

## Antworten zu Kapitel 6: Univariate Deskriptivstatistik

- (1) **Wie ist der Modalwert definiert? Welches Skalenniveau setzt die Anwendung des Modalwertes voraus?**  
Der Modalwert oder Modus ist der Wert derjenigen Merkmalsausprägung (bzw. derjenigen Kategorie), der die meisten Merkmalsträger angehören. Die Anwendung des Modalwertes setzt Variablen auf Nominalskalenniveau voraus.
- (2) **Was ist der Unterschied zwischen singulären Daten und kategorialen Variablen mit geordneten Antwortkategorien?**  
Ordinalskalierte Variablen, bei denen jedem Untersuchungsobjekt der Messwert zugewiesen wird, der seinem Rangplatz entspricht, nennt man singuläre Daten. Bei kategorialen Variablen mit geordneten Kategorien werden die Messwerte aller Untersuchungsobjekte nach einem bestimmten Kriterium einer von mehreren ordinalen Kategorien zugeteilt.
- (3) **Was bedeutet ein Prozentrangwert von 35?**  
Ein Prozentrangwert von 35 bedeutet, dass 35 % der Merkmalsträger eine gleich große oder eine kleinere Merkmalsausprägung aufweisen.
- (4) **Was versteht man unter »verbundenen Rängen«?**  
Wenn sich mehrere Personen einen Rangplatz teilen (etwa weil sie im Rahmen einer Gruppierung der Daten der gleichen Kategorie angehören), spricht man von verbundenen Rängen oder Rangbindungen.
- (5) **Wie ist der Median definiert? Welches Skalenniveau setzt die Anwendung des Medians voraus?**  
Der Median ist der Wert, für den gilt, dass mindestens 50 % der Daten kleiner oder gleich dem Wert sind und mindestens 50 % der Daten größer oder gleich sind. Die Anwendung des Medians setzt ordinalskalierte Daten voraus.
- (6) **Was versteht man unter dem Begriff »Medianklasse«?**  
Bei gruppierten Daten nennt man die Kategorie, in die der Median fällt, Medianklasse.
- (7) **Was ist der Unterschied zwischen einer primären und einer sekundären Häufigkeitsverteilung?**  
Primäre Häufigkeitsverteilungen werden aus der Urliste der Rohwerte gebildet. Sekundäre Häufigkeitsverteilungen werden hingegen durch Kategorisieren und Gruppieren der Messwerte gebildet. Sekundäre Häufigkeitsverteilungen sind meist anschaulicher als primäre Häufigkeitsverteilungen.
- (8) **Was ist der Unterschied zwischen einem Histogramm und einem Säulendiagramm?**  
Das Histogramm unterscheidet sich vom Säulendiagramm dadurch, dass bei ihm die Breite der Säulen sinnvoll interpretierbar ist, da die Säulen auf dem Zahlenstrahl angeordnet sind und der Abstand zwischen zwei Zahlen bei metrischen Variablen bedeutsam ist. Dies hat zur Folge, dass auch die Fläche des Histogramms sinnvoll interpretierbar ist. Dies ist beim Säulendiagramm nicht so, denn dort werden typischerweise diejenigen Kategorien, die nicht besetzt sind, auf der X-Achse überhaupt nicht dargestellt.
- (9) **Wie sind Ausreißer- und Extremwerte in einer Häufigkeitsverteilung definiert?**  
Einen Ausreißerwert definieren wir dadurch, dass er kleiner ist als der Wert  $Q_1 - 1,5 \cdot IQA$  bzw. größer ist als der Wert  $Q_3 + 1,5 \cdot IQA$ . In Worten: Ausreißer sind jene Werte, die mehr als das 1,5-Fache (aber nicht mehr als das 3-Fache) des Interquartilsabstands oberhalb des 3. Quartils oder unterhalb des 1. Quartils liegen. Extremwerte sind Ausreißer, die besonders weit nach unten von  $Q_1$  oder besonders weit nach oben von  $Q_3$  abweichen. Ein Extremwert ist dadurch definiert, dass er

kleiner ist als der Wert  $Q_1 - 3 \cdot IQA$  bzw. größer ist als der Wert  $Q_3 + 3 \cdot IQA$ . In Worten: Extremwerte sind jene Werte, die mehr als das 3-Fache des Interquartilsabstands oberhalb des 3. Quartils oder unterhalb des 1. Quartils liegen.

**(10) Was ist eine Fünf-Punkte-Zusammenfassung?**

Ergänzt man die drei Quartile, die im Box-Whisker-Diagramm dargestellt werden, durch den geringsten vorkommenden Wert ( $x_{\min}$ ) und den höchsten vorkommenden Wert ( $x_{\max}$ ), so erhält man die Fünf-Punkte-Zusammenfassung, die wesentliche Informationen über eine Verteilung enthält.

**(11) Erläutern Sie die vier Eigenschaften des arithmetischen Mittels.**

1) Die Summe der Abweichungen aller Messwerte vom Mittelwert beträgt stets 0.

$$\sum_{m=1}^n (x_m - \bar{x}) = 0$$

2) Die Summe der quadrierten Abweichungen der Messwerte vom Mittelwert ist stets kleiner als die Summe der quadrierten Abweichungen von irgendeinem anderen Wert.

$$\sum_{m=1}^n (x_m - \bar{x})^2 = \min$$

3) Wird zur Variablen  $X$  (d. h. zu jedem Messwert  $x_m$ ) eine Konstante  $a$  addiert, verändert sich der Mittelwert additiv um eben diese Konstante  $a$ .

$$y_m = x_m + a \Rightarrow \bar{y} = \bar{x} + a$$

4) Wird die Variable  $X$  (d. h. jeder Messwert  $x_m$ ) mit einer Konstanten  $b$  multipliziert, verändert sich der Mittelwert multiplikativ um eben diese Konstante  $b$ .

$$y_m = b \cdot x_m \Rightarrow \bar{y} = b \cdot \bar{x}$$

**(12) Welche Vorteile haben robuste Lagekennwerte?**

Robuste Kennwerte werden von Ausreißerwerten gar nicht oder in geringem Umfang beeinflusst.

**(13) Was ist der Unterschied zwischen dem getrimmten Mittel und dem winsorisierten Mittel?**

Bei der Berechnung des  $\delta$ -winsorisierten Mittels  $\bar{x}_w$  werden die Extremwerte nicht wie beim getrimmten Mittel entfernt, sondern auf einen bestimmten Wert festgelegt. Die unteren Extremwerte werden dabei auf den niedrigsten »gezählten« (d. h. nicht entfernten) Wert gesetzt; die oberen Extremwerte werden auf den höchsten »gezählten« Wert gesetzt.

- (14) Wie ist der Semiquartilsabstand definiert? Welches Skalenniveau setzt die Anwendung des Semiquartilsabstands voraus?**

Der Semiquartilsabstand (SQA) ist definiert als der halbe Interquartilsabstand. Er gibt an, in welchem Abstand zum Verteilungszentrum das obere und untere Viertel der Verteilung durchschnittlich liegen:

$$SQA = \frac{Q_3 - Q_1}{2}$$

Die Anwendung des Semiquartilsabstands setzt intervallskalierte Variablen voraus.

- (15) Erläutern Sie die Eigenschaften der Varianz.**

1) Die Varianz ist das arithmetische Mittel der quadratischen Abweichungen und weist daher die gleichen Stärken und Schwächen auf wie das arithmetische Mittel. Die Quadrierung der Differenzen zwischen Messwert und Mittelwert hat zur Folge, dass größere Abweichungen überproportional stärker ins Gewicht fallen als kleinere Abweichungen. Daraus ergibt sich eine hohe Sensibilität für Ausreißer und Extremwerte.

2) Wird zur Variablen  $X$  (d. h. zu jedem Messwert  $x_m$ ) eine Konstante  $a$  addiert, bleiben die Varianz und die Standardabweichung davon gänzlich unberührt. Formal:

$$y_m = x_m + a \Rightarrow s_Y^2 = s_X^2$$

3) Wird die Variable  $X$  (d. h. jeder Messwert  $x_m$ ) mit einer Konstanten  $b$  multipliziert, verändert sich die Varianz um den Faktor  $b^2$ , die Standardabweichung um den Faktor  $b$ . Formal:

$$\begin{aligned} y_m = b \cdot x_m &\Rightarrow s_Y^2 = b^2 \cdot s_X^2 \\ s_Y &= b \cdot s_X \end{aligned}$$

- (16) Wie funktioniert eine z-Standardisierung? Zu welchem Zweck führt man eine z-Standardisierung durch?**

Eine z-Standardisierung bietet einen Referenzrahmen, um Einzelwerte sinnvoll interpretieren und mit anderen Einzelwerten vergleichen zu können. Standardwerte (oder z-Werte) sind als Standardabweichungen vom Mittelwert zu interpretieren. Man erhält sie, indem man die zentrierten Werte durch die Standardabweichung der Verteilung teilt:

$$z_m = \frac{x_m - \bar{x}}{s_X}$$