KAPITEL 2

Ordnungsstatistiken und Quantile

Um robuste Lage- und Streuungsparameter einführen zu können, benötigen wir Ordnungsstatistiken und Quantile.

2.1. Ordnungsstatistiken und Quantile

DEFINITION 2.1.1. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Wir können die Elemente der Stichprobe aufsteigend anordnen:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Wir nennen $x_{(i)}$ die i-te *Ordnungsstatistik* der Stichprobe.

Zum Beispiel ist $x_{(1)} = \min\limits_{i=1,\ldots,n} x_i$ das Minimum und $x_{(n)} = \max\limits_{i=1,\ldots,n} x_i$ das Maximum der Stichprobe.

DEFINITION 2.1.2. Der *Stichprobenmedian* ist gegeben durch

$$\text{med}_n = \text{med}_n(x_1, \ldots, x_n) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{falls } n \text{ ungerade,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & \text{falls } n \text{ gerade.} \end{cases}$$

Somit befindet sich die Hälfte der Stichprobe über dem Stichprobenmedian und die andere Hälfte der Stichprobe darunter.

BEISPIEL 2.1.3. Der Median ist ein robuster Lageparameter. Als Beispiel dafür betrachten wir zwei Stichproben mit Stichprobenumfang $n = 8$.
Die erste Stichprobe sei

$$(x_1, \ldots, x_8) = (1, 2, 2, 2, 1, 1, 1, 2).$$

Somit sind die Ordnungsstatistiken gegeben durch

$$(x_{(1)}, \ldots, x_{(8)}) = (1, 1, 1, 1, 2, 2, 2, 2).$$

Daraus lässt sich der Median berechnen und dieser ist $\text{med}_8 = \frac{1+2}{2} = 1.5$.
Als zweite Stichprobe betrachten wir

$$(y_1, \ldots, y_8) = (1, 2, 2, 2, 1, 1, 1, 20).$$

Die Ordnungsstatistiken sind gegeben durch

$$(y_{(1)}, \ldots, y_{(n)}) = (1, 1, 1, 1, 2, 2, 2, 20),$$

und der Median ist nach wie vor $\text{med}_8 = 1.5$. Dies zeigt, dass der Median robust ist.

BEMERKUNG 2.1.4. Im Allgemeinen gilt $\text{med}_n \neq \bar{x}_n$.

Ein weiterer robuster Lageparameter ist das getrimmte Mittel.

DEFINITION 2.1.5. Das *getrimmte Mittel* einer Stichprobe $(x_1, \ldots, x_n)$ ist definiert durch

$$\frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}.$$

Die Wahl von k entscheidet, wie viele Daten nicht berücksichtigt werden. Man kann zum Beispiel $k = [0.05 \cdot n]$ wählen, dann werden 10% aller Daten nicht berücksichtigt. In diesem Fall spricht man auch vom 5%-getrimmten Mittel.

Anstatt des getrimmten Mittels betrachtet man oft das *winsorisierte Mittel*:

$$\frac{1}{n} \left(\sum_{i=k+1}^{n-k} x_{(i)} + k\, x_{(k+1)} + k\, x_{(n-k)} \right).$$

Nachdem wir nun einige robuste Lageparameter konstruiert haben, wenden wir uns den robusten Streuungsparametern zu. Dazu benötigen wir die empirischen Quantile.

DEFINITION 2.1.6. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe und $\alpha \in (0, 1)$. Das *empirische α-Quantil* ist definiert durch

$$q_\alpha = \begin{cases} x_{([n\alpha]+1)}, & \text{falls } n\alpha \notin \mathbb{N}, \\ \frac{1}{2}\big(x_{([n\alpha])} + x_{([n\alpha]+1)}\big), & \text{falls } n\alpha \in \mathbb{N}. \end{cases}$$

Hierbei steht $[\cdot]$ für die Gaußklammer.

Der Median ist somit das $\frac{1}{2}$-Quantil.

DEFINITION 2.1.7. Die *empirischen Quartile* sind die Zahlen

$$q_{0,25}, \quad q_{0,5}, \quad q_{0,75}.$$

Die Differenz $q_{0,75} - q_{0,25}$ nennt man den *empirischen Interquartilsabstand*.

Der empirische Interquartilsabstand ist ein robuster Streuungsparameter.

Die empirischen Quantile können als Schätzer für die theoretischen Quantile betrachtet werden, die wir nun einführen werden.

DEFINITION 2.1.8. Sei X eine Zufallsvariable mit Verteilungsfunktion $F(t)$ und sei $\alpha \in (0, 1)$. Das "theoretische" α-Quantil $Q(\alpha)$ von X ist definiert als die Lösung der Gleichung

$$F(Q(\alpha)) = \alpha.$$

Leider kann es passieren, dass diese Gleichung keine Lösungen hat (wenn die Funktion F den Wert α überspringt) oder dass es mehrere Lösungen gibt (wenn die Funktion F auf einem Intervall konstant und gleich α ist). Deshalb benutzt man die folgende Definition, die auch in diesen Ausnahmefällen Sinn ergibt:

$$Q(\alpha) = \inf \{t \in \mathbb{R} : F(t) \geq \alpha\}.$$

BEISPIEL 2.1.9. Weitere Lageparameter, die in der Statistik vorkommen:

(1) Das Bereichsmittel $\frac{x_{(n)}+x_{(1)}}{2}$ (nicht robust).
(2) Das Quartilsmittel $\frac{q_{0,25}+q_{0,75}}{2}$ (robust).

BEISPIEL 2.1.10. Weitere Streuungsparameter:

(1) Die Spannweite $x_{(n)} - x_{(1)}$.

(2) Die mittlere absolute Abweichung vom Mittelwert $\frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}_n|$.

(3) Die mittlere absolute Abweichung vom Median $\frac{1}{n}\sum_{i=1}^{n} |x_i - \mathrm{med}_n|$.

Alle drei Parameter sind nicht robust.

2.2. Verteilung der Ordnungsstatistiken

SATZ 2.2.1. *Seien $X_1, X_2, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen, die absolut stetig sind mit Dichte f und Verteilungsfunktion F. Es seien*

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

die Ordnungsstatistiken. Dann ist die Dichte der Zufallsvariable $X_{(i)}$ gegeben durch

$$f_{X_{(i)}}(t) = \frac{n!}{(i-1)!(n-i)!} f(t) F(t)^{i-1} (1 - F(t))^{n-i}.$$

ERSTER BEWEIS. Damit $X_{(i)} = t$ ist, muss Folgendes passieren:

1. Eine der Zufallsvariablen, z.B. X_k, muss den Wert t annehmen. Es gibt n Möglichkeiten, das k auszuwählen. Die "Dichte" des Ereignisses $X_k = t$ ist $f(t)$.

2. Unter den restlichen $n-1$ Zufallsvariablen müssen genau $i-1$ Zufallsvariablen Werte unter t annehmen. Wir haben $\binom{n-1}{i-1}$ Möglichkeiten, die $i-1$ Zufallsvariablen auszuwählen. Die Wahrscheinlichkeit, dass die ausgewählten Zufallsvariablen allesamt kleiner als t sind, ist $F(t)^{i-1}$.

3. Die verbliebenen $n-i$ Zufallsvariablen müssen allesamt größer als t sein. Die Wahrscheinlichkeit davon ist $(1 - F(t))^{n-i}$.

Indem wir nun alles ausmultiplizieren, erhalten wir das Ergebnis:

$$f_{X_{(i)}}(t) = n f(t) \cdot \binom{n-1}{i-1} F(t)^{i-1} \cdot (1 - F(t))^{n-i}.$$

Das ist genau die erwünschte Formel, denn $n\binom{n-1}{i-1} = \frac{n(n-1)!}{(i-1)!(n-i)!} = \frac{n!}{(i-1)!(n-i)!}$. □

ZWEITER BEWEIS.

SCHRITT 1. Die Anzahl der Elemente der Stichprobe, die unterhalb von t liegen, bezeichnen wir mit

$$N = \#\{i \in \{1, \ldots, n\} : X_i \leq t\} = \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}.$$

Dabei steht # für die Anzahl der Elemente in einer Menge. Die Zufallsvariablen $X_1, \ldots, X_n$ sind unabhängig und identisch verteilt mit $\mathbb{P}[X_i \leq t] = F(t)$. Somit ist die Zufallsvariable N binomialverteilt:

$$N \sim \mathrm{Bin}(n, F(t)).$$

SCHRITT 2. Es gilt $\{X_{(i)} \leq t\} = \{N \geq i\}$. Daraus folgt für die Verteilungsfunktion von $X_{(i)}$, dass

$$F_{X_{(i)}}(t) = \mathbb{P}[X_{(i)} \leq t] = \mathbb{P}[N \geq i] = \sum_{k=i}^{n} \binom{n}{k} F(t)^k (1 - F(t))^{n-k}.$$

SCHRITT 3. Die Dichte ist die Ableitung der Verteilungsfunktion. Somit erhalten wir

$$\begin{aligned} f_{X_{(i)}}(t) &= F'_{X_{(i)}}(t) \\ &= \sum_{k=i}^{n} \binom{n}{k} \left\{ kF(t)^{k-1} f(t)(1 - F(t))^{n-k} - (n-k)F(t)^k (1 - F(t))^{n-k-1} f(t) \right\} \\ &= \sum_{k=i}^{n} \binom{n}{k} kF(t)^{k-1} f(t)(1 - F(t))^{n-k} - \sum_{k=i}^{n} \binom{n}{k} (n-k)F(t)^k (1 - F(t))^{n-k-1} f(t). \end{aligned}$$

Wir schreiben nun den Term mit $k = i$ in der ersten Summe getrennt, und für alle anderen Terme in der ersten Summe führen wir den neuen Summationsindex $l = k - 1$ ein. Die zweite Summe lassen wir unverändert, ersetzen aber den Summationsindex k durch l:

$$\begin{aligned} f_{X_{(i)}}(t) = \binom{n}{i} iF(t)^{i-1} f(t)(1 - F(t))^{n-i} \\ + \sum_{l=i}^{n-1} \binom{n}{l+1} (l+1)F(t)^l f(t)(1 - F(t))^{n-l-1} \\ - \sum_{l=i}^{n} \binom{n}{l} (n-l)F(t)^l f(t)(1 - F(t))^{n-l-1}. \end{aligned}$$

Der Term mit $l = n$ in der zweiten Summe ist wegen des Faktors $n - l$ gleich 0, somit können wir in der zweiten Summe bis $n - 1$ summieren. Nun sehen wir, dass die beiden Summen gleich sind, denn

$$\binom{n}{l+1}(l+1) = \frac{n!}{l!(n-l-1)} = \binom{n}{l}(n-l).$$

Die Summen kürzen sich und somit folgt

$$f_{X_{(i)}}(t) = \binom{n}{i} iF(t)^{i-1} f(t)(1 - F(t))^{n-i}. \qquad \square$$

AUFGABE 2.2.2. Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte f und Verteilungsfunktion F. Man zeige, dass für alle $1 \leq i < j \leq n$ die gemeinsame Dichte der Ordnungsstatistiken $X_{(i)}$ und $X_{(j)}$ durch die folgende Formel gegeben ist:

$$f_{X_{(i)},X_{(j)}}(t,s) = f(t)f(s)\binom{n}{2}\binom{n}{i-1, j-1-i, n-j} F(t)^{i-1}(F(s)-F(t))^{j-1-i}(1-F(s))^{n-j}.$$

Im nächsten Satz bestimmen wir die gemeinsame Dichte *aller* Ordnungsstatistiken.

SATZ 2.2.3. *Seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte f. Seien $X_{(1)} \leq \ldots \leq X_{(n)}$ die Ordnungsstatistiken. Dann ist die gemeinsame Dichte des Zufallsvektors $(X_{(1)}, \ldots, X_{(n)})$ gegeben durch*

$$f_{X_{(1)},\ldots,X_{(n)}}(t_1, \ldots, t_n) = \begin{cases} n! \cdot f(t_1) \cdot \ldots \cdot f(t_n), & \text{falls } t_1 \leq \ldots \leq t_n, \\ 0, & \text{sonst.} \end{cases}$$

BEWEIS. Da die Ordnungsstatistiken per Definition aufsteigend sind, ist die Dichte gleich 0, wenn die Bedingung $t_1 \leq \ldots \leq t_n$ nicht erfüllt ist. Sei nun die Bedingung $t_1 \leq \ldots \leq t_n$ erfüllt. Damit $X_{(1)} = t_1, \ldots, X_{(n)} = t_n$ ist, muss eine der Zufallsvariablen (für deren Wahl es n Möglichkeiten gibt) gleich t_1 sein, eine andere (für deren Wahl es $n-1$ Möglichkeiten gibt) gleich t_2, usw. Wir haben also $n!$ Möglichkeiten für die Wahl der Reihenfolge der Variablen. Zum Beispiel tritt für $n = 2$ das Ereignis $\{X_{(1)} = t_1, X_{(2)} = t_2\}$ genau dann ein, wenn entweder $\{X_1 = t_1, X_2 = t_2\}$ oder $\{X_1 = t_2, X_2 = t_1\}$ eintritt, was 2 Möglichkeiten ergibt. Da alle Möglichkeiten sich nur durch Permutationen unterscheiden und somit die gleiche "Dichte" besitzen, betrachten wir nur eine Möglichkeit und multiplizieren dann das Ergebnis mit $n!$. Die einfachste Möglichkeit ist, dass $\{X_1 = t_1, \ldots, X_n = t_n\}$ eintritt. Diesem Ereignis entspricht die "Dichte" $f(t_1) \cdot \ldots \cdot f(t_n)$, da die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind. Multiplizieren wir nun diese Dichte mit $n!$, so erhalten wir das gewünschte Ergebnis. □

BEISPIEL 2.2.4. Seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf dem Intervall $[0, 1]$. Die Dichte von X_i ist $f(t) = \mathbb{1}_{[0,1]}(t)$. Somit gilt für die Dichte der i-ten Ordnungsstatistik

$$f_{X_{(i)}}(t) = \begin{cases} \binom{n}{i} i \cdot t^{i-1}(1-t)^{n-i}, & \text{falls } t \in [0, 1], \\ 0, & \text{sonst.} \end{cases}$$

Diese Verteilung ist ein Spezialfall der Betaverteilung, die wir nun einführen.

DEFINITION 2.2.5. Eine Zufallsvariable Z heißt *betaverteilt* mit Parametern $\alpha, \beta > 0$, falls

$$f_Z(t) = \begin{cases} \frac{1}{\mathrm{B}(\alpha,\beta)} \cdot t^{\alpha-1}(1-t)^{\beta-1}, & \text{falls } t \in [0, 1], \\ 0, & \text{sonst.} \end{cases}$$

Bezeichnung: $Z \sim \mathrm{Beta}(\alpha, \beta)$. Hierbei ist $\mathrm{B}(\alpha, \beta)$ die Eulersche *Betafunktion*, gegeben durch

$$\mathrm{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt.$$

Indem wir nun die Dichte von $X_{(i)}$ im gleichverteilten Fall mit der Dichte der Betaverteilung vergleichen, erhalten wir, dass

$$X_{(i)} \sim \mathrm{Beta}(i, n - i + 1).$$

Dabei muss man gar nicht nachrechnen, dass $\frac{1}{\mathrm{B}(i,n-i+1)} = \binom{n}{i} i$ ist, denn in beiden Fällen handelt es sich um eine Dichte. Wären die beiden Konstanten unterschiedlich, so wäre das Integral einer der Dichten ungleich 1, was nicht möglich ist.

AUFGABE 2.2.6. Seien $X_1, \ldots, X_n$ unabhängig und gleichverteilt auf dem Intervall $[0, 1]$. Man zeige, dass

$$\mathbb{E}[X_{(i)}] = \frac{i}{n+1}.$$