week5: Wahrscheinlichkeitsverteilungen und Zufallszahlen in R

Allgemeine Logik: Die Wahrscheinlichkeitsverteilung $p = p(x)$ einer kontinuierlichen Zufallszahl x mit $p(x) \geq 0$ und $\int_{-\infty}^{+\infty} p(x)\,dx = 1$ hat die R-Syntax

$$p(x) \quad = \quad \texttt{dVert(}\,\texttt{x}\,\texttt{,}\,\cdots\texttt{)}$$

wobei `Vert` der R-Name der Verteilung ist und die Punkte $\cdots$ stehen für optionale Parameter der Verteilung. Dabei steht das `d` für 'density'. Eine Übersicht über die in R eingebauten Wahrscheinlichkeitsverteilungen findet man hier:

`https://hsrm-mathematik.de/WS2425/Wirtschaftsmathematik3/W'keitsverteilungen-in-R.pdf`

Neben der Dichte gibt es noch die R-Funktionen

$$\texttt{pVert(}\,\texttt{x}\,\texttt{,}\,\cdots\texttt{)} \quad := \quad \int_{-\infty}^{x} p(y)\,dy$$

wobei das `p` etwa für probability steht und es gibt die Umkehr-Funktion von dieser Funktion,

$$\begin{aligned} \texttt{pVert(}\,\texttt{x}\,\texttt{,}\,\cdots\texttt{)} \quad &= \quad \texttt{y} \\ \overset{\text{Def. qVert}}{\Leftrightarrow} \qquad \texttt{x} \quad &= \quad \texttt{qVert(}\,\texttt{y}\,\texttt{,}\,\cdots\texttt{)} \end{aligned}$$

wobei das `q` dann für Quantil steht. Das `pVert` wird auch als Verteilungsfunktion bezeichnet und das `qVert` als Quantil-Funktion. Schliesslich gibt es noch die Funktion

$$\texttt{rVert(}\,\texttt{n}\,\texttt{,}\,\cdots\texttt{)}$$

die n Zufallszahlen mit Verteilung $p(x)$ generiert. Dabei steht das `r` dann für random numbers. Die Funktionen `dVert`, `pVert` und `qVert` sind deterministische Funktionen, nur das `rVert` liefert zufällige Resultate.

Beispiel 1, uniforme Verteilung: Die uniforme oder auch Gleichverteilung auf dem Intervall $[a, b]$ hatten wir im week2.pdf definiert durch die Wahrscheinlichkeitsdichte

$$
\begin{aligned}
p_{a,b}(x) \quad &:= \quad \begin{cases} 1/(b-a) & \text{falls} \quad x \in [a,b] \\ 0 & \text{sonst} \end{cases} \\
&\overset{\text{R-Befehl}}{=} \quad \texttt{dunif(}\,\texttt{x}\,\texttt{, min = a, max = b}\,\texttt{)}
\end{aligned}
$$

Daraus ergibt sich

$$
\begin{aligned}
\texttt{punif(}\,\texttt{x}\,\texttt{, min = a, max = b}\,\texttt{)} \quad &:= \quad \int_{-\infty}^{x} p_{a,b}(y)\,dy \\
&= \quad \begin{cases} 0 & \text{falls} \quad x \le a \\ \int_a^x \frac{1}{b-a}\,dy & \text{falls} \quad a < x < b \\ 1 & \text{falls} \quad x \ge b \end{cases}
\end{aligned}
$$

oder, wir können hier das Integral explizit berechnen,

$$
\texttt{punif(}\,\texttt{x}\,\texttt{, min = a, max = b}\,\texttt{)} \quad = \quad \begin{cases} 0 & \text{falls} \quad x \le a \\ \frac{x-a}{b-a} & \text{falls} \quad a < x < b \\ 1 & \text{falls} \quad x \ge b \end{cases}
$$

Die Umkehrfunktion, die Quantil-Funktion $\texttt{qunif(}\,\texttt{y}\,\texttt{, min = a, max = b}\,\texttt{)}$ ergibt sich dann folgendermassen: Gegeben sei ein y zwischen 0 und 1, $y \in [0,1]$. Das y hat die Bedeutung einer Wahrscheinlichkeit. Wir müssen die Gleichung

$$
y \quad = \quad \int_{-\infty}^{x} p(\tilde{x})\,d\tilde{x} \tag{1}
$$

nach x auflösen, das x hat dann die folgende Bedeutung: Erzeugen wir n Zufallszahlen mit Verteilung `Vert`, dann ist der Anteil an Zufallszahlen $n_{\le x}$ die kleiner oder gleich x sind, gegeben durch

$$
\frac{n_{\le x}}{n} \quad = \quad y
$$

Hier in unserem konkreten Beispiel reduziert sich die Gleichung (1) auf

$$
\begin{aligned}
y \quad &= \quad \frac{x-a}{b-a} \\
\Leftrightarrow \qquad (b-a)\cdot y \quad &= \quad x-a \\
\Leftrightarrow \qquad x \quad &= \quad (b-a)\cdot y \;+\; a
\end{aligned}
$$

Wenn wir etwa $y = 0$ setzen, ist $x = a$ und der Anteil n_a/n an Zufallszahlen kleiner oder gleich a ist 0. Setzen wir $y = 1$, bekommen wir $x = b$ und der Anteil n_b/n an Zufallszahlen kleiner oder gleich b ist 1. Da alle Zufallszahlen nach Definition von $p_{a,b}(x)$ zwischen a und b liegen, ist klar, dass das dann so sein muss. Also für die uniforme Verteilung ist

$$
\texttt{qunif(}\,\texttt{y}\,\texttt{, min = a, max = b}\,\texttt{)} \quad = \quad (b-a)\cdot y \;+\; a
$$

mit $0 \le y \le 1$.

Beispiel 2, Normalverteilung: Die Normalverteilung mit Mittelwert μ und Standardabweichung σ ist definiert durch die Wahrscheinlichkeitsdichte

$$\begin{aligned} \varphi_{\mu,\sigma}(x) \quad &:= \quad \tfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &\overset{\text{R-Befehl}}{=} \quad \texttt{dnorm(}\, \texttt{x}\,, \texttt{mean} = \mu\,, \texttt{sd} = \sigma\, \texttt{)} \end{aligned}$$

Die Verteilungsfunktion ist dann gegeben durch

$$\begin{aligned} \Phi_{\mu,\sigma}(x) \;:=\; \textstyle\int_{-\infty}^{x} \varphi_{\mu,\sigma}(y)\, dy \quad &= \quad \textstyle\int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, \frac{dy}{\sqrt{2\pi\sigma^2}} \\ &\overset{\text{R-Befehl}}{=} \quad \texttt{pnorm(}\, \texttt{x}\,, \texttt{mean} = \mu\,, \texttt{sd} = \sigma\, \texttt{)} \end{aligned}$$

In diesem Fall lässt sich das Integral nicht weiter vereinfachen. Die Verteilungsfunktion braucht man etwa bei der folgenden typischen Aufgabenstellung:

Aufgabe: Eine Zufallszahl X sei normalverteilt mit Mittelwert $\mu = 5$ und Standardabweichung $\sigma = 1$.

a) Mit welcher Wahrscheinlichkeit liegt X zwischen 3 und 4 ?

b) Überprüfen Sie Ihr Resultat aus (a), indem Sie $n = 10000$ solcher Zufallszahlen generieren und dann schauen, wie viele davon zwischen 3 und 4 liegen.

Lösung: a) Wir bekommen

$$\begin{aligned} \mathsf{Prob}[\, 3 \le X \le 4\,] \quad &= \quad \textstyle\int_{3}^{4} \varphi_{\mu=5,\sigma=1}(y)\, dy \\ &= \quad \textstyle\int_{-\infty}^{4} \varphi_{\mu=5,\sigma=1}(y)\, dy \;-\; \int_{-\infty}^{3} \varphi_{\mu=5,\sigma=1}(y)\, dy \\ &= \quad \Phi_{\mu=5,\sigma=1}(4) \;-\; \Phi_{\mu=5,\sigma=1}(3) \\ &\overset{\text{R-Befehl}}{=} \quad \texttt{pnorm(}\, \texttt{4}\,, \texttt{mean} = \texttt{5}\,, \texttt{sd} = \texttt{1}\, \texttt{)} \;-\; \texttt{pnorm(}\, \texttt{3}\,, \texttt{mean} = \texttt{5}\,, \texttt{sd} = \texttt{1}\, \texttt{)} \\ &\approx \quad 0.1587 \;-\; 0.0228 \;=\; 0.1359 \end{aligned}$$

b) $\rightarrow$ siehe `week5.txt` ■

Die Quantil-Funktion `qnorm` lässt sich in diesem Fall ebenfalls nicht mehr explizit angeben. Sie hat die folgende anschauliche Bedeutung, machen wir gleich konkret im `week5.txt`: Wir generieren etwa $N = 10000$ (oder $N = 10001$, der Grund wird gleich klarer) (μ, σ)-normalverteilte Zufallszahlen `z` und ordnen sie dann der Grösse nach mit dem Befehl

$$\texttt{sz} \;=\; \texttt{sort(z)}$$

so dass also

$$\texttt{sz[1]} \;<\; \texttt{sz[2]} \;<\; \texttt{sz[3]} \;<\; \cdots \;<\; \texttt{sz[9999]} \;<\; \texttt{sz[10000]} \;<\; \texttt{sz[10001]}$$

Das Element `sz[5001]` hat dann die folgende Eigenschaft: Genau 5000 Zahlen sind kleiner als `sz[5001]` und 5000 Zahlen sind grösser als `sz[5001]`, deshalb heisst es dann auch der **Median** von der Stichprobe `z`. Entsprechend heissen `sz[2501]` das **erste Quartil**, genau ein viertel aller Zahlen sind kleiner als `sz[2501]` und 3/4 aller Zahlen sind grösser als `sz[2501]`, und `sz[7501]` heisst dann das **dritte Quartil**. Es gilt dann für $1 \leq i \leq N$:

$$\texttt{sz[i]} \overset{N\to\infty}{\approx} \texttt{qnorm(i/N, mean} = \mu, \texttt{sd} = \sigma)$$

Insbesondere gilt also:

$$\begin{aligned} \text{erstes Quartil} &= \texttt{qnorm(0.25, mean} = \mu, \texttt{sd} = \sigma) \\ \text{Median} &= \texttt{qnorm(0.5, mean} = \mu, \texttt{sd} = \sigma) \\ \text{drittes Quartil} &= \texttt{qnorm(0.75, mean} = \mu, \texttt{sd} = \sigma) \end{aligned}$$

Allgemein, für ein $0 < \alpha < 1$ heisst die Zahl

$$\texttt{qnorm(}\ \alpha\ \texttt{, mean} = \mu, \texttt{ sd} = \sigma)$$

ein α-Quantil. Diese Bezeichnungen gelten nicht nur für die Normalverteilung, sondern für beliebige W'keitsverteilungen.