Materialien der Statistik der Bundesagentur für Arbeit für Universitäten

Berechnung des Median für klassierte Daten

Niveau statistischer Einführungsveranstaltungen
Empfohlen für wirtschafts- und sozialwissenschaftliche Fachrichtungen

Bundesagentur für Arbeit
Statistik

Impressum

Produktlinie/Reihe: Materialien der Statistik der Bundesagentur für Arbeit für Universitäten

Titel: Berechnung des Median für klassierte Daten

Veröffentlichung: Juni 2023

Herausgeberin: Bundesagentur für Arbeit

Rückfragen an: Thorsten Espenkotte
Statistik-Service West
Josef-Gockeln-Straße 7
40474 Düsseldorf

E-Mail: Statistik-Service-West@arbeitsagentur.de

Telefon: 0211 4306-331

Fax: 0911 179-470

Internet: Statistik-Service-West@arbeitsagentur.de

Zitierhinweis: Statistik der Bundesagentur für Arbeit,
Materialien der Statistik der Bundesagentur für Arbeit für Universitäten – Berechnung des Median für klassierte Daten, Düsseldorf, Juni 2023

Berechnung des Median für klassierte Daten – Lernziele

Das vorliegende Arbeitsblatt erörtert am Beispiel der Entwicklung sozialversicherungspflichtiger Bruttoarbeitsentgelte sowohl das theoretische als auch in der Statistik der Bundesagentur für Arbeit 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.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Formelverzeichnis

Symbolsammlung

N = Grundgesamtheit

n = Stichprobe

x_i = i-te Ausprägung der nach Größe geordneten Datenreihe

X_i = Variable

$\bar{x}$ = Arithmetische Mittel

$\tilde{x}$ = Median

$\mathbb{R}$ = Reelle Zahlen

$\in$ = Element von

j = Laufindex

$\sum$ = (Sigma) Summe

Δ = (Delta) Abweichung oder Differenz

k = Laufindex

h = Häufigkeit

f = Dichte

[ ] = abgeschlossenes Intervall

( ) = offenes Intervall

Begriffliche Klärungen

Merkmalsausprägungen: So werden diejenigen Werte bezeichnet, die eine Variable annehmen kann. Z.B. hat die Variable "Geschlecht" die beiden Merkmalsausprägungen "männlich" und "weiblich".

Klassierte Daten: Oft ist es nicht sinnvoll, Merkmale mit stetiger Ausprägung beliebig genau anzugeben, so dass diese klassiert ausgewertet werden. Außerdem ist jeder Messung aufgrund der Möglichkeiten der Messmethodik Grenzen gesetzt, was die Genauigkeit der Angabe betrifft. Es werden somit Klassen aus den Daten gebildet, die unterschiedlich große aneinandergrenzende Intervalle bilden.

Urliste: Dabei handelt es sich um eine Liste oder Tabelle, in der die jeweilige Merkmalsausprägung eines Merkmals für jede untersuchte statistische Einheit einer statistischen Erhebung enthalten ist. Die Urliste ist der Ausgangspunkt für die weiteren statistischen Analysen, sie enthält jedoch noch keine Zusammenfassungen (z. B. in Klassen) oder Sortierungen. Es handelt sich um eine Datentabelle in ihrem ursprünglichen Zustand. Alternativ spricht man auch von Rohdaten.

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".

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 Merkmalsauprägungen bezeichnet (z. B. "Einkommen").

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").

Variablen unterschieden nach Skalen- bzw. Messniveau

Nominalskalierte Variablen: Einzelne Merkmalsausprägungen können nicht rangmäßig unterschieden werden. Ebenso wenig können sie in eine 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".

1 Was ist der Median und wie erfolgt die Berechnung?

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.

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

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:

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

Formel 1: Median bei ordinalskalierten Werten

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

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

Formel 2: Median bei intervall- oder ratioskalierten Werten

2 Welche Eigenschaften besitzt der Median?

Der Median weist eine Reihe besonderer Eigenschaften auf. Unter anderem 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:

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

Formel 3: Minimumfunktion

Voraussetzung für die obige Aussage ist ein Vorliegen metrischer Merkmale.

3 Wie erfolgt in der Theorie die Berechnung des Median für klassierte Daten?

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 statistischen 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 sogenannte 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
⋮	⋮	⋮	⋮
k	$(x_{k-1}^*, x_k^*]$	n_k	h_k

Tabelle 1: Allgemeiner Aufbau einer Häufigkeitstabelle mit klassierten Daten

Anwendungsbeispiel 1:

Gegeben sei die stetige Variable Bruttojahresarbeitsentgelt (in Tausend Euro) von Berufsanfängern. Die zugehörige Urliste (n = 25) 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 Formel 2 auf Seite 8 berechnet. Er lautet: $x_{0,5} = 28$ (Tausend 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 diese hier 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 $\triangle_i$ gewählt.

Die zugehörige Funktion wird als empirische Dichtefunktion $\widehat{f_n^*}: \mathbb{R} \to \mathbb{R}$ mit

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

Formel 4: Empirische Dichtefunktion

bezeichnet.

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

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

Beispiel 1: Empirische Dichtefunktion

Abbildung 1 zeigt das zugehörige Histogramm:

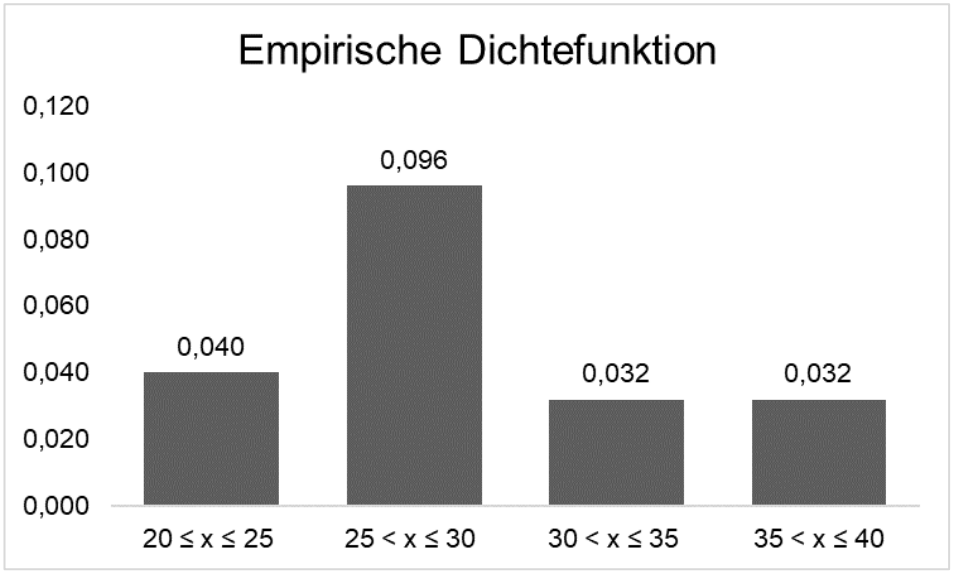

Abbildung 1: Histogramm am Beispiel der empirischen Dichtefunktion

Daraus wird ersichtlich, dass fast die Hälfte der Berufsanfänger ein Bruttojahresarbeitsentgelt zwischen mehr als 25 Tausend und 30 Tausend 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** $\widehat{\mathbf{F}_{\mathbf{n}}^{*}}(\mathbf{x})$. Der Median ist die Lösung der Gleichung: $\widehat{F_n^*}(x) = 0{,}5$.

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 Formel 4 auf Seite 10 als Ausgangspunkt herangezogen. Der Wert der approximierenden empirischen Verteilungsfunktion $\widehat{F_n^*}(x)$ an der Stelle x entspricht 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:

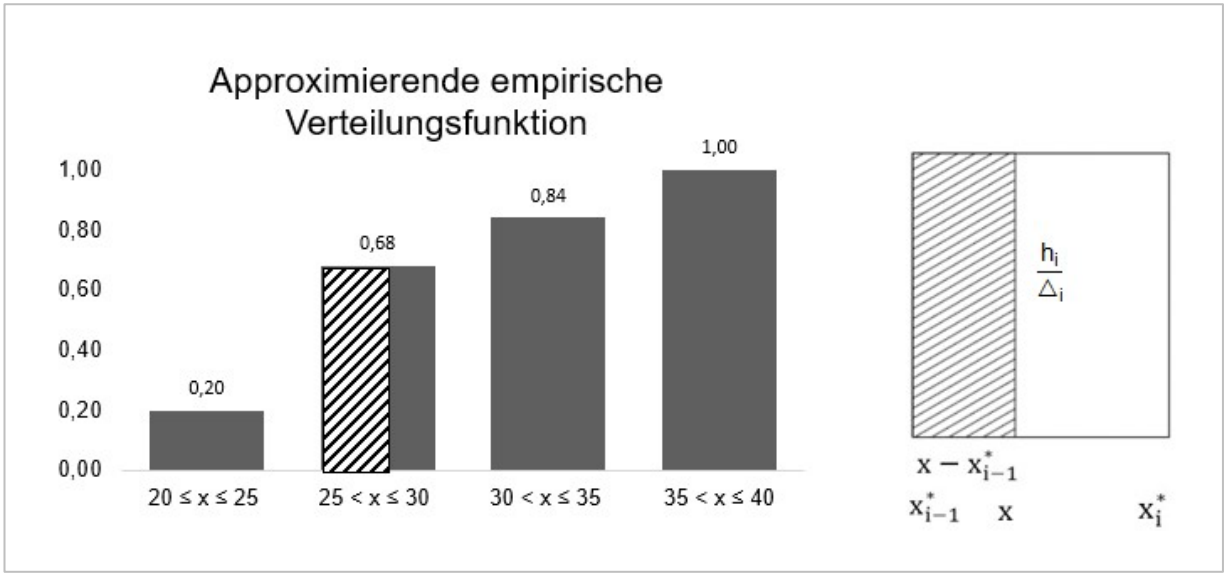

Abbildung 2: Ermittlung der approximierenden empirischen Verteilungsfunktion

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

Folglich gilt für die approximierende empirische Verteilungsfunktion:

$$\widehat{F^*_n}(x) = \begin{cases} 0, & \text{für } x \le x^*_0 \\ \widehat{F^*_n}(x^*_{i-1}) + \dfrac{(x - x^*_{i-1})}{\triangle_i} \cdot h_i, & \text{für } x^*_{i-1} < x \le x^*_i, \ i = 1, \dots, k \\ 1, & \text{für } x \ge x^*_k. \end{cases}$$

Formel 5: Approximierende empirische Verteilungsfunktion

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

$$\begin{aligned} \widehat{F^*_n}(x) &= \widehat{F^*_n}(x^*_{i-1}) + \frac{(x - x^*_{i-1})}{\triangle_i} \cdot h_i \\ &= \left(\widehat{F^*_n}(x^*_{i-1}) - \frac{x^*_{i-1}}{\triangle_i} \cdot h_i \right) + \frac{h_i}{\triangle_i} \cdot x \end{aligned}$$

Formel 6: Alternative Darstellung der approximierenden empirischen Verteilungsfunktion

Die approximierende empirische Verteilungsfunktion für das obige Beispiel lautet:

$$\widehat{F_n^*}(x) = \begin{cases} 0, & \text{für } x < 20 \\ -0{,}8 + 0{,}04 \cdot x, & \text{für } 20 \leq x \leq 25 \\ -2{,}2 + 0{,}096 \cdot x, & \text{für } 25 < x \leq 30 \\ -0{,}28 + 0{,}032 \cdot x, & \text{für } 30 < x \leq 35 \\ -0{,}28 + 0{,}032 \cdot x, & \text{für } 35 < x \leq 40 \\ 1, & \text{für } x > 40. \end{cases}$$

Beispiel 2: Approximierende empirische Verteilungsfunktion

Tabelle 3 beinhaltet die berechneten zugehörigen Werte:

i	$(x_{i-1}^*, x_i^*]$	h_i	$\triangle_i$	$\widehat{F_n^*}(x_{i-1}^*)$	$\widehat{F_n^*}(x_i^*)$
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:

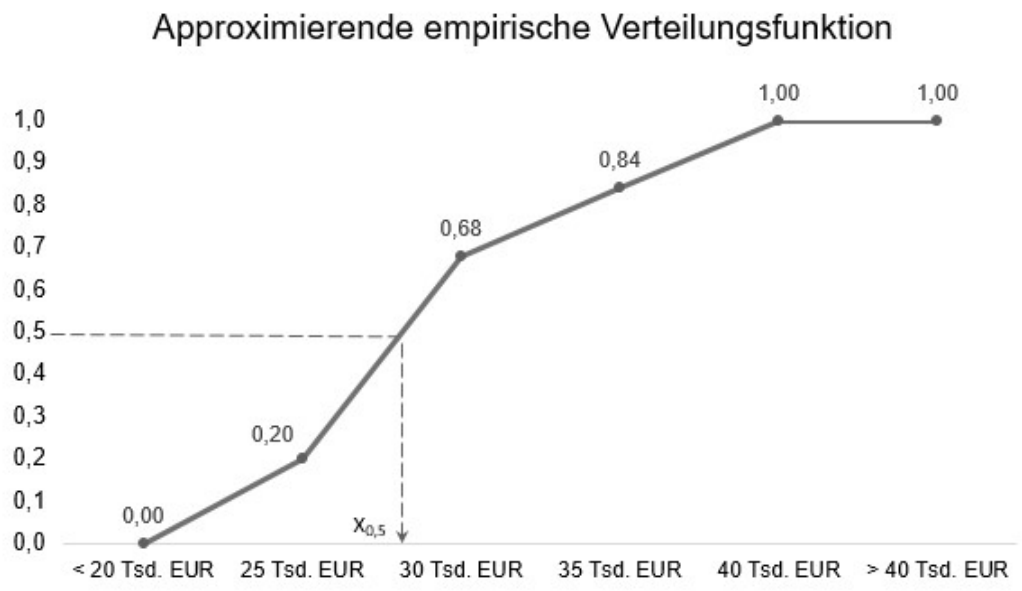

Abbildung 3: Approximierende empirische Verteilungsfunktion des Anwendungsbeispiels

Bei Vorliegen einer Häufigkeitstabelle von klassierten Daten lässt sich der Median gemäß Formel 5 auf Seite 12 bestimmen, indem diese Gleichung nach der Variablen x aufgelöst wird. Nehmen wir nun an, dass der Median in Klasse $i \in \{1, \ldots, k\}$ liegt, dann lässt sich diese Gleichung wie folgt umformen:

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

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

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

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

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

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

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

Formel 7: Mathematische Umformungen

Da der Median laut Tabelle 3 auf Seite 13 in der Klasse 2, das heißt folglich im Intervall $(25{,}30]$ liegen muss, ergibt sich unter Anwendung von Formel 7 auf dieser Seite:

$$x_{0{,}5} = x_{2-1}^* + \frac{\triangle_2 \cdot \left[0{,}5 - \widehat{F_n^*}(x_{2-1}^*)\right]}{h_2}$$

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

$$= 28{,}125.$$

Beispiel 3: Berechneter Median auf Basis der klassierten Daten

Der Median auf Basis der klassierten Häufigkeitsverteilung beträgt somit für das obige Anwendungsbeispiel $x_{0{,}5} = 28{,}125$ (Tausend Euro).

Erkennbar ist der geringfügige Unterschied zu dem berechneten Wert auf Seite 10, der sich bei Vorliegen als geordnete Liste ergibt $(x_{0{,}5} = 28)$. Der Werteunterschied von 0,125 (Tausend Euro) resultiert aufgrund der Verwendung von klassierten Daten und des damit einhergehenden Informationsverlustes aufgrund des Einsatzes der approximativen empirischen Verteilungsfunktion $\widehat{F_n^*}(x)$.

4 Wie erfolgt in der Statistik der Bundesagentur für Arbeit die Berechnung des Median 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 in Höhe von 500 heruntergesetzt wird. Denn je größer die Fallzahl, desto stabiler und unverzerrter sind die resultierenden statistischen Ergebnisse. Begründung für die 500er-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 Median zu erzielen.

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 Median 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", Seite 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 Median beträgt 9.877.003.
3. Unter der Annahme, dass in dieser Entgeltklasse eine Gleichverteilung vorliegt, gilt nachfolgende Berechnungsformel für die Ermittlung des Median:

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

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

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

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

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

$$x_{0,5} = UG_{MKL} + \frac{0{,}5 \cdot B_{insg} - B_{uMKl}}{B_{MKL}} \cdot \triangle$$

Es folgt demnach:

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

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

5 Übungsaufgabe

Erstellen und bewerten Sie

a. gemäß dem Ansatz in Kapitel 3, Seite 9 bis 14, die approximative empirische Verteilungsfunktion und die Medianentgelte für Deutschland und die Bundesländer

sowie

b. gemäß dem Ansatz der Statistik der Bundesagentur für Arbeit in Kapitel 4, Seite 15 bis 16, die Medianentgelte für Deutschland und die Bundesländer.

Übungsdatei barrierefrei | Übungsdatei

Hinweise:

- Die "Keine Angabe-Fälle" in obiger Excel-Dateien gehen nicht in die Medianberechnung ein.
- Zum Öffnen der Übungsdatei speichern Sie diese lokal ab und öffnen Sie sie anschließend mit Acrobat Reader.

Diskutieren Sie anschließend die Ergebnisse für Deutschland und die Bundesländer:

- Sind Unterschiede oder Gemeinsamkeiten erkennbar?
- Welche Ursachen könnten hierfür ausschlaggebend sein?
- Warum gehen "Keine Angabe-Fälle" nicht in den Medianberechnung ein?

Literaturverzeichnis

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

Frank, T., Grimm, C. Beschäftigungsstatistik: Sozialversicherungspflichtige Bruttoarbeitsentgelte, 2010, (Zugriff am 12.08.2021).

Handl, A., Kuhlenkasper, T. Einführung in die Statistik: Theorie und Praxis mit R. Berlin: Springer, 2018.

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

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

Weiterführende Produkte zu dem Thema "sozialversicherungspflichtige Bruttoarbeitsentgelte" auf den Seiten der Statistik der Bundesarbeitsagentur:

Entgeltstatistik

Interaktive Visualisierungen

Methodenberichte

Statistik-Infoseite

Im Internet stehen statistische Informationen unterteilt nach folgenden Themenbereichen zur Verfügung:

Fachstatistiken:
- Arbeitsuche, Arbeitslosigkeit und Unterbeschäftigung
- Ausbildungsmarkt
- Beschäftigung
- Einnahmen/Ausgaben
- Förderung und berufliche Rehabilitation
- Gemeldete Arbeitsstellen
- Grundsicherung für Arbeitsuchende (SGB II)
- Leistungen SGB III

Themen im Fokus:
- Berufe
- Bildung
- Corona
- Demografie
- Eingliederungsbilanzen
- Entgelt
- Fachkräftebedarf
- Familien und Kinder
- Frauen und Männer
- Jüngere
- Langzeitarbeitslosigkeit
- Menschen mit Behinderungen
- Migration
- Regionale Mobilität
- Transformation
- Ukraine-Krieg
- Wirtschaftszweige
- Zeitarbeit

Die Methodischen Hinweise der Statistik bieten ergänzende Informationen.

Die Qualitätsberichte der Statistik erläutern die Entstehung und Aussagekraft der jeweiligen Fachstatistik.

Das Glossar enthält Erläuterungen zu allen statistisch relevanten Begriffen, die in den verschiedenen Produkten der Statistik der BA Verwendung finden.

Abkürzungen und Zeichen, die in den Produkten der Statistik der BA vorkommen, werden im Abkürzungsverzeichnis bzw. der Zeichenerklärung der Statistik der BA erläutert.