Aufgabe 1

Musikfan Paul K. hat sich einen neuen MP3-Player zugelegt, welcher über den sogenannten Shuffle-Wiedergabemodus verfügt. In diesem Modus werden die Musiktitel, die auf dem Player gespeichert sind, in einer zufälligen Reihenfolge wiedergegeben. Ein bereits abgespielter Titel wird dabei im weiteren Ablauf von einer erneuten Wiedergabe ausgeschlossen. Paul K. lässt seinen MP3-Player fortan nur noch im Shuffle-Modus laufen.

1. Von den 20 Titeln, die Paul K. auf seinem Player gespeichert hat, sind 5 von seiner Lieblingsband. Berechnen Sie davon ausgehend folgende Wahrscheinlichkeiten.
   (a) Wie groß ist die Wahrscheinlichkeit, dass genau einer von 5 abgespielten Titeln von seiner Lieblingsband ist?
   (b) Wie groß ist die Wahrscheinlichkeit, dass mindestens einer und weniger als 3 von 5 abgespielten Titeln von seiner Lieblingsband sind?
2. Paul K. hat mittlerweile 500 Titel auf seinem MP3-Player gespeichert, davon sind nun 50 von seiner Lieblingsband.
   (a) Wie groß ist die Wahrscheinlichkeit, dass mindestens 9 von 25 abgespielten Titeln von seiner Lieblingsband sind? (**Hinweis:** Nehmen Sie zur Berechnung eine geeignete Approximation vor.)
   (b) Geben Sie einen **R**-Befehl an, mit dem die Wahrscheinlichkeit aus a) berechnet werden kann.
3. Angenommen, im Shuffle-Modus werden bereits abgespielte Titel nicht von einer erneuten Wiedergabe ausgeschlossen.
   (a) Berechnen Sie die Wahrscheinlichkeit, dass an 5. Stelle zum ersten Mal ein Titel von Pauls Lieblingsband abgespielt wird, die wieder mit 50 von 500 Titeln vertreten ist. Verwenden Sie dafür eine Zufallsvariable Z: "Anzahl der abgespielten Titel bis zum ersten Lieblingsband-Titel".
   (b) Wie groß ist die Wahrscheinlichkeit, dass Z im 2-fachen zentralen Schwankungsintervall liegt?
4. Schreiben Sie in **R** eine Funktion *chebyshev*, die Ihnen die Mindestwahrscheinlichkeit für das k-fache Schwankungsintervall nach der Ungleichung von Chebyshev wiedergibt.

Aufgabe 2

1. Geben Sie die Anforderungen an, die die Dichte einer stetigen Wahrscheinlichkeitsverteilung erfüllen muss.

2. Skizzieren Sie die Dichtefunktion einer Normalverteilung mit den Parametern $(\mu = 0, \sigma^2 = 1)$ auf dem Intervall $[-2, 2]$. Berechnen Sie dazu $f_N(-1; \mu, \sigma)$, $f_N(0; \mu, \sigma)$ und $f_N(1; \mu, \sigma)$. Die Dichte einer Normalverteilung ist gegeben durch:

$$f_N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Die Zufallsvariable X sei normalverteilt mit Mittelwert μ und Varianz σ^2.

3. Zeigen Sie, dass die Dichtefunktion der Normalverteilung symmetrisch um den Mittelwert ist, d.h.
$$f_N(\mu + x; \mu, \sigma) = f_N(\mu - x; \mu, \sigma).$$

4. Leiten Sie die Dichte der Normalverteilung nach x ab.

5. An welcher Stelle nimmt die Dichte der Normalverteilung ein Extremum an?

6. Berechnen Sie die Wahrscheinlichkeit, dass die normalverteilte Zufallsvariable X im einfachen zentralen Schwankungsintervall liegt?

7. Zeigen Sie, dass für die normalverteilte Zufallsvariable X Mittelwert und Median übereinstimmen.

8. Schreiben Sie eine Funktion SI in `R`, die die Wahrscheinlichkeit angibt, dass eine normalverteilte Zufallsvariable im k-fachen Schwankungsintervall liegt:

   (a) Geben Sie den Befehl zur Erstellung der Funktion SI an, der das Command-Window aufruft.

   (b) Nachdem sich das Command-Window geöffnet hat, sehen Sie nachfolgendes Schema. Vervollständigen Sie dieses.

```
function(k,mu,sd)
{
...
return(SI)
}
```

Aufgabe 3

Im Laufe einer **Bouldermeisterschaft** (das Klettern ohne Kletterseil und Klettergurt) wurde die erbrachte Leistung von 100 Teilnehmern notiert. Jeder Teilnehmer hat je nach erreichter Höhe 0, 1 oder 2 Punkte bekommen. Zusätzlich wurden die Teilnehmer danach befragt, wie oft sie in der Woche trainieren. Die Ergebnisse der Befragung erbrachten folgende Häufigkeitsfunktion:

$X \downarrow \backslash Y \rightarrow$	0	1	2
1	0.05	0.10	0.05
2	0.01	0.30	0.15
3	0.04	0.20	0.10

Definiert seien das Merkmal X : "Anzahl der Trainingseinheiten pro Woche" und das Merkmal Y : "Im Wettbewerb erreichte Punkte".

1. Bestimmen Sie die Randverteilung des Merkmals X. Geben Sie deren Median an und interpretieren Sie diesen.
2. Berechnen Sie die Werte der zweidimensionalen Verteilungsfunktion $F_{X,Y}(x, y)$ aus dem Träger der Merkmale X und Y und interpretieren Sie den Wert $F_{X,Y}(2, 1)$.
3. Bestimmen Sie die Häufigkeitsfunktion und die Varianz der im Wettbewerb erreichten Punkte unter der Bedingung, dass mindestens zweimal in der Woche trainiert wurde. Zusätzlich wurde errechnet, dass die durchschnittliche Anzahl erreichter Punkte unter der oben genannten Bedingung 1.25 beträgt.
4. Sind die beiden Merkmale X und Y statistisch unabhängig? Begründen Sie Ihre Antwort.
5. Nehmen Sie Stellung zu folgenden Aussagen. Begründen Sie kurz Ihre Antwort, wenn Sie die Aussagen als falsch bewerten.
   (a) Als Varianz bezeichnet man den Erwartungwert $E[g(X)]$ mit $g(X) = [(X - E(X))^2]$.
   (b) Das Merkmal X : "Anzahl der Trainingseinheiten pro Woche" ist verhältnisskaliert, da das Merkmal einen natürlichen Nullpunkt, jedoch keine natürliche Einheit besitzt.
6. Gegeben seien Körpergröße und Körpergewicht der fünf besten Teilnehmer:

Teilnehmer i	Körpergröße in m	Körpergewicht in kg
1	1.83	90
2	1.70	70
3	1.82	86
4	1.90	88
5	1.72	71

   (a) Geben Sie die Befehle in **R** an zum Einlesen der Daten aus der Tabelle in eine Matrix M, so dass sich die Größe und das Gewicht jeweils in den Spalten befinden.
   (b) Geben Sie einen Befehl zur Berechnung der Varianz-Kovarianz-Matrix beider Merkmale in **R** an.

Aufgabe 4

Im Rahmen einer Untersuchug werden Kasinowürfel mit der Kantenlänge 19 mm gewogen. Man betrachtet dabei acht Standardwürfel aus Acrylglas und acht Würfel, bei deren Herstellung ein spezieller manipulierbarer Polymer beigemischt wurde. Folgende **Gewichte in Milligramm**(= 1/1000 Gramm) ergaben sich bei der einfachen Stichprobe:

Beobachtung i	1	2	3	4	5	6	7	8
standard x_i (in mg)	9870	9970	10190	10000	9930	9830	10080	10130
manipuliert y_i (in mg)	9600	10180	9890	10260	10180	9500	10520	9870

Die Zufallsvariablen X: "Das Gewicht eines Würfels aus Acrylglas" und Y: "Das Gewicht eines mit einem speziellen Polymer manipulierten Würfels" sind in der Grundgesamtheit mit (μ_X, σ_X^2) und (μ_Y, σ_Y^2) verteilt.

1. Wie ist das Stichprobenmittel $\overline{X}_n$ für großen Stichprobenumfang approximativ verteilt? Womit lässt sich die approximaive Verteilung begründen? Leiten Sie die Parameter der approximativen Verteilung des Stichprobenmittels $\overline{X}_n$ her.

Nun seien X und Y lognormalverteilt, d.h., dass $U = \ln X$ und $V = \ln Y$ normalverteilt sind mit $U \sim N(\mu_U, \sigma_U^2)$ und $V \sim N(\mu_V, \sigma_V^2)$. Anhand erhobener Stichprobe wurde bereits errechnet, dass $\overline{u}_8 = 9.2103$, $\overline{v}_8 = 9.2098$, $s_U^2 = 0.0002$, $\sum_{i=1}^{8}(\ln y - \overline{\ln y})^2 = 0.0085$ und $\sum_{i=1}^{8} \ln[(y - \overline{y})^2] = 87.949$.

2. Berechnen Sie die Stichprobenvarianz von V.
   Hinweis: Wenn Sie die Varianz nicht berechnen konnten, rechnen Sie mit $s_V^2 = 0.01$ weiter.

3. Testen Sie die Vermutung mit einer Irrtumswahrscheinlichkeit $\alpha = 0.04$, dass sich die manipulierten Würfel durch ein durchschnittlich geringeres (logarithmiertes) Gewicht erkennen lassen, wenn bekannt ist, dass $\sigma_U^2 = \sigma_V^2$.

4. Es sei bekannt, dass die wahre Varianz von V 0.01 beträgt.
   (a) Bestimmen Sie das 96%ge Konfidenzintervall für den Mittelwert von V.
   (b) Wie groß muss der Stichprobenumfang gewählt werden, damit das Stichprobenmittel $\overline{V}_n$ mit einer Vertrauenswürdigkeit von 96% maximal um 0.01 von dem wahren Mittelwert μ_V abweicht?

5. Die Werte für $u = \ln x$ und $v = \ln y$ wurden in R eingelesen. Interpretieren Sie folgenden **R-Output**:

```
> var.test(u, v, ratio=1, alternative="less", conf.level=0.95)
  data:  u and v
  F = 0.1311, num df = 7, denom df = 7, p-value = 0.0078
  alternative hypothesis: true ratio of variances is less than 1
  95 percent confidence interval: 0.0000000 0.4963995
  sample estimates: ratio of variances 0.1310784
```

   (a) Welcher Test wird durchgeführt?
   (b) Wie lautet die Nullhypothese?
   (c) Wie lautet die Testentscheidung? Begründen Sie Ihre Entscheidung.

Lösung 1

1. Zufallsvariable X: "Anzahl der Titel von Pauls Lieblingsband (bei einer Wiedergabe von 5 Titeln)"

   $$X \sim Hyp(n = 5, N = 20, M = 5)$$

   (a)

   $$P(X = 1) = f_{Hyp}(1; 5, 20, 5) = \frac{\binom{5}{1} \cdot \binom{20-5}{5-1}}{\binom{20}{5}} = \frac{5 \cdot 1365}{15504} = 0.4402$$

   (b)

   $$\begin{aligned} P(1 \leq X < 3) &= f_{Hyp}(1; 5, 20, 5) + f_{Hyp}(2; 5, 20, 5) \\ &= 0.4402 + \frac{\binom{5}{2} \cdot \binom{20-5}{5-2}}{\binom{20}{5}} = 0.4402 + \frac{10 \cdot 455}{15504} = 0.7337 \end{aligned}$$

2. (a) Zufallsvariable Y: "Anzahl der Titel von Pauls Lieblingsband (bei einer Wiedergabe von 25 Titeln)"

   $$Y \sim Hyp(n = 25, N = 500, M = 50)$$

   Gesucht:

   $$P(Y \geq 9) = 1 - P(Y \leq 8) = 1 - F_{Hyp}(8; 25, 500, 50)$$

   Approximation durch Binomialverteilung, da $n/N = 25/500 = 0.05 \leq 0.05$

   $$\rightarrow Y \sim Bin(n = 25, p = \frac{M}{N} = \frac{50}{500} = 0.1)$$

   $$P(Y \geq 9) = 1 - F_{Bin}(8; 25, 0.1) = 1 - 0.9995 = 0.0005$$

   (b) `> 1-phyper(8,50,450,25)` (oder `> 1-pbinom(8,25,0.1)`)

3. Zufallsvariable Z: " Anzahl abgespielter Titel bis zum ersten der Lieblingsband"

   $$Z \sim Geo(p = \frac{M}{N} = \frac{50}{500} = 0.1)$$

   (a)

   $$P(Z = 4) = f_{Geo}(4; 0.1) = 0.1 \cdot (1 - 0.1)^4 = 0.06561$$

(b) Gesucht:

$$P(\mu - k \cdot \sigma < Z < \mu + k \cdot \sigma)$$

mit

$$\mu = E[Z] = \frac{1-p}{p} = \frac{0.9}{0.1} = 9; \qquad \sigma^2 = Var[Z] = \frac{1-p}{p^2} = \frac{0.9}{0.1^2} = 90; \qquad k = 2$$

$$\begin{aligned} P(9 - 2 \cdot \sqrt{90} < Z < 9 + 2 \cdot \sqrt{90}) &= F_{Geo}(27.9737) - F_{Geo}(-9.9737) \\ &= 1 - (1-0.1)^{28} - 0 = 0.9477 \end{aligned}$$

4. `> chebyshev=function(k){1-1/k^2}`
   oder mit Öffnen des Command-Windows:
   `> chebyshev=fix(chebyshev)`
   im Command-Window: `function(k){1-1/k^2}`

Lösung 2

1 $f_X(x) \geq 0 \; \forall x \in \mathbb{R}$ und $\int_{\mathbb{R}} f(x)dx = 1$

2 $f_X(-1) = 0.242 \; f_X(0) = 0.399, \; f_X(1) = 0.242$

3

$$\begin{aligned} f(\mu + x) =& \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{((\mu+x)-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(-x)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(-x+\mu-\mu)^2}{2\sigma^2}} \\ =& \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{((\mu-x)-\mu)^2}{2\sigma^2}} = f(\mu - x) \end{aligned}$$

4 Die Ableitung ergibt sich aus:

$$f'(x) = f(x) \cdot \frac{(-2(x-\mu))}{2\sigma^2}$$

5 Für ein Extremum muss gelten: $f'(y) = 0$

$$\begin{aligned} & f(x)\frac{(-2(x-\mu))}{2\sigma^2} \overset{!}{=} 0 | : f(x), \\ & \frac{(-(x-\mu))}{\sigma^2} = 0 | \cdot \sigma^2 \\ & (-(x-\mu)) = 0 \\ & x = \mu \end{aligned}$$

6

$$\begin{aligned} & P(\mu - \sigma \leq x \leq \mu + \sigma) = P(-\sigma \leq x - \mu \leq \sigma) = P(x - \mu \leq \sigma) - P(x - \mu \leq -\sigma) \\ & = P\left(\frac{x-\mu}{\sigma} \leq 1\right) - P\left(\frac{x-\mu}{\sigma} \leq -1\right) \\ & = \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 2 \cdot \Phi(1) - 1 = 2 \cdot 0.8413 - 1 = 0.6826 \end{aligned}$$

7 Für den Erwartungswert gilt $E[X] = \mu$ und der Median ergibt sich aus:

$$\begin{aligned} & F(x_{0.5}) = 0.5 \Leftrightarrow P(x \leq x_{0.5}) = 0.5 \\ & P(x \leq x_{0.5}) = P\left(\frac{x-\mu}{\sigma} \leq \frac{x_{0.5}-\mu}{\sigma}\right) = \Phi\left(\frac{x_{0.5}-\mu}{\sigma}\right) \\ & \Phi\left(\frac{x_{0.5}-\mu}{\sigma}\right) = 0.5 \Rightarrow \frac{x_{0.5}-\mu}{\sigma} = \Phi^{-1}(0.5) \\ & x_{0.5} - \mu = 0 \Rightarrow x_{0.5} = \mu \end{aligned}$$

Oder direkt aus der Symmetrie der Verteilungsfunktion um den Mittelwert μ:

$$P(x \leq \mu) = 0.5$$

8 (a) SI=fix(SI)

(b)

```
function (k){
    SI=2*pnorm(k)-1
    return(SI)
}
oder
function (k, mu, sd){
    SI=pnorm(mu+k*sd,mu,sd)-pnorm(mu-k*sd,mu,sd)
    return(SI)
}
```

Lösung zu Aufgabe 3

1. Die Randverteilung von Y:

x_i	1	2	3
$f_X(x)$	0.2	0.46	0.34
F_Y	0.2	0.66	1.00

   Der Median von X ist $x_{(0.5)} = 2$, da $F_X(x_2) \geq 0.5$. Mindestens 50 % aller Teilnehmer haben höchstens 2 mal in der Woche trainiert.

2. Die zweidimensionale Verteilungsfunktion $F_{X,Y}(x,y) = f(X \leq x, Y \leq y)$ lautet:

$X \downarrow \backslash Y \rightarrow$	0	1	2
1	0.05	0.15	0.20
2	0.06	0.46	0.66
3	0.10	0.70	1.00

   Der Wert $F_{X,Y}(2,1) = 0.46$ bedeutet, dass 46% der Befragten höchstens zwei Mal in der Woche trainierten und maximal 1 Punkt im Wettbewerb erreicht haben.

3.
   - Die bedingte Häufigkeitsfunktion von Y $f_{(Y|X)}(y|X \geq 2) = P(Y = y|X \geq 2) = \frac{P(Y=y, X\geq 2)}{P(X \geq 2)}$:
     Bedingte Häufigkeitsfunktion an der Stelle $Y = 0$

$$P(Y = 0|X \geq 2) = \frac{P(Y = 0, X = 2) + P(Y = 0, X = 3)}{P(X = 2) + P(X = 3)} =$$
$$= \frac{0.01 + 0.04}{0.46 + 0.34} = 0.05/0.8 = 0.0625$$

$$P(Y = 1|X \geq 2) = 0.5/0.8 = 0.625$$

$$P(Y = 2|Y \geq 2) = 0.25/0.8 = 0.3125$$

y_i	0	1	2
$f_{Y\|X}(y\|X \geq 2)$	0.0625	0.625	0.3125

   - Die bedingte Varianz von X:

$$E(Y|X \geq 2) = 1.25$$

$$Var(Y|X \geq 2) = 0^2 \cdot 0.0625 + 1^2 \cdot 0.625 + 2^4 \cdot 0,3125 - (1.25)^2 = 1.875 - 1.25^2 = 0.3125$$

4. Statistische Unabhängigkeit von X und Y liegt vor, wenn für jedes Paar gilt:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

   Z.B. für

$$0.15 = P(X = 2, Y = 2) \neq P(X = 2) \cdot P(Y = 2) = 0.46 \cdot 0.3 = 0.138$$

   $\Rightarrow$ X und Y sind statistisch abhängig.

5. (a) Richtig.
   (b) Falsch. Das Merkmal X : "Anzahl Trainings pro Woche" ist absolutskaliert, da das Merkmal einen natürlichen Nullpunkt, und eine natürliche Einheit besitzt.

6.
```
> M=matrix(c(1.83, 1.70, ..., 1.72, 90, 70, ..., 71), ncol=2, byrow=F)
> n=5
> (n-1)/n*var(M)
```

Lösung zu Aufgabe 4

1. (a) Approximative Verteilung des Stichprobenmittels: $\overline{X}_n \sim N\left(\mu_X; \frac{\sigma_X^2}{n}\right)$.
   (b) Begründung der approximativen Verteilung: zentraler Grenzwertsatz.
   (c) Parameter der Verteilung von $\overline{X}$ ergeben sich wie folgt:

$$\text{Mittelwert:} \quad E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n}E(X_i) = \frac{1}{n}n\mu_X = \mu_X,$$

$$\text{Varianz:} \quad VAR(\bar{X}_n) = VAR\left(\frac{1}{n}\sum_{i=1}^{n}(X_i)\right) = \frac{1}{n^2}\sum_{i=1}^{n}VAR(X_i) = \frac{1}{n^2}n\sigma_X^2 = \sigma_X^2/n$$

2. Berechnung der Stichprobenvarianz von V: $s_V^2 = \frac{1}{8-1}\left(\sum_{i=1}^{8}(\ln y - \overline{\ln y})^2\right) = 0.0012$.
   [$s_V^2 = 0.01$, wenn diese Teilaufgabe nicht gelöst wurde.]

3. Mittelwertdifferenztest bei unverbundenen Stichproben:

   $H_0 : \mu_U - \mu_V \leq 0$ gegen $H_A : \mu_U - \mu_V > 0$

   Kritischer Bereich:

$$\frac{\overline{U}_n - \overline{V}_n - \delta_0}{\overline{S}}\sqrt{\frac{mn}{m+n}} > t_{1-\alpha;n+m-2}$$

   Teststatistik:

$$\overline{S}^2 = \frac{(n-1)S_U^2 + (m-1)S_V^2}{m+n-2} = \frac{7 \cdot 0.0002 + 7 \cdot 0.0012[0.01]}{8+8-2} = 0.0007[0.0051]$$

$$t = \frac{9.2103 - 9.2098 - 0}{\sqrt{0.0007[0.0051]}}\sqrt{\frac{64}{16}} = 0.0377[0.014]$$

   Kritische Schranke: $t_{1-\alpha;n+m-2} = t_{0,96;14} = 1.887$
   Testentscheidung: Da $t = 0.0377[0.014] \not> 1.887 = t_{0.96;14}$, wird H_0 nicht abgelehnt.

4. (a) KI für den Mittelwert bei bekannter Varianz:

$$P\left(\overline{V} - \lambda_{1-\alpha/2}\frac{\sigma_V}{\sqrt{n}} \leq \mu_V \leq \overline{V} + \lambda_{1-\alpha/2}\frac{\sigma_V}{\sqrt{n}}\right) = 1 - \alpha$$

$$\alpha = 1 - 0.96 = 0.04; \; \lambda_{1-\alpha/2} = \lambda_{0.98} = 2.0537$$

$$KI = [9.2098 - 2.0537\sqrt{\frac{0.01}{8}}; 9.2098 + 2.0537\sqrt{\frac{0.01}{8}}] = [9.1371; 9.2825]$$

   (b) $2.0537\sqrt{\frac{0.01}{n}} \leq 0.01$
   $n \geq \frac{0.01 \cdot 2.0537^2}{0.01^2} = 421.77$ Die Stichprobengröße muss mindestens n=422 sein.

5. (a) Es wird ein Varianzqoutiententest bei unverbundenen Stichproben durchgeführt.
   (b) Die Nullhypothese ist $H_0 : \sigma_U \geq \sigma_V$.
   (c) Die Nullhypothese wird mit einer Irrtumswahrscheinlichkeit $\alpha = 0.05$ abgelehnt, da *p-value*=0.0078 kleiner ist als $\alpha = 0.05$.