

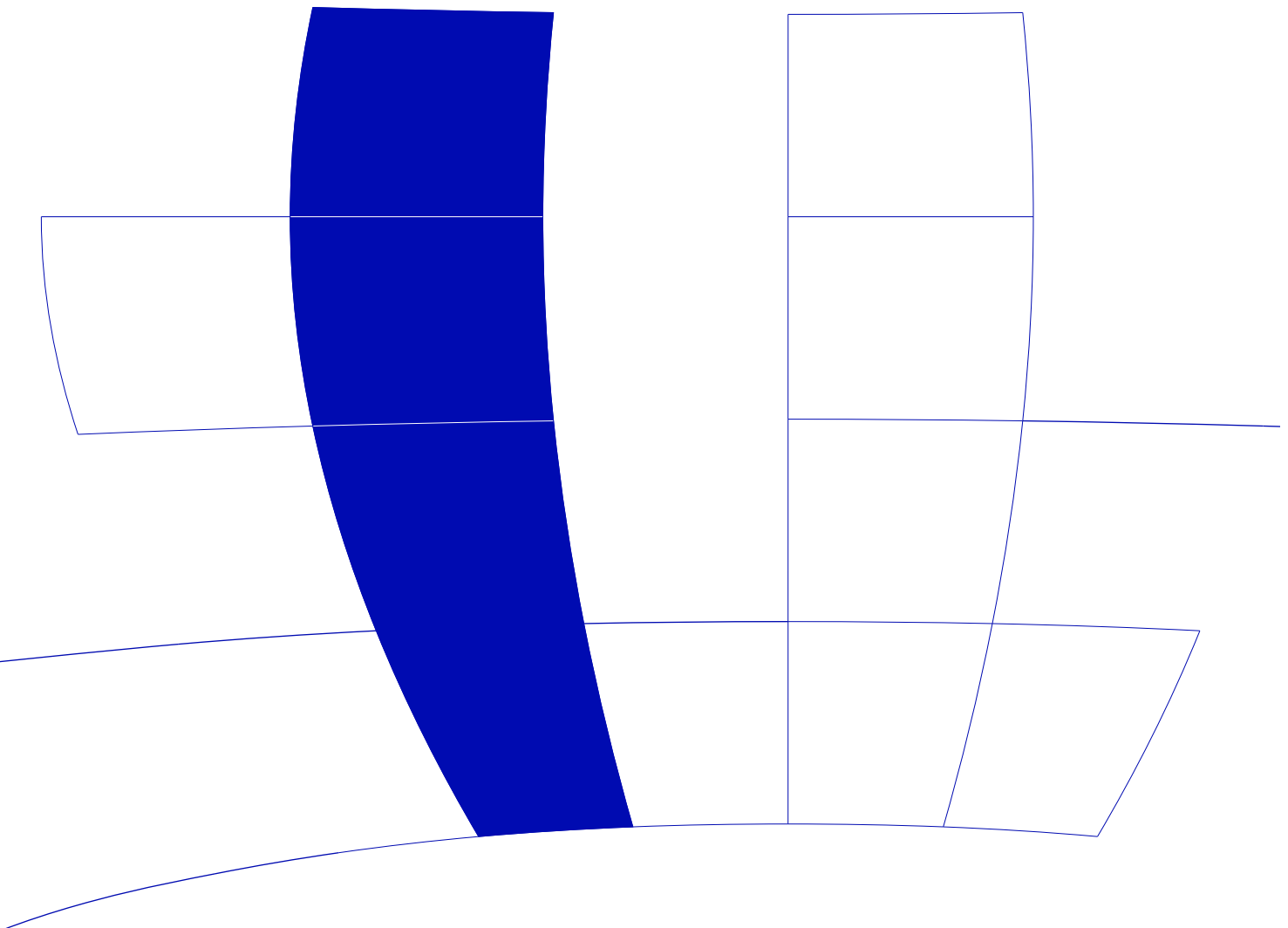
Statistik

Musterlösungen



Regina Tüchler & Achim Zeileis

Institut für Statistik & Mathematik
Wirtschaftsuniversität Wien



1 Grundbegriffe

(1.23) Skript

Reaktionen auf Videofilm. Aussagen M , E , P , S

$$\begin{aligned}
 h(M) &= h(\text{Mann} \cup \text{Junge}) \\
 &= h(\text{Mann}) + h(\text{Junge}) \\
 &= 19 + 15 \\
 &= 34 \\
 h(P) &= h(\text{positiv} \cup \text{sehr positiv}) \\
 &= h(\text{positiv}) + h(\text{sehr positiv}) \\
 &= 21 + 20 \\
 &= 41 \\
 h(P') &= h(\Omega) - h(P) \\
 &= 67 - 41 \\
 &= 26 \\
 h((M \cap E)') &= h(\Omega) - h(M \cap E) \\
 &= 67 - h(\text{Mann}) \\
 &= 67 - 19 \\
 &= 48 \\
 h(M' \cap E \cap P) &= h(\text{Frau} \cap P) \\
 &= 6 + 8 \\
 &= 14
 \end{aligned}$$

- Beispiel:

$$\begin{aligned}
 f(\text{kein Kredit}) &= f(\text{alle Kunden}) - f(\text{Wohnungskredit} \cup \text{Konsumkredit}) \\
 &= 1 - [f(\text{Wohnungskredit}) + f(\text{Konsumkredit}) - f(\text{Wohnungskredit} \cap \text{Konsumkredit})] \\
 &= 1 - [0.28 + 0.23 - 0.13] \\
 &= 0.62
 \end{aligned}$$

2 Beschreibung von Datenlisten

- Beispiel:

$x_{(i)}$	0.8	$\overbrace{4.3 \quad 4.3}$		4.5	$\overbrace{4.8 \quad 4.8}$		$\overbrace{5.2 \quad 5.2}$		5.8	6.6	8.6	10.3	15.9
		Bindung			Bindung		Bindung						
i	1	2	3	4	5	6	7	8	9	10	11	12	13
$r(x)$	1	2	2	4	5	5	7	7	9	10	11	12	13

AbleSEN: $x_{(3)} = 4.3$, $r(4.8) = 5$.

Median: bei ungeradem n (hier $n = 13$ Beobachtungen) ist der Median $Q_{0.5} = x_{(n+1)/2} = x_{(7)} = 5.2$.

Angenommen wir hätten nur $n = 12$ Beobachtungen (ohne $x_{(13)} = 15.9$), dann liegt der Median zwischen $x_{n/2}$ und $x_{n/2+1}$. Hier: $Q_{0.5} = (x_{(6)} + x_{(7)})/2 = (4.8 + 5.2)/2 = 5$.

- Das 75%-Quantil $Q_{0.75}$ der *Verkaufszahlen* ist 62 und wurde an 25% der Tage überschritten. Erlös = Preis · Verkaufszahl. Damit ist das 75%-Quantil der *Erlöse* $25 \cdot 62 = 1550$. Die Interquartilsdistanz QD der Verkaufszahlen ist $QD = 62 - 25 = 37$. Daher ist die Interquartilsdistanz der Erlöse $QD = 25 \cdot 37 = 925$.

3 Die Rolle des Zufalls

46. Binomialverteilung mit $n = 10$, $p = 0.41$, $k = 3$.

$$\begin{aligned}
 P(h_{10}(A) = 3) &= \binom{10}{3} p^3 (1-p)^7 \\
 &= \frac{10!}{3! 7!} 0.41^3 0.59^7 \\
 &= 120 \cdot 0.069 \cdot 0.025 \\
 &= 0.206
 \end{aligned}$$

52. Binomialverteilung mit $n = 10$, $p = 0.15$, $k \geq 2$.

$$\begin{aligned}
 P(h_{10}(A) = 0) &= \binom{10}{0} p^0 (1-p)^{10} \\
 &= 1 \cdot 1 \cdot 0.85^{10} \\
 &= 0.197 \\
 P(h_{10}(A) = 1) &= \binom{10}{1} 0.15^1 0.85^9 \\
 &= 10 \cdot 0.15 \cdot 0.85^9 \\
 &= 0.347 \\
 P(h_{10}(A) \geq 2) &= 1 - P(h_{10}(A) = 0) - P(h_{10}(A) = 1) \\
 &= 1 - 0.197 - 0.347 \\
 &= 0.456
 \end{aligned}$$

51. Hypergeometrische Verteilung mit $N = 25$, $M = 10$, $n = 3$, $k = 0$.

$$\begin{aligned}
 P(h_3(A) = 0) &= \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \\
 &= \frac{\binom{10}{0} \binom{15}{3}}{\binom{25}{3}} \\
 &= \frac{1 \cdot 455}{2300} \\
 &= 0.198
 \end{aligned}$$

49. Hypergeometrische Verteilung mit $N = 15$, $M = 3$, $n = 5$, $k \geq 1$.

$$\begin{aligned}
 P(h_5(A) = 0) &= \frac{\binom{3}{0} \binom{12}{5}}{\binom{15}{5}} \\
 &= \frac{1 \cdot 792}{3003} \\
 &= 0.264 \\
 P(h_5(A) \geq 1) &= 1 - P(h_5(A) = 0) \\
 &= 1 - 0.264 \\
 &= 0.736
 \end{aligned}$$

53. Hypergeometrische Verteilung mit $N = 10$, $M = 5$, $n = 5$, $k \geq 4$.

$$\begin{aligned}
 P(h_5(A) \geq 4) &= P(h_5(A) = 4) + P(h_5(A) = 5) \\
 &= \frac{\binom{5}{4} \binom{5}{1}}{\binom{10}{5}} + \frac{\binom{5}{5} \binom{5}{0}}{\binom{10}{5}} \\
 &= \frac{5 \cdot 5 + 1 \cdot 1}{252} \\
 &= 0.103
 \end{aligned}$$

(3.42) Skript

Zweiseitiges Prognoseintervall für $p = 0.8$, $n = 100$.

$$\begin{aligned}
 SD &= \sqrt{\frac{p(1-p)}{n}} \\
 &= \sqrt{\frac{0.8 \cdot 0.2}{100}} \\
 &= 0.04 \\
 p \pm N_{0.975} SD &= 0.8 \pm 1.96 \cdot 0.04 \\
 &= 0.8 \pm 2 \cdot 0.04 \quad (\text{Faustregel}) \\
 0.95 &\approx P(0.72 \leq f_{100}(A) \leq 0.88)
 \end{aligned}$$

(3.43) Skript

Zweiseitiges Prognoseintervall für $p = 0.4$, $n = 1000$.

Beobachtet: $h_{1000}(A) = 367$, d.h. $f_{1000}(A) = 0.367$.

$$\begin{aligned}
 SD &= \sqrt{\frac{0.4 \cdot 0.6}{1000}} \\
 &\approx 0.0155 \\
 p \pm 2SD &= 0.4 \pm 0.031 \\
 0.95 &\approx P(0.369 \leq f_{1000}(A) \leq 0.431) \\
 &= P(369 \leq h_{1000}(A) \leq 431)
 \end{aligned}$$

Die beobachtete Häufigkeit liegt also knapp am Rande des Prognoseintervalls.

60. Zweiseitiges Prognoseintervall für $p = 0.3$, $n = 850$.

$$\begin{aligned}
 SD &= \sqrt{\frac{0.3 \cdot 0.7}{850}} \\
 &\approx 0.0157 \\
 p \pm 2SD &= 0.3 \pm 0.0314 \\
 \text{Prognoseintervall für Anteil} &= [0.2686, 0.3314] \\
 \text{Prognoseintervall für Anzahl} &= [0.2686, 0.3314] \cdot 850 \\
 &= [228.3, 281.7] \\
 &= [228, 282]
 \end{aligned}$$

$$\begin{aligned}
 2 \cdot SD &\leq 0.02 \\
 \Leftrightarrow SD &\leq 0.01 \\
 \Leftrightarrow \frac{\sqrt{0.3 \cdot 0.7}}{0.01} &\leq \sqrt{n} \\
 \Leftrightarrow \frac{0.3 \cdot 0.7}{0.0001} &\leq n
 \end{aligned}$$

$$\Leftrightarrow 2100 \leq n$$

59. Zweiseitiges Prognoseintervall für $p = 0.19$, $n = 85$.

$$\begin{aligned} SD &= \sqrt{\frac{0.19 \cdot 0.81}{85}} \\ &\approx 0.0426 \\ p \pm 2SD &= 0.19 \pm 0.0852 \\ &= [0.105, 0.275] \\ 2 \cdot SD &\leq 0.05 \\ \Leftrightarrow SD &\leq 0.025 \\ \Leftrightarrow \frac{0.19 \cdot 0.81}{0.025^2} &\leq n \\ \Leftrightarrow 246.24 &\leq n \\ \Leftrightarrow 247 &\leq n \end{aligned}$$

58. Einseitiges Prognoseintervall für $p = 0.44$, $n = 50$.

$$\begin{aligned} SD &= \sqrt{\frac{0.44 \cdot 0.56}{50}} \\ &\approx 0.0702 \\ \text{Prognoseintervall} &= [0, p + N_{0.95}SD] \\ &= [0, p + 1.645 \cdot SD] \\ &= [0, 0.555] \\ p + N_{0.99} \cdot SD &\leq 0.5 \\ \Leftrightarrow N_{0.99} \cdot SD &\leq 0.5 - 0.44 \\ \Leftrightarrow 2.326 \cdot SD &\leq 0.06 \\ \Leftrightarrow \frac{0.44 \cdot 0.56}{0.0258^2} &\leq n \\ \Leftrightarrow 370.17 &\leq n \\ \Leftrightarrow 371 &\leq n \end{aligned}$$

57. Einseitiges Prognoseintervall für $p = 1 - 2.7/2.9 = 0.069$, $n = 70$.

$$\begin{aligned} SD &= \sqrt{\frac{0.069 \cdot 0.931}{70}} \\ &\approx 0.0303 \\ \text{Prognoseintervall für Anteil} &= [p - N_{0.95}SD, 1] \\ &= [0.069 - 1.645 \cdot 0.0303, 1] \\ &= [0.0191, 1] \\ \text{Prognoseintervall für Anzahl} &= [1.341, 70] \\ p - 1.645 \cdot SD &\geq 0.03 \\ \Leftrightarrow 0.069 - 1.645 \cdot SD &\geq 0.03 \end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow -1.645 \cdot \frac{\sqrt{0.069 \cdot 0.931}}{\sqrt{n}} \geq -0.039 \\
&\Leftrightarrow 1.645 \cdot \frac{\sqrt{0.069 \cdot 0.931}}{0.039} \leq \sqrt{n} \\
&\Leftrightarrow 10.691 \leq \sqrt{n} \\
&\Leftrightarrow 114.29 \leq n \\
&\Leftrightarrow 115 \leq n
\end{aligned}$$

55. Um zu überprüfen ob die empirisch beobachtete relative Häufigkeit $\hat{p} = 0.568$ zu der hypothetisch unterstellten Wahrscheinlichkeit $p_0 = 0.6$ "paßt", muß ein einseitiges Prognoseintervall unter der Annahme von $p_0 = 0.6$ berechnet werden. Hier wählen wir dabei die Stichprobengröße so groß, dass die untere Schranke des Intervalls mindestens so groß ist wie 0.568.

$$\begin{aligned}
p_0 - 1.645 \cdot SD &\geq \hat{p} \\
\Leftrightarrow 0.6 - 0.568 &\geq 1.645 \cdot SD \\
\Leftrightarrow \frac{0.6 - 0.568}{1.645} &\geq 1.645 \cdot \frac{\sqrt{0.6 \cdot 0.4}}{\sqrt{n}} \\
\Leftrightarrow \sqrt{n} &\geq 1.645 \cdot \frac{\sqrt{0.6 \cdot 0.4}}{0.6 - 0.568} \\
\Leftrightarrow \sqrt{n} &\geq 25.18382 \\
\Leftrightarrow n &\geq 634.2248 \\
\Leftrightarrow n &\geq 635
\end{aligned}$$

4 Testen und Schätzen

61. Zweiseitiger Test: Nullhypothese $p = 0.5$ vs. Alternative $p \neq 0.5$, d.h. $p_0 = 0.5$.

Ärzte	$n = 65$	$\hat{p} = \frac{4}{65} = 0.062$	$SD = \sqrt{\frac{p_0(1-p_0)}{n}} = 0.062$	$T = \frac{\hat{p}-p_0}{SD} = \frac{-0.438}{0.062} \approx -7.07$
Rechtsanwälte	$n = 55$	$\hat{p} = \frac{30}{55} = 0.545$	$SD = \frac{0.5}{\sqrt{55}} = 0.067$	$T = \frac{0.045}{0.067} \approx 0.674$
Krankenpfleger	$n = 68$	$\hat{p} = \frac{32}{68} = 0.471$	$SD = 0.061$	$T = -0.485$
Verkäufer	$n = 55$	$\hat{p} = \frac{43}{55} = 0.782$	$SD = 0.067$	$T = 4.18$

Die Faustregel besagt, daß die Nullhypothese verworfen wird, wenn $|T|$ den Wert 2 überschreitet. Das heißt der Anteil der Ärzte, die das Pensionssystem (sehr) positiv beurteilen, ist signifikant kleiner als 50% und der Anteil der Verkäufer ist signifikant größer als 50%.

76. Einseitiger Test: Nullhypothese $p \geq 0.25$ vs. Alternative $p < 0.25$, d.h. $p_0 = 0.25$.

Die einseitige Nullhypothese wird verworfen, falls T den Wert -1.645 unterschreitet. Dies ist der Fall für den Alfa Romeo und den Golf Rabbit. Dort ist die Wahrscheinlichkeit für eine Reparatur im 1. Jahr signifikant kleiner als 25%.

67. Einseitiger Test: Nullhypothese $p \leq 0.1$ vs. Alternative $p > 0.1$, d.h. $p_0 = 0.1$.

Die einseitige Nullhypothese wird verworfen, falls T den Wert 1.645 überschreitet. Dies ist der Fall für die Privathaftpflicht und die Hausversicherung. Dort ist die Wahrscheinlichkeit für einen betrügerischen Versicherungsfall signifikant größer als 10%. Antwort: (a)

56. Die Fragestellung entspricht einem rechtsseitigem Testproblem der Hypothese $p \geq p_0 = 0.25$ gegen die Alternative $p < 0.25$. Die beobachtete relative Häufigkeit ist $\hat{p} = 0.21$. Um die Nullhypothese zu verwerfen und damit den Politiker zum Signifikanzniveau 0.95 zu widerlegen,

muß die Teststatistik T kleiner als -1.645 gewesen sein. Also

$$\begin{aligned}
 T &< -N_{0.95} \\
 \Leftrightarrow \frac{\hat{p}-p_0}{SD} &< -1.645 \\
 \Leftrightarrow 0.21 - 0.25 &< -1.645 \cdot SD \\
 \Leftrightarrow -0.04 &< -1.645 \cdot \sqrt{\frac{0.25 \cdot 0.75}{n}} \\
 \Leftrightarrow 0.04 &> 1.645 \frac{0.433}{\sqrt{n}} \\
 \Leftrightarrow \sqrt{n} &> \frac{1.645 \cdot 0.433}{0.04} \\
 \Leftrightarrow n &> \left(\frac{1.645 \cdot 0.433}{0.04} \right)^2 \\
 \Leftrightarrow n &> 317.11
 \end{aligned}$$

Wenn also n grösser sein muß als 317.11, dann ist die nächstgrößte ganzzahlige Zahl 318. Und das muß mindestens der Stichprobenumfang gewesen sein.

125. Parameter: $n = 125$, $\hat{p} = \frac{42}{125} \approx 0.336$.

Faustregel: Der kritische Wert ist $c = c(0.95) = 1.96 \approx 2$.

Konfidenzintervall mit Bootstrapethode: Die Standardabweichung ist *ungefähr*

$$\begin{aligned}
 \widehat{SD} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
 &\approx 0.0422
 \end{aligned}$$

Konfidenzintervall mit robuster Methode: Die Standardabweichung ist *maximal*

$$\begin{aligned}
 SD_{\max} &= \frac{1}{2\sqrt{n}} \\
 &= 0.0447
 \end{aligned}$$

$$\begin{aligned}
 0.95 &\approx P(\hat{p} - c \cdot \widehat{SD} \leq p \leq \hat{p} + c \cdot \widehat{SD}) \\
 &= P(0.336 - 2 \cdot 0.0422 \leq p \leq 0.336 + 2 \cdot 0.0422) \\
 &= P(0.252 \leq p \leq 0.420)
 \end{aligned}$$

$$\begin{aligned}
 0.95 &\approx P(\hat{p} - c \cdot SD_{\max} \leq p \leq \hat{p} + c \cdot SD_{\max}) \\
 &= P(0.336 - 0.0894 \leq p \leq 0.336 + 0.0894) \\
 &= P(0.247 \leq p \leq 0.425)
 \end{aligned}$$

61. Parameter: $n = 243$, $\hat{p} = \frac{109}{243} = 0.449$.

Die Schrittweite im robusten Konfidenzintervall darf höchstens 5% sein, d.h.

$$\begin{aligned}
 2 \cdot SD_{\max} &\leq 0.05 \\
 \Leftrightarrow \frac{1}{\sqrt{n}} &\leq 0.05 \\
 \Leftrightarrow 20 &\leq \sqrt{n} \\
 \Leftrightarrow 400 &\leq n
 \end{aligned}$$

5 Verteilungsmaßzahlen

- Boxplot, Beispiel 1

falsch $R = 40 - 30 = 10$, $L = 50 - 40 = 10$

Da $R = L$ ist die Box des Boxplots symmetrisch und die Verteilung daher nicht schief:

$$SK = (10 - 10)/(10 + 10) = 0$$

falsch $SK = 0$

falsch Die Verteilung ist nicht linksschief ($SK = 0$).

richtig 30 ist $Q_{0.25}$. Damit sind 25% der Daten kleiner oder gleich 30.

richtig $QD = Q_{0.75} - Q_{0.25} = 50 - 30 = 20 < 30$

123. Boxplot, Beispiel 2

richtig 200 ist 25% Quantil

richtig $R = 400 - 250 = 150$, $L = 250 - 200 = 50$

$$SK = (150 - 50)/(150 + 50) = 0.5 > 0.4$$

falsch 400 ist nur 75% Quantil, Maximum ist 700

falsch 250 ist $Q_{0.5}$, 400 ist $Q_{0.75}$

\Rightarrow im Intervall $[250, 400]$ liegen 25

richtig $SK > 0$

falsch denn Mittelwert $>$ Median (weil rechtsschief) $= 250$

112. Aus Tabelle ablesen: 1.25 ist $N_{0.894}$, d.h. das 0.894 Quantil.

Lösungsweg 1: Aus Tabelle ablesen: -1.25 ist $N_{0.106}$, d.h. das 0.106 Quantil. Dazwischen liegen $0.894 - 0.106 = 0.788$.

Lösungsweg 2: Zwischen $[-N_{(1+\alpha)/2}, N_{(1+\alpha)/2}]$ liegt der Anteil α .

$$N_{(1+\alpha)/2} = 1.25 = N_{0.894}$$

Lösen nach α ergibt $\alpha = 0.789$.

132.

$$\begin{aligned} P(\text{Stromverbrauch} \leq 8) &= P\left(\frac{\text{Stromverbrauch} - 6.5}{\sqrt{4.6}} \leq \frac{8 - 6.5}{\sqrt{4.6}}\right) \\ &= P\left(\frac{\text{Stromverbrauch} - 6.5}{\sqrt{4.6}} \leq 0.699\right) \\ &= 0.756 \end{aligned}$$

Aus Tabelle ablesen: 0.699 ist das 0.756 Quantil der Standardnormalverteilung.

109. Die Wahrscheinlichkeit α liegt zwischen $[-N_{(1+\alpha)/2}, N_{(1+\alpha)/2}]$.

Einsetzen für $\alpha = 0.9$: a ist $N_{0.95} = 1.645$

- Beispiel:

$$\begin{aligned} 0.9 &= P(\bar{x} - c \cdot s_x \leq X \leq \bar{x} + c \cdot s_x) \\ &= P\left(-c \leq \frac{X - \bar{x}}{s_x} \leq c\right) \end{aligned}$$

Daher muß $c = N_{0.95} = 1.645$ (siehe 75.).

133. Zuerst bestimmen wir das 75% Quantil der Müllmenge die normalverteilt ist mit Mittel 2.5 und Standardabweichung 0.4. Das 75% Quantil der Standardnormalverteilung ist 0.6745, deshalb:

$$\begin{aligned} 0.75 &= P\left(\frac{\text{Menge} - 2.5}{0.4} \leq 0.6745\right) \\ &= P(\text{Menge} \leq 2.5 + 0.4 \cdot 0.6745) \\ &= P(\text{Menge} \leq 2.7698) \end{aligned}$$

An 75% der Tage wird also die Müllmenge von 2.7698 Tonnen nicht überschritten, die zugehörigen Kosten betragen:

$$\begin{aligned} \text{Kosten} &= \text{FixeKosten} + \text{VariableKosten} \cdot \text{Menge} \\ &= 2500 + 1600 \cdot 2.7698 \\ &= 6931.68 \end{aligned}$$

134. Gesucht ist das 30% Quantil der Einkommensverteilung, einer Normalverteilung mit Mittelwert 28000 und Standardabweichung 12000. Das 30% Quantil der Standardnormalverteilung ist -0.5244, deshalb:

$$\begin{aligned} 1 - 0.7 = 0.3 &= P\left(\frac{\text{Einkommen} - 28000}{12000} \leq -0.5244\right) \\ &= P(\text{Einkommen} \leq 28000 - 12000 \cdot 0.5244) \\ &= P(\text{Einkommen} \leq 21707.2) \end{aligned}$$

135. Bei der Normalverteilung ist Mittelwert = Median (symmetrisch). Gesucht wird der Median der Verteilung, hier also der Mittelwert = 2.5 Millionen.

6 Der Erwartungswert

(6.19) Skript

Die Qualitätskontrolle akzeptiert jeweils 20 Packungen, deren mittleres Gewicht \bar{X} zwischen $4.9/20 = 0.245\text{kg}$ und $5.1/20 = 0.255\text{kg}$ liegt. Mit $\mu = 250$ und $\sigma = 10$ ergibt sich

$$\begin{aligned} SD &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{10}{\sqrt{20}} \\ &= 2.236 \end{aligned}$$

$$\begin{aligned} P(245 \leq \bar{X} \leq 255) &= P\left(\frac{245 - 250}{2.236} \leq \frac{\bar{X} - \mu}{SD} \leq \frac{255 - 250}{2.236}\right) \\ &= P(-2.236 \leq Z \leq 2.236) \\ &= \Phi(2.236) - \Phi(-2.236) \\ &= 0.9873 - 0.0126 \\ &= 0.975 \end{aligned}$$

Es ist also mit 97.5% Kartons zu rechnen, die die Qualitätskontrolle passieren. Daher werden etwa 2.5% der Kartons die Qualitätskontrolle nicht passieren.

167. Beidseitiges Prognoseintervall fuer \bar{X} bei $E(X) = \mu = 9.5$ und $V(X) = \sigma^2 = 9.2$. Das Signifikanzniveau ist $\alpha = 0.95$. Dann ist das Prognoseintervall für \bar{X} bei $n = 50$:

$$\begin{aligned}\mu \pm N_{(1+\alpha)/2} \cdot SD &= \mu \pm N_{(1+\alpha)/2} \cdot \sqrt{\frac{V(X)}{n}} \\ &= 9.5 \pm N_{0.975} \cdot \sqrt{\frac{9.2}{50}} \\ &\approx 9.5 \pm 2 \cdot 0.429 \\ &= [8.642, 10.358]\end{aligned}$$

Der Wert 10l liegt innerhalb des Prognoseintervalls, eine Fehleinschätzung kann somit nicht ausgeschlossen werden.

168. Beidseitiges Prognoseintervall fuer \bar{X} bei $E(X) = \mu = 12$ und $V(X) = \sigma^2 = 10.2$. Das Signifikanzniveau ist $\alpha = 0.95$. Dann ist das Prognoseintervall für \bar{X} bei $n = 50$:

$$\begin{aligned}\mu \pm N_{(1+\alpha)/2} \cdot SD &= 12 \pm 2 \cdot \sqrt{\frac{10.2}{50}} \\ &= [11.097, 12.903]\end{aligned}$$

Der Wert 10l liegt außerhalb des Prognoseintervalls, eine Fehleinschätzung kann somit ausgeschlossen werden.

169. Beidseitiges Prognoseintervall fuer \bar{X} bei $E(X) = \mu = 8$ und $V(X) = \sigma^2 = 16.4$. Das Signifikanzniveau ist $\alpha = 0.95$. Dann ist das Prognoseintervall für \bar{X} bei $n = 50$:

$$\begin{aligned}\mu \pm N_{(1+\alpha)/2} \cdot SD &= 8 \pm 2 \cdot \sqrt{\frac{16.4}{50}} \\ &= [6.855, 9.145]\end{aligned}$$

Der Wert 10l liegt außerhalb des Prognoseintervalls, eine Fehleinschätzung kann somit ausgeschlossen werden.

170. Beidseitiges Prognoseintervall fuer \bar{X} bei $E(X) = \mu = 11$ und $V(X) = \sigma^2 = 16.5$. Das Signifikanzniveau ist $\alpha = 0.95$. Dann ist das Prognoseintervall für \bar{X} bei $n = 50$:

$$\begin{aligned}\mu \pm N_{(1+\alpha)/2} \cdot SD &= 11 \pm 2 \cdot \sqrt{\frac{16.5}{50}} \\ &= [9.851, 12.149]\end{aligned}$$

Der Wert 10l liegt innerhalb des Prognoseintervalls, eine Fehleinschätzung kann somit nicht ausgeschlossen werden.

165. Beidseitiges Konfidenzintervall fuer $E(X) = \mu$ bei $\bar{x} = 35.95$ und $s^2 = 54.2$. Beim Konfidenzintervall wird die Varianz σ^2 mit der **Stichprobenvarianz** s_{n-1}^2 und nicht der Varianz der Daten σ^2 approximiert. Es gilt $s_{n-1}^2 = s^2 \cdot n/(n-1)$. Das Signifikanzniveau ist $\alpha = 0.95$. Dann ist das Konfidenzintervall für μ bei $n = 36$:

Gastgewerbe

$$\begin{aligned}\bar{x} \pm N_{(1+\alpha)/2} \cdot \widehat{SD} &= \bar{x} \pm N_{(1+\alpha)/2} \cdot \sqrt{\frac{s_{n-1}^2}{n}} \\ &= \bar{x} \pm 2 \cdot \sqrt{\frac{s^2}{n-1}} \\ &= 35.95 \pm 2 \cdot \sqrt{\frac{54.2}{35}}\end{aligned}$$

	=	[33.461, 38.439]
Bauwirtschaft		[48.518, 56.262]
Metall		[35.273, 42.547]
Uhren und Schmuck		[2.20, 7.14]
Leder und Pelze		[9.646, 15.614]

(6.28) Skript, siehe auch (6.32)

Folgende Hypothese H soll überprüft werden: der Erwartungswert μ des IQ der Bewerber ist identisch mit dem Erwartungswert $\mu_0 = 100$ des IQ in der Gesamtbevölkerung $\mu = \mu_0$.

Es liegt eine Stichprobe von Beobachtungen x_i vom Umfang $n = 20$ vor. Aus dieser können folgende Größen berechnet werden: der Mittelwert $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 105.2$ und die Stichprobenstandardabweichung $s_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 12.54$.

Als Signifikanzniveau wird $\alpha = 0.05$ verwendet.

Es gibt verschiedene **äquivalente** Lösungswege:

1. Prognose/Konfidenzintervall

Mit 95% Wahrscheinlichkeit gilt unter H :

$$\begin{aligned}
 |\bar{x} - \mu_0| &\leq N_{(1+\alpha)/2} \cdot \widehat{SD} \\
 |105.2 - 100| &\leq 2 \cdot \frac{12.54}{\sqrt{20}} \\
 5.2 &\leq 2 \cdot \frac{12.54}{\sqrt{20}} \\
 5.2 &\leq 5.608
 \end{aligned}$$

Dies steht nicht im Widerspruch zu H .

2. Teststatistik

Der Absolutbetrag der Teststatistik

$$\begin{aligned}
 z &= \frac{\bar{x} - \mu_0}{\widehat{SD}} \\
 &= \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{s_{n-1}} \\
 &= \sqrt{20} \cdot \frac{105.2 - 100}{12.54} \\
 &= 1.854
 \end{aligned}$$

überschreitet den kritischen Wert

$$c = N_{(1+\alpha)/2} \approx 2 \text{ nicht.}$$

3. P -Wert

Da unter H

$$\begin{aligned}
 P(|Z| \leq z) &= P(|Z| \leq 1.854) \\
 &= \Phi(1.854) - \Phi(-1.854) \\
 &= 0.936
 \end{aligned}$$

beträgt der P -Wert

$$1 - 0.936 = 0.064.$$

Die Nullhypothese kann zum 95% Niveau nicht verworfen werden.

142. Einseitiger Test der Hypothese $\mu \leq \mu_0 = 30$.

Werte der Stichprobe sind $\bar{x} = 35.95$ und $s^2 = 54.2$. Beim Testen wird die Varianz σ^2 mit der **Stichprobenvarianz** s_{n-1}^2 und nicht der Varianz der Daten σ^2 approximiert. Es gilt $s_{n-1}^2 = s^2 \cdot n/(n-1)$. Die Anzahl der Beobachtungen ist $n = 36$:

Gastgewerbe	
	$z = \frac{\bar{x} - \mu_0}{\widehat{SD}}$ $= \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{s_{n-1}}$ $= \sqrt{n-1} \cdot \frac{\bar{x} - \mu_0}{s}$ $= \sqrt{35} \cdot \frac{35.95 - 30}{\sqrt{54.2}}$ $= 4.781$
Bauwirtschaft	1.234
Metall	-0.599
Uhren und Schmuck	-4.315
Leder und Pelze	1.763

Es sollen folgende Aussagen überprüft werden (Achtung: die Aussagen in den Folien sind anders als die Formulierung und Reihenfolge in Aufgabe 103.):

- (a) Im Gastgewerbe kann nachgewiesen werden, dass der Erwartungswert von X kleiner als 30 GE ist.

Um eine Aussage (statistisch) nachzuweisen muß die Nullhypothese verwerfen, weil nur dann die Wahrscheinlichkeit für einen Fehler (1. Art) kontrolliert wird. In diesem Fall muß also das Testproblem formuliert werden für die Hypothese $H : \mu \geq 30$ gegen die Alternative $\mu < 30$. Der empirische Mittelwert von $\bar{x} = 35.95 > 30$ *kann nie* Hinweise darauf enthalten, dass der wahre Erwartungswert $\mu < 30$ ist. Formal: die Teststatistik $z = 4.781$ unterschreitet den kritischen Wert $-N_\alpha = -N_{0.95} = -1.645$ nicht (Fehler 1. Art ist als 5% gegeben, damit ist das Signifikanzniveau $\alpha = 0.95$). Die Nullhypothese kann nicht verworfen werden.

Damit ist Aussage (a) **falsch**.

- (b) In der Bauwirtschaft kann keine Entscheidung darüber getroffen werden, ob der Erwartungswert ungleich 50 GE ist.

Die Formulierung unterstellt ein zweiseitiges Testproblem, d.h. die Hypothese $H : \mu = 50$ vs. die Alternative $\mu \neq 50$. Der kritische Wert zum 95% Niveau ist daher $N_{(1+\alpha)/2} = N_{0.975} \approx 2$ (Faustregel). Die Teststatistik $z = 1.234$ unterschreitet weder -2 noch überschreitet sie 2. Also kann die Nullhypothese nicht verworfen werden.

Daher ist Aussage (b) **richtig**.

- (c) Beim Metall kann nachgewiesen werden, dass der Erwartungswert von X größer als 40 GE ist.

Um dies nachzuweisen muß $\mu > 40$ die Alternative sein, die Hypothese ist also $H : \mu \leq 40$. Die Teststatistik $z = -0.599$ überschreitet den kritischen Wert $N_\alpha = N_{0.95} = 1.645$ nicht, die Nullhypothese kann daher nicht verworfen werden. (Ähnlich wie in (a) kann wieder ein Mittelwert von $\bar{x} = 38.91 < 40$ keine Beweise beinhalten, daß μ in Wahrheit > 40 ist.)

Daher ist die Aussage (c) **falsch**.

- (d) Bei Uhren und Schmuck kann nachgewiesen werden, dass der Erwartungswert von X größer als 10 GE ist.

Hier muss also die Hypothese $H : \mu \leq 10$ sein, damit die Alternative $\mu > 10$ ist. Die

Teststatistik $z = -4.315$ überschreitet jedoch den kritischen Wert $N_\alpha = N_{0.95} = 1.645$ nicht, die Nullhypothese kann nicht verworfen werden. (Ähnlich wie in (a) kann wieder ein Mittelwert von $\bar{x} = 4.67 < 10$ keine Beweise beinhalten, daß μ in Wahrheit > 10 ist.)

Daher ist die Aussage (d) **falsch**.

- (e) Bei Leder und Pelze kann keine Entscheidung darüber getroffen werden, ob der Erwartungswert ungleich 10 GE ist.

Die Formulierung unterstellt ein zweiseitiges Testproblem, d.h. die Hypothese $H : \mu = 10$ vs. die Alternative $\mu \neq 10$. Der kritische Wert zum 95% Niveau ist daher $N_{(1+\alpha)/2} = N_{0.975} \approx 2$ (Faustregel). Die Teststatistik $z = 1.763$ unterschreitet weder -2 noch überschreitet sie 2. Daher kann die Nullhypothese nicht verworfen werden.

Daher ist Aussage (e) **richtig**.

- (6.38) Skript, siehe auch (6.32) und (6.28)

Folgende Hypothese H soll überprüft werden: der Erwartungswert μ des IQ der Bewerber ist identisch mit dem Erwartungswert $\mu_0 = 100$ des IQ in der Gesamtbevölkerung $\mu = \mu_0$.

Es liegt eine Stichprobe vom Umfang $n = 20$ vor mit $\bar{x} = 105.2$ und $s_{n-1} = 12.54$.

Als Signifikanzniveau wird $\alpha = 0.05$ verwendet. Anders als in Aufgabe (6.28) verwenden wir nun die F -Test Version auf Basis der ANOVA-Tabelle. Die ist wiederum **äquivalent** zu den drei Lösungswegen in Aufgabe (6.28).

Die ANOVA-Tabelle

	SS	df	MSS
$*$	$\sum_{i=1}^n (\bar{x} - \mu_0)^2$	1	$n(\bar{x} - \mu_0)^2$
R	$\sum_{i=1}^n (x_i - \bar{x})^2$	$n - 1$	s_{n-1}^2
	SS_T	n	

ist in diesem Fall also

	SS	df	MSS
$*$	540.80	1	$20 \cdot (105.2 - 100)^2 = 540.8$
R	$19 \cdot 157.25 = 2987.75$	19	$12.54^2 = 157.25$
	3528.58	20	

Damit ergibt sich die F -Größe als

$$F = \frac{MSS^*}{MSS_R} = \frac{540.8}{157.25} = 3.439.$$

Diese überschreitet den kritischen Wert $c = (N_{(1+\alpha)/2})^2 \approx 4$ (Faustregel) nicht. Damit kann wie in Aufgabe (6.28) die Hypothese nicht verworfen werden.

Der Anteil der systematischen Abweichung an der Gesamtstreuung ist das Bestimmtheitsmaß

$$\frac{SS^*}{SS_T} = \frac{540.80}{3528.58} = 0.153.$$

Nur 15.3% der Gesamtstreuung der Daten kann durch einen systematisch höheren IQ der Bewerber erklärt werden.

152. Die Anzahl der Beobachtungen ist $n = 50$, der Mittelwert $\bar{x} = 180$, die Varianz der Daten $s_x^2 = 446$, damit ist die Stichprobenvarianz $s_{x,n-1}^2 = \frac{n}{n-1} s_x^2 = \frac{50}{49} 446 = 455.102$. Der vorgegebene Wert μ_0 ist $\mu_0 = 200$.

Die ANOVA-Tabelle ergibt sich als

	SS	df	MSS
$*$	20000	1	$50 \cdot (180 - 200)^2 = 20000$
R	22300	49	455.102
	42300	50	

Das Signifikanzproblem liesse sich durch die F -Größe lösen:

$$F = \frac{MSS^*}{MSS_R} = \frac{20000}{455.102} = 43.946.$$

Diese überschreitet den kritischen Wert 4 deutlich, somit weicht der mittlere Rechnungswert der Filiale Capri signifikant von 200 ab.

Gefragt war jedoch nach dem Relevanzproblem, das sich durch das Bestimmtheitsmaß lösen läßt:

$$\frac{SS^*}{SS_T} = \frac{20000}{42300} = 0.473.$$

Es lassen sich also 47.3% der Abweichung vom vorgegebenen Wert 200 durch systematische Abweichung erklären.

7 Regression und Korrelation

186. Ablesen: Die jährliche Fahrleistung (in 1000km) hat einen Mittelwert von $\bar{x} = 72.3$ und eine Varianz von $s_x^2 = 236.5$. Die Schadenssumme (in GE) hat den Mittelwert $\bar{y} = 12.5$ und eine Varianz von $s_y^2 = 6.4$. Die Korrelation der beiden Merkmale beträgt $r = 0.8$.

Berechne die Parameter der Geradengleichung der empirischen Regressionsgeraden $y = \hat{a} + \hat{b}x$.

Die Steigung \hat{b} ergibt sich als

$$\hat{b} = r \frac{s_y}{s_x} = 0.8 \sqrt{\frac{6.4}{236.5}} = 0.1316.$$

Damit ist der Achsenabschnitt \hat{a} bestimmt durch

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 12.5 - 0.1316 \cdot 72.3 = 2.985$$

Nun soll der Schätzwert \hat{y} berechnet werden für $x = 90$:

$$\hat{y} = \hat{a} + \hat{b}x = 2.985 + 0.1316 \cdot 90 = 14.829$$

Die geschätzte Schadenssumme ist also 14.8 GE.

189. (siehe auch 188.)

Ablesen: der Stichprobenumfang ist $n = 20$, die Anzahl der Mitarbeiter hat einen Mittelwert von $\bar{x} = 150$ bei einer Varianz von $s_x^2 = 625$. Die Höhe der Kosten hat einen Mittelwert von $\bar{y} = 2000$ bei einer Varianz von $s_y^2 = 250000$. Die Korrelation der beiden Merkmale ist $r = 0.65$.

Berechne die Parameter der Geradengleichung der empirischen Regressionsgeraden $y = \hat{a} + \hat{b}x$.

Die Steigung \hat{b} ergibt sich als

$$\hat{b} = r \frac{s_y}{s_x} = 0.65 \sqrt{\frac{250000}{625}} = 13.$$

Damit ist der Achsenabschnitt \hat{a} bestimmt durch

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 2000 - 13 \cdot 150 = 50$$

Für diesen Datensatz lautet also die Gleichung der Regressionsgeraden $y = 50 + 13x$.

188. Ablesen: der Stichprobenumfang ist $n = 20$, die Anzahl der Mitarbeiter hat einen Mittelwert von $\bar{x} = 150$ bei einer Varianz von $s_x^2 = 625$. Die Höhe der Kosten hat einen Mittelwert von $\bar{y} = 2000$ bei einer Varianz von $s_y^2 = 250000$. Die Korrelation der beiden Merkmale ist $r = 0.65$.

Überprüfe, ob diese Korrelation signifikant ist (oder äquivalent ob $b \neq 0$) mit Hilfe der F -Größe aus der ANOVA-Tabelle:

	SS	df	MSS
$*$	$r^2 ns_y^2$	1	$r^2 ns_y^2$
R	$(1 - r^2) ns_y^2$	$n - 2$	$\frac{r^2 ns_y^2}{n - 2}$
	ns_y^2	$n - 1$	

Für die Daten ergibt sich dafür

	SS	df	MSS
$*$	2112500	1	2112500
R	2887500	18	160416.7
	5000000	19	

Die F -Größe ergibt sich also als

$$F = \frac{MSS^*}{MSS_R} = \frac{2112500}{160416.7} = 13.169$$

und ist hochsignifikant, da sie deutlich größer als 4 ist.

Äquivalent und bei den gegebenen Daten viel leichter läßt sich die F -Größe auch bestimmen als

$$F = (n - 2) \frac{r^2}{1 - r^2} = 18 \frac{0.4225}{0.5775} = 13.169.$$

Der Test verwirft also die Nullhypothese $b = 0$ bzw. äquivalent “ X und Y sind nicht gekoppelt” und es kann daher geschlossen werden, daß eine Koppelung zwischen X und Y vorliegt.

187. Ablesen: der Stichprobenumfang ist $n = 13$, die Anzahl der Passagiermeilen hat einen Mittelwert von $\bar{x} = 49.4$ bei einer Varianz von $s_x^2 = 670$. Die Gewinne haben einen Mittelwert von $\bar{y} = 0.475$ bei einer Varianz von $s_y^2 = 0.15$. Die Korrelation der beiden Merkmale ist $r = 0.48$.

Überprüfe, ob diese Korrelation signifikant ist (oder äquivalent ob $b \neq 0$) mit Hilfe der F -Größe aus der ANOVA-Tabelle. Für die Daten ergibt sich dafür

	SS	df	MSS
$*$	0.45	1	0.45
R	1.50	11	0.1364
	1.95	12	

Die F -Größe ergibt sich also als

$$F = \frac{MSS^*}{MSS_R} = \frac{0.45}{0.1364} = 3.293$$

und ist nicht signifikant, da sie kleiner als 4 ist.

Äquivalent und bei den gegebenen Daten viel leichter läßt sich die F -Größe auch bestimmen als

$$F = (n - 2) \frac{r^2}{1 - r^2} = 11 \frac{0.2304}{0.7696} = 3.293.$$

Die Nullhypothese $b = 0$ bzw. äquivalent “ X und Y sind nicht gekoppelt” kann also zum 95% Niveau nicht verworfen werden.

8 Das Zweistichprobenproblem

214. Wir haben zwei Stichproben mit $n_1 = 50$ und $n_2 = 80$. Dabei ist bei Sparag der Mittelwert der Wartezeit $\bar{x} = 11.2$ und die Stichprobenvarianz $s_{x,n-1}^2 = 16.4$. Bei Consumo ist der Mittelwert der Wartezeit $\bar{y} = 7.3$ bei einer Stichprobenvarianz von $s_{y,n-1}^2 = 10.2$.

Die Standardabweichung der Differenz der Mittelwerte wird geschätzt mit:

$$\widehat{SD} = \sqrt{\frac{s_{x,n-1}^2}{n_1} + \frac{s_{y,n-1}^2}{n_2}} = \sqrt{\frac{16.4}{50} + \frac{10.2}{80}} = 0.675$$

Damit ist die Teststatistik

$$T = \frac{\bar{x} - \bar{y}}{\widehat{SD}} = \frac{11.2 - 7.3}{0.675} = 5.778$$

Da der Absolutbetrag der Teststatistik $|T| = 5.778$ den Wert 2 überschreitet, kann zum 95% Niveau nachgewiesen werden, daß die Wartezeit bei Sparag und Consumo unterschiedlich ist.

(8.9) Skript

Wir haben zwei Stichproben mit $n_1 = 50$ und $n_2 = 50$. Dabei ist bei den Männern $\bar{x} = 104.72$ und $s_{x,n-1} = 12.387$ und bei den Frauen $\bar{y} = 110.36$ und $s_{y,n-1} = 11.133$.

Berechne die Einträge der ANOVA-Tabelle:

$$\begin{aligned} MSS_{ZW} &= \frac{n_1 n_2}{n_1 + n_2} (\bar{x} - \bar{y})^2 \\ &= \frac{50 \cdot 50}{50 + 50} (104.72 - 110.36)^2 \\ &= 795.24 \\ SS_{IN} &= n_1 s_x^2 + n_2 s_y^2 \\ &= (n_1 - 1) s_{x,n-1}^2 + (n_2 - 1) s_{y,n-1}^2 \\ &= 49 \cdot 12.387^2 + 49 \cdot 11.133^2 \\ &= 13591.6 \\ SS_T &= SS_{ZW} + SS_{IN} \\ &= 14386.84 \\ MSS_{IN} &= \frac{SS_{IN}}{n_1 + n_2 - 2} \\ &= \frac{13591.6}{98} \\ &= 138.69 \end{aligned}$$

Damit ist die F -Größe

$$F = \frac{MSS_{ZW}}{MSS_{IN}} = \frac{795.24}{138.69} = 5.734$$

signifikant, da sie den kritischen Wert 4 (Faustregel) überschreitet.

Das Bestimmtheitsmaß ist

$$\frac{SS_{ZW}}{SS_T} = \frac{795.24}{14386.84} = 0.055$$

Also sind nur etwa 5.5% der Streuung auf den Unterschied zwischen den Geschlechtern zurückzuführen.

- Beispiel:

Wir haben zwei Stichproben mit $n_1 = 10$ und $n_2 = 11$. Dabei ist bei den Calcium-Patienten $\bar{x} = 5$ und $s_x^2 = 68.8$ und bei den Placebo-Patienten $\bar{y} = -0.273$ und $s_y^2 = 31.653$.

Berechne die Einträge der ANOVA-Tabelle:

$$\begin{aligned}
 MSS_{ZW} &= \frac{n_1 n_2}{n_1 + n_2} (\bar{x} - \bar{y})^2 \\
 &= \frac{10 \cdot 11}{10 + 11} (5 + 0.273)^2 \\
 &= 145.628 \\
 SS_{IN} &= n_1 s_x^2 + n_2 s_y^2 \\
 &= 10 \cdot 68.8 + 11 \cdot 31.653 \\
 &= 1036.182 \\
 SS_T &= SS_{ZW} + SS_{IN} \\
 &= 1181.81 \\
 MSS_{IN} &= \frac{SS_{IN}}{n_1 + n_2 - 2} \\
 &= \frac{1036.182}{19} \\
 &= 54.536
 \end{aligned}$$

Damit ist die F -Größe

$$F = \frac{MSS_{ZW}}{MSS_{IN}} = \frac{145.628}{54.536} = 2.67$$

nicht signifikant, da sie den kritischen Wert 4 (Faustregel) überschreitet.

Das Bestimmtheitsmaß ist

$$\frac{SS_{ZW}}{SS_T} = \frac{145.628}{1181.81} = 0.123$$

Also sind nur etwa 12.3% der Streuung auf die Wirksamkeit des Calciums zurückzuführen.

221. Wir haben zwei Stichproben mit $n_1 = 50$ und $n_2 = 80$. Dabei ist bei Sparag der Mittelwert der Wartezeit $\bar{x} = 11.2$ und die Stichprobenvarianz $s_{x,n-1}^2 = 16.4$. Bei Consumo ist der Mittelwert der Wartezeit $\bar{y} = 7.3$ bei einer Stichprobenvarianz von $s_{y,n-1}^2 = 10.2$.

Berechne die Einträge der ANOVA-Tabelle:

$$\begin{aligned}
 MSS_{ZW} &= \frac{n_1 n_2}{n_1 + n_2} (\bar{x} - \bar{y})^2 \\
 &= \frac{50 \cdot 80}{50 + 80} (11.2 - 7.3)^2 \\
 &= 468 \\
 SS_{IN} &= n_1 s_x^2 + n_2 s_y^2 \\
 &= (n_1 - 1) s_{x,n-1}^2 + (n_2 - 1) s_{y,n-1}^2 \\
 &= 49 \cdot 16.4 + 79 \cdot 10.2 \\
 &= 1609.4 \\
 SS_T &= SS_{ZW} + SS_{IN} \\
 &= 2077.4 \\
 MSS_{IN} &= \frac{SS_{IN}}{n_1 + n_2 - 2} \\
 &= \frac{1609.4}{128} \\
 &= 12.57344
 \end{aligned}$$

Damit ist die ANOVA-Tabelle

	SS	df	MSS
ZW	468	1	468
IN	1609.4	128	12.573
	2077.4	129	

Damit ist die F -Größe

$$F = \frac{MSS_{ZW}}{MSS_{IN}} = \frac{468}{12.57344} = 37.221$$

und signifikant, da sie den kritischen Wert 4 (Faustregel) überschreitet.

Das Bestimmtheitsmaß ist

$$\frac{SS_{ZW}}{SS_T} = \frac{468}{2077.4} = 0.225$$

Also sind etwa 22.5% der Streuung auf die unterschiedlichen Wartezeiten bei Sparag und Consumo zurückzuführen.

- Beispiel:

Es müssen zunächst die fehlenden Größen der ANOVA-Tabelle ergänzt werden. Die Freiheitsgrade df in der Zeile ZW sind immer 1. Daher ist die $MSS_{ZW} = SS_{ZW} = 63.42$ wie angegeben. Für die totale Abweichung gilt: $SS_T = SS_{ZW} + SS_{IN} = 63.42 + 71.77 = 135.19$. Und zuletzt ist als $MSS_{IN} = \frac{SS_{IN}}{df_{IN}} = \frac{71.77}{5} = 14.354$. Damit ist die ANOVA-Tabelle gefüllt und es können F -Statistik und Bestimmtheitsmaß berechnet werden:

$$F = \frac{MSS_{ZW}}{MSS_{IN}} = \frac{63.42}{14.354} = 4.418$$

$$\frac{SS_{ZW}}{SS_T} = \frac{63.42}{135.19} = 0.469$$

Die F -Größe überschreitet ihren kritischen Wert 4 (Signifikanzniveau 95%) und daher kann die Hypothese der Gleichheit der Mittelwerte verworfen werden.

9 Kontingenztafeln

(9.5) Skript

Aus der Aufgabenstellung kann entnommen werden, daß $P(\text{Sorte 2}) = 0.26$, $P(\text{nicht einwandfrei}) = 0.04$ und daß $P(\text{Sorte 1} \mid \text{einwandfrei}) = 0.75$.

Gesucht sind $P(\text{Sorte 1} \cap \text{einwandfrei})$ und $P(\text{einwandfrei} \mid \text{Sorte 1})$.

Wir berechnen zunächst alle Randwahrscheinlichkeiten: $P(\text{einwandfrei}) = 1 - 0.04 = 0.96$ und $P(\text{Sorte 1}) = 1 - 0.26 = 0.74$.

Dann können wir die gemeinsame Wahrscheinlichkeit berechnen:

$$\begin{aligned} P(\text{Sorte 1} \cap \text{einwandfrei}) &= P(\text{Sorte 1} \mid \text{einwandfrei}) \cdot P(\text{einwandfrei}) \\ &= 0.75 \cdot 0.96 \\ &= 0.72 \end{aligned}$$

Und damit kann dann die gesamte Kontingenztafel der Wahrscheinlichkeiten hingeschrieben werden als.

	einwandfrei	nicht einwandfrei	
Sorte 1	0.72	0.02	0.74
Sorte 2	0.24	0.02	0.26
	0.96	0.04	1

Gefragt war dann nach:

$$\begin{aligned}
 P(\text{einwandfrei} \mid \text{Sorte 1}) &= P(\text{Sorte 1} \cap \text{einwandfrei}) / P(\text{Sorte 1}) \\
 &= 0.72 / 0.74 \\
 &= 0.973
 \end{aligned}$$

- Beispiel

Aus der Aufgabenstellung kann entnommen werden, daß $P(\text{einwandfrei} \cap \text{Sorte 1}) = 0.18$, $P(\text{einwandfrei}) = 0.42$ und $P(\text{Sorte 2} \mid \text{nicht einwandfrei}) = 0.68$.

$$\begin{aligned}
 P(\text{Sorte 2} \mid \text{einwandfrei}) &= 1 - P(\text{Sorte 1} \mid \text{einwandfrei}) \\
 &= 1 - P(\text{Sorte 1} \cap \text{einwandfrei}) / P(\text{einwandfrei}) \\
 &= 1 - 0.18 / 0.42 \\
 &= 0.571
 \end{aligned}$$

238. Sei P_- das Ereignis, daß das Produkt wirklich mangelhaft ist, und T_- das Ereignis, daß der Test es als mangelhaft einstuft. Gegeben ist, daß $P(P_-) = 0.2$, $P(T_- \mid P_-) = 0.8$ und $P(T_- \mid P_+) = 0.1$.

$$\begin{aligned}
 P(P_- \mid T_+) &= \frac{P(T_+ \mid P_-)P(P_-)}{P(T_+ \mid P_-)P(P_-) + P(T_+ \mid P_+)P(P_+)} \\
 &= \frac{0.2 \cdot 0.2}{0.2 \cdot 0.2 + 0.9 \cdot 0.8} \\
 &= 0.0526
 \end{aligned}$$

- Beispiel

Sei S_- das Ereignis, daß das Student zum WU-Studium wirklich nicht geeignet ist, und T_- das Ereignis, daß ein Student im Test durchfällt.

Gegeben ist, daß $P(S_-) = 0.05$, $P(T_- \mid S_-) = 0.9$ und $P(T_+ \mid S_+) = 0.9$.

$$\begin{aligned}
 P(S_- \mid T_-) &= \frac{P(T_- \mid S_-)P(S_-)}{P(T_- \mid S_-)P(S_-) + P(T_- \mid S_+)P(S_+)} \\
 &= \frac{0.9 \cdot 0.05}{0.9 \cdot 0.05 + 0.1 \cdot 0.95} \\
 &= 0.321
 \end{aligned}$$

Die Wahrscheinlichkeit für das Studium nicht geeignet zu sein, obwohl man die Prüfung nicht bestanden hat, liegt also nur bei 32.1%, ist also unwahrscheinlicher als doch für das Studium geeignet zu sein.

229. Die einfachste Lösungsstrategie für diese Art Aufgaben ist es, sich zunächst die gesamte Kontingenztafel zu rekonstruieren und damit dann die Aufgabe zu lösen. Zum Rechnen reicht dabei in der Regel die Formel für bedingte Wahrscheinlichkeiten $P(A \cap B) = P(A|B) \cdot P(B)$.

Zunächst können einige Wahrscheinlichkeiten abgelesen werden. Wenn M das Ereignis ist, daß ein Student männlich ist und R das Ereignis das er/sie raucht, dann

$$\begin{aligned} P(M \cap R) &= 0.18 \\ P(R) &= 0.42 \\ P(M'|R') &= 0.68 \end{aligned}$$

und damit

$$\begin{aligned} P(R') &= 1 - P(R) = 0.58 \\ P(M' \cap R') &= P(M'|R') \cdot P(R') = 0.3944 \end{aligned}$$

Diese Wahrscheinlichkeiten können dann in die Kontingenztafel eingetragen werden:

	R	R'	
M	0.18		
M'		0.3944	
	0.42	0.58	1

Durch einfache Summen- bzw. Differenzenbildung in Zeilen und Spalten können dann die übrigen Einträge ergänzt werden:

	R	R'	
M	0.18	0.1856	0.3656
M'	0.24	0.3944	0.6344
	0.42	0.58	1

und damit kann man dann leicht die gesuchte Wahrscheinlichkeit berechnen:

$$P(R|M') = \frac{P(R \cap M')}{P(M')} = \frac{0.24}{0.6344} = 0.3783.$$

231. Ablesen: $n = 580$, $h(A) = 330$, $h(B) = 120$, $h(A \cap B) = 30$.

Daraus läßt sich die gesamte Kontingenztafel konstruieren

	B	B'	
A	30	300	330
A'	90	160	250
	120	460	580

$$\begin{aligned} r &= \frac{f(A \cap B) - f(A)f(B)}{\sqrt{f(A)f(A')f(B)f(B')}} \\ &= \frac{h(A \cap B)/n - h(A)/n h(B)/n}{\sqrt{h(A)h(A')h(B)h(B')/n^2}} \\ &= \frac{30/580 - 330/580 \cdot 120/580}{\sqrt{330 \cdot 250 \cdot 120 \cdot 460/336400}} \\ &= -0.329 \end{aligned}$$

Die beiden Merkmale sind also negativ gekoppelt. Die Teststatistik beträgt $\sqrt{580} \cdot r = -7.923$, damit ist die Kopplung hochsignifikant.

- Die Werbekampagne ergab folgendes Ergebnis:

	Würden Sie das Produkt kaufen?	
	Ja	Nein
Stadt A	74	26
Stadt B	19	31

Um festzustellen, ob die Kampagne Erfolg hatte, muss man also feststellen, ob das Marktsegment des Produktes in Stadt A höher ist als das in B. Dies kann man durch einen Test auf Unabhängigkeit durchführen: wenn die Präferenz für das Produkt nicht von der Stadt abhängt, hatte die Kampagne offensichtlich keinen Erfolg.

Der Marktanteil in Stadt A ist $f(\text{Ja}|\text{Stadt A}) = 74/(74 + 26) = 0.74$, in Stadt B hingegen nur $f(\text{Ja}|\text{Stadt B}) = 19/(19 + 31) = 0.38$.

Die Koppelung der beiden Merkmale beträgt

$$\begin{aligned}
 r &= \frac{f(\text{Ja} \cap \text{Stadt A}) - f(\text{Ja}) f(\text{Stadt A})}{\sqrt{f(\text{Ja}) f(\text{Nein}) f(\text{Stadt A}) f(\text{Stadt B})}} \\
 &= \frac{h(\text{Ja} \cap \text{Stadt A})/n - h(\text{Ja})/n h(\text{Stadt A})/n}{\sqrt{h(\text{Ja}) h(\text{Nein}) h(\text{Stadt A}) h(\text{Stadt B})/n^2}} \\
 &= \frac{74/150 - 93/150 \cdot 100/150}{\sqrt{93 \cdot 57 \cdot 100 \cdot 50/22500}} \\
 &= 0.3496
 \end{aligned}$$

Die Merkmale sind also positiv gekoppelt, der Marktanteil in Stadt A (mit Werbung) ist also höher als in Stadt B (ohne Werbung). Anhand der Teststatistik $\sqrt{150} \cdot r = 4.282$ läßt sich nachweisen, daß es einen signifikanten Erfolg der Werbekampagne gegeben hat.

232. Für die Daten aus Aufgabe 231:

$$\begin{aligned}
 h(C) &= h(A' \cap B) + h(A \cap B') \\
 &= 90 + 300 \\
 &= 390
 \end{aligned}$$

und damit

$$\begin{aligned}
 f(A|C) &= \frac{h(A \cap C)}{h(C)} \\
 &= \frac{h(A \cap B')}{h(C)} \\
 &= 300/390 = 0.769
 \end{aligned}$$

Die Teststatistik berechnet sich dann als

$$\begin{aligned}
T &= \frac{f(A|C) - 0.5}{\frac{0.5}{\sqrt{h(C)}}} \\
&= \frac{300/390 - 0.5}{\frac{0.5}{\sqrt{390}}} \\
&= 10.634
\end{aligned}$$

- Die Werbekampagne ergab folgendes Ergebnis:

		Würden Sie das Produkt kaufen?	
		Vorher:	
Nachher:	Ja	116	44
	Nein	31	65

Die Kunden, die sowohl vor als auch nach der Werbekampagne das Produkt kaufen würden enthalten offensichtlich keine Information über den Werbeerfolg, da sie das Produkt immer kaufen würden. Analog enthalten die Kunden, die das Produkt weder vorher noch nachher kaufen würden keine Information. Die Frage ist also, ob die 44 Kunden, die von Nicht-Kaufen zu Kaufen gewechselt sind, signifikant mehr sind als die 31 Kunden, die von Kaufen zu Nicht-Kaufen gewechselt sind. Insgesamt haben also $44 + 31 = 75$ Kunden ihre Präferenz geändert.

Wir müssen also die Symmetrie der Kontingenztafel testen. Die Teststatistik berechnet sich dann als

$$\begin{aligned}
T &= \frac{44/75 - 0.5}{\frac{0.5}{\sqrt{75}}} \\
&= 1.501
\end{aligned}$$

und ist somit nicht signifikant. Es kann nicht ausgeschlossen werden, daß die Werbekampagne nutzlos war.