74\right\} = 1 - 0.1,$$

d_2) unter den Annahmen von c)

Das Konfidenzintervall für μ, wenn die Varianz unbekannt ist:

$$P\left\{\bar{X} - t\frac{S}{\sqrt{N}} \leq \mu \leq \bar{X} + t\frac{S}{\sqrt{N}}\right\} = 1 - \alpha,$$

mit dem t-Quantil $t = t(1 - \alpha/2, N - 1)$.

$t = t(1 - 0.01/2, 10 - 1) = t(0.95, 9) = 1.833.$

$$P\left\{69.3 - 1.833\frac{10.31}{\sqrt{10}} \leq \mu \leq 69.3 + 1.833\frac{10.31}{\sqrt{10}}\right\} = 1 - 0.1,$$
$$P\left\{63.325 \leq \mu \leq 75.274\right\} = 1 - 0.1,$$

und vergleichen Sie die beiden Konfidenzintervalle.

Die Konfidenzintervalle unterscheiden sich voneinander in den für die Berechnung verwendeten Quantilen und den Varianzwerten. Bei der bekannten Varianz wird das Konfidenzintervall mit dem z-Quantil und der angegebenen bekannten Varianz berechnet, bei der unbekannten Varianz mit dem t-Quantil und der geschätzten Varianz. Bei der bekannten Varianz ist das KI schmaler, als KI bei der unbekannten Varianz, da es für die Berechnung mehr Information vorliegt und deshalb ist das KI genauer.

Aufgabe 20

Bei empirischen Untersuchungen wurde festgestellt, dass das Merkmal X „Körpergröße (gemessen in cm)“ approximativ normalverteilt ist. In einer Stichprobe von $N = 26$ ergab sich ein Mittelwert von $\bar{X} = 181$ und für die Stichprobenvarianz $S^2 = 5.1^2$.

a) Man prüfe die Hypothese, die durchschnittliche Körpergröße betrage höchstens 180 cm unter der Voraussetzung, dass das Merkmal zwar normalverteilt, die Varianz jedoch nicht bekannt ist ($\alpha = 0.05$).

$H_0 : \mu \leq \mu_0 = 180$
$H_1 : \mu > \mu_0 = 180$

Die Standardabweichung $S = 5.1$.

Als Prüfgröße wird $T = \frac{\bar{X}-\mu_0}{S_{\bar{X}}}$ verwendet, welche approximativ standardnormalverteilt ist ($N = 26, \alpha = 0.05$). Mit $S_{\bar{X}} = \frac{S}{\sqrt{N}} = \frac{5.1}{\sqrt{26}} = 1$ ergibt sich:

$$T = \frac{181 - 180}{1} = 1$$

Wegen $1 < 1.708 = t(0.95, 25) = t(1 - \alpha, N - 1)$ kann die Nullhypothese nicht abgelehnt werden. Die Behauptung, dass durchschnittliche Körpergröße betrage höchstens 180 cm, kann nicht statistisch abgelehnt werden.

b) Wie wirken sich eine Verringerung des Signifikanzniveaus α bzw. eine Vergrößerung des Stichprobenumfangs N auf die Wahrscheinlichkeit eines Fehlers zweiter Art bei einem statistischen Test aus (kurze Begründung erforderlich)?

Die Verkleinerung des Fehlers 1. Art (kleines α) führt zu einer Vergrößerung der kritischen Werte $\mu_0 \pm z(1-\alpha/2) \times \sigma/\sqrt{N}$ und somit bei gleicher Konstellation zu einem vergrößerten Fehler 2. Art (H_0 wird also eher beibehalten).

Die Vergrößerung des Stichprobenumfangs N führt zu einer Verkleinerung der kritischen Werte $\mu_0 \pm z(1-\alpha/2) \times \sigma/\sqrt{N}$ und somit bei gleicher Konstellation zu einem verkleinerten Fehler 2. Art (H_0 wird also eher abgelehnt). Dies ist auch sinnvoll, da ja Abweichungen der Daten vom hypothetischen Wert bei großen Stichproben schneller detektiert werden sollten

Aufgabe 21

Angenommen, ein Untersucher zieht 3 Hypothesen in Erwägung. Nur eine der 3 kann wahr sein:

$$\begin{aligned} H_0: &\quad \mu = 200 \\ H_1: &\quad \mu = 210 \\ H_2: &\quad \mu = 220 \end{aligned}$$

Er weiß, dass die Population normalverteilt ist mit $\sigma_X = 20$. Er zieht eine Zufallsstichprobe vom Umfang $N = 25$ und berechnet $\bar{X}$. Dann formuliert er die Entscheidungsregel: Wenn $\bar{X} \leq 205$ nehme H_0 an; wenn $205 < \bar{X} < 215$ nehme H_1 an; wenn $215 \geq \bar{X}$ nehme H_2 an. Finden Sie die Wahrscheinlichkeiten einer richtigen bzw. falschen Entscheidung unter dieser Regel.

Es muss die Wahrscheinlichkeit berechnet werden, die Nullhypothese nicht abzulehnen, obwohl sie falsch ist, für jeden einzelnen Fall.

Wenn $\bar{X} \leq 205$ ist, wird H_0 ($\mu = 200$) angenommen.

In diesem Fall werden β_{01} (die Wahrscheinlichkeit H_0 beibehalten, obwohl H_1 richtig ist) und β_{02} (die Wahrscheinlichkeit H_0 beibehalten, obwohl H_2 richtig ist) gesucht.

$S_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}} = \frac{20}{\sqrt{25}} = 4$

$\beta_{01} = P(\bar{X} \leq 205|\mu = 210) = P(Z \leq \frac{205-210}{4}) = P(Z \leq -1.25) = 1 - P(Z \leq 1.25) = 1 - F(1.25) = 1 - 0.8944 = 0.1056$

$\beta_{02} = P(\bar{X} \leq 205|\mu = 220) = P(Z \leq \frac{205-220}{4}) = P(Z \leq -3.75) = 1 - P(Z \leq 3.75) = 1 - F(3.75) = 1 - 1 = 0$

Wenn $205 < \bar{X} < 215$ ist, wird H_1 ($\mu = 210$) angenommen.

In diesem Fall werden β_{10} (die Wahrscheinlichkeit H_1 beibehalten, obwohl H_0 richtig ist) und β_{12} (die Wahrscheinlichkeit H_1 beibehalten, obwohl H_2 richtig ist) gesucht.

$\beta_{10} = P(205 < \bar{X} < 215|\mu = 200) = P(\frac{205-200}{4} < Z < \frac{215-200}{4}) = P(\frac{5}{4} < Z < \frac{15}{4}) = F(3.75) - F(1.25) = 1 - 0.8944 = 0.1056$

$\beta_{12} = P(205 < \bar{X} < 215|\mu = 220) = P(\frac{205-220}{4} < Z < \frac{215-220}{4}) = P(-\frac{15}{4} < Z < -\frac{5}{4}) = F(3.75) - F(1.25) = 1 - 0.8944 = 0.1056$

Wenn $215 \geq \bar{X}$ ist, wird H_2 ($\mu = 220$) angenommen.

In diesem Fall werden β_{20} (die Wahrscheinlichkeit H_2 beibehalten, obwohl H_0 richtig ist) und β_{21} (die Wahrscheinlichkeit H_2 beibehalten, obwohl H_1 richtig ist) gesucht.

$\beta_{20} = P(\bar{X} \geq 215|\mu = 200) = P(Z \geq \frac{215-200}{4}) = P(Z \geq \frac{15}{4}) =$

$1 - F(3.75) = 1 - 1 = 0$

$\beta_{21} = P(\bar{X} \geq 215|\mu = 210) = P(Z \geq \frac{215-210}{4}) = P(Z \geq \frac{5}{4}) = 1 - F(1.25) = 1 - 0.8944 = 0.1056$

	$\bar{X} \leq 205$ (H_0)	$205 < \bar{X} < 215$ (H_1)	$215 \geq \bar{X}$ (H_2)
$\mu = 200$		$\beta_{10} = 0.1056$	$\beta_{20} = 0$
$\mu = 210$	$\beta_{01} = 0.1056$		$\beta_{21} = 0.1056$
$\mu = 220$	$\beta_{02} = 0$	$\beta_{12} = 0.1056$	

Aufgabe 22

In einer Firma soll festgestellt werden, wieviel Zeit die Mitarbeiter pro Monat aufzuwenden bereit sind, um sich über betriebliche Belange zu informieren. Die Untersuchung soll unter anderem Aufschluss bringen über eventuelle Unterschiede zwischen männlichen und weiblichen Mitarbeitern. Bei $N_1 = 45$ zufällig ausgewählten männlichen Beschäftigten ergaben sich $\bar{x} = 4.3$ (Std.) und $s_x^2 = 1.1^2$, bei $N_2 = 50$ zufällig ausgewählten weiblichen Beschäftigten ergaben sich $\bar{y} = 3.5$ (Std.) und $s_y^2 = 0.9^2$

a) Man prüfe die Hypothese, die Erwartungswerte der Untersuchungsmerkmale unterscheiden sich nicht ($\alpha = 0.05$).

$H_0 : \mu_X - \mu_Y = 0 \qquad H_1 : \mu_X - \mu_Y \neq 0$

Da $N, M > 30$ sind, wird zur Berechnung der Prüfgröße $T = \frac{\bar{x}-\bar{y}}{S}$ die Standardabweichung

$S = \sqrt{\frac{s_x^2}{N_1} + \frac{s_y^2}{N_2}} = \sqrt{\frac{1.1^2}{45} + \frac{0.9^2}{50}} = 0.2076$

benötigt.Der Wert der Prüfgröße T ergibt sich zu

$T = \frac{\bar{x} - \bar{y}}{S} = \frac{4.3 - 3.5}{0.2076} = 3.85.$

Da $T = 3.85 > z(0.975) = 1.96$ kann die Nullhypothese auf 5%-Signifikanzniveau abgelehnt werden.

b) Man prüfe die Hypothese, männliche Mitarbeiter informieren sich im Durchschnitt um mindestens 1 Stunde pro Monat länger über betriebliche Belange als weibliche Belegschaftsmitglieder ($\alpha = 0.05$).

$H_0 : \mu_X - \mu_Y \geq 1 \qquad H_1 : \mu_X - \mu_Y < 1$

Da $N, M > 30$ sind, wird zur Berechnung der Prüfgröße $T = \frac{\bar{x}-\bar{y}}{S}$ die Standardabweichung

$$S = \sqrt{\frac{s_x^2}{N_1} + \frac{s_y^2}{N_2}} = \sqrt{\frac{1.1^2}{45} + \frac{0.9^2}{50}} = 0.2076$$

benötigt.Der Wert der Prüfgröße T ergibt sich zu

$$T = \frac{\bar{x} - \bar{y} - 1}{S} = \frac{4.3 - 3.5 - 1}{0.2076} = -0.96.$$

Da $T = -0.96 > z(0.05) = -1.65$ kann die Nullhypothese (Männliche Mitarbeiter informieren sich im Durchschnitt um mindestens 1 Stunde pro Monat länger über betriebliche Belange als weibliche Belegschaftsmitglieder'") auf 5%-Signifikanzniveau nicht abgelehnt werden.

c) Genügen die obigen Angaben, um Teilaufgabe a) mit einem verteilungsfreien Verfahren zu lösen?

Für die nichtparametrische Methode für den Vergleich der Mittelwerte (Wilcoxon-Rangsummen-Test) braucht man Ränge beider Stichproben (bzw. Werte für die einzelnen Personen) , die in dieser Aufgabe nicht angegeben sind.

Aufgabe 23

Zur Überprüfung der Gedächtnisleistung von Mitarbeitern wurden 30 Begriffe den Personen jeweils zweimal vorgelesen und dann das Untersuchungsmerkmal X: „Anzahl der reproduzierten Begriffe" registriert. Bei $N = 10$ gleichaltrigen Personen ergaben sich folgende Werte:

$$17, 12, 13, 16, 9, 19, 21, 12, 18, 17.$$

An weiteren 12 Mitarbeitern wurde dasselbe Experiment durchgeführt, wobei die Personen jedoch während des gesamten Experiments einer ständigen Lärmbelästigung ausgesetzt waren. Hier ergaben sich die Werte:

$$10, 6, 15, 9, 8, 11, 8, 16, 13, 7, 5, 14.$$

a) Man überprüfe mit Hilfe des Wilcoxon-Rangsummen-Tests die Hypothese, die Lärmbeeinflussung habe keinen Einfluss auf die Anzahl der reproduzierten Begriffe ($\alpha = 0.05$; zweiseitig).

$H_0 : \mu_x = \mu_y \qquad H_1 : \mu_x \neq \mu_y$

Als Prüfgröße des Wilcoxon-Rangsummen-Tests wird

$$T_W = \sum_{n=1}^{N} \mathrm{Rg}(X_n)$$

Die Rangzahlen werden für die gepoolte Stichprobe vergeben, wobei Bindungen innerhalb derselben Stichprobe nicht von Interesse sind. Liegen Bindungen zwischen beiden Stichproben vor, werden Durchschnittsränge gebildet ($N + M = 22$).

	# Anzahl der reproduzierten Begriffe												$\sum$
X_n	17	12	13	16	9	19	21	12	18	17			
$Rg(X_n)$	18.5	10.5	12.5	16.5	6.5	21	22	10.5	20	18.5			**156.5**
Y_n	10	6	15	9	8	11	8	16	13	7	5	14	
$Rg(Y_n)$	8	2	15	6.5	4.5	9	4.5	16.5	12.5	3	1	14	96.5

Mit $\alpha = 0.05$ ergeben sich die kritischen Werte zu $w(0.025) = 85$ und $w(0.975) = N(N + M + 1) - w(0.025) = 10 * (10 + 12 + 1) - 85 = 145$. Da $T_W = 156.5 > 145$ gilt, kann die Nullhypothese, Lärm hat keinen Einfluss auf die Anzahl der reproduzierten Begriffe, abgelehnt werden.

b) Prüfen Sie dieselbe Hypothese unter Normalverteilungsannahme.

$$H_0 : \mu_X - \mu_Y = 0 \qquad \mu_X - \mu_Y \neq 0$$

$$\hat{\mu}_X = \bar{X} = \frac{1}{N} \sum_{n=1}^{N} X_n = 15.4$$
$$\hat{\mu}_Y = \bar{Y} = \frac{1}{M} \sum_{m=1}^{M} Y_m = 10.17$$

$$S_x^2 = \frac{1}{N-1} \sum_{n=1}^{N} (X_n - \bar{X})^2 = 14.05$$
$$S_y^2 = \frac{1}{M-1} \sum_{m=1}^{M} (Y_m - \bar{Y})^2 = 13.24$$

Unter den getroffenen Annahmen (Normalverteilungsannahme, Varianzhomogenität, σ_X und σ_Y unbekannt, $N, M < 30$, $\alpha = 0.05$) wird zur Berechnung der Prüfgröße $T = \dfrac{\bar{x} - \bar{y}}{S}$ die Varianz benötigt:

$$S^2 = \sqrt{\frac{s_x^2}{N} + \frac{s_y^2}{M}} = \sqrt{\frac{14.05}{10} + \frac{13.24}{12}} = 2.5083$$

$$k \approx \frac{\left(\frac{S_x^2}{N} + \frac{S_y^2}{M}\right)^2}{\left(\frac{S_x^2}{N}\right)^2/(N-1)) + \left(\frac{S_y^2}{M}\right)^2/(M-1)} = \frac{\left(\frac{14.05}{10} + \frac{13.24}{12}\right)^2}{\left(\frac{14.05}{10}\right)^2/(10-1)) + \left(\frac{13.24}{12}\right)^2/(12-1)} = 19$$

Da sich der Wert der Prüfgröße T zu

$$T = \frac{\bar{x} - \bar{y}}{S} = \frac{15.4 - 10.17}{\sqrt{2.5083}} = 3.302$$

ergibt, mit $T = 3.302 > t(0.975, 19) = 2.093$ kann die Nullhypothese abgelehnt werden.

Aufgabe 24

Für den Vergleich zweier Verteilungen sind verschiedene Tests besprochen worden. Diskutieren Sie Gemeinsamkeiten und Unterschiede dieser Tests. Geben Sie für jeden Test eine typische Datensituation an, für die der Test angemessen erscheint.

Der Zweistichproben-t-Test ist ein Signifikanztest , der anhand der Mittelwerte zweier Stichproben prüft, ob die Mittelwerte zweier Grundgesamtheiten einander gleich sind, ggf. gegen die Alternative, dass einer der Mittelwerte kleiner ist als der andere.
Es gibt mehrere Varianten des Zweistichproben-t-Tests:

- den für zwei unabhängige Stichproben mit gleichen und unbekannten Standardabweichungen in beiden Grundgesamtheiten
- den für zwei unabhängige Stichproben mit ungleichen und unbekannten Standardabweichungen in beiden Grundgesamtheiten

- den für zwei abhängige Stichproben.

Die Voraussetzung für diesen Test ist die Normalverteilungsannahme der Variablen.
Bei Mittelwertvergleichen normalverteilter Stichproben mit bekannter Standardabweichung können Gauß-Tests verwendet werden.
Liegt keine Normalverteilung vor, können als Ersatz für den t-Test nichtparametrische Tests angewendet werden, etwa ein Wilcoxon-Rangsummentest für unabhängige Stichproben oder ein Wilcoxon-Vorzeichen-Rang-Test für gepaarte Stichproben.

Mit einem F-Test kann man prüfen, ob die Varianzen zweier Grundgesamtheiten einander gleich sind, ggf. gegen die Alternative, dass einer der Mittelwerte kleiner ist als der andere. Um zwei Varianzen v1, und v2 zu vergleichen, muss man das Verhältnis dieser zwei Varianzen berechnen. Dieses Verhältnis wird F-Wert genannt. Der F-Test basiert auf zwei Annahmen: (1) die Stichproben sind normalverteilt und (2) die Stichproben sind voneinander unabhängig.

Aufgabe 25

$N = 10$ zufällig ausgewählte Mitarbeiter sollen zwei Geschicklichkeitsaufgaben A und B vom gleichen Schwierigkeitsgrad lösen. Zuerst wird jedem Mitarbeiter die Aufgabe A vorgelegt und jeweils die Lösungszeit (in Minuten) gemessen. Danach erhält die gesamte Gruppe ein Training, in dessen Verlauf sie lernt, mit ähnlichen, jedoch nicht identischen Geschicklichkeitsaufgaben fertig zu werden. Es soll untersucht werden, ob das Vortraining einen auf die nachfolgende Arbeitsleistung (Lösung der Geschicklichkeitsaufgabe B) begünstigenden Einfluss hat, so dass eine Übertragung der durch Training erworbenen Fähigkeiten auf die nachfolgende Aufgabe B möglich ist (positiver Transfer). Die gemessenen Werte sind in der folgenden Tabelle festgehalten:

Person	Lösungszeit für Aufgabe A (vor dem Training)	Lösungzeit für Aufgabe B (nach dem Training)
1	6.24	6.10
2	5.80	5.75
3	5.57	5.60
4	6.04	5.91
5	5.56	5.30
6	5.92	5.62
7	6.17	6.29
8	5.48	5.28
9	6.11	5.70
10	6.20	6.01

a) Man prüfe mit Hilfe des Tests von Wilcoxon die Hypothese, das Training habe keinen Einfluss auf das Merkmal Lösungszeit ($\alpha = 0.05$).

Da es sich in dieser Aufgabe um die abhängigen Stichproben handelt, wird Wilcoxon-Vorzeichen-Rang-Test für die Zufallsvariable $D_n = X_n - Y_n$ angewendet.

$$H_0 : x(0.5) = 0 \qquad H_1 : x(0.5) \neq 0$$

Person	Lösungszeit für Aufgabe A X_n	Lösungzeit für Aufgabe B Y_n	$D_n = X_n - Y_n$	Rg_n^+	Rg_n^-
1	6.24	6.10	0.14	5	
2	5.80	5.75	0.05	2	
3	5.57	5.60	-0.03		1
4	6.04	5.91	0.13	4	
5	5.56	5.30	0.26	8	
6	5.92	5.62	0.30	9	
7	6.17	6.29	-0.12		3
8	5.48	5.28	0.20	7	
9	6.11	5.70	0.41	10	
10	6.20	6.01	0.19	6	
				51	4

Die Prüfgröße des Wilcoxon-Vorzeichen-Rang-Tests

$$W^+ = \sum_{n=1}^{N} R_n Z_n \text{mit} Z_n = \begin{cases} 1 & \text{falls } D_n > 0 \\ & n = 1, ..., N \\ 0 & \text{falls } D_n < 0 \end{cases}$$

und $R_n = rg|D_n| = rg|X_n - Y_n|$ nimmt den Wert 51 an. Für $N = 10$ und $\alpha = 0.05$ ergeben sich der untere kritische Wert zu $w(\alpha/2) = w(0.025) = 9$ und der obere Kritische Wert zu $w(1-\alpha/2) = w(0.975) = 46$. Die Nullhypothese muss abgelehnt werden, denn es gilt $W^+ = 51 > 46$.

b) Prüfen Sie die gleiche Hypothese mit dem t-Test.

Da es sich in dieser Aufgabe um die abhängigen Stichproben handelt, wird t-Test für die Zufallsvariable $D_n = X_n - Y_n$ angewendet.

$$H_0 : \mu_x - \mu_y = 0 \qquad H_1 : \mu_x - \mu_y \neq 0$$

Person	Lösungszeit für Aufgabe A (vor dem Training) X_n	Lösungszeit für Aufgabe B (nach dem Training) Y_n	$D_n = X_n - Y_n$
1	6.24	6.10	0.14
2	5.80	5.75	0.05
3	5.57	5.60	-0.03
4	6.04	5.91	0.13
5	5.56	5.30	0.26
6	5.92	5.62	0.30
7	6.17	6.29	-0.12
8	5.48	5.28	0.20
9	6.11	5.70	0.41
10	6.20	6.01	0.19
	$\bar{X} = 5.909$	$\bar{Y} = 5.756$	$\bar{D} = 0.153$

$S_D^2 = \frac{1}{N-1} \sum_{n=1}^{N} (D_n - \bar{D})^2 = 0.0247$

Führt man für die Differenzen einen t-Test der Form

$$T = \frac{\bar{D}}{S_D/\sqrt{N}}$$

durch, so ergibt sich:

$$t = \frac{0.153}{\sqrt{0.0247}/\sqrt{10}} = 3.0785$$

Die kritischen Werte der T-Statistik sind $t(\alpha/2, N-1) = t(0.025, 9) = -2.262$ und $t(1-\alpha/2, N-1) = t(0.975, 9) = 2.262$

Die Nullhypothese ($\mu_x - \mu_y = 0$) muss abgelehnt werden, denn es gilt $T = 3.0785 > t(0.975, 9) = 2.262$.

Aufgabe 26

Zu einer Diskussion über die Einführung einer neuen Verkaufsmethode wurden $N = 70$ zufällig ausgewählte Kundenberater eingeladen und vor bzw. nach der Diskussion zu ihrer Einstellung (dafür oder dagegen) über die neue Methode befragt. Es ergab sich, dass vorher 40, nachher 56 Personen für die neue Methode waren, während 36 Berater vor und nach der Diskussion für die neue Methode waren.

a) Man ermittle mit diesen Angaben eine Kontingenztabelle der Struktur

		Einstellung nach der Diskussion	
		dafür	dagegen
Einstellung vor	dafür	**36**	40-36= **4**
der Diskussion	dagegen	56-36=**20**	70-36-4-20=**10**

b) Man prüfe mit einem nichtparametrischen Test die Hypothese, die Einstellung zur neuen Verkaufsmethode habe sich durch die Diskussion nicht geändert ($\alpha = 0.05$).

In dieser Aufgabe handelt es sich um die abhängigen Stichproben und es wird McNemar-Test durchgeführt (Achtung: Der Test wurde noch nicht im Skript beschrieben).

		Einstellung nach der Diskussion	
		dafür	dagegen
Einstellung vor	dafür	**a=36**	**b=4**
der Diskussion	dagegen	**c=20**	**d=10**

McNemar-Test:

$$H_0 : p_a + p_b = p_a + p_c \qquad H_1 : p_a + p_b \neq p_a + p_c$$

Die Prüfgröße des McNemar-Tests

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Diese Teststatistik ist approximativ χ^2 verteilt mit einem Freiheitsgrad.

$$\chi^2 = \frac{(b-c)^2}{b+c} = \frac{(4-20)^2}{4+20} = \frac{256}{24} = 10.67$$

Der kritische Wert des McNemar-Tests ist $\chi^2(1, 0.95) = 3.841$.

Da die Prüfgröße größer als der kritische Wert ist, wird die Nullhypothese abgelehnt, so kann man davon ausgehen, dass ein statistisch signifikanter Unterschied zwischen den beiden Stichproben besteht, d.h. es gibt eine signifikante Veränderung in der Einstellungen der Kundenberater gegenüber der neuen Verkaufsmethode nach der Diskussion.

16.1.5. Zusammenhangsanalyse

Aufgabe 27

Bei $N = 10$ Kunden wurde die Zufriedenheit mit den Produkten (Variable X) und die Loyalität mit der Firma (Variable Y) ermittelt. Man erhielt die Wertepaare:

X	124	79	118	102	86	89	109	128	114	95
Y	100	94	101	112	76	98	91	73	90	84

a) Testen Sie die Hypothese der Unabhängigkeit von X und Y unter der Verwendung des Bravais-Pearsonschen Korrelationskoeffizienten ($\alpha = 0.05$).
Hinweis: $\sum X_i^2 = 111548$, $\sum Y_i^2 = 85727$, $\sum X_i Y_i = 95929$.

Der Bravais-Pearsonsche Korrelationskoeffizient:

$$R = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{\sqrt{(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2)(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2)}}$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = 104.4$$
$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = 91.9$$

$$R = \frac{95929 - 10 \cdot 104.4 \cdot 91.9}{\sqrt{(111548 - 10 \cdot 104.4^2)(85727 - 10 \cdot 91.9^2)}} = -0.0081$$

Die Nullhypothese und Alternativhypothese sind:

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0$$

Test der Korrelationskoeffizienten:

$$T = \sqrt{N-2}\frac{R}{\sqrt{1-R^2}}$$

Diese Teststatistik ist t-verteilt mit N-2 Freiheitsgraden.

So ergibt sich die Prüfgröße zu
$T = \sqrt{N-2}\frac{R}{\sqrt{1-R^2}} = \sqrt{10-2}\frac{-0.0081}{\sqrt{1-(-0.0081)^2}} = -0.0229.$

Die kritischen Werte des Tests sind $t(\alpha/2, N-2) = t(0.025, 8) = -2.306$ und $t(1-\alpha/2, N-2) = t(0.975, 8) = 2.309$.

Da die Prüfgröße zwischen den kritischen Werten ist, kann die Nullhypothese nicht abgelehnt werden, so kann man davon ausgehen, dass es zwischen der Zufriedenheit mit den Produkten und Loyalität mit der Firma keinen Zusammenhang gibt.

b) Lösen Sie Aufgabe a) unter Verwendung des Rangkorrelationskoeffizienten von Spearman.

Der Korrelationskoeffizient von Spearman wird von Rangdaten berechnet. Ohne Bindungen (bzw. gleiche Ränge) kann die Prüfgröße mit Hilfe der Differenzwerte berechnet:

$$R_{Sp} = 1 - 6\frac{\sum D_i^2}{N(N^2-1)}$$

wo $D_i = Rg(X_i) - Rg(Y_i)$ ist.

X	124	79	118	102	86	89	109	128	114	95
$Rg(X_i)$	9	1	8	5	2	3	6	10	7	4
Y	100	94	101	112	76	98	91	73	90	84
$Rg(Y_i)$	8	6	9	10	2	7	5	1	4	3
D_i	1	-5	-1	-5	0	-4	1	9	3	1

$$R_{Sp} = 1 - 6\frac{\sum D_i^2}{N(N^2-1)} = 1 - 6\frac{160}{10 \cdot 99} = 0.0303$$

$$T = \sqrt{N-2}\frac{R_{Sp}}{\sqrt{1-R_{Sp}^2}} = \sqrt{10-2}\frac{0.0303}{\sqrt{1-(0.0303)^2}} = 0.08574$$

Die kritischen Werte des Tests sind $t(\alpha/2, N-2) = t(0.025, 8) = -2.306$ und $t(1-\alpha/2, N-2) = t(0.975, 8) = 2.309$.

Da $T = 0.08574 < t(0.975, 8) = 2.309$, kann die Nullhypothese nicht abgelehnt werden.

Aufgabe 28

Vergleichen Sie in den folgenden Teilaufgaben den

- χ^2-Unabhängigkeitstest
- Test auf $\rho_{XY} = 0$ unter Verwendung des Stichprobenkorrelationskoeffizienten.

a) Geben Sie für jeden der beiden Tests Null- und Alternativhypothesen an. Unter welcher Bedingung sind die beiden Nullhypothesen äquivalent?

χ^2-Unabhängigkeitstest
Die Nullhypothese und Alternativhypothese sind:

$$H_0: \quad \pi_{ij} = \pi_{i.} \cdot \pi_{.j} \text{für alle } i, j$$
$$H_1: \quad \pi_{ij} \neq \pi_{i.} \cdot \pi_{.j} \text{für mindestens ein } i, j$$

χ^2-Unabhängigkeitstest wird zur Überprüfung der Unabhängigkeit von zwei kategorialen Variablen angewendet.

Test auf $\rho_{XY} = 0$ unter Verwendung des Stichprobenkorrelationskoeffizienten
Die Nullhypothese und Alternativhypothese sind:

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0$$

Test auf $\rho_{XY} = 0$ unter Verwendung des Stichprobenkorrelationskoeffizienten wird durchgeführt, um die Unabhängigkeit von zwei metrischen Variablen zu überprüfen. Wenn die beiden Variable normalverteilt sind, kann der Person-Korrelationskoeffizient gerechnet und der Test auf $\rho_{XY} = 0$ durchgeführt. Bei nichtnormalverteilten Variablen wird die Unabhängigkeit durch Spearmans Rangkorrelation überprüft.

Wenn beide Variable geordnete Werte enthalten, kann die Unabhängigkeit dieser Variablen sowohl mit χ^2-Unabhängigkeitstest, als auch mit dem Test auf $\rho_{XY} = 0$ unter Verwendung des Stichprobenkorrelationskoeffizienten geprüft werden.

b) Charakterisieren Sie hinsichtlich des Messniveaus der Daten die unterschiedlichen Anwendungsmöglichkeiten der beiden Tests. Lassen sich bei intervallskalierten Daten beide Tests verwenden? Geben Sie einen fiktiven Datensatz an, der sich mit dem einen Test behandeln ließe, mit dem anderen nicht.

Test auf $\rho_{XY} = 0$ unter Verwendung des Stichprobenkorrelationskoeffizienten wird durchgeführt, um die Unabhängigkeit von zwei metrischen Variablen zu überprüfen. χ^2-Unabhängigkeitstest wird zur Überprüfung der Unabhängigkeit von zwei kategorialen Variablen angewendet.

Bei der intervallskalierten Daten können beide Tests verwendet werden, aber der Test auf $\rho_{XY} = 0$ unter Verwendung des Stichprobenkorrelationskoeffizienten setzt voraus, dass benachbarte Ränge immer den gleichen Abstand haben und einen linearen Zusammenhang besteht. Z.B. um den Zusammenhang zwischen den Anzahl der Arbeitsstunden (1 - von 0 bis 5 Stunden der Woche, 2 - von 6 bis 10, 3 - von 11 bis 15 etc, bis 8 - von 35 bis 40) und Alter (1 - von 1 bis 14, 2 von 15 bis 35, 3 von 35 bis 65 und 4- älter als 65) zu überprüfen ist nur χ^2-Unabhängigkeitstest geeignet, da benachbarte Ränge der Variable Älter'nicht den gleichen Abstand haben und höchstwahrscheinlich ein nichtlinearer Zusammenhang zwischen den Variablen besteht.

c) Charakterisieren Sie informell den Unterschied zwischen exakter und asymptotischer Verteilung einer Prüfgröße. Welchen Sinn haben hier Faustregeln wie „$n > 30$“, „$np_j \geq 1$ für alle Klassen j“ etc.?

Exakte statistische Tests legen der statischen Entscheidung die exakte Verteilung der Prüfgröße zugrunde, asymptotische Tests eine approximative Verteilung. In großen Stichproben gilt der zentraler Grenzwertsatz: Summen von beliebig verteilten Zufallsvariablen sind nämlich approximativ notmalverteilt. Die T-Statistik ist in großen Stichproben approximativ normalverteilt. Die Summenvariable von normalverteilten Variablen in Quadrat ist χ^2-verteilt. Die Summenvariable von approximativ normalverteilten Variablen in Quadrat ist approximativ χ^2-verteilt. Faustregeln wie „$n > 30$“, „$np_j \geq 1$ werden genutzt, um die Stichprobe auf die Größe zu kontrollieren.

Aufgabe 29

Illustrieren Sie anhand einiger selbstgewählter Beispiele, dass man Kausalzusammenhänge nicht allein auf signifikante Korrelationskoeffizienten stützen kann.

Abhängigkeiten und Korrelationen sind notwendige, aber nicht hinreichende Bedingungen für einen kausalen Zusammenhang. Wenn es um einen nichtlinearen Zusammenhang geht, ist der Korrelationskoeffizient nicht signifikant, obwohl ein Zusammenhang besteht. Außerdem ist ein signifikanter Korrelationskoeffizient nicht immer sinnvoll interpretierbar, da die Korrelation durch die Wirkung einer dritten Variable erzeugt wird.
Beispiel: Es lässt sich ein hoher Zusammenhang zwischen Männern mit wenig Haaren und hohem Einkommen nachweisen. Tatsächlich besteht aber eher ein Zusammenhang zwischen dem Alter der Männer und ihrem Einkommen und mit zunehmendem Alter nimmt auch die Zahl der Haare ab.

Aufgabe 30

Mit einer Befragungsaktion in Holstein-Unterwasser soll die politische Durchsetzbarkeit eines Kraftwerkbaus überprüft werden. In dieser Region wurden $N = 2000$ zufällig ausgewählte Personen befragt. Für den Kernkraftwerksbau sprachen sich 400 Befragte aus, von denen 350 parteilos waren. Unter den insgesamt 200 parteigebundenen Befragten waren 130 Kernkraftsgegner und 20 hatten „keine Meinung“. Von allen Befragten sprachen sich 1400 gegen das Projekt aus. Es soll statistisch bewiesen werden, daß eine Abhängigkeit

zwischen den statistischen Größen

X : Parteizugehörigkeit mit den Ausprägungen x_1: „parteigebunden“ und x_2: „parteilos“ und

Y : Einstellung zum Kernkraftwerksbau mit mit den Ausprägungen y_1: „dafür“, y_2: „dagegen“ und y_3: „keine Meinung“

besteht.

a) Stellen Sie das Befragungsergebnis in einer Kontingenztafel dar.

		Parteizugehörigkeit	
		parteigebunden	parteilos
Einstellung zum	dafür	400-350= **50**	**350**
Kernkraft-	dagegen	**130**	1400-130=**1270**
werksbau	keine Meinung	**20**	2000-400-1400-20=**180**

n_{ij}	parteigebunden (n_{i1})	parteilos (n_{i2})	$n_{i.}$
dafür (n_{1j})	50	350	400
dagegen (n_{2j})	130	1270	1400
keine Meinung (n_{3j})	20	180	200
$n_{.j}$	200	1800	2000

b) Formulieren Sie das Hypothesenpaar und begründen Sie die Wahl der Nullhypothese.

H_0 : $\pi_{ij} = \pi_{i.} \cdot \pi_{.j}$ für alle $i = 1, 2, 3,\ j = 1, 2$

H_1 : $\pi_{ij} \neq \pi_{i.} \cdot \pi_{.j}$ für mindestens ein i, j

c) Führen Sie den Test mit einer Irrtumswahrscheinlichkeit von $\alpha = 0.01$ durch und interpretieren Sie das Ergebnis.

Für den χ^2-Unabhängigkeitstest wird $\chi^2 = \sum_{i,j=1}^{I,J} \frac{(n_{ij} - N\hat{\pi}_{ij})^2}{N\hat{\pi}_{ij}}$ als Testgröße verwendet mit $N\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{N}$ ($N = 2000$).

Die erwarteten Häufigkeiten unter Unabhängigkeit $N\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{N}$ sind :

n_{ij}	parteigebunden ($j=1$)	parteilos ($j=2$)
dafür ($i=1$)	400*200/2000= 40	400*1800/2000=360
dagegen ($i=2$)	1400*200/2000=140	1400*1800/2000=1260
keine Meinung ($i=3$)	200*200/2000=20	200*1800/2000=180

$$\chi^2 = \sum_{i,j=1}^{I,J} \frac{(n_{ij} - N\hat{\pi}_{ij})^2}{N\hat{\pi}_{ij}} = \frac{(50-40)^2}{40} + \frac{(350-360)^2}{360} + \frac{(130-140)^2}{140} + \frac{(1270-1260)^2}{1260} + \frac{(20-20)^2}{20} + \frac{(180-180)^2}{180} = 2.5 + 0.27778 + 0.7142 + 0.0793 + 0 + 0 = 3.57128$$

Zum Signifikanzniveau $\alpha = 0.01$ ergibt sich der obere kritische Wert c_o zu $\chi^2(1-\alpha, (I-1)(J-1)) = \chi^2(0.99, 2) = 9.21$. Wegen $\chi^2 = 3.57128 < 9.21$ kann die Nullhypothese (Unabhängigkeit der Merkmale „Parteizugehörigkeit“ und „Einstellung zum Kernkraftwerksbau“) nicht abgelehnt werden.

16.1.6. Regressionsanalyse

Aufgabe 31

Zur Überprüfung der Güte eines Schuleignungstests X zur Vorhersage der Schulreife ermittelt ein Schulpsychologe an $N = 500$ Vorschulkindern die Werte $\bar{x} = 40$ und $s_x = 5$. Nach Ablauf des ersten Schuljahres werden mit einem Schulleistungstest Y die tatsächlichen Leistungen der Kinder gemessen und man erhielt die folgenden Kennwerte: $\bar{y} = 30$, $s_y = 4$. Die Kovarianz in der Stichprobe zwischen dem Schuleignungstest und dem Schulleistungstest betrug:

$$s_{xy} := \widehat{\mathrm{Cov}}(x,y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = 10$$

a) Berechnen Sie einen Korrelationskoeffizienten zwischen Schuleignungs- und Schulleistungstest.

$$R = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{10}{5 \cdot 4} = 0.5$$

b) Prüfen Sie die Alternativhypothese, der Korrelationskoeffizienten sei größer als 0.6 ($\alpha = 0.05$).

$$H_0 : \rho > \rho_0 = 0.6 \qquad H_1 : \rho \leq \rho_0 = 0.6$$

Zuerst muss auf Fishers Z-Statistik transformiert werden.

$$Z = \frac{\sqrt{N-3}}{2}(\ln\frac{1+R}{1-R} - \ln\frac{1+\rho_0}{1-\rho_0})$$

$$Z = \frac{\sqrt{500-3}}{2}(ln\frac{1+0.5}{1-0.5} - ln\frac{1+0.6}{1-0.6}) = -3.2067$$

Das $z(0.05)$-Quantil ist -1.65. Da $Z = -3.2067 < z(0.05) = -1.65$, muss die Nullhypothese abgelehnt werden.

c) Man ermittle die Stichproben-Regressionsgerade zur Vorhersage der tatsächlichen Leistung nach dem ersten Schuljahr aufgrund des Schuleignungstests.

Für das obige Modell gilt:

$$\begin{aligned} \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\ \hat{\beta} &= \frac{\sum_{n=1}^{N} X_n Y_n - N\bar{X}\bar{Y}}{\sum_{n=1}^{N} X_n^2 - N\bar{X}^2} = R\frac{s_y}{s_x} \end{aligned}$$

$$\begin{aligned} \hat{\beta} &= 0.5\frac{4}{5} = 0.4 \\ \hat{\alpha} &= 30 - 0.4 \cdot 40 = 14 \end{aligned}$$

$$\hat{Y} = 14 + 0.4 \cdot x$$

d) Man prüfe, ob der Regressionskoeffizient signifikant von 0 verschieden ist ($\alpha = 0.05$; $\hat{\sigma} = 1.75$).

Für die Hypothese $H_0 : \beta = 0$ wird die Testgröße

$$T_\beta = \frac{\hat{\beta} - 0}{\widehat{\sigma_\beta}}$$

verwendet, welche t-verteilt ist mit $(N-2)$ Freiheitsgraden.

$$\widehat{\sigma_\beta} = \frac{\hat{\sigma}}{\sqrt{\sum_{n=1}^{N} X_n^2 - N\bar{X}^2}} = \frac{\hat{\sigma}}{\sqrt{(N-1)s_x^2}}$$

$$\widehat{\sigma_\beta} = \frac{1.75}{\sqrt{(500-1)5^2}} = 0.01567$$

Es ergibt sich $T = 0.4/0.01567 = 25.53$. Wegen $N = 500 > 30$ ist die t-Statistik approximativ normalverteilt, somit $z(1-\alpha/2) = z(0.975) = 1.96$. Wegen $T = 25.53 > z(0.975) = 1.96$ muss die Nullhypothese abgelehnt werden. So ist der Regressionskoeffizient signifikant von 0 verschieden.

e) Ein Kind erreicht im Eignungstest einen Wert von $x_0 = 48$. Mit welcher schulischen Leistung ist bei ihm zu rechnen?

$\hat{Y_0} = 14 + 0.4 \cdot X_0 = 14 + 0.4 \cdot 48 = 33.2$

f) Man ermittle ein 95%-Prognoseintervall zu Vorhersage der schulischen Leistung des Kindes.

Das Konfidenzintervall für die Regressionsgerade wird mit

$$\hat{E}[Y|X] \pm t(1-\alpha/2, N-2)\sqrt{\widehat{Var}(\hat{E})}$$

angegeben, wobei

$$\sqrt{\widehat{Var}(\hat{E})} = \hat{\sigma}\sqrt{\frac{1}{N} + \frac{(X-\bar{X})^2}{\sum_n X_n^2 - N\bar{X}^2}}$$

$$\sqrt{\widehat{Var}(\hat{E})} = \hat{\sigma}\sqrt{\frac{1}{N} + \frac{(X-\bar{X})^2}{(N-1)s_x^2}}$$

gilt.

$$\sqrt{\widehat{Var}(\hat{E})} = 1.75\sqrt{\frac{1}{500} + \frac{(x-40)^2}{(500-1)5^2}} = 1.75\sqrt{0.002 + 0.00008(x-40)^2} =$$

$$= \sqrt{0.006 + 0.00024(x-40)^2}$$

Mit $z(1-\alpha/2) = z(0.975) = 1.96$ lautet das zweiseitige 95%-Konfidenzintervall

$$(14 + 0.4x) \pm 1.96\sqrt{0.006 + 0.00024(x - 40)^2}$$

An der Stelle $X_0 = 48$ ist das 95%-Prognoseintervall zu Vorhersage der schulischen Leistung des Kindes

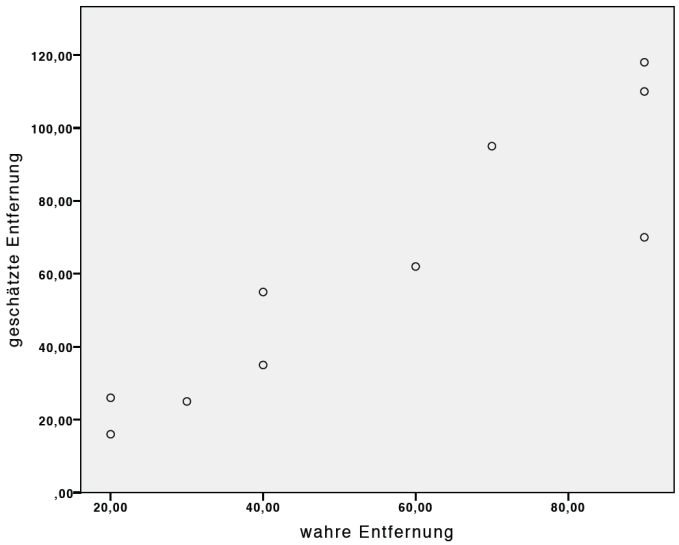
$$33.2 \pm 0.1$$

Aufgabe 32

$N = 10$ Vpn sollen verschiedene Entfernungen schätzen. Die wahren Entfernungen seien X_i, die geschätzten Entfernungen Y_i, $i = 1, ..., 10$. Man erhielt folgende Wertpaare:

x_i	20	20	30	40	40	60	70	90	90	90
y_i	16	26	25	35	55	62	95	70	110	118

a) Zeichnen Sie die Wertpaare (x_i,y_i) in ein Streudiagramm ein.



b) Schätzen Sie die Parameter α und β des Regressionsmodells

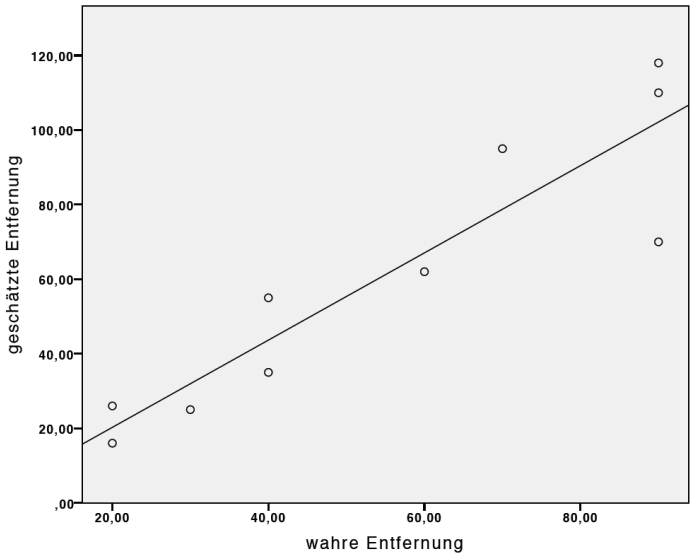
$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, ..., 10$$

und tragen Sie die Stichproben-Regressionsgerade in das Streudiagramm ein.
Hinweis: $\sum x_i^2 = 37700$, $\sum y_i^2 = 49600$, $\sum x_i y_i = 42380$

$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i = 55$
$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i = 61.2$

$$\begin{aligned}
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\
\hat{\beta} &= \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{\sum_{i=1}^{N} X_i^2 - N\bar{X}^2} \\
\hat{\beta} &= \frac{42380 - 10 \cdot 55 \cdot 61.2}{37700 - 10 \cdot 55^2} = 1.17 \\
\hat{\alpha} &= 61.2 - 1.17 \cdot 55 = -3.15
\end{aligned}$$

$\hat{Y} = -3.15 + 1.17 \cdot x$



c) Erläutern Sie an diesem Beispiel den Unterschied zwischen Stichprobengrößen und Grundgesamtheitsgrößen.

Als Stichprobe wird eine Untermenge der Grundgesamtheit bezeichnet. In diesem Beispiel besteht die Stichprobe aus 10 Vpn, wobei die GG alle VPN sind.

d) Ist in diesem Beispiel die übliche Annahme plausibel, dass die Varianz der Abweichung ε_i für alle Entfernungen X_i gleich ist?

Wie auf dem Bild zu sehen ist, nimmt die Varianz der Abweichung (Fehler der Prognose) mit der Entfernung zu. Es ist nicht zu erwarten, dass die Varianz der Abweichung ε_i für alle Entfernungen X_i gleich ist. Es kann erwartet werden, dass bei den großen Distanzen die geschätzte Entfernung mehr von wahrem Wert abweicht, als bei kleinen Distanzen.

e) Was bedeutet die Hypothese $\alpha = 0$, $\beta = 1$, also

$$Y_i = X_i + \varepsilon_i, E(\varepsilon_i) = 0, i = 1, ..., 10?$$

Die Hypothese $\alpha = 0$, $\beta = 1$ bedeutet, dass die Prognose perfekt die wahre Werte abbildet, wobei die geschätzte Entfernung gleich der wahren Entfernung plus Fehler ist.

Aufgabe 33

Es soll überprüft werden, ob die sensomotorische Koordinationsfähigkeit von Arbeitnehmern durch Training verbessert werden kann. Das unabhängige Merkmal Trainingszeit an einem Reaktionsgerät wird variiert von 0 Std. bis 5 Std. Bei $N = 30$ Vpn wurden in einem abschließenden Test folgende Fehlerzahlen registriert:

X_i (Std.)	Y_i	X_i (Std.)	Y_i
0	8	3	5
0	10	3	6
0	10	3	6
0	11	3	6
0	9	3	4
1	11	4	6
1	9	4	3
1	8	4	3
1	9	4	4
1	7	4	2
2	8	5	4
2	6	5	2
2	4	5	3
2	6	5	3
2	7	5	2

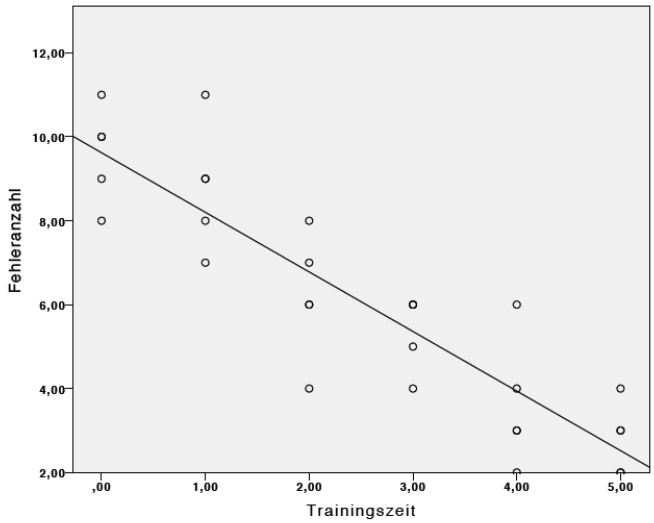
Rechenhilfen: $\sum x_i = 75, \sum x_i^2 = 275, \sum y_i = 182, \sum x_i y_i = 331, \hat{\sigma} = 1.23.$

a) Man ermittle die Stichprobenregressionsgerade.

$$\begin{aligned}
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\
\hat{\beta} &= \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{\sum_{i=1}^{N} X_i^2 - N\bar{X}^2} \\
\hat{\beta} &= \frac{331 - 30 \cdot \frac{75}{30} \cdot \frac{182}{30}}{275 - 30 \cdot \frac{75}{30}^2} = -1.42 \\
\hat{\alpha} &= \frac{182}{30} - (-1.42) \cdot \frac{75}{30} = 9.62
\end{aligned}$$

$$\hat{Y} = 9.62 - 1.42 \cdot x$$

b) Man zeichne die Stichprobenregressionsgerade.



c) Man ermittle ein 95%-Konfidenzintervall für β. Ist β signifikant von Null verschieden? (Begründung!)

Das Konfidenzintervall für β:

$$P\left\{\beta \in \hat{\beta} \pm t(1-\alpha/2, N-2)\widehat{\sigma_\beta}\right\} = 1-\alpha,$$

mit Quantil der t-Verteilung $t(1-\alpha/2, N-2) = t(0.975, 28) = 2.048$ und mit

$$\begin{aligned}\widehat{\sigma_\beta} &= \frac{\hat{\sigma}}{\sqrt{\sum_{n=1}^{N} X_n^2 - N\bar{X}^2}} \\ \widehat{\sigma_\beta} &= \frac{1.23}{\sqrt{275 - 30 \cdot \frac{75}{30}^2}} = 0.132\end{aligned}$$

Das Konfidenzintervall für β:

$$\begin{aligned}P\left\{\beta \in \hat{\beta} \pm t(1-\alpha/2, N-2)\widehat{\sigma_\beta}\right\} &= 1-\alpha \\ P\left\{\beta \in -1.42 \pm 2.048 \cdot 0.132\right\} &= 1-0.05,\end{aligned}$$

Das 95%-Konfidenzintervall für β ist [-1.69; -1.15]. Da das zweiseitige Konfidenzintervall den Wert $\beta = 0$ nicht überdeckt, ist der Regressionskoeffizient signifikant von 0 verschieden.

d) Man ermittle ein 95%-Konfidenzintervall (Prognoseintervall) für die Fehlerzahl eines Probanden, der 2.5 Std. trainiert.

Das Konfidenzintervall für die Regressionsgerade wird mit

$$\hat{E}[Y|X] \pm t(1-\alpha/2, N-2)\sqrt{\widehat{Var}(\hat{E})}$$

angegeben, wobei

$$\sqrt{\widehat{Var}(\hat{E})} = \hat{\sigma}\sqrt{\frac{1}{N} + \frac{(X-\bar{X})^2}{\sum_n X_n^2 - N\bar{X}^2}}$$

gilt.

$$\sqrt{\widehat{Var}(\hat{E})} = 1.23\sqrt{\frac{1}{30} + \frac{(x-\frac{75}{30})^2}{275 - 30\cdot\frac{75}{30}^2}} = 1.23\sqrt{0.002 + 0.01143(x-2.5)^2} =$$

$$= \sqrt{0.003 + 0.0173(x-2.5)^2}$$

Mit $t(1-\alpha/2; N-2) = t(0.975, 28) = 2.048$ lautet das zweiseitige 95%-Konfidenzintervall

$$(9.62 - 1.42x) \pm 2.048\sqrt{0.003 + 0.0173(x-2.5)^2}$$

An der Stelle $X_0 = 2.5 = \bar{X}$ ergibt sich der minimale Wert des 95%-Prognoseintervalls

$$6.07 \pm 0.11$$

e) Geben Sie das Bestimmtheitmaß an.

Gesucht ist das Bestimmtheitsmaß in der Form

$$R^2_{XY} = PRE = \frac{SQE}{SQT} = \frac{SQE}{SQE + SQR}$$

mit

$$\begin{aligned} SQE &= \hat{\beta}^2 \sum_{n=1}^{N}(X_n - \bar{X})^2 = \hat{\beta}^2(\sum_{n=1}^{N} X_n^2 - N\bar{X}^2) \\ SQR &= (N-2)\hat{\sigma}^2 \end{aligned}$$

Obige Daten eingesetzt ergibt

$$\begin{aligned} SQE &= (-1.42)^2(275 - 30 \cdot \frac{75^2}{30^2}) = 176.435 \\ SQR &= (30-2)1.23^2 = 42.36 \\ R^2_{XY} &= PRE = \frac{176.435}{176.435 + 42.36} = 0.81 \end{aligned}$$

Es ergibt sich der Wert $R^2_{xy} = 0.81$. Das Bestimmtheitsmaß gibt das Verhältnis von erklärter zu totaler Streuung an. Im vorliegenden Modell werden lediglich 81% der Streuung von Y durch X erklärt.

Aufgabe 34

Sechs Austauschstudenten, die einen Studienaufenthalt in Deutschland verbrachten, wurden vor und nach dem Aufenthalt über die Einschätzung der Arbeitsintensität der Deutschen im akademischen Bereich gefragt. Angegeben sind im folgenden Testergebnisse, die eine Zusammenfassung diverser Einzelaussagen enthalten. Man erhielt folgende Testergebnisse, wobei X für die Befragungsergebnisse vorher, Y für die Ergebnisse nachher steht:

X	10	15	20	19	30	26
Y	8	10	15	12	25	20

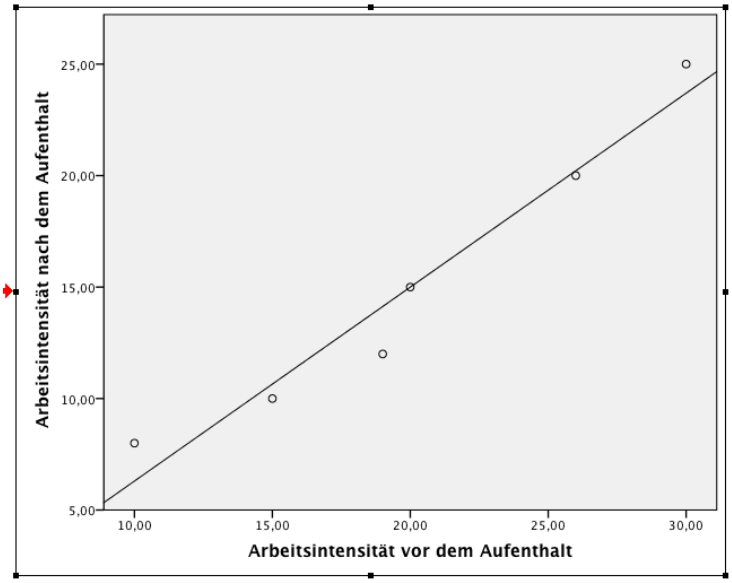
a) Schätzen Sie die Koeffizienten des linearen Regressionsmodells $Y_i = \alpha + \beta X_i + \epsilon_i$.
Zeichnen Sie die Gerade und die Daten in ein Diagramm.

$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = 20$
$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = 15$

$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 = \frac{1}{N-1}\sum_{i=1}^{N} X_i^2 - N\bar{X}^2) = 52.4$
$s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N-1}\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2) = 41.6$
$s_{xy} := \widehat{Cov}(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{N-1}\sum_{i=1}^{N} X_iY_i - N\bar{X}\bar{Y}) = 45.6$

$$
\begin{aligned}
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} \\
\hat{\beta} &= \frac{s_{xy}}{s_x^2} \\
\hat{\beta} &= \frac{45.6}{52.4} = 0.87 \\
\hat{\alpha} &= 15 - 0.87 \cdot 20 = -2.4
\end{aligned}
$$

$$\hat{Y} = -2.4 + 0.87 \cdot x$$



b) Überprüfen Sie die Nullhypothese $H_0 : \beta = 1$ in einer zweiseitigen Fragestellung. (Benutzen Sie $\hat{\sigma}^2 = 2.4$).

Man prüfe, ob der Regressionskoeffizient $\beta = 1$ ist, mit $\alpha = 0.05$; $\hat{\sigma}^2 = 2.4$.

Für die Hypothese $H_0 : \beta = 1$ wird die Testgröße

$$T_\beta = \frac{\hat{\beta} - 1}{\widehat{\sigma_\beta}}$$

verwendet, welche t-verteilt ist mit $(N - 2)$ Freiheitsgraden.

$$\widehat{\sigma_\beta} = \frac{\hat{\sigma}}{\sqrt{\sum_{n=1}^{N} X_n^2 - N\bar{X}^2}} = \frac{\hat{\sigma}}{\sqrt{(N-1)s_x^2}}$$

$$\widehat{\sigma_\beta} = \frac{\sqrt{2.4}}{\sqrt{(6-1)52.4}} = 0.0957$$

Es ergibt sich $T = (0.87 - 1)/0.0957 = -1.3582$. Die t-Quantile sind $t(1-\alpha/2, N-2) = t(0.975, 4) = 2.776$ und $t(\alpha/2, N-2) = t(0.025, 4) = -2.776$. Wegen $T = -1.3582 > t(0.025, 4) = -2.776$ muss die Nullhypothese ($\beta = 1$) beibehalten werden.

c) Berechnen Sie aus den Resultaten in a) die Korrelation zwischen X und Y.

$\hat{\beta} = \frac{s_{xy}}{s_x^2} = R_{xy}\frac{s_y}{s_x}$

$R_{xy} = \hat{\beta}\frac{s_x}{s_y}$

$R_{xy} = 0.87\frac{\sqrt{52.4}}{\sqrt{41.6}} = 0.9764$

d) Wie viel Prozent der Varianz von Y lässt sich durch Kenntnis der Werte in der 1. Befragung erklären?

Gesucht ist das Bestimmtheitsmaß in der Form

$$PRE = \frac{SQE}{SQT} = R_{XY}^2$$

Es ergibt sich der Wert $R_{xy}^2 = 0.9764^2 = 0.953$. Das Bestimmtheitsmaß gibt das Verhältnis von erklärter zu totaler Streuung an. Im vorliegenden Modell werden lediglich 95.3% der Streuung von Y durch X erklärt.

e) Prognostizieren Sie das Resultat Y eines weiteren Austauschstudenten, wenn er vor dem Studienaufenthalt einen Wert $X = 25$ angegeben hat.

$\hat{Y_0} = -2.4 + 0.87 \cdot X_0$

$\hat{Y_0} = -2.4 + 0.87 \cdot 25 = 19.35$

f) Wie groß ist das 95%-Prognoseintervall für diesen vorhergesagten Testwert?

Das Konfidenzintervall für die Regressionsgerade wird mit

$$\hat{E}[Y|X] \pm t(1-\alpha/2, N-2)\sqrt{\widehat{Var}(\hat{E})}$$

angegeben, wobei

$$\sqrt{\widehat{Var}(\hat{E})} = \hat{\sigma}\sqrt{\frac{1}{N} + \frac{(X-\bar{X})^2}{\sum_n X_n^2 - N\bar{X}^2}}$$

gilt.

$$\sqrt{\widehat{Var}(\hat{E})} = \sqrt{2.4}\sqrt{\frac{1}{6} + \frac{(x-20)^2}{5 \cdot 52.4}} = \sqrt{0.4 + 0.0038(x-20)^2}$$

Mit $t(1-\alpha/2; N-2) = t(0.975, 4) = 2.776$ lautet das zweiseitige 95%-Konfidenzintervall

$$(-2.4 + 0.87x) \pm 2.776\sqrt{0.4 + 0.0038(x-20)^2}$$

An der Stelle $X_0 = 25$ ergibt sich das 95%-Prognoseintervalls

$$19.35 \pm 1.95$$

16.1.7. Varianzanalyse

Aufgabe 35

Es soll untersucht werden, ob die Arbeitsleistung in einem Betrieb durch den am Arbeitsplatz herrschenden Lärm beeinflusst wird. Der Faktor „Lärm“ als unabhängige Variable wird durch drei Bedingungskategorien (Faktorstufen) repräsentiert. Eine Stichprobe von $n = 24$ Personen wurden zufällig auf die 3 Faktorstufen aufgeteilt, so dass jede Gruppe aus 8 Personen bestand. Für die abhängige Variable „Arbeitsleistung“ wurde ein Leistungsindex entwickelt, welcher in guter Näherung normalverteilte Werte liefert. Die Ergebnisse des Versuchs sind in der folgenden Tabelle enthalten:

Faktor „Lärm“		
I	II	III
54	40	18
46	30	27
42	35	18
50	38	22
46	36	20
45	32	22
40	43	21
48	40	15

Übt der Faktor „Lärm“ einen Einfluss auf die abhängige Variable „Arbeitsleistung“ aus ($\alpha = 0.05$)?

Mittels der Streuungszerlegung $SQT = SQR + SQE$ reicht es aus, zwei Quadratsummen zu berechnen.

$$\begin{aligned} SQE &= \sum_{ij}(\bar{Y}_i - \bar{Y})^2 = J\sum_i \bar{Y}_i^2 - IJ\bar{Y}^2 \\ SQT &= \sum_{ij}(Y_{ij} - \bar{Y})^2 = \sum_{ij} Y_{ij}^2 - IJ\bar{Y}^2 \end{aligned}$$

Zuerst ermittelt man die Gruppenmittelwerte:
$\bar{y}_1 = \frac{1}{8}(54 + 46 + 42 + 50 + 46 + 45 + 40 + 48) = 46.375$
$\bar{y}_2 = \frac{1}{8}(40 + 30 + 35 + 38 + 36 + 32 + 43 + 40) = 36.625$
$\bar{y}_3 = \frac{1}{8}(18 + 27 + 18 + 22 + 20 + 22 + 21 + 15) = 20.375$
$\bar{y}_i = \{46.376, 36.625, 20.375\}$ und den Gesamtmittelwert
$\bar{y} = \frac{1}{3}(46.376 + 36.625 + 20.375) = 34.46.$

$\sum_{ij} Y_{ij}^2 = (54^2 + 46^2 + ... + 21^2 + 15^2) = 31605$

Daraus findet man

$$\begin{aligned} SQE &= 8(46.376^2 + 36.625^2 + 20.375^2) - 3 \cdot 8 \cdot 34.46^2 = 2760.333 \\ SQT &= 31605 - 24 \cdot 34.46^2 = 3107.96. \end{aligned}$$

Die residuale Quadratsumme ist somit $SQR = SQT - SQE = 3107.96 - 2760.33 = 347.627$

Für den F-Test ergibt sich

$$F = \frac{SQE/(I-1)}{SQR(I(J-1))} = \frac{2760.333/2}{347.627/21} = 83.37$$

mit dem Quantil $f(0.95, 2, 21) = 3.47$.

Somit lautet die Tabelle:

SQ	Wert	df	F-Statistik
SQE (zwischen)	2760.333	2	
SQR (innerhalb)	347.627	21	83.37
SQT (total)	3107.96	23	

Das Quantil $F(0.95, 2, 21)$ nimmt den Wert 3.47 an, wegen $F = 83.37 > F(0.95, 2, 21) = 3.47$ muß H_0 abgelehnt werden: der am Arbeitsplatz herrschenden Lärm übt einen Einfluss auf die Arbeitsleistung.

Aufgabe 36

Bei einem Medikament hat man den Verdacht, dass es als Nebenwirkung zu verlangsamten Reaktionen führen könnte. Um dies zu prüfen, werden $N = 30$ Vpn in drei Gruppen A, B, C eingeteilt. Gruppe A erhält das Medikament, Gruppe B ein Placebo, Gruppe C gar nichts. In einem Experiment erhielt man als Reaktionszeit (in sec.):

A	1.2	0.7	0.8	0.9	0.6	0.7	0.8	0.9	1.0	0.7
B	1.0	0.5	0.6	0.4	0.7	0.5	0.8	0.4	0.4	0.8
C	0.7	0.4	0.5	0.5	1.0	0.6	0.5	0.5	0.8	0.3

a) Zunächst beschränken wir uns auf den Vergleich der Gruppen A und C.
Vergleichen Sie mit einem einseitigen Test die Mittelwerte dieser beiden Gruppen ($\alpha = 0.05$).
Hinweis: $s_1^2 = 0.031$, $s_2^2 = 0.042$
Unterstellen Sie Varianzhomogenität und Normalverteilung. Geben Sie Null- und Alternativ-Hypothese an und legen Sie diese so, dass Sie einen sinnvollen Test zur Klärung des obigen Verdachts erhalten.

Die Nullhypothese lautet :
$H_0 : \mu_A - \mu_C \leq 0$
$H_1 : \mu_X - \mu_Y > 0$ (Medikament führt zu verlangsamten Reaktionen)

$$\hat{\mu_A} = \bar{Y_A} = \frac{1}{N} \sum_{n=1}^{N} Y_A = 0.83$$
$$\hat{\mu_C} = \bar{Y_C} = \frac{1}{N} \sum_{n=1}^{N} Y_C = 0.58$$

Unter den getroffenen Annahmen (Normalverteilungsannahme, Varianzhomogenität, σ_X und σ_Y unbekannt, $N, M < 30$, $\alpha = 0.05$) wird zur Berechnung der Prüfgröße $T = \frac{\bar{Y_A} - \bar{Y_C}}{S}$ die Varianz benötigt:

$$S^2 = (\frac{1}{N} + \frac{1}{N}) \frac{(N-1)s_{YA}^2 + (N-1)s_{Yc}^2}{N+N-2} =$$

$$= (\frac{1}{10} + \frac{1}{10}) \frac{9 \cdot 0.031 + 9 \cdot 0.042}{18} = 0.0073$$

Da sich der Wert der Prüfgröße T zu

$$T = \frac{\bar{Y_A} - \bar{Y_C}}{S} = \frac{0.83 - 0.58}{\sqrt{0.0073}} = 2.926$$

ergibt, mit $T = 2.926 > t(0.95, 18) = 1.734$ kann die Nullhypothese (das Medikament wirkt nicht auf die Reaktionsfähigkeit) abgelehnt werden: das Medikament führt zu verlangsamten Reaktionen .

b) Prüfen Sie die Hypothese, die Populations-Mittelwerte der Reaktionszeiten seien für alle drei Gruppen gleich ($\alpha = 0.05$).

Mittels der Streuungszerlegung $SQT = SQR + SQE$ reicht es aus, zwei Quadratsummen zu berechnen.

$$
\begin{aligned}
SQE &= \sum_{ij}(\bar{Y}_i - \bar{Y})^2 = J\sum_i \bar{Y}_i^2 - IJ\bar{Y}^2 \\
SQT &= \sum_{ij}(Y_{ij} - \bar{Y})^2 = \sum_{ij} Y_{ij}^2 - IJ\bar{Y}^2
\end{aligned}
$$

Zuerst ermittelt man die Gruppenmittelwerte
$\bar{y}_i = \{0.83, 0.61, 0.58\}$, den Gesamtmittelwert
$\bar{y} = \frac{1}{3}(0.83 + 0.61 + 0.58) = 0.6733$, und $\sum_{ij} Y_{ij}^2 = 15.02$

Daraus findet man

$$
\begin{aligned}
SQE &= 10(0.83^2 + 0.61^2 + 0.58^2) - 3 \cdot 10 \cdot 0.6733^2 = 0.3727 \\
SQT &= 15.02 - 30 \cdot 0.6733^2 = 1.4187.
\end{aligned}
$$

Die residuale Quadratsumme ist somit $SQR = SQT - SQE = 1.4187 - 0.3727 = 1.046$

Für den F-Test ergibt sich

$$
F = \frac{SQE/(I-1)}{SQR(I(J-1))} = \frac{0.3727/2}{1.046/27} = 4.81
$$

mit dem Quantil $f(0.95, 2, 27) = 3.35$.

Mit $F = 4.81 > F(0.95, 2, 27) = 3.35$ muß H_0 abgelehnt werden: die Mittelwerte der Reaktionszeiten sind nicht für alle drei Gruppen gleich.

Aufgabe 37

Bei einer Waldschadensuntersuchung wird in 15 Waldstücken gleicher Größe die Anzahl stark geschädigter Bäume erhoben. Die Waldstücke sind so gewählt, dass sie sich hinsichtlich der klimatischen Bedingungen unterscheiden. Man erhielt folgende Daten:

Klima I	7	8	12	13	10
Klima II	10	7	8	13	12
Klima III	14	18	13	19	16

a) Überprüfen Sie mit einem geeigneten Testverfahren, ob die klimatischen Bedingungen einen Einfluss auf die Anzahl stark geschädigter Bäume haben ($\alpha = 0.05$). Geben Sie an, welche Voraussetzungen in Ihre Analyse eingehen.

Mit der Varianzanalyse wird geprüft, ob die klimatischen Bedingungen einen Einfluss auf die Anzahl stark geschädigter Bäume haben.

$$\begin{aligned} SQE &= \sum_{ij}(\bar{Y}_i - \bar{Y})^2 = J\sum_i \bar{Y}_i^2 - IJ\bar{Y}^2 \\ SQT &= \sum_{ij}(Y_{ij} - \bar{Y})^2 = \sum_{ij} Y_{ij}^2 - IJ\bar{Y}^2 \end{aligned}$$

Zuerst ermittelt man die Gruppenmittelwerte
$\bar{y}_i = \{10, 10, 16\}$, den Gesamtmittelwert
$\bar{y} = \frac{1}{3}(10 + 10 + 16) = 12$, und $\sum_{ij} Y_{ij}^2 = 2358$

Daraus findet man

$$\begin{aligned} SQE &= 5(10^2 + 10^2 + 16^2) - 3 \cdot 5 \cdot 12^2 = 120 \\ SQT &= 2358 - 15 \cdot 12^2 = 198. \end{aligned}$$

Die residuale Quadratsumme ist somit $SQR = SQT - SQE = 198 - 120 = 78$

Für den F-Test ergibt sich

$$F = \frac{SQE/(I-1)}{SQR(I(J-1))} = \frac{120/2}{78/12} = 9.23$$

mit dem Quantil $f(0.95, 2, 12) = 3.89$.

Mit $F = 9.23 > F(0.95, 2, 12) = 3.89$ muß H_0 abgelehnt werden: Klima übt einen Einfluss auf die Anzahl geschädigter Bäume.

b) Prüfen Sie bei Einhaltung eines gemeinsamen Signifikanzniveaus von ($\alpha = 0.05$), welche Klimabedingungen sich voneinander unterscheiden.

Der F-Test hat bestätigt, dass es einen signifikanten Unterschied zwischen den Mittelwerten der Anzahl der geschädigter Bäume gibt. Aber es ist noch nicht geklärt, welche Mittelwertunterschiede zu der Ablehnung von H_0 geführt haben. Um es zu prüfen, führt man Post-Hoc-Test durch.

Als Teststatistiken verwendet man die T-Brüche

$$T_{ii'} = \frac{\bar{Y}_i - \bar{Y}_{i'}}{S\sqrt{2/J}} \sim t(IJ - I). \quad (1)$$

mit

$$S^2 = \widehat{\sigma^2} = SQR/(I(J-1)) = MQR, \quad (2)$$

.

Im Beispiel ist $s^2 = SQR/(I(J-1)) = 78/12 = 6.5; s = 2.55$. Der Nenner der T-Statistik ist also $s\sqrt{2/J} = 2.55\sqrt{2/5} = 1.61$.

Die 3 T-Statistiken lauten (die signifikanten Vergleiche sind durch * gekennzeichnet)

$T_{ii'}$	10	10	16
10		0	3.7267*
10			3.7267*
16			

mit dem kritischen t-Wert $t(1-\alpha/(2k), df) \approx 2.87; k = 3, df = 12, \alpha = 0.05$.

Allerdings würde auch hier der Vergleich 1–2 nicht signifikant. Der Mittelwert geschädigter Bäume in Klima III unterscheidet sich signifikant von den Mittelwerten geschädigter Bäume in Klima I und Klima II.

c) Angenommen, Sie wissen nur, ob die Anzahl geschädigter Bäume an den drei Standorten jeweils unter 12 oder mindestens 12 beträgt. Wie lässt sich ein möglicher Unterschied der Standorte untersuchen? Formulieren Sie eine Nullhypothese und geben Sie eine geeignete Teststatistik an. Welcher Nachteil entsteht im Verhältnis zur Auswertungsmethode in a)?

In diesem Fall kann man für jede Gruppe testen, ob der Mittelwert unter 12 oder mindestens 12 beträgt. Dafür eignet sich der Test für Mittelwert:
$H_0 : \mu < \mu_0 = 12$
$H_1 : \mu \geq \mu_0 = 12$

Als Prüfgröße wird $Z = \frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}$ verwendet, welche approximativ standardnormalverteilt ist ($N = 5, \alpha = 0.05$), mit $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{N}}$.

Wenn die Nullhypothese abgelehnt wird, handelt es sich um den Standort mit mindestens 12 geschädigter Bäume. Andernfalls sind am Standort weniger als 12 geschädigter Bäume. Nachteil in dieser Statistik ist es, dass man für jeden Standort einen eigenen Test durchführen muss.

16.2. Aufgaben KE 2

Aufgabe 1

a) Ein Fragebogen besteht aus 3 items A, B, C. Wie viele verschiedene Fragebögen lassen sich daraus bilden?
Bei der Fragebogenkonstruktion spielt eine große Rolle die Reihenfolge der Items. Betrachtet man einen Fragebogen mit den Items A,B und C, so sind die Versionen $F_1 = (A, B, C)$, $F_2 = (A, C, B)$, $F_3 = (B, A, C)$, $F_4 = (B, C, A)$, $F_5 = (C, A, B)$, $F_6 = (C, B, A)$ unterschiedlich in der Reihenfolge. So besteht 6 Möglichkeiten den Fragebogen mit 3 Items zu konstruieren.

b) Ergeben diese das gleiche Ergebnis, wenn sie jeweils einer vergleichbaren Stichprobe von Probanden vorgelegt werden?
Man bekommt nicht immer das gleiche Ergebnis aus den Fragebögen, in denen die gleichen Items in verschiedenen Reihenfolgen platziert sind. Der Grund dazu ist es, dass die Randverteilungen des Items A unterscheiden können, wenn das Item A vor oder nach dem Item B platziert wurde.

c) Diskutieren Sie das Problem der Kontextabhängigkeit von Antworten (responses) sowie das Problem der Reaktivität von Messungen. Sind analoge Phänomene auch in der Naturwissenschaft bekannt?

Kontexabhängigkeit von Antworten bedeutet, dass die Antwort auf eine Frage von derselben Person eventuell anders ausfallen kann, je nachdem in welches gedankliche Umfeld man die Person durch die vorherige Fragen geführt hat.

Reaktivität von Messungen bedeutet, dass die Messung auch dadurch beeinflusst wird, dass sich Personen in der Testsituationen anders verhalten, als sie es unter normalen Umständen tun würden.

Analoge Phänomene sind auch in der Naturwissenschaft bekannt: Orts- und Impulsmessungen, Ortsmessung von Elektronen.

Aufgabe 2

a) Welche Phasen werden beim Forschungsprozess der (quantifizierenden) empirischen Forschung unterschieden?

Man unterscheidet im Rahmen der quantifizierenden Forschung zwischen folgenden Phasen des Forschungsprozesses (nach Bortz, 2005):

1. Erkundungsphase,
2. Theoretische Phase
3. Planungsphase
4. Untersuchungsphase
5. Auswertungsphase
6. Entscheidungsphase

Die Phasen können möglicherweise mehrmals durchlaufen werden.

b) Erklären Sie die Begriffe Indikatoren, Operationalisierung, Falsifikatoren, Tautologie und Kontradiktion.

Indikatoren sind die beobachtbare Sachverhalte, die mit den theoretischen Begriffen möglichst übereinstimmen.
Operationalisierung ist der Oberbegriff von Messung, Skalierung und Indexbildung. Sie legt fest, mit welchen Indikatoren ein theoretisches Konstrukt gemessen werden soll.
Falsifikatoren sind Aussagen oder experimentelle Ergebnisse, die Ungültigkeit einer Aussage, Methode, These, Hypothese oder Theorie nachweisen können.
Tautologie ist eine allgemein gültige Aussage, d.h. eine Aussage, die aus logischen Gründen immer wahr ist.
Kontradiktion ist es, wenn zwei Aussagen einander widersprechen. So wird eine Beziehung zweier Aussagen genannt, bei der sowohl von der Wahrheit der einen Aussage auf die Falschheit der anderen geschlossen werden kann als auch von der Falschheit der einen Aussage auf die Wahrheit der anderen.

c) Was ist der Unterschied zwischen einer Beobachtungs(Survey)-Studie und einem geplanten Experiment. Was ist bei der Analyse von Wirkungszusammenhängen zu beachten, insbesondere bei der Modellierung der unabhängigen Variablen (UV).

Im Gegensatz zu einem Experiment liegen bei reinen Beobachtungs(Survey)-Studien allerdings die Werte der Variablen in ihrem natürlichen Umfeld faktisch vor, so daß die Unterscheidung in unabhängig/abhängig entweder durch theoretische Überlegungen oder statistische Modelle (etwa Strukturgleichungsmodelle) erfolgen muß.
Sehr wichtig ist, bei Analysen von Beobachtungs-Studien zu beachten, daß

auch die UV (X) zufällig sind. Dann lassen sich die y-Werte nicht bei festen x-Werten wiederholen, sondern es gibt nur zufällige Wertepaare $(x_n, y_n), n = 1, \ldots, N$. Die üblichen Resultate sind dann nur **bedingt** auf die gemessenen $X_n = x_n$-Werte zu interpretieren.

d) Was sind Störvariablen und wie wirken sie sich aus?

Die potentiell wichtigen, jedoch nicht miterhobenen Variablen werden oft als Störvariablen bezeichnet. Dies umfaßt auch Größen, deren Einfluß überhaupt nicht bekannt ist. Störvariablen werden häufig als latente stochastische Größen modelliert (etwa Personeneffekte), da sie nicht erhoben wurden, jedoch auch nicht vernachlässigt werden können.

e) Kann die sog. Validität einer Variable auch empirisch überprüft bzw. quantifiziert werden?
Die Validität, ist ein Maß dafür, ob die bei der Messung erzeugten Daten wie beabsichtigt die zu messende Größe repräsentieren. Die Validität wird durch Experten-Schätzung festgelegt und kann nicht empirisch überprüft bzw. quantifiziert werden.

Aufgabe 3

Welche der folgenden Aussagen sind richtig?

A Der empirische Gehalt einer Hypothese hängt von dem Grad ihrer Falsifizierbarkeit ab.

B Die Falsifizierbarkeit einer Hypothese ist unabhängig von ihrem Wahrheitsgehalt.

C Die Operationalisierung ist eine Methode, die durch Ausprobieren aller oder vieler möglicher Fälle Lösungen von Problemen sucht.

D Ein Interviewer erhält eine nach seiner zugrundegelegten Theorie unerwartete Antwort. Mittels des Exhaustionsprinzips kann der Interviewer eine nach seiner Theorie erwartete Antwort erhalten, indem er Zusatzbedingungen einfügt.

E Keine der Aussagen A - D ist richtig.

Lösung: A, B, D

Aufgabe 4

a) Erläutern Sie bitte das Kontrapositionsgesetz

$$(A \Rightarrow B) \Longleftrightarrow (\bar{B} \Rightarrow \bar{A})$$

an einem Beispiel.
Aus A folgt B, dann gilt auch: wenn B nicht gilt, gilt auch A nicht. Aus A="Die Sonne geht unter" folgt B=Ës wird dunkel". Dann stimmt auch die Aussage: Wenn es nicht dunkel wird ($\bar{B}$) , geht die Sonne nicht unter ($\bar{A}$).

b) Berechnen Sie $\overline{A \wedge B}$ ($\bar{A}$ = nicht A).

$\overline{A \wedge B} = \bar{A} \vee \bar{B}$

c) Was ergibt sich aus der Falsifikation der Aussage eines Satzsystems

$$T \wedge A \wedge R \Rightarrow O,$$

wobei T die Theorie, A Zusatzhypothesen und R Randbedingungen sind.

Falls O nicht gilt, ist entweder T, A oder R falsch. Wenn man empirisch überprüfbare Resultate nicht bekommen hat, ist entweder die Theorie, oder Zusatzhypothesen oder Randbedingungen falsch sind.

d) Kann ein Theorie wirklich falsifiziert werden?

Nein. s. Duhem-Quine-These. Die Duhem-Quine-These behauptet die Unterbestimmtheit einer Theorie durch Beobachtungsdaten. Demnach besteht eine Theorie aus vielen miteinander verknüpften Aussagen, die zusammen ein möglichst kohärentes Ganzes bilden. Dementsprechend kann eine Theorie nicht durch einzelne empirische Beobachtungen und Experimente verifiziert oder falsifiziert werden: es stehen immer eine Reihe weiterer Theorien mit zur Debatte.

Aufgabe 5

Gegeben sei folgendes Gesetz:

Wenn ein Verhalten X_1 unterdrückt wird, dann erhöht sich die Auftretenswahrscheinlichkeit eines bestimmten Ereignisses Y_1.

Welche der folgenden Aussagen widerlegt das Gesetz? (x aus 5)

A Ein Verhalten X_2 wird nicht unterdrückt und die Auftretenswahrscheinlichkeit des Ereignisses Y_1 erhöht sich.

B Ein Verhalten X_2 wird nicht unterdrückt und die Auftretenswahrscheinlichkeit des Ereignisses Y_1 erhöht sich nicht.

C Das Verhalten X_1 wird unterdrückt und die Auftretenswahrscheinlichkeit des Ereignisses Y_1 erhöht sich nicht.

D Das Verhalten X_1 wird nicht unterdrückt und die Auftretenswahrscheinlichkeit des Ereignisses Y_1 erhöht sich nicht.

E Keine der Aussagen A - D ist richtig.

Lösung: C.

Aufgabe 6

Gegeben sind drei Hypothesen:

I Je höher der Schulabschluss ist, desto höher ist das Einkommen und die Firmenposition.

II Je höher der Schulabschluss ist, desto höher ist das Einkommen.

III Je höher der Schulabschluss und das Engagement ist, desto höher ist das Einkommen und die Firmenposition.

Welche der folgenden Aussagen sind richtig?

A Allgemein gilt: Der Informationsgehalt einer Hypothese steigt, wenn weitere Elemente in die Dann-Komponente aufgenommen werden (Wenn A, dann B und C).

B Allgemein gilt: Der Informationsgehalt einer Hypothese steigt, wenn weitere Elemente in die Wenn-Komponente aufgenommen werden (Wenn A und B, dann C).

C Hypothese III besitzt den größten Informationsgehalt.

D Hypothese II besitzt den größten Informationsgehalt.

E Hypothese I besitzt den größten Informationsgehalt.

Lösung: A, E

Aufgabe 7

Welche der folgenden Aussagen sind richtig?

A Die Reliabilität ist der Anteil der geschätzten Varianz an der Gesamtvarianz.

B Die Reliabilität setzt die Varianz des wahren Wertes mit der Varianz des beobachteten Wertes ins Verhältnis.

C Die Reliabilität entspricht dem Quadrat der Korrelation des beobachteten Wertes mit dem wahren Wert.

D Da der wahre Wert nicht bekannt ist, kann die Reliabilität nicht berechnet, sondern lediglich geschätzt werden.

E Keine der Aussagen A - D ist richtig.

Lösung: B, C

Aufgabe 8

Welche der folgenden Aussagen sind richtig?

A Die Reliabilität stellt neben der Validität und der Repräsentativität eines der drei wichtigsten Gütekriterien für empirische Untersuchungen dar.

B Die Höhe der Validität ist unabhängig von der Höhe der Reliabilität.

C Unter der Voraussetzung von gleichen Bedingungen ist die Reliabilität ein Maß für die Replizierbarkeit von Messwerten.

D Die Reliabilität nimmt Werte zwischen –1 und 1 an.

E Keine der Aussagen A - D ist richtig.

Lösung: C

Aufgabe 9

Welche der folgenden Aussagen sind richtig?

A Die Validität ist ein Maß für die inhaltliche Gültigkeit einer Untersuchung.

B Eine hohe Validität impliziert eine hohe Reliabilität und umgekehrt.

C Mit Hilfe der Reliabilität können systematische Fehler identifiziert werden.

D Nimmt die Reliabilität den Wert 1 an, so liegt kein Messfehler vor.

E Keine der Aussagen A - D ist richtig.

Lösung: A, D

Aufgabe 10

Welche der folgenden Aussagen sind richtig?

A Ein statistisches Maß für die Interrater-Reliabilität ist der Kappa(κ)-Koeffizient von Cohen, der die Vergleichbarkeit verschiedener Probanden misst.

B Kappa ist die Differenz der Nichtübereinstimmungen der Beurteiler (zufällige minus tatsächliche) bezogen auf die zufällige Nichtübereinstimmung.

C
$$\kappa = \frac{G(+) - G(-)}{1 - G(-)} = \frac{\sum_{i=1}^{I} f_{ii} - \sum_{i=1}^{I} f_{i.} \cdot f_{.i}}{1 - \sum_{i=1}^{I} f_{i.} \cdot f_{.i}}$$

D Kappa nimmt Werte in dem Bereich von –1 bis 1 an.

E Keine der Aussagen A - D ist richtig.

Lösung: C

16.3. Aufgaben KE 3

Aufgabe 1

a) Was ist der Unterschied zwischen der Datenansicht und der Variablenansicht eines SPSS-Datensatzes Daten.sav?

Datenansicht: 1 Fall = 1 Zeile, in den Spalten stehen die Werte für die Variablen. Daten können direkt eingegeben oder bearbeitet werden, eine Berechnung (=Eingabe von Formeln) ist hier nicht möglich.
Variablenansicht: In jeder Zeile wird eine Variable definiert (Name, Typ, vordefinierte Werte, etc.)

b) Erklären Sie den Zweck der Spalteneinträge:
Name, Typ, Variablenlabel, Spaltenlabel, fehlende Werte, Messniveau
in der Variablenansicht.

Der Name der Variable ist eine Zeichenkette, die zur eindeutigen Identifikation der Variable führt.
Typ der Variable kann numerisch oder string sein. String Typ wird für die Variable mit expliziten Zeichenketten, die z.B. Buchstaben enthalten, benutzt.
Spaltenformat steht für die Anzahl der Zeichen, die maximal verwendet werden sollen.
Variablenlabel dient der Beschreibung einer Variable.
Fehlende Werte: Hat man, wie im Beispiel der Variable Geschlecht, mit der 2 einen Wert festgelegt, der eigentlich einen fehlenden Wert darstellt, so kann man diesen explizit als solchen deklarieren. Fehlende Werte werden bei der Berechnung von Mittelwerten nicht berücksichtigt.
Messniveau: Es wird zwischen Nominalskala (Nominal), Ordinalskala (Ordinal) und metrischer Skala (Skala/Metrisch) unterschieden.

Aufgabe 2

a) Was ist eine Datenmatrix, welche Dimensionen hat sie und wie wird sie in Statistik-Programmen repräsentiert?

Formal ausgedrückt ist eine Datenmatrix ein rechteckiges (zweidimensionales) Schema, das aus Zeilen und Spalten zur Aufnahme von Daten besteht. Die Datenmatrix besteht aus den Zeilen $i = 1, 2, ..., n$, den Spalten $j = 1, 2, ..., m$, und den Elementen (Daten) x_{ij}. Üblicherweise werden die Objek-

te zeilenweise angeordnet, die einzelnen Variablen (Merkmale) spaltenweise . Die Datenmatrix aus den n Zeilen und den j Spalten hat $n \times m$ Dimensionen. In SPSS ist die Datenmatrix in Datenansicht anschaulich.

b) Welche statistische Annahme wird im allgemeinen für verschiedene Zeilen der Datenmatrix gemacht?
In der Datenmatrix entspricht jede Zeile einer Untersuchungseinheit (Person, Unternehmen etc.). Eine statistische Annahme ist es, dass die Zeilen voneinander unabhängig sein sollen (Stichprobe ohne Messwiederholung).

c) Wie kann ein SPSS-Datensatz durch manuelle Eingabe erzeugt werden?
In der Variablenansicht kann man eine neue Variable definieren und dann in der Datenansicht die Werte der Variable eingeben.

d) Die Variable Alter wurde als stetige (kontinuierliche Variable) erhoben. Sie interessieren sich aber nur für die Altersklassen (jung/mittel/alt). Wie gehen Sie vor, um eine neue Variable Alterklassiert zu erzeugen und im Datensatz abzuspeichern?

In bestimmten Fällen ergeben sich die Werte von Variablen aus den Werten anderer Variablen. Man gibt dann für die Variable nur die Berechnungsformel ein, anstatt für jeden Fall selbst Wert auszurechnen und einzutippen. SPSS stellt dafür einen eigenen Formeleditor mit sehr umfangreichen Funktionen zur Verfügung. Ein Beispiel, was bei der Analyse gesprochener Sprache häufig benötigt wird, ist der Stimmumfang (Range). Man würde also den minimalen und maximalen Grundfrequenzwert messen, diese Werte in SPSS einlesen und eine neue Variable Range erstellen, indem man aus dem Menü **Transformieren** den Befehl **Berechnen** auswählt, dann den Namen und Typ der Variable festlegt und die entsprechende Berechnungsformel eingibt. Das Dialogfeld Variable berechnen zeigt auf der linken Seite die vorhandenen Variablen an. Aus dieser Liste können die für die Berechnung erforderlichen Variablen in das rechte Feld für die Formel übernommen werden.

Aufgabe 3
a) Welche graphischen Darstellungen und Tabellen werden bei einer explorativen Analyse erstellt?

Bei einer univariater Analyse werden deskriptive Statistiken (Mittelwert, Standardfehler, Median, Modus etc.) berechnet, eine Häufigkeitstabelle mit den Merkmalsausprägungen und den zugehörigen absoluten und relativen Häufigkeiten gerechnet, Histogramme, Balken- und Kreisdiagramme können zur Veranschaulichung genutzt werden.

Bei einer bivariaten Analyse wird der Zusammenhang von Variablen untersucht. Dabei werden Kreuztabelle und Streudiagramme erstellt.

b) Was ist der Unterschied zwischen einem Stabdiagramm und einem Histogramm, auch im Hinblick auf das Skalenniveau der Variable.

Auf einem Stabdiagramm werden relative bzw. absolute Häufigkeiten der Beobachtungswerte einer *nominalskalierten* Variable dargestellt.
Ein Histogramm ist eine graphische Darstellung der Häufigkeitsverteilung *metrisch skalierter* Merkmale.

c) Warum ist es sinnvoll, dichotome Variablen mit den Werten 0/1 zu codieren (Indikatoren, Dummy-Variablen, 0-1-Variablen).

Bei den numerisch codierten Variablen sind deskriptive Statistiken (Mittelwert, Standardabweichung, etc.) nicht sinnvoll interpretierbar, da die Ordnung der Ausprägungen nicht interpretierbar ist. Deshalb muss man aus den numerisch skalierten Variablen binäre Variablen mit 0-1 Codierung erstellen, was zu einer verbesserten Interpretierbarkeit führt.

d) Wie läßt sich der Mittelwert einer Indikatorvariable interpretieren?

Der Mittelwert einer Indikatorvariable ist als die relative Häufigkeit der Variabel für die Ausprägung 1 zu interpretieren.

Aufgabe 4

a) Nennen Sie Zusammenhangsmaße für nominale Variablen X und Y.

χ^2 Koeffizient, K- Kontingenzkoeffizient, τ-Kendall-Tau und Goodman-Kruskal.

b) Wie wird der λ-Koeffizient (Goodman-Kruskal) in einer Kreuztabelle berechnet. Warum ist dies ein a posteriori-Maß?

Goodman-Kruskal $\lambda(x \to y)$
$G(+) = \sum_i f_{i\hat{j}(i)}$ = Summe der Modalhäufigkeiten der Zeilen i
$G(-) = f_{\cdot\hat{j}}$ = Maximum der Randverteilung (Spalten j)
$\hat{j}$ = Modalwert der Randverteilung (Spalten j)
$\hat{j}(i)$ = Modalwert in Zeile i
$PRE = \frac{F(-)-F(+)}{F(-)} = \frac{G(+)-G(-)}{1-G(-)}$

Goodman-Kruskal $\lambda(y \to x)$
$G(+) = \sum_j f_{\hat{i}(j)j}$ = Summe der Modalhäufigkeiten der Spalten j
$G(-) = f_{\hat{i}.}$ = Maximum der Randverteilung (Zeilen i)
$\hat{i}$ = Modalwert der Randverteilung (Zeilen i)
$\hat{i}(j)$ = Modalwert in Spalte j = Modalregel in den Spalten (Ausprägungen von Y)
$PRE = \frac{F(-)-F(+)}{F(-)} = \frac{G(+)-G(-)}{1-G(-)}$

Es handelt sich um ein **a-posteriori-Maß**, da die Modalwerte aus den Daten (empirische Kreuztabellen) gewonnen werden.

c) Was sind a priori-Zusammenhangsmaße?

Cohen κ ist ein a priori-Zusammenhangsmaß.

$PRE = \frac{F(-)-F(+)}{F(-)} = \frac{G(+)-G(-)}{1-G(-)}$

mit
$G(+) = \sum_{ij,H} f_{ij}$
$G(-) = \sum_{ij,H} f_{i.}f_{.j}$

In diesem Fall wird eine **a-priori-Hypothese** H formuliert, welche Ausprägungen von X und Y zusammenhängen (etwa Diagonale: Beobachterübereinstimmung). In diesem Fall ist $H = \text{diag}(1, \ldots, 1)$. Die Summation erfolgt nur über die Zellen der Kreuztabelle, wo $H_{ij} = 1$ ist.

d) Berechnen Sie Cohen-κ aus der folgenden Kreuztabelle f_{ij} der Urteile A, B, C von 2 Beobachtern, die ein Ereignis klassifizieren sollten (Hypothese: exakte Übereinstimmung der Beobachter):

f_{ij}	A	B	C
A	.3	.1	0
B	.2	.2	0
C	0	0	.1

f_{ij}	A	B	C	A+B+C
A	.3	.1	0	0.4
B	.2	.2	0	0.4
C	0	0	.1	0.1
A+B+C	0.5	0.3	0.1	0.9

$$
\begin{aligned}
G(-) &= \frac{0.5 \cdot 0.4 + 0.3 \cdot 0.4 + 0.1 \cdot 0.1}{0.9 \cdot 0.9} = 0.4074 \\
G(+) &= \frac{0.3 + 0.2 + 0.1}{0.9} = 0.6666 \\
\kappa &= \frac{0.6667 - 0.4074}{1 - 0.4074} = 0.4373
\end{aligned}
$$

e) Berechnen Sie aus obiger Kreuztabelle die Koeffizienten $\lambda(x \to y), \lambda(y \to x)$ sowie $\lambda(x \leftrightarrow y)$. Vergleichen Sie mit κ.

$$
\begin{aligned}
\lambda(x \to y) &= \frac{max(0.3; 0.1; 0) + max(0.2; 0.2; 0) + max(0; 0; 0.1) - max(0.5; 0.3; 0.1)}{0.9 - max(0.5; 0.3; 0.1)} = \\
&= \frac{0.3 + 0.2 + 0.1 - 0.5}{0.9 - 0.5} = 0.25 \\
\lambda(y \to x) &= \frac{max(0.3; 0.2; 0) + max(0.1; 0.2; 0) + max(0; 0; 0.1) - max(0.4; 0.4; 0.1)}{0.9 - max(0.4; 0.4; 0.1)} = \\
&= \frac{0.3 + 0.2 + 0.1 - 0.4}{0.9 - 0.4} = 0.4 \\
\lambda(x \leftrightarrow y) &= = \frac{0.3 + 0.2 + 0.1 - 0.5 + 0.3 + 0.2 + 0.1 - 0.4}{0.9 - 0.5 + 0.9 - 0.4} = 0.33333
\end{aligned}
$$

Aufgabe 5

a) Wie lassen sich die Parameter der Regressionsgleichung

$$Y_n = \alpha + \beta X_n + \epsilon_n$$

interpretieren, wenn die Variable X_n als Indikator (0/1) codiert wird. Setzen Sie die Werte ein und schreiben Sie die beiden Gleichungen getrennt auf.

$$Y_n = \alpha + \beta X_n + \epsilon_n, \qquad 1 \leq n \leq N,$$

mit

$$X_n = \begin{cases} 1 & \text{für } n = 1, ...N_1, \\ 0 & \text{für } n = N_1 + 1, ..., N. \end{cases}$$

Nach der Umformung sieht die Gleichung so aus:

$$Y_n = \begin{cases} \alpha + \beta + \epsilon_n, & \text{für } n = 1, ...N_1, \\ \alpha + \epsilon_n & \text{für } n = N_1 + 1, ..., N. \end{cases}$$

Im Falle einer dichotomen nominalskalierten Variablen X kann β als Mittelwertunterschied interpretiert werden.

b) Interpretieren Sie die Parameter der Interaktionsterme in

$$Y_n = \alpha + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_{12} X_{1n} X_{2n} + \epsilon_n$$

wenn X_{1n} eine Indikatorvariable und X_{2n} eine stetige Regressorvariable ist. Stellen Sie sich vor, daß die Variable X_{1n} zwei Gruppen unterscheidet.

Wenn die Indikatorvariable X_{1n} gleich 0 ist, dann ist die Gleichung:

$$Y_n = \alpha + \beta_2 X_{2n} + \epsilon_n$$

Wenn Wenn die Indikatorvariable X_{1n} gleich 1 ist, sieht die Gleichung so aus:

$$Y_n = \alpha + \beta_1 + (\beta_2 + \beta_{12}) X_{2n} + \epsilon_n$$

Aus der zweiten Gleichung ist zu sehen, dass der Regressionskoeffizient des Interaktionsterms β_{12} als Änderung des Einflusses der Variable X_{2n} auf Y_n, wenn die binäre Variable gleich 1 ist, zu interpretieren. So zeigt β_{12} den Unterschied im Zusammenhang zwischen den Y_N und X_{2n} für beide Gruppen ($X_{1n} = 1$ und $X_{1n} = 0$).

c) Kann das Problem auch mit einem t-Test bearbeitet werden (warum)?
Wie der Zusammenhang zwischen den Y_N und $X_2 n$ in beiden Gruppen unterscheidet, kann man mit dem T-Test nicht prüfen. Mit einem T-Test kann man prüfen, ob sich Mittelwerte von Y_N (oder von $X_2 n$) in beiden Gruppen signifikant voneinander unterscheiden.

Aufgabe 6

a) Geben Sie Beispiele für die Skalenniveaus *Nominal, Ordinal, Intervall, Verhältnis.*

Nominal: Geschlecht (Mann/Frau) Raucher (Raucher/Nicht Raucher)
Ordinal: Schulnoten, Tabellenplätze der Bundesliga
Intervall:Temperaturen (Celsius), Zeitskala (Datum)
Verhältnis: Temperaturen (Kelvin), Alter, Größe, Gewicht

b) Welche Transformationen lassen die Skala invariant?

Nominalskala ist invariant gegenüber eineindeutigen, bijektiven Transformationen. Ordinalskala ist nur invariant gegenüber echt monoton steigender Transformationen, die die Ordnung der Klassen unverändert lässt. Die Intervallskala ist invariant gegenüber linearen Transformationen der Form $y = kx + d$ $(k > 0)$. Und die Verhältnissskala ist invariant gegenüber linearen Transformationen der Form $y = kx$ $(k > 0)$.

c) Welche mathematischen Operationen sind für derartige Skalen sinnvoll?
Nominal: Modalwert
Ordinal: Medianwert
Intervall: Mittelwert
Verhältnis: Multiplikationen, geometrisches Mittel

Aufgabe 7

a) Wie unterscheiden sich die Gesichtspunkte *qualitativ / quantitativ, diskret / metrisch* und *kategorial /kontinuierlich* voneinander?

Diskrete Variablen sind qualitative Variablen, die Variablen sind nominal- oder ordinalskaliert. Solche Variable nennt man auch kategorial.
Metrische Variablen sind quantitativ, die sind intervall- oder verhältnissskaliert und die sind kontinuierlich.

b) Welche Statistiken lassen sich bei nominalem Skalenniveau *sinnvollerweise* berechnen?
Bei nominalem Skalenniveau ist sinnvoll das Bestimmen von Auftrittshäufigkeiten der Kategorien in einer Menge von Untersuchungseinheiten, die dann Gegenstand der Statistik sind. Als Lageparameter einer solchen Häufigkeitsverteilung kann lediglich der häufigste Wert bestimmt werden, der sogenannte *Modalwert*.

c) Schulnoten werden üblicherweise zu Durchschnittswerten gemittelt. Unter welchen Voraussetzungen ist dies sinnvoll?

Auch wenn Kategorien durch Zahlen kodiert werden, sind mathematische Operationen mit diesen Zahlen nicht sinnvoll, da sie keinen numerischen Wert, sondern eine Kategorie darstellen. Da es sich bei Schulnoten in der Re-

gel um ordinalskalierte Merkmale handelt, ist die Bildung von Durchschnittsnoten eigentlich nicht sinnvoll. Wenn die Schulnoten aber anders skaliert sind (z.B. mit 100 Punkten - das beste Note und 0 die schlechteste Note) kann auch Durchschnitt gebildet werden.

Aufgabe 8

a) Erläutern Sie das Kontrapositionsgesetz

$(A \Rightarrow B) \Longleftrightarrow (\bar{B} \Rightarrow \bar{A})$

anhand der Aussage:

Korrelierte Variablen sind abhängig.

A: Variablen sind korreliert.
B: Variablen sind abhängig.
$A \Rightarrow B$: Aus der Korrelation der Variablen folgt, dass die Variablen abhängig sind.
$\bar{A}$: Variablen sind unkorreliert.
$\bar{B}$:: Variablen sind unabhängig.
$\bar{A} \Rightarrow \bar{B}$: Wenn die Variablen unabhängig sind, sind sie auch unkorreliert.

b) Was gilt bei unkorrelierten Variablen? Sind diese unabhängig?

Die unkorrelierte Variablen sind nicht unbedingt unabhängig. Z.B. kann ein nichtlinearer Zusammenhang zwischen den Variablen vorliegen. In diesem Fall liegt der Korrelationskoeffizient nahe zu Null, obwohl die Variablen voneinander abhängig sind.

Aufgabe 9

a) Klassifizieren Sie das asymmetrische Modell $X \rightarrow Y$ nach dem Skalenniveau der Variablen.

Symmetrische Maße geben einen Zusammenhang (Assoziation), lassen sich jedoch **nicht** als Wirkungsweisen interpretieren. Dazu sind asymmetrische Ansätze in der Lage, wo zwischen abhängigen (AV = Y) und unabhängigen Variablen (UV = X) unterschieden wird:

Asymmetrische Verfahren $Y = f(X)$		
	AV $= Y$	
UV $= X$	diskret	metrisch
diskret	Kreuztabellen	Varianz-Analyse
metrisch	kategoriale Regression	Regression

b) Wie läßt sich der Quotient aus erklärter und totaler Streuung SQE/SQT beim linearen Regressionsmodell alternativ berechnen und wie kann er interpretiert werden?
Der Quotient aus erklärter und totaler Streuung SQE/SQT beim linearen Regressionsmodell heißt Determinationskoeffizient und kann als quadrierte Korrelation R^2_{XY} berechnet werden. Dieser Koeffizient zeigt der prozentuale Anteil der Streuung Y, die durch X erklärt wird.

c) Was wird beim globalen F-Test der Regressionsanalyse überprüft, insbesondere bei mehreren Regressoren $X_1, \ldots, X_q$?
Mit dem F-Test wird es geprüft, ob der Regressor X (unabhängige Variabel) etwas zur Erklärung von Y beiträgt (man prüft , ob der Regressionskoeffizient β signifikant von 0 verschieden). Bei einem Regressor ist der F-Test äquivalent zum t-Test für β, bei mehreren Regressoren wird es getestet, ob es mindestens ein Regressionskoeffizient β_1 β_2, β_n signifikant von 0 verschieden.

Aufgabe 10

a) Was versteht man unter einer Scheinkorrelation zwischen den Variablen X und Y?

Scheinkorrelation ist eine zweifelsfreie, nicht sinnvoll interpretierbare Korrelation, die durch die Wirkung einer dritten Variable erzeugt wird.

b) Wie kann man den 'tatsächlichen' Zusammenhang zwischen den Variablen (näherungsweise) ermitteln?
Instrumentvariablen-Schätzung, MANOVA mit Kontrollvariable

c) Was ist eine Kontrollvariable und eine partielle Korrelation?

Die Kontrollvariable ist eine Variable, die zur Interpretation oder Erklärung eines Zusammenhangs zwischen zwei anderen Variablen herangezogen wird und mit diesen Variablen hoch korreliert.
Die partielle Korrelation ist die Korrelation zwischen zwei Variablen X und

Y , die übrig bleibt, wenn man den Einfluss einer oder mehrerer anderer Variablen (Kontrollvariable Z) ausgeschaltet hat.

d) Berechnen Sie die partielle Korrelation zwischen X und Y bei Kenntnis der Korrelationen $r_{xy} = .7, r_{xz} = .6, r_{yz} = .3$ mit der Kontrollvariablen Z.

$$\begin{aligned} r_{\tilde{x}\tilde{y}} &= \frac{r_{xy} - r_{xz}r_{zy}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}} \\ &= \frac{0.7 - 0.6 \cdot 0.3}{\sqrt{1 - 0.6^2}\sqrt{1 - 0.3)^2}} = 0.68 \end{aligned}$$

e) Wie kann die Variable Y bei Kenntnis von Z optimal (im Quadratmittel) prognostiziert werden?

Die optimale lineare Prognose im quadratischen Mittel ist durch das lineare Regressionsmodell

$$Y = \alpha + \beta Z + \epsilon : p \times 1$$

mit Parametern

$$\begin{aligned} \alpha &= E[Y] - \beta E[Z] : p \times 1 \\ \beta &= \mathrm{Cov}(Y, Z)\mathrm{Cov}(Z, Z)^{-1} : p \times q \end{aligned}$$

gegeben.

f) Wann ist die lineare Prognose optimal?
Prognostizieren Sie Y linear aus der Kenntnis von $E[Y] = 1, (X, Y) = 0.7$ und $S_X = S_Y = 1$.

Eine lineare Prognose ist optimal, wenn Prognose und Prognosefehler orthogonal (nicht korreliert) sind.

$$\begin{aligned} \hat{Y} &= E[Y] + \mathrm{Cov}(X, Y)\mathrm{Cov}(X, X)^{-1}(X - E[X]) \\ &= 1 + 0.7(X - \bar{X}) \end{aligned}$$

Aufgabe 11

a) Wie hängen die allgemeine Spearman-Brown-Formel und Cronbachs α zusammen?

Die allgemeine Spearman-Brown-Formel:

$$rel = \frac{k\rho_{ij}}{1+(k-1)\bar{\rho}}$$

Cronbachs α:

$$rel = \frac{k\bar{\rho}}{1+(k-1)\bar{\rho}}. \quad (3)$$

Cronbach's Alpha ist die Spearman-Brown-Formel für die durchschnittliche Subtest-Korrelation.

b) Berechnen Sie aus der durchschnittlichen Korrelation $\bar{r} = 0.5$ von 5 Items die interne Konsistenz der Skala. Welchen Wert hätte man bei 10 Items erhalten?

Für $k = 5$: Cronbachs α:

$$rel = \frac{k\bar{\rho}}{1+(k-1)\bar{\rho}} = \frac{5 \cdot 0.5}{1+(5-1)0.5} = 0.8333.$$

Für $k = 10$: Cronbachs α:

$$rel = \frac{k\bar{\rho}}{1+(k-1)\bar{\rho}} = \frac{10 \cdot 0.5}{1+(10-1)0.5} = 0.91$$

c) Was ist bei Berechnung der Retest-Reliabilität zu berücksichtigen?

Bei Berechnung der Retest-Reliabilität ist zu berücksichtigen, dass der true score im Zeitablauf nicht ändert.
Das Messmodell:

$$X_1 = T + \epsilon_1$$
$$X_2 = T + \epsilon_2.$$

d) Was verstehen Sie unter der Abschwächungskorrektur?
Berechnen Sie die Korrelation der Konstrukte T_1 und T_2, wenn die Korrelation der Messungen 0.3 beträgt und eine Reliabilität von 0.7 vorliegt.
Wenn im Modell der true score ändert, sieht das Messmodell so aus:

$$X_1 = T_1 + \epsilon_1$$
$$X_2 = T_2 + \epsilon_2.$$

Aufgrund der Messfehler ist die Korrelation von Indikatoren geringer als die der zugrundeliegenden true scores. Daher muss Abschwächungskorrektur vorgenommen werden.

Berechnen Sie die Korrelation der Konstrukte T_1 und T_2, wenn die Korrelation der Messungen 0.3 beträgt und eine Reliabilität von 0.7 vorliegt.

$$\begin{aligned} \text{Corr}(T_1, T_2) &= \text{Corr}(X_1, X_2)/\sqrt{rel_1 rel_2} \\ &= 0.3/0.7 = 0.4285 \end{aligned}$$

Aufgabe 12

a) Berechnen Sie die Item-Schwierigkeit bei einer 5er-Skala, die zwischen 1 und 5 codiert ist und deren Item-Mittelwert 4.7 beträgt.

Item-Schwierigkeit:
$$0 \leq p_i = \frac{\bar{X}_i - X_{i,min}}{X_{i,max} - X_{i,min}} = \frac{4.7 - 1}{5 - 1} = 0.925$$

b) Welche Trennschärfe hat ein Item, das mit dem Gesamtwert eine Korrelation von 0.71 aufweist?

Trennschärfe ist $\text{Corr}(X_i, X)$. Ein Item, das mit dem Gesamtwert eine Korrelation von 0.71 aufweist, hat Trennschärfe gleich 0.71.

c) Was kann man tun, wenn ein Item eine negative Trennschärfe aufweist?

Items mit geringer Trennschärfe sind schlechte Indikatoren des Konstrukts und werden aus dem Test entfernt.

d) Wie kann die Dimensionalität einer Item-Batterie untersucht werden? Wann kann ein Gesamtwert sinnvoll berechnet werden?

Mit einer Faktorenanalyse oder Clusteranalyse kann die Dimensionalität untersucht werden.

Aufgabe 13

a) Was ist der Unterschied zwischen dem Meßmodell der klassischen Testtheorie und dem Modell der Faktorenanalyse?

Das Messmodell der klassischen Theorie ist ein einfaches Modell, das die gemessenen Werte X durch die wahren Variablen (oder latente Variablen) T und Messfehler ϵ erklärt. $X = T + \epsilon$

Das Modell der Faktorenanalyse analysiert gleichzeitig mehrere Items (Faktoren). Das Modell erklärt auch die gemessenen Werte durch die latenten Faktoren. Das Messmodell der klassischen Theorie ist eine einfache Version der Faktorenanalyse.

b) Schreiben Sie das Modell der Faktorenanalyse auf und interpretieren Sie die Bestandteile (Bezeichnungen, stochastische Annahmen).

Das Faktorenanalyse-Modell ist in allgemeiner Form :
$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon}$
Die Gewichts-Matrix $\boldsymbol{\Lambda} : p \times q$ wird als Faktor-Ladungs-Matrix bezeichnet und $\boldsymbol{\xi} : q \times 1$ ist ein Vektor von Faktoren (latenten Variablen, true scores).
Die Annahmen des Modells sind:

- $\text{Cov}(\boldsymbol{\xi}, \boldsymbol{\epsilon}) = \mathbf{O} : q \times p$ (Faktoren und Fehler sind orthogonal)
- $E[\boldsymbol{\xi}] = 0; E[\boldsymbol{\epsilon}] = E[\mathbf{x}] = 0$ (zentrierte Variablen)
- $\text{Cov}(\mathbf{x}) = \boldsymbol{\Sigma} : p \times p$
- $\text{Cov}(\boldsymbol{\xi}) = \boldsymbol{\Phi} : q \times q$
  ($\boldsymbol{\Phi} = \boldsymbol{I}$: Faktoren sind rechtwinklig und standardisiert)
- $\text{Cov}(\boldsymbol{\epsilon}) = \mathbf{V} : p \times p$ (Fehler-Kovarianzmatrix).

c) Wie können die (latenten) Faktorwerte geschätzt werden?

Die (latenten) Faktorwerte können mit Hauptkomponentenanalyse, Maximum-Likelihood- Methode oder Methode der kleinsten Quadrate geschätzt werden.

d) Berechnen Sie die Kommunalitäten aus der Ladungsmatrix

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0.5 & 1 \end{pmatrix}$$

$$h_1^2 = 1^2 + 0^2 = 1$$

$$h_2^2 = 0^2 + 1^2 = 1$$

$$h_3^2 = 0.5^2 + 1^2 = 1.25$$

e) Was wird bei einer Hauptkomponentenanalyse durchgeführt? Erklären Sie die Begriffe Eigenwert, Eigenvektor und Spektral(Eigenwert)zerlegung.

Die Hauptkomponentenanalyse ist ein Verfahren zur Extraktion der latenten Faktoren. Mittels der Hauptkomponentenanalyse werden unkorrelierte Linearkombinationen der beobachtbaren Variablen gebildet. Die erste Komponente besitzt den größten Varianzanteil. Nachfolgende Komponenten erklären stufenweise kleinere Anteile der Varianz. Bei der HKA wird die beobachtete Kovarianzmatrix so zerlegt, dass eine geringere Zahl der Faktoren ausreicht, um die Beobachtungen zu erklären.

Eigenwertzerlegung:

$$\begin{aligned} \mathbf{\Sigma} &= \sum_{i=1}^{p} \mu_i \psi_i \psi_i' = \mathbf{P\ M\ P'} \\ \mathbf{P} &= [\psi_1, \ldots, \psi_p] : p \times p \\ \mathbf{M} &= \text{Diag}(\mu_1, \ldots, \mu_p) : p \times p \end{aligned}$$

wobei die reellen Zahlen $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_p$ als Eigenwerte und die Vektoren ψ_i als Eigenvektoren bezeichnet werden.

Aufgabe 14

a) Was verstehen Sie unter einem Mosaik-Diagramm und welche Größen werden tabelliert?

Mosaik Plot ist eine graphische Darstellung der bedingten Häufigkeiten.

b) Was ist ein stem-leaf-Diagramm und wie unterscheidet es sich von einem Histogramm?

stem-leaf-Diagramm dient der Visualisierung von Häufigkeitsverteilungen. Das Diagramm besteht aus zwei Spalten. Die linke Spalte enthält als "Stämme" die Äquivalenzklassen, in die die auf der rechten Seite als "Blätter" dargestellten Merkmale eingeteilt werden. Aus einem stem-leaf-Diagramm lassen sich statistische Kennzahlen wie Modalwert, Median und Quantile ablesen. Allerdings stößt diese Art der Darstellung bei einer großen Zahl von Merkmalen an ihre Grenzen. Anders als bei Histogramm bleibt die Darstellung der Werte jeder einzelnen Beobachtung mit gewünschter Genauigkeit erhalten.

c) Wie lassen sich gruppierte Daten (etwa nach Geschlecht) übersichtlich graphisch darstellen?
Die gruppierte Daten (etwa nach Geschlecht) kann man übersichtlich graphisch mit Stabdiagrammen bzw. Histogrammen darstellen.

Aufgabe 15

a) Warum können 0/1-codierte und stetige Variablen in einer gemeinsamen Korrelationsmatrix dargestellt werden?

Die Korrelationen der dichotomen Variablen, die als Indikatoren codiert wurden, lassen sich sinnvoll interpretieren. Produkt-Moment-Korrelationen von X mit anderen Indikatoren Y als ϕ-Koeffizient oder als biseriale Korrelation (Y = Intervallskaliert) gelesen werden können.

b) Wie ist die Korrelation $\mathrm{Corr}(X, Y)$ zwischen einer Indikatorvariablen $X = (0/1)$ und einer stetigen Variablen Y zu interpretieren?

Produkt-Moment-Korrelationen von X mit anderen Indikatoren Y als ϕ-Koeffizient oder als biseriale Korrelation zu interpretieren.

c) Lassen sich Indikatorvariablen und stetige Variablen gemeinsam in einer Regressionsanalyse verwenden?
Ja. In Regressionsgleichung $Y = \alpha + \beta X + \epsilon$, wo X eine unabhängige binäre Variable ist, der Regressionskoeffizient $\hat{\beta}$ als Mittelwertunterschied $= \bar{Y_1} - \bar{Y_0}$ der durch X definierten Gruppen interpretiert werden kann. Außerdem ist die Korrelation eine Maß für den standardisierten Mittelwertunterschied.

d) Wie gehen Sie vor, wenn die nominale Variable X mehr als 2 Ausprägungen hat (etwa Berufsgruppe mit den Ausprägungen Schüler, Arbeiter/Angestellte, Hausfrau/mann,. . .)
Ordinale Variable, d.h. Variable mit 3 und mehr Ausprägungen, wird zunächst in soviele dichotome Variable, d.h. 0-1 kodierten Dummies, aufgelöst, wie sie Ausprägungen besitzen. Erst dann werden die gebildeten binären Variablen in die Gleichung genommen.