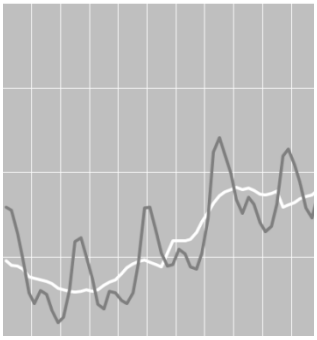
Materialien der Statistik der Bundesagentur für Arbeit für Universitäten

Niveau statistischer Einführungsveranstaltungen



Berechnung des Medians für klassierte Daten

„Die Löhne und Gehälter in Deutschland sind 2013 deutlich stärker gestiegen als in den Vorjahren. Im Mittel hatten sozialversicherungspflichtige Vollzeitbeschäftigte im vergangenen Jahr 2960 Euro pro Monat auf ihrem Gehaltskonto, wie aus der Entgeltstatistik der Bundesagentur für Arbeit (BA) hervorgeht."[1]

Lernziele

Das vorliegende Arbeitsblatt erörtert am Beispiel der Entwicklung sozialversicherungspflichtiger Bruttoarbeitsentgelte sowohl das theoretische als auch in der BA-Statistik verwendete Berechnungsverfahren für den Median bei klassierten Daten. Ziel ist es, sowohl die Anwendungsvoraussetzungen und Berechnungsschritte als auch die Bedeutung herauszustellen. Der Leser ist anschließend in der Lage, eigenständig Medianberechnungen bei vorliegendem klassiertem Datenmaterial vorzunehmen und die Berechnungsergebnisse inhaltlich zu interpretieren.

[1] Spiegel-Online vom 23.07.2014. Zu beachten ist bei diesem Zitat, dass es sich bei dem genannten Euro-Wert um das Bruttoarbeitsentgelt handelt. URL im Internet: http://www.spiegel.de/wirtschaft/unternehmen/deutsche-haben-laut-arbeitsagentur-ba-statistik-2013-mehr-verdient-a-982575.html. Stand: 03.08.2017.

1. Warum ist die Analyse sozialversicherungspflichtiger Bruttoarbeitsentgelte sinnvoll?

Die statistische Berichterstattung über sozialversicherungspflichtige Bruttoarbeitsentgelte ist Bestandteil der von der Statistik der BA erstellten Beschäftigungsstatistik und basiert auf den Angaben aus dem Meldeverfahren zur Sozialversicherung. Damit werden andere Quellen zu Verdiensten und Einkommen wie z. B. die Verdienststrukturerhebung, der Mikrozensus oder das Sozioökonomische Panel (SOEP) um eine wichtige Quelle ergänzt.

Die Entgeltstatistik der Bundesagentur für Arbeit zeichnet sich dadurch aus, dass sie als Vollerhebung regional tief differenzierte Ergebnisse nach Arbeitsort und Wohnort vorlegen kann, die mit anderen Merkmalen aus der Beschäftigungsstatistik kombiniert werden können. Die Ergebnisse aus der Entgeltstatistik, darunter der Median sozialversicherungspflichtiger Bruttoarbeitsentgelte, ermöglichen Aussagen über die Verteilung und Streuung der Bruttoarbeitsentgelte sowie über den Einfluss wichtiger, die individuelle Entgelthöhe bestimmender Faktoren auf unterschiedlichen regionalen Ebenen. Auf Grundlage der Entgeltstatistik sind vielfältige sozioökonomische Analysen möglich. Damit dienen die Ergebnisse der Wirtschaftsbeobachtung und bilden eine der Grundlagen für wirtschafts-, sozial- und konjunkturpolitische Entscheidungen.

2. Was ist der Median und wie erfolgt die Berechnung?

Variablen

Als Variable wird das vom Forscher an der Untersuchungseinheit erhobene Merkmal und damit die interessierende Eigenschaft an der Untersuchungseinheit bezeichnet.

Z. B. durch Befragung oder Beobachtung werden diese Eigenschaften erhoben. Konkrete Variablen sind u. a. „Lebensalter“, „Arbeitszufriedenheit“, „Geschlechtszugehörigkeit“.

Durch Tabellen und Diagramme lassen sich Verteilungen von Merkmalen bzw. Variablen ohne Informationsverlust darstellen. Treffende Maßzahlen tragen dazu bei, Informationen bewusst zu verdichten, um spezifische Eigenschaften zu betonen und die Vergleichbarkeit von Verteilungen zu gewährleisten. Bei den statistischen Maßzahlen unterscheidet man zwischen Schiefemaßen, Streuungsmaßen und Lagemaßen. Letztgenannte geben an, wo sich die Zentren der Verteilung befinden. Der Median ist ein solches Lagemaß.

Dabei stellt der Median $(x_{0,5})$ denjenigen Merkmalswert eines mindestens ordinalskalierten Merkmals X dar, den mindestens 50 Prozent aller Merkmalswerte einer geordneten Stichprobe vom Umfang n unterschreiten und den mindestens 50 Prozent aller Merkmalswerte überschreiten.

Bei ordinalskalierten und geordneten vorliegenden Messwerten $(x_1, x_2, \dots, x_n)$, auch als Urliste bezeichnet, ist der Median wie folgt definiert:

$$x_{0,5} = \begin{cases} x_{\frac{(n+1)}{2}}, & \text{für } n \text{ ungerade} \\ x_{\frac{n}{2}} \quad \text{sowie} \quad x_{\frac{(n+1)}{2}}, & \text{für } n \text{ gerade.} \end{cases} \tag{2.1}$$

Ist das Merkmal hingegen intervall- oder ratioskaliert, d. h. metrisch, so wird der Median wie folgt berechnet:

$$x_{0,5} = \begin{cases} x_{\frac{(n+1)}{2}}, & \text{für } n \text{ ungerade} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}, & \text{für } n \text{ gerade.} \end{cases} \tag{2.2}$$

3. Welche Eigenschaften besitzt der Median?

Der Median weist eine Reihe besonderer Eigenschaften auf. U.a. ist er derjenige Wert, welcher die Summe der Absolutbeträge der Abstände zu den Messwerten $(x_1, x_2, \dots, x_n)$ minimiert. Damit erfüllt der Median die mathematische Bedingung:

$$x_{0,5} = \underset{x \in \mathbb{R}}{\operatorname{argmin}}\, g(x) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \sum_{j=1}^{n} |x - x_j|. \qquad (3.1)$$

Voraussetzung für die obige Aussage ist ein Vorliegen metrischer Merkmale.

Eine vorteilhafte Eigenschaft des Medians ist die ausgeprägte Robustheit gegen extreme Werte (sog. „Ausreißer“). Im Gegensatz zum arithmetischen Mittel haben sehr kleine oder sehr große Werte (fast) keinen Einfluss auf den Wert des Medians.

4. Wie erfolgt in der Theorie die Berechnung des Medians für klassierte Daten?

Merkmalsausprägungen

So werden diejenigen Werte bezeichnet, die eine Variable annehmen kann.

Z. B. hat die Variable „Geschlecht“ die beiden Merkmalsausprägungen „männlich“ und „weiblich“.

Im Gegensatz zu diskreten Merkmalen, die nur bestimmte Werte annehmen und zwischen den Werten Lücken oder Sprungstellen aufweisen (z. B. Anzahl sozialversicherungspflichtige Beschäftigte), können stetige Merkmale alle Werte aus einem Intervall annehmen. In der Praxis werden quantitative Merkmale als stetig behandelt, wenn sie sehr viele Merkmalsausprägungen besitzen (z. B. sozialversicherungspflichtige Bruttoarbeitsentgelte). Analog zu einem diskreten Merkmal bildet die Urliste $(x_1, x_2, \dots, x_n)$ bei einem stetigen Merkmal den Ausgangspunkt der stat. Analyse. Der zugehörige geordnete Datensatz lautet $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$, wobei $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ gilt.

Da ein stetiges Merkmal sehr viele Merkmalsausprägungen besitzt, wäre eine Häufigkeitstabelle, wie sie im Fall eines diskreten Merkmals gebildet würde, sehr unübersichtlich. Daher werden sog. Klassen gebildet. Darin sind mehrere Werte zusammengefasst. Die Untergrenze der i-ten Klasse wird mit x_{i-1}^* und die Obergrenze mit x_i^* bezeichnet. Bis auf die erste Klasse gehört die Obergrenze zur Klasse, die Untergrenze hingegen nicht. Das Intervall $[x_0^*, x_1^*]$ bildet folglich die erste Klasse, während die i-te Klasse für $i > 1$ von der Form $(x_{i-1}^*, x_i^*]$ ist, die auch als links offene oder rechts abgeschlossene Klasse bezeichnet wird.

Sei für $i = 1, 2, \dots, k$ die absolute Häufigkeit n_i und die relative Häufigkeit h_i der i-ten Klasse gegeben. Dann lautet der allgemeine Aufbau einer Häufigkeitstabelle mit klassierten Werten:

Klasse	Intervall	absolute Häufigkeit	relative Häufigkeit
1	$[x_0^*, x_1^*]$	n_1	h_1
2	$(x_1^*, x_2^*]$	n_2	h_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$(x_{k-1}^*, x_k^*]$	n_k	h_k

Tabelle 1: Allgemeiner Aufbau einer Häufigkeitstabelle mit klassierten Daten

Variablen unterschieden nach Wertebereich

- *Qualitative Variablen*:

Die Merkmalsausprägungen werden hinsichtlich ihrer unterschiedlichsten Art differenziert; sie sind immer diskret (z. B. „Parteipräferenz“).

- *Quantitative Variablen:*

Die Merkmalsausprägungen werden hinsichtlich ihrer unterschiedlichen Größe unterschieden; sie können entweder diskret oder stetig sein (z. B. „Alter“, „Noten“).

- *Stetige Variablen:*

Innerhalb eines bestimmten Bereichs kann eine stetige Variable jeden beliebigen Wert annehmen. Es gibt keine Lücken oder Sprungstellen. Zwischen zwei Messwerten sind beliebig viele Zwischenwerte möglich (z. B. „Einkommen“).

- *Diskrete Variablen:*

Eine diskrete Variable kann lediglich bestimmte Werte annehmen. Es existieren Lücken bzw. Sprungstellen zwischen den Werten. In der Praxis werden oftmals diskrete Variablen als quasi-stetige Variablen aufgefasst (z. B. „Alter“).

- *Dichotome Variablen:*

So werden Variablen mit lediglich zwei Merkmalsausprägungen bezeichnet (z. B. „Geschlecht“).

- *Trichotome Variablen:*

So werden Variablen mit drei Merkmalsausprägungen bezeichnet (z. B. „Unterschicht“, „Mittelschicht“, „Oberschicht“).

- *Polytome Variablen:*

So werden Variablen mit mehr als drei Merkmalsausprägungen bezeichnet (z. B. „Einkommen“).

Anwendungsbeispiel 1:

Gegeben sei die stetige Variable Bruttojahresarbeitsentgelt (in Tsd. Euro) von Berufsanfängern. Die zugehörige Urliste sehe folgendermaßen aus:

27, 27, 38, 28, 28, 28, 29, 38, 37, 26, 31, 25, 29,
32, 23, 26, 24, 23, 31, 28, 37, 33, 26, 23, 30.

Der zugehörige geordnete Datensatz lautet:

23, 23, 23, 24, 25, 26, 26, 26, 27,27, 28, 28, **28**,
28, 29, 29, 30, 31, 31, 32, 33, 37, 37, 38, 38.

Liegen die Daten - wie oben dargestellt - als geordnete Liste vor, so wird der Median gemäß Gleichung 2.2 auf Seite 2 berechnet. Er lautet: $x_{0,5} = 28$ (Tsd. Euro).

Die Häufigkeitstabelle mit den absoluten und relativen Häufigkeiten sieht für das obige Anwendungsbeispiel bei Bildung von vier Klassen wie folgt aus:

Klasse i	Intervall $(x_{i-1}^*, x_i^*]$	absolute Häufigkeit n_i	relative Häufigkeit h_i
1	$[20, 25]$	5	0,20
2	$(25, 30]$	12	0,48
3	$(30, 35]$	4	0,16
4	$(35, 40]$	4	0,16

Tabelle 2: Häufigkeitstabelle mit vier Klassen

Bei Scott (1992), Heiler & Michels (1994) finden sich eine Vielzahl von Vorschlägen zur Bestimmung der Klassenanzahl. Aufgrund des inhaltlichen Umfangs werden sie hier allerdings nicht diskutiert.

Um grafisch einen Überblick über die Verteilung der relativen Häufigkeiten zu bekommen, erfolgt die Darstellung in einem Histogramm. Dabei wird in einem rechtwinkligen Koordinatensystem über jede Klasse ein Rechteck abgetragen, so dass dessen Fläche der relativen Häufigkeit der Klasse entspricht. Als Höhe des Rechtecks wird der Quotient aus relativer Häufigkeit h_i und Klassenbreite Δ_i gewählt. Die zugehörige Funktion wird als empirische Dichtefunktion $\widehat{f_n^*}: \mathbb{R} \to \mathbb{R}$ mit

$$\widehat{f_n^*}(x) = \begin{cases} \frac{h_i}{\Delta_i}, & \text{für } x_{i-1}^* < x \leq x_i^*,\ i = 1, \dots, k \\ 0, & sonst. \end{cases} \qquad (4.1)$$

bezeichnet.

Die empirische Dichtefunktion $\widehat{f_n^*}(x)$ bezogen auf das obige Beispiel lautet:

$$\widehat{f_n^*}(x) = \begin{cases} 0{,}040, & \text{für } 20 \leq x \leq 25 \\ 0{,}096, & \text{für } 25 < x \leq 30 \\ 0{,}032, & \text{für } 30 < x \leq 35 \\ 0{,}032, & \text{für } 35 < x \leq 40 \\ 0, & sonst. \end{cases} \qquad (4.2)$$

Variablen unterschieden nach Beobachtbarkeit

- *Manifeste bzw. empirische Variablen:*

Diese sind direkt beobachtbar bzw. direkt messbar (z. B. „Altersangaben“).

- *Latente bzw. theoretische Variablen:*

Sie sind nicht direkt beobachtbar und können nur durch relevante Indikatoren messbar gemacht werden (z. B. „Arbeitszufriedenheit“).

Abbildung 1 zeigt das zugehörige Histogramm:

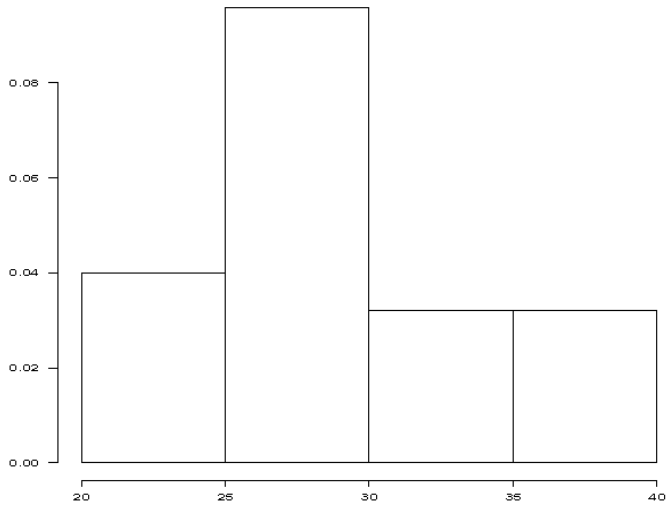


Abbildung 1: Histogramm für das obige Anwendungsbeispiel

Daraus wird ersichtlich, dass fast die Hälfte der Berufsanfänger ein Bruttojahresarbeitsentgelt zwischen 25 Tsd. und 30 Tsd. Euro verdient.

Doch wie lässt sich der Median bei Vorliegen einer Häufigkeitstabelle von klassierten Daten annähernd genau ermitteln? Dies geschieht mithilfe der **approximierenden empirischen Verteilungsfunktion** $\widehat{F_n^*}(x)$. Der Median ist die Lösung der Gleichung:

$$\widehat{F_n^*}(x) = 0{,}5. \qquad (4.3)$$

Dadurch das x als stetig angenommen wird, existiert stets ein $i \in \{1, \dots, k\}$ mit $x \in (x_{i-1}^*, x_i^*]$, so dass $\widehat{F_n^*}(x) = 0{,}5$ gilt. Diese Lösung bezeichnen wir mit $x_{0,5}$.

Zur Bestimmung von $\widehat{F_n^*}(x)$ wird die empirische Dichtefunktion $\widehat{f_n^*}(x)$ aus Gleichung 4.1 auf Seite 4 als Ausgangspunkt herangezogen. An der Stelle x entspricht der Wert der approximierenden empirischen Verteilungsfunktion $\widehat{F_n^*}(x)$ der Fläche unter der empirischen Dichtefunktion $\widehat{f_n^*}(x)$ bis zur Stelle x.

Falls nun der gesuchte Wert x in der i-ten Klasse mit den Klassengrenzen x_{i-1}^*und x_i^* liegt, erhält man diesen Wert, indem die Fläche unter dem Histogramm bis zu der Stelle x bestimmt wird.

Der Wert von $\widehat{F_n^*}(x)$ an der Untergrenze x_{i-1}^* beträgt $\widehat{F_n^*}(x_{i-1}^*)$. Hinzu kommt die Fläche innerhalb der Klasse $(x_{i-1}^*, x]$.

In Abbildung 2 ist diese Fläche schraffiert dargestellt:

Variablen unterschieden nach Skalen- bzw. Messniveau

- *Nominalskalierte Variablen:*

Einzelne Merkmalsausprägungen können nicht rangmäßig unterschieden werden. Ebenso wenig können sie in einen Reihenfolge gebracht werden. Sie stellen Benennungen von Kategorien dar; diese wiederum müssen vollständig sein und sich gegenseitig ausschließen. Die Nominalskala stellt das niedrigste Messniveau dar. Beispiele: „Geschlecht", „Nationalität".

- *Ordinalskalierte Variablen:*

Sie besitzen die gleichen Eigenschaften wie nominalskalierte Variablen. Zusätzlich können „größer/kleiner"-Aussagen zwischen den Merkmalsausprägungen getroffen werden und die jeweiligen Merkmalsausprägungen können rangmäßig der Reihenfolge nach geordnet werden. Jedoch können keine exakten Abstände zwischen den einzelnen Merkmalsausprägungen ausgemacht werden. Beispiele: „Noten", „Lebenszufriedenheit".

- *Intervallskalierte Variablen:*

Hier können Merkmalsausprägungen nicht nur rangmäßig geordnet werden, sondern man kann auch die exakten Abstände zwischen den Ausprägungen angeben. Diese Abstände sind immer gleich groß. Ein Nullpunkt ist willkürlich festlegbar und hat keine inhaltliche Bedeutung; daher sind Aussagen über Verhältnisse unzulässig. Beispiele: „Intelligenzmessung", „Temperatur in Celsius oder in Fahrenheit".

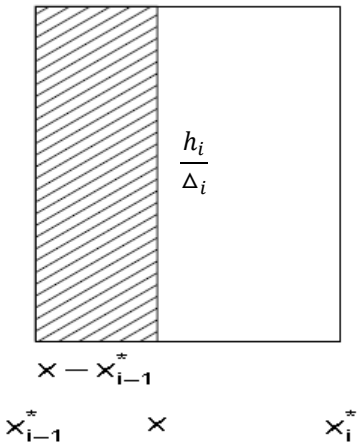


Abbildung 2: Ermittlung der approximierenden empirischen Verteilungsfunktion Mittels des Histogramms

Die Höhe der schraffierten Fläche beträgt $\frac{h_i}{\Delta_i}$ und die Breite $(x - x^*_{i-1})$, so dass damit die schraffierte Fläche den Wert $(x - x^*_{i-1}) \cdot \frac{h_i}{\Delta_i}$ annimmt.

Folglich gilt für die approximierende empirische Verteilungsfunktion:

$$\widehat{F_n^*}(x) = \begin{cases} 0, & f\ddot{u}r\ x \le x_0^* \\ \widehat{F_n^*}(x_{i-1}^*) + \frac{(x - x_{i-1}^*)}{\Delta_i} \cdot h_i, & f\ddot{u}r\ x_{i-1}^* < x \le x_i^*,\ \ i = 1, \dots, k \\ 1, & f\ddot{u}r\ x \ge x_k^*. \end{cases} \tag{4.4}$$

Dabei ist die approximierende empirische Verteilungsfunktion innerhalb jeder Klasse eine in x lineare Funktion der Form $a + b \cdot x$, da

$$\begin{aligned} \widehat{F_n^*}(x) &= \widehat{F_n^*}(x_{i-1}^*) + \frac{(x - x_{i-1}^*)}{\Delta_i} \cdot h_i \\ &= \left(\widehat{F_n^*}(x_{i-1}^*) - \frac{x_{i-1}^*}{\Delta_i} \cdot h_i\right) + \frac{h_i}{\Delta_i} \cdot x \end{aligned} \tag{4.5}$$

gilt.

Die approximierende empirische Verteilungsfunktion für das obige Beispiel lautet:

$$\widehat{F_n^*}(x) = \begin{cases} 0, & f\ddot{u}r\ x < 20 \\ -0{,}8 + 0{,}04 \cdot x, & f\ddot{u}r\ 20 \le x \le 25 \\ -2{,}2 + 0{,}096 \cdot x, & f\ddot{u}r\ 25 < x \le 30 \\ -0{,}28 + 0{,}032 \cdot x, & f\ddot{u}r\ 30 < x \le 35 \\ -0{,}28 + 0{,}032 \cdot x, & f\ddot{u}r\ 35 < x \le 40 \\ 1, & f\ddot{u}r\ x > 40. \end{cases} \tag{4.6}$$

Fortsetzung...

- *Ratioskalierte Variablen:*

  Stellt das höchste Messniveau dar. Es gibt einen absoluten (natürlichen) Nullpunkt im Wertebereich. Daher sind Aussagen über Verhältnisse zulässig. Beispiele: „Temperatur in Kelvin“, „Lebensalter“, „Einkommen“.

  Hinweis:

  Die Unterscheidung zwischen Intervall- und Ratioskala ist in der Praxis für viele Analysezwecke entbehrlich. Zusammenfassend werden beide Messniveaus auch als Variablen auf metrischem Messniveau bezeichnet, da es die Durchführung arithmetischer Rechenoperationen erlaubt.

Tabelle 3 beinhaltet die berechneten zugehörigen Werte:

i	$(x_{i-1}^*, x_i^*]$	h_i	Δ_i	$\widehat{F_n^*}(x_{i-1}^*)$	$\widehat{F_n^*}(x_i^*)$
1	$[20, 25]$	0,20	5	0	0,20
2	$(25, 30]$	0,48	5	0,20	0,68
3	$(30, 35]$	0,16	5	0,68	0,84
4	$(35, 40]$	0,16	5	0,84	1

Tabelle 3: Häufigkeitstabelle für das obige Anwendungsbeispiel mit allen Werten

Abbildung 3 zeigt den Graphen der zugehörigen approximierenden empirischen Verteilungsfunktion und den Median auf der Abszisse:

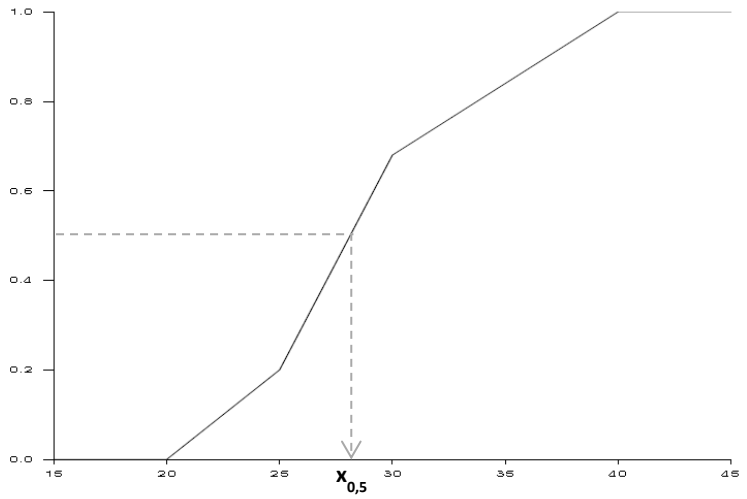


Abbildung 3: Approximierende empirische Verteilungsfunktion des Anwendungsbeispiels

Bei Vorliegen einer Häufigkeitstabelle von klassierten Daten lässt sich der Median gemäß Gleichung 4.3 auf Seite 5 bestimmen, indem diese Gleichung nach der Variablen x aufgelöst wird.Nehmen wir nun an, dass der Median in Klasse $i \in \{1, \dots, k\}$ liegt, dann lässt sich diese Gleichung wie folgt umformen:

$$\widehat{F_n^*}(x) = 0{,}5$$

$$\Leftrightarrow \widehat{F_n^*}(x_{i-1}^*) + \frac{(x - x_{i-1}^*)}{\Delta_i} \cdot h_i = 0{,}5$$

$$\Leftrightarrow \frac{(x - x_{i-1}^*)}{\Delta_i} \cdot h_i = 0{,}5 - \widehat{F_n^*}(x_{i-1}^*)$$

$$\Leftrightarrow (x - x_{i-1}^*) \cdot h_i = \Delta_i \cdot [0{,}5 - \widehat{F_n^*}(x_{i-1}^*)]$$

$$\Leftrightarrow (x - x_{i-1}^*) = \frac{\Delta_i \cdot [0{,}5 - \widehat{F_n^*}(x_{i-1}^*)]}{h_i}$$

$$\Leftrightarrow x = x_{i-1}^{*} + \frac{\Delta_i \cdot [0{,}5 - \widehat{F_n^*}(x_{i-1}^{*})]}{h_i}$$

$$\Rightarrow x_{0,5} = x_{i-1}^{*} + \frac{\Delta_i \cdot [0{,}5 - \widehat{F_n^*}(x_{i-1}^{*})]}{h_i}. \qquad (4.7)$$

Da der Median laut Tabelle 3 in der Klasse 2, d. h. folglich im Intervall (25,30] liegen muss, ergibt sich unter Anwendung von Gleichung 4.7 auf dieser Seite:

$$x_{0,5} = x_{2-1}^{*} + \frac{\Delta_2 \cdot \left[0{,}5 - \widehat{F_n^*}(x_{2-1}^{*})\right]}{h_2}$$

$$= 25 + \frac{5 \cdot [0{,}5 - 0{,}2]}{0{,}48}$$

$$= 28{,}125.$$

Der Median auf Basis der klassierten Häufigkeitsverteilung beträgt für das obige Anwendungsbeispiel somit $\mathrm{x}_{0,5} = 28{,}125$ (Tsd. Euro).

Erkennbar ist der geringfügige Unterschied zu dem berechneten Wert auf Seite 4, der sich bei Vorliegen als geordnete Liste ergibt ($x_{0,5} = 28$). Der Werteunterschied i. H. v. 0,125 (Tsd. Euro) resultiert aufgrund der Verwendung von klassierten Daten und des damit einhergehenden Informationsverlustes aufgrund des Einsatzes der approximativen empirischen Verteilungsfunktion $\widehat{F_n^*}(x)$.

5. Wie erfolgt in der BA-Statistik die Berechnung des Medians für klassierte Daten?

In der Entgeltstatistik der Bundesagentur für Arbeit liegen nach erfolgter Revision im August 2014 die Ergebnisse zu den Bruttomonatsentgelten in 50-Euro-Schritten vor, im Gegensatz zu 100-Euro-Schritten in der nicht-revidierten Statistik. Dabei führt die Verringerung der Klassenbreite nicht dazu, dass die Mindestfallzahl in der amtlichen Berichterstattung i. H. v. 1.000 heruntergesetzt wird. Denn je größer die Fallzahl, desto stabiler und unverzerrter sind die resultierenden statistischen Ergebnisse. Begründung für die 1000er-Grenze ist die Überlegung, dass die Medianklasse mit ausreichend vielen Beobachtungen versehen ist, um bei theoretisch angenommener Gleichverteilung der Werte innerhalb dieser Klasse eine ausreichende Genauigkeit bei der Schätzung des Medians zu erzielen. Bei insgesamt 1000 Fällen liegt durchschnittlich in der Medianklasse eine Besetzungszahl von mindestens 10 Beobachtungen vor. Bei gleichmäßiger Verteilung dieser 10 Beobachtungen auf die Klassenbreite von 50 Euro erzielt man eine Schätzung, die auf 5 Euro genau ist.

Da es in der Praxis einen hohen Aufwand bedeutet, die klassenspezifischen relativen Häufigkeiten $\mathrm{h_i}$ sowie die empirische approximative Verteilungsfunktion an den unteren Klassengrenzen $\widehat{F_n^*}(x_{i-1}^{*})$ zu bilden, kommt ein vereinfachtes Annäherungsverfahren zum Einsatz.

Der Median wird approximativ mit klassierten Daten für Gruppen von Beschäftigten mit Entgeltangaben ermittelt. Genau wie in der nicht-revidierten Statistik, kann in der im August 2014 revidierten Statistik der Mittelwert nicht berechnet werden, da für viele sozialversicherungspflichtig Beschäftigte in der obersten, offenen Entgeltklasse, die jeweilige Höhe des tatsächlich erzielten Entgelts unbekannt ist.

Anwendungsbeispiel 2:

Anhand der Berechnung des Medians für Deutschland vollzieht sich die Vorgehensweise bei der Bestimmung der relevanten Quantilsgrenzen nach dem folgenden Schema:

1.
Die 20.048.103 sozialversicherungspflichtig Vollzeitbeschäftigten der Kerngruppe (vgl. zur Definition der Kerngruppe die Ausführungen im Methodenbericht „Bruttomonatsentgelte von Beschäftigten nach der Revision 2014“, S. 8) am 31.12.2014 mit Entgeltangaben nach Höhe des Entgelts (gemessen an der Zugehörigkeit zu einer Entgeltklasse) werden der Größe nach in zwei Hälften sortiert.

2.
Der Beschäftigte im Mittelpunkt der bundesweiten Verteilung fällt dabei in die Entgeltklasse über 3.000 Euro bis 3.050 Euro. In dieser Klasse gibt es 307.965 sozialversicherungspflichtig Vollzeitbeschäftigte der Kerngruppe. Die korrespondierende Anzahl sozialversicherungspflichtig Beschäftigter in den Klassen unterhalb des Medians beträgt 9.877.003.

3.
Unter der Annahme, dass in dieser Entgeltklasse eine Gleichverteilung vorliegt, gilt nachfolgende Berechnungsformel für die Ermittlung des Medians:

$$B_{insg} = Anzahl\ der\ svB\ insgesamt\ (in\ der\ Kerngruppe)$$

$$B_{uMKl} = Anzahl\ der\ svB\ in\ den\ Klassen\ unterhalb\ der\ Klasse\ des\ Medians$$

$$B_{MKL} = Anzahl\ der\ svB\ in\ der\ Klasse\ des\ Medians$$

$$UG_{MKL} = Untergrenze\ (in\ Euro)\ der\ Klasse\ des\ Medians$$

$$\Delta = Klassenbreite\ (in\ Euro)$$

$$x_{0,5} = UG_{MKL} + \frac{0{,}5 \cdot B_{insg} - B_{uMKl}}{B_{MKL}} \cdot \Delta \qquad (5.1)$$

Es folgt demnach:

$$x_{0,5} = 3.000{,}50\ Euro + \frac{0{,}5 \cdot 20.048.103 - 9.877.003}{307.965} \cdot 50\ Euro$$ = 3.024,37 Euro.

Damit ergibt sich zum 31.12.2014 ein Medianentgelt auf Bundesebene von gerundet 3.024 Euro.

Literatur

Fahrmeier, L., Künstler, R., Pigeot, I., Tutz, G. (2000): Statistik: Der Weg zur Datenanalyse, Springer, Berlin.

Frank, T., Grimm, C. (2010): Beschäftigungsstatistik: Sozialversicherungspflichtige Bruttoarbeitsentgelte. Internetlink Stand: 01.08.2017

Handl, A. (2006): Einführung in die Statistik mit R. Internetlink.
Stand: 02.08.2017.

Heiler, S., Michels, P. (1994): Deskriptive und explorative Datenanalyse. Oldenbourg, München.

Scott, D. W. (1992): Multivariate Density Estimation. Wiley, New York.

Übungen:

1.
Erstellen und bewerten Sie

a)
gemäß des Ansatzes in Abschnitt 4, S. 3-8, die approximative empirische Verteilungsfunktion und die Medianentgelte für Deutschland und die Bundesländer

sowie

b)
gemäß des Ansatzes der BA-Statistik in Abschnitt 5, S. 8-9, die Medianentgelte für Deutschland und die Bundesländer.

Die Ausgangsdaten finden Sie in beigefügter Excel-Datei:


Daten.xlsx

Hinweis: Die „Keine Angabe-Fälle“ in obiger Excel-Datei gehen NICHT in die Medianberechnung ein.

2.
Diskutieren Sie anschließend die Ergebnisse für Deutschland und die Bundesländer. Sind Unterschiede oder Gemeinsamkeiten erkennbar? Welche Ursachen könnten hierfür ausschlaggebend sein?

Weiterführende Produkte zu dem Thema „sozialversicherungspflichtige Bruttoarbeitsentgelte“ finden Sie auf den Seiten der Statistik der Bundesarbeitsagentur: http://statistik.arbeitsagentur.de/.

Entgeltstatistik:
https://statistik.arbeitsagentur.de/Navigation/Statistik/Statistik-nach-Themen/Beschaeftigung/Entgeltstatistik/Entgeltstatistik-Nav.html

Interaktive Visualisierungen:
https://statistik.arbeitsagentur.de/Navigation/Statistik/Statistische-Analysen/Interaktive-Visualisierung/Interaktive-Visualisierung-Nav.html

Methodenberichte:
https://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Methodenberichte/Beschaeftigungsstatistik/Methodeberichte-Beschaeftigungsstatistik-Nav.html

Impressum

Herausgeber: Bundesagentur für Arbeit
Statistik

Erstellungsdatum: Oktober 2017

Statistik-Service West
Tel.: 0211/4306-331
Fax: 0211/4306-470
E-Mail: Statistik-Service-West@arbeitsagentur.de

Weiterführende statistische Informationen:

Internet: http://statistik.arbeitsagentur.de

Statistik-Service Nordost
Tel.: 0511/919-3455
Fax: 0511/919-4103456
E-Mail: Statistik-Service-Nordost@arbeitsagentur.de

Statistik-Service Ost
Tel.: 030/555599-7373
Fax: 030/555599-7375
E-Mail: Statistik-Service-Ost@arbeitsagentur.de

Statistik-Service Südost
Tel.: 0911/179-8001
Fax: 0911/179-908001
E-Mail: Statistik-Service-Suedost@arbeitsagentur.de

Statistik-Service Südwest
Tel.: 069/6670-601
Fax: 069/6670-910307
E-Mail: Statistik-Service-Suedwest@arbeitsagentur.de

Statistik-Service West
Tel.: 0211/4306-331
Fax: 0211/4306-470
E-Mail: Statistik-Service-West@arbeitsagentur.de