Datenanalyse für Naturwissenschaftler*Innen
Statistische Methoden in Theorie und Praxis **Uni Freiburg**
WiSe 2022/23
Fernandez Lebron, Andres (RZ: af231), Mikaya, Maxim (RZ: mm1242)

Blatt 01 $\Sigma = 22$ 27. Oktober 2022

Aufgabe 1 (SARS-COV-2 Antigenschnelltests)

Die aktuellen Mindestkriterien für SARS-COV-2 Antigenschnelltests in Deutschland lauten

- Sensitivität: > 82 %
- Spezifität: > 98 %

Nehmen wir an, dass ein Schnelltest eine Sensitivität von 82 % und eine Spezifität von 98 % erfüllt. Und nehmen wir zusätzlich an, dass 3 % aller Menschen in Deutschland derzeit mit SARS-COV-2 infiziert sind und damit weitere Menschen anstecken könnten. Nehmen Sie der Einfachheit halber an, Deutschland hätte eine Bevölkerung von 80 Millionen.

(a) Wie groß ist die Wahrscheinlichkeit, das Virus zu haben, wenn der Schnelltest negativ ist (man sagt, das Testresultat wäre dann "falsch negativ")? (2P)

(b) Wie groß ist die Wahrscheinlichkeit, das Virus nicht zu haben, wenn der Schnelltest positiv ist ("falsch positiv")? (2P)

(c) Stellen Sie sich vor, es nähmen 5000 Menschen an einem Konzert teil, die alle negativ getestet sind. Wie viele Personen sind trotzdem erkrankt und können andere Personen bei dem Konzert anstecken? (2P)

Lösung

T+: Test positiv V+: Virus vorhanden
T-: Test negativ V-: Virus nicht vorhanden

Sensitivität (SEN): P(T+|V+) > 82 % = 0.82
Spezifität (SPE): P(T-|V-) > 98 % = 0.98
Prävalenz (PRV): P(V+) = 3 % = 0.03 ⇒ P(V-) = 97 % = 0.97

(a) Falsch-Negativ Rate (FNR) = (V+|T-)

$$
\begin{aligned}
P(V+|T-) &= \frac{P(T-|V+)\cdot P(V+)}{P(T-|V+)\cdot P(V+) + P(T-|V-)\cdot P(V-)} \\
&= \frac{(1-SEN)\cdot PRV}{(1-SEN)\cdot PRV + SPE\cdot(1-PRV)} \\
&= \frac{0.18\cdot 0.03}{0.18\cdot 0.03 + 0.98\cdot 0.97} \\
&= \frac{0.0054}{0.0054 + 0.9506} \\
&= \frac{0.0054}{0.956} \\
&= 0.0056 \quad +2
\end{aligned}
$$

(b) Falsch-Positiv Rate (FPR) = (V-|T+)

$$\begin{aligned}
P(V-|T+) &= \frac{P(T+|V-)\cdot P(V-)}{P(T+|V-)\cdot P(V-)+P(T+|V+)\cdot P(V+)} \\
&= \frac{(1-SPE)\cdot(1-PRV)}{(1-SPE)\cdot(1-PRV)+SEN\cdot PRV} \\
&= \frac{0.02\cdot 0.97}{0.02\cdot 0.97+0.82\cdot 0.03} \\
&= \frac{0.019}{0.019+0.024} \\
&= \frac{0.019}{0.044} \\
&= 0.44 \quad +2
\end{aligned}$$

(c) Infizierte im Konzert trotz negativem Test (N=5000)

$$\begin{aligned}
|(V+|T-)| \in N &= N\cdot FNR \\
&= 5000\cdot 0.0056 \\
&= 28 \quad +2
\end{aligned}$$

+6

Aufgabe 2 (Varianz und Kovarianz)

(a) Die Varianz einer Stichprobe x vom Umfang N, mit den Elementen $\{x_1, x_2, \ldots, x_N\}$, ist gegeben durch

$$V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2,$$

wobei

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

der Mittelwert ist. Zeigen Sie, dass dies zu

$$V(x) = \overline{x^2} - \overline{x}^2$$

äquivalent ist. Mit anderen Worten: Die Varianz ergibt sich aus dem Mittelwert der Quadrate minus dem Quadrat des Mittelwertes. (3P)

(b) Die Kovarianz der N Paare von Messobservablen $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$ ist gegeben durch

$$\mathrm{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

(i) Zeigen Sie, dass dies äquivalent zu

$$\mathrm{cov}(x, y) = \overline{xy} - \overline{x}\,\overline{y}$$

ist. (3P)

(ii) Zeigen Sie, dass unter der Variablentransformation $x \to x + C$, wobei C eine beliebige Konstante ist, die Kovarianz $\mathrm{cov}(x, y)$ erhalten ist. (3P)

Lösung

(a)

$$\begin{aligned}
Var(x) &= \frac{1}{n}\sum_{i=1}^{N}(x_i - \overline{x})^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x\overline{x} + \overline{x}^2) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i^2) + \frac{1}{N}\sum_{i=1}^{N}(-2x_i\overline{x} + \overline{x}^2) \\
&= \overline{x^2} + \frac{1}{N}\sum_{i=1}^{N}(-2x_i\overline{x}) + \frac{N}{N}\overline{x}^2 \\
&= \overline{x^2} - \frac{2\overline{x}}{N}\sum_{i=1}^{N}(x_i) + \overline{x}^2 \\
&= \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 \\
&= \overline{x^2} - \overline{x}^2 \qquad +3
\end{aligned}$$

□

(b) (i)

$$\begin{aligned}
cov(x,y) &= \frac{1}{n}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_iy_i - x_i\overline{y} - \overline{x}y_i + \overline{x}\,\overline{y}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_iy_i) - \overline{y}\frac{1}{N}\sum_{i=1}^{N}(x_i) - \overline{x}\frac{1}{N}\sum_{i=1}^{N}(y_i) + \frac{N}{N}\overline{x}\,\overline{y} \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_iy_i) - \overline{y}\,\overline{x} - \overline{x}\,\overline{y} + \overline{x}\,\overline{y} \\
&= \overline{xy} - \overline{x}\,\overline{y} \qquad +3
\end{aligned}$$

□

(ii)

$$\begin{aligned}
cov(x,y) &\Leftrightarrow cov(x+c,y) \\
\overline{x+C} &= \frac{1}{N}\sum_{i=1}^{N}(x_i+C) \\
&= \overline{x} + \frac{N}{N}C \\
&= \overline{x} + C \\
\rightarrow cov(x+C,y) &= \frac{1}{N}\sum_{i=1}^{N}(x_i+C-\overline{x}-C)(y_i-\overline{y}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y}) \\
&= cov(x,y) \qquad +3
\end{aligned}$$

□

+9

Aufgabe 3 (Autounfälle)

Man hört oft, dass starker Regen Autounfälle verursachen kann. Unten steht die Anzahl der Autounfälle in Freiburg an einem spezifischen Tag zusammen mit der Niederschlagsmenge für diesen Tag.

Autounfälle	Niederschlagsmenge [mm]
7	0
16	10,2
4	0
12	0,4
11	2,8
9	3,0
18	7,5
14	7,6
20	15,8
24	25,5
23	24,1
35	30,4

(a) Berechnen Sie den Mittelwert und die Standardabweichung für die Anzahl der Autounfälle und der Niederschlagsmenge pro Tag. (2.5P)

(b) Was ist die Korrelation zwischen den beiden Größen in diesem Datensatz? (2.5P)

(c) Zeichnen Sie die Anzahl der Autounfälle in ein Histogram. Wählen Sie ein sinnvolles Binning. Berechnen Sie Mittelwert und Standardabweichung erneut aus den Binmittelpunkten und der Binhöhe. Was fällt Ihnen auf? (6P)

Lösung

(a) Berechnung

+0

	$\overline{x}$	σ
Niederschlag (x)	5.25	5.43
Autounfälle (y)	15.91	7.98

$\overline{x}$	σ
10.61 mm	8.23 mm
16.08	10.37

(b) Korrelation

$$\rho(x,y) \quad = \quad \frac{Cov(x,y)}{\sigma(x)\sigma(y)} = \frac{\overline{xy} - \overline{x}\;\overline{y}}{\sigma(x)\sigma(y)} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i \cdot y_i) - \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{var(x)} \cdot \sqrt{var(x)}}$$

$$= \frac{\sum_{i=1}^{N}(x_i \cdot y_i) - \frac{1}{\cancel{n}N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{\cancel{n}N}\left(\sum_{i=1}^{N} x_i\right)^2\right) \cdot \left(\sum_{i=1}^{N} y_i^2 - \frac{1}{\cancel{n}N}\left(\sum_{i=1}^{N} y_i\right)^2\right)}}$$

$$= \frac{3013.5 - \left(127.3 \cdot \frac{193}{12}\right)}{\sqrt{\left(2639.91 - \frac{127.3^2}{12}\right) \cdot \left(3917 - \frac{193^2}{12}\right)}}$$

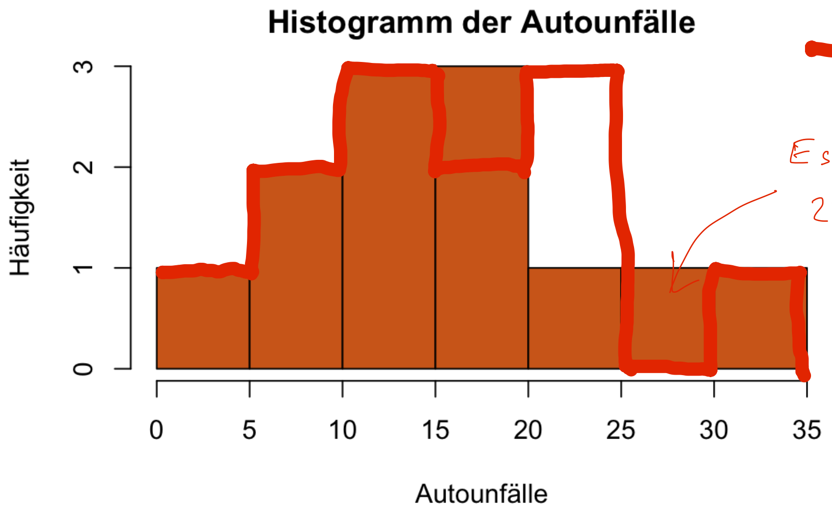
$$= \frac{966.092}{\sqrt{1289.469 \cdot 812.917}}$$

$$= \frac{966.092}{\sqrt{1048231.271}}$$

$$= \frac{966.092}{1023.831}$$

$$= 0.943$$ +2.5

(c) Histogramm



Korrigierung

Es gibt keine Tage mit $25 \leq N_{Autounf.} < 30$!

?

+0.5

Klassen (bins): [0-5) [5-10) [10-15) [15-20) [20-25) [25-30) [30-35)
Mittelpunkte nach bin: 2.5, 7.5, 12.5, 17.5, 22.5, 27.5, 32.5
Binhöhe nach bin: 1 2 3 3 1 1 1 → 1 2 3 2 3 0 1

	$\overline{x}$	σ
Binmittelpunkten	17	10.80
Binhöhe	1.7	0.95

Mittelwert und Standardabweichung für die Anzahl der Autounfälle werden hier gefragt!

Bemerkenswert:

(i) Differenz zwischen den Binmittelpunkten und Binhöhe bzgl. $\overline{x}$ und σ.

(ii) Dieser Unterschied beträgt eine Zehnerposition bei den $\overline{x}$ und ca. eine Zehnerposition bei den σ.

$\bar{y} = \frac{1}{S} \sum_{i=1}^{N} b_i \cdot h_i$, $S = \sum_{i=1}^{N} h_i$

b_i – Mittelpunkt des Bins i

h_i – Binhöhe des Bins i

$\bar{y} = 16.67$

$\sigma(y) = \left(\frac{1}{S} \sum_{i=1}^{N} (b_i - \bar{y})^2\right)^{1/2}$,

$\sigma(y) = 8.98$

+3

Vergleichen $\bar{y}$ und $\sigma(y)$ mit Ergebnissen von (a).

Aufgabe 4 (Wechseln oder nicht?)

In einer Spielshow gibt es drei Türen. Hinter einer Tür wartet ein Gewinn, hinter den beiden anderen Türen befinden sich Nieten. Nachdem eine der Kandidat*Innen eine Tür ausgewählt hat, wird eine der anderen beiden Türen geöffnet hinter der sich eine Niete befindet. Nun darf erneut zwischen den beiden übrigen Türen gewählt werden.

Sollte man bei der ursprünglichen Tür bleiben oder besser die andere Tür wählen, um die Gewinnchance zu erhöhen? Benutzen Sie explizit den Satz von Bayes, um Ihre Antwort zu begründen.

Lösung

Situation:
Kandidat*In wählt die 1.Tür
Moderator öffnet die 2.Tür

Legende:
A: Moderator öffnet Tür 2.
B: Gewinn in Tür 3.
C: Gewinn in Tür 1.
D: Gewinn in Tür 2.

$P(B) = \frac{1}{3}$, $P(C) = \frac{1}{3}$, $P(D) = \frac{1}{3}$,

$P(A|B) = 1$, $P(A|C) = \frac{1}{2}$, $P(A|D) = 0$

$$\begin{aligned} P(B|A) &= +\frac{P(B) \cdot P(A|B)}{P(B) \cdot P(A|B) + P(C) \cdot P(A|C) + P(D) \cdot P(A|D)} \\ &= \frac{\frac{1}{3} \cdot 1}{\frac{1}{3} \cdot 1 + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 0} \\ &= \frac{2}{3} \end{aligned}$$

Berechnet man die W-keit, den Gewinn zu treffen, unter der Bedingungen, (1) dass eine Tür ohne den Gewinn nach der 1. Türwahl geöffnet wurde und dass (2) die türwählende Person nach der Kenntnisnahme über (1) von Tür wechseln, ergibt sich eine W-keit von $\frac{2}{3}$.

Die W-keit, den Gewinn zu treffen, unter der Bedingungen (1) aber nicht (2) ist $\frac{1}{3}$.

Da $\frac{2}{3} > \frac{1}{3}$, lohnt es sich von Tür zu wechseln, wenn man den Gewinn entschiedend treffen will. +4