„Verdienen Männer mehr als Frauen?“ – Reale Daten im Stochastikunterricht mit der Software TinkerPlots erforschen

Rolf Biehler & Daniel Frischemeier, Paderborn

Zusammenfassung: *„Männer verdienen mehr als Frauen!“, „…“ solche Statistiken werden in der deutschen Medienlandschaft oft aufgegriffen und verbreitet. Die Untersuchung solcher Schlagzeilen und daran anknüpfend die Exploration realer Daten eignet sich auch für den Mathematikunterricht der Sekundarstufe 1. Für ein Drehen und Wenden der Daten und für ein Untersuchen multivariater Daten ist es unerlässlich, dass man auf die Unterstützung einer geeigneten Software zurückgreift. Wir wollen im Folgenden exemplarisch das Explorieren von realen Daten mit der Software TinkerPlots auf zwei möglichen Wegen beschreiben.*

1 Einleitung

„Erziehung der Schülerinnen und Schüler zu mündigen Bürgern“ „Schülerinnen und Schüler sollen die Mathematik selbst als Akteure erleben“ „mit Daten umgehen, die uns etwas angehen“ – das sind Zitate die man findet, wenn es um die Behandlung von Daten im Stochastikunterricht geht. Viele Findungs- und Endscheidungsprozesse, sei es in der Wirtschaft oder in der Politik, beruhen auf Statistiken. Daher – um eines der oben genannten Zitate aufzugreifen – muss die Erziehung und Ausbildung zu „mündigen Bürgern“ auch eine statistische Komponente enthalten. Krüger (2012, S. 8 f.) hat bereits das „unterrichtliche Potential amtlicher Statistik am Beispiel „Haushaltsnettoeinkommen“ untersucht. Engel (2014) hat am Datensatz der Verdienststrukturerhebung 2006 des statistischen Bundesamtes Unterschiede zwischen den Einkommen von männlichen und weiblichen Arbeitnehmern analysiert. Diesen Datensatz verwenden wir auch in unserem Aufsatz. Eine Stichprobe aus diesem umfangreichen Datensatz wurde bereits im Rahmen der Dissertation des zweitgenannten Autors zu Studien mit Lehramtsstudierenden eingesetzt (vgl. Frischemeier, 2014). In den Empfehlungen zu Zielen und zur Gestaltung des Stochastikunterrichts (AK Stochastik 2002)[1] wird u. a. gefordert, dass „Schüler lernen, Daten sachgerecht zu interpretieren und nach Zusammenhängen in Daten zu suchen.“ Schließlich sollte „der Stochastikunterricht ferner durch einen hohen Stellenwert experimenteller Arbeiten und durch selbstständige Datenerhebungen charakterisiert sein. Dabei sind oft Computer zur Darstellung und Auswertung von Daten oder zur Simulation sinnvoll einsetzbar.“ Wir wollen in diesem Artikel dort anknüpfen und auf ein elementares Aufgabenformat im Zyklus der Datenanalyse eingehen – den Verteilungsvergleich. Dieses Vorgehen stellen wir nun anhand eines Beispieldatensatzes aus der Genesis-Datenbank[2] des statistischen Bundesamtes vor, bei dem wir exemplarisch der Frage nachgehen wollen, inwiefern sich weibliche Arbeitnehmer und männliche Arbeitnehmer hinsichtlich ihres Gehalts unterscheiden. Dabei wollen wir fundamentale Ideen zum Blick auf Verteilungen und zum Verteilungsvergleich diskutieren sowie dabei das Unterstützungspotential der Datenanalysesoftware TinkerPlots (Konold & Miller 2011) aufzeigen.

2 Verteilungsvergleiche mit realen Daten

Generell lässt sich festhalten, dass Gruppenvergleiche im Kern einer jeden Datenanalyse stecken. Dieses belegen auch Konold und Higgins (2003, S. 206): „We might think of it [the ability of conducting group comparisons] as the place where instruction in the early years is headed on a foundation from which further statistics will arise. Making such comparisons is the heart of statistics.” Was genau aber ist eigentlich ein Verteilungsvergleich? Darunter versteht man einen Vergleich von Verteilungen einer Variablen in verschiedenen Teilgruppen, die durch ein qualitatives Merkmal definiert werden, im einfachsten Fall eines mit zwei Ausprägungen. Ein Beispiel aus dem Muffins-Datensatz[3] (Biehler et al. 2003) ist: „Inwiefern unterscheiden sich Schülerinnen und Schüler hinsichtlich ihrer Zeit am Computer (in Stunden pro Woche)?“[4] Hier lassen sich Unterschiede bezüglich des „Computer-Verhaltens“ auf vielfältige Art und Weise herausarbeiten (Biehler 2007c). Auch im alltäglichen Leben und in den Medien sind Verteilungsvergleiche präsent. Allerdings sind diese oft nur auf den Vergleich der arithmetischen Mittelwerte beschränkt, die der Problemstellung oft nicht gerecht werden. Dieses „Zurückfallen auf Mittelwerte“ ist eine Beobachtung, die oftmals auch in empirischen Studien, die Lösungsprozesse von Lernenden beim Vergleich von Verteilungen untersuchen, gemacht wird (u. a. Frischemeier & Biehler 2011). Diesem Habitus, Verteilungen nur anhand ihrer Mittelwerte zu unterscheiden und anzunehmen, dass sie sonst im

Wesentlichen gleich sind bzw. weitere Aspekte als nicht relevant angesehen werden, wollen wir entgegenwirken, unter anderem mit den hier aufgezeigten Grundvorstellungen zum Verteilungsvergleich angelehnt an Biehler (2001) und Biehler (2007c). Kurz und prägnant lassen sich folgende Tipps zum Verteilungsvergleich formulieren:

- „Nicht nur Unterschiede in Mittelwerten herausarbeiten“,
- „Verteilungen als Ganzes betrachten“,
- „Streuungsunterschiede zwischen den Verteilungen herausarbeiten und interpretieren (größere Streuung kann ein heterogeneres Verhalten innerhalb der Gruppe bedeuten),
- „Vergleich von Anteilen (relativen Häufigkeiten h) unterhalb oder oberhalb einer bestimmten Grenze in den verschiedenen Teilgruppen“[5],
- „Unterschiede q(uantil)-basiert herausarbeiten“[6],
- „verschiedene Darstellungen (Histogramme mit verschiedenen Klassenbreiten, Boxplots) betrachten und diese nutzen, um Muster und Unterschiede zwischen den Verteilungen zu entdecken“.

Diese Tipps sollen helfen, die Komplexität eines Verteilungsvergleichs zu beschränken und Hinweise für mögliche Explorationen innerhalb der Daten zu geben. Bei einer realen Datenanalyse können diese Aspekte mit Gesichtspunkten kombiniert werden, die sich im Laufe der Datenanalyse aus den Daten selber ergeben bzw. aus dem Kontext, dem die Daten entstammen. Zum einen geben Explorationen in den Daten Anlass, bestimmte Fragen an den Kontext zu stellen (z. B. Suche nach Erklärungen für Ausreißer), zum anderen kann ein spezifischer Kontext zu Fragen und Erkundungen in den Daten verleiten. Dabei lassen sich unserer Auffassung nach in Anlehnung an Makar & Confrey (2014, S. 357) zwei verschiedene Ansätze der Datenanalyse ausmachen: sogenannte „Wanderer“ und „Wunderer“. Während die „Wanderer“ ihren (Analyse-)„Weg“ kennen, einen Analyseplan vor Augen haben und zielorientiert auf die Daten zusteuern, gehen die „Wunderer“ eher unvoreingenommen an die Daten heran und lassen sich von den Auffälligkeiten und Mustern in den Daten inspirieren und lenken und bauen darauf ihre Theorien auf. Idealerweise mischen sich die beiden Vorgehensweisen.

Zunächst stellen wir nun kurz die Datenanalysesoftware TinkerPlots vor, die uns bei unseren Explorationen im Reich der Daten unterstützen soll.

3 Die Software TinkerPlots[7]

Will man mit komplexen, multivariaten Datensätzen arbeiten und diese explorieren – im Sinne eines „Drehen und Wenden nach ausgewählten Fragestellungen“ –, so ist der Einsatz von Software unumgänglich. Eine in unserem Sinne geeignete Software, die beim Vergleichen von Verteilungen unterstützen kann, ist – wie wir im weiteren Verlauf dieses Artikels anregen werden – die Software TinkerPlots 2.0 (Konold & Miller 2011). Diese zeichnet sich durch ihre „konstruktivistische Philosophie“ und „schnelle Erlernbarkeit“ aus und wird in Biehler (2007a), Biehler (2007b) und Biehler et al. (2013) ausführlich beschrieben. Unterrichtspraktisch lässt sich die Software sowohl als Demonstrationsmedium für den Lehrer als auch als Lernsoftware für den Lerner einsetzen. Ein wesentliches Charakteristikum ist neben der Verwaltung der Daten in Form von Datenkarten das Erstellen von Graphen und Darstellungen, in denen die Datenkarten einem Symbol zugeordnet werden und diese Symbole anhand der Operationen „Stapeln“, „Trennen“ und „Ordnen“ sortiert werden. Dadurch können neben selbst gewählten Graphiken auch konventionelle (wie Kreis-, Balken-, Säulendiagramm, Histogramm, Boxplot, etc.) erstellt werden. Exemplarisch wollen wir nun das Potential der Software bei Verteilungsvergleichen aufzeigen sowie gleichzeitig herausarbeiten, was man bei Verteilungsvergleichen in der deskriptiven Statistik alles herausarbeiten kann.

4 Verteilungsvergleiche mit TinkerPlots

Die Importfunktion in TinkerPlots ermöglicht das leichte Importieren von Datensätzen, die auf der Homepage des statistischen Bundesamtes bereitgestellt werden, indem man die Datensätze als csv- oder txt-Datei herunterlädt und dann in Arbeitsfläche von TinkerPlots per „Drag & Drop“ einfügt. Der in diesem Beispiel verwendete Datensatz „Verdienststrukturerhebung 2006“ enthält „absolut anonymisierte Daten für Wissenschaft und Lehre, generiert aus den Daten der Verdienststrukturerhebung (kurz: VSE) 2006. Diese wurde als Stichprobe bei knapp 28.700 Betrieben mit 10 und mehr Beschäftigten erhoben. […]“[8] Wir haben den Datensatz mit einen ursprünglichen Umfang von 60551 Fällen und 32 Merkmalen in TinkerPlots importiert und seitens der Merkmale reduziert und bereinigt. Der Datensatz, mit dem wir nun arbeiten, umfasst 59504 Fälle mit den Merkmalen Bundesland, Wirtschaftszweig-Gruppe, Leistungsgruppe, Geschlecht, Alter, Geburtsjahr, Beruf, Stellung_Beruf, Ausbildung, Arbeitsvertrag,

Bruttomonatsverdienst, bezahlte Stunden, Bruttojahresverdienst und Tarif. Ein Merkmal wie „Stundenlohn“ kommt im Datensatz zwar nicht explizit vor, lässt sich aber anhand des Quotienten aus Bruttomonatsverdienst und der Anzahl der bezahlten Stunden pro Monat leicht errechnen. Dabei wurden riesige Ausreißer (Stundenlohn größer als 1000 € etc.) hinsichtlich dieses Merkmals festgestellt. Diese sind vor allem durch eine sehr geringe Anzahl an „bezahlten Stunden“ entstanden und wurden im oben erwähnten „Bereinigungsprozess“ ohne weitere Recherche aus dem Datensatz eliminiert.

Man kann nun die Explorationen im Datensatz mit einer Zeitungsmeldung z. B. zu Verdienstunterschieden (Onlineausgabe der Wirtschaftszeitung „Handelsblatt“[9]) motivieren:

HANDELSBLATT BERLIN. Frauen hinken beim Gehalt ihren männlichen Kollegen weit hinterher. Pro Stunde verdiente eine Frau 2009 im Schnitt 23 Prozent weniger als ein Mann, wie das Statistische Bundesamt am Montag mitteilte. Je Stunde lag der Bruttoverdienst einer Frau im Schnitt bei 14,90 Euro und bei einem Mann bei 19,40 Euro. Die Kluft bei den Einkommen erweise sich dabei seit mehreren Jahren als stabil, schrieben die Statistiker. […]

Bevor wir die Angaben zum Mittelwert überprüfen, schauen wir uns die Verteilung der gesamten Daten an. Die Arbeit mit einem solchen nicht-aggregierten Datensatz soll eine erste Annäherung an das Thema „Umgang mit Big Data“ sein und erfordert andere Analyseinstrumente als z. B. bei aggregierten Daten.

Betrachten wir aber zunächst einmal die Verteilung des Merkmals „Stundenlohn“ im Einzelnen (Abb. 1a).

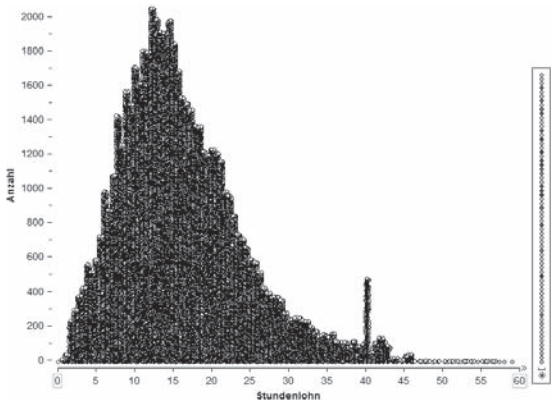


Abb. 1a: Punktdiagramm in TinkerPlots

Dieses „gestapelte Punktdiagramm“ ist eine gängige Darstellung, die mit der Software erzeugt werden kann und liefert einen ersten Eindruck von der Verteilung des Merkmals „Stundenlohn“. Wie auch andere Einkommensverteilungen ist auch diese Verteilung „rechtsschief“. Im Punktdiagramm sind ferner mehrere Gipfel zu erkennen so zum Beispiel der „Hauptgipfel“ links (bei ca. 12,50 €) und rechts (bei ca. 40 € und bei 42 €) kleinere, aber sehr auffällige Gipfel. Wie kann man sich diese erklären? Betrachtet man den Fragebogen zur Befragung wird dieses klar. Es gibt keine Gehaltsobergrenze, sondern es wird nach dem Einkommen von 7000 € oder mehr gefragt. Dieses kann man auf „Stundenlohn“ umrechnen, indem man durch die bezahlten Stunden pro Monat dividiert (diese schwanken bis auf Ausreißer zwischen 152 und 187 Stunden, daher kommt man so zu verschiedenen Häufungen im Intervall [40 €; 45 €]). Die Datenpunkte die jenseits von 60 € und größer liegen (insgesamt 49 von 59504 Fällen) haben wir aus dieser Darstellung herausgenommen, damit der Rest der Verteilung besser sichtbar wird. Diese sind rechts neben der Verteilung aufgestapelt (siehe Kasten rechts in Abb. 1a). In einem weiteren Schritt lässt sich in TinkerPlots ein Boxplot nach der Definition von Tukey[10] erstellen – die Punkte der einzelnen Daten haben wir „unsichtbar“ gemacht, um einen besseren Blick auf den Boxplot zu haben.

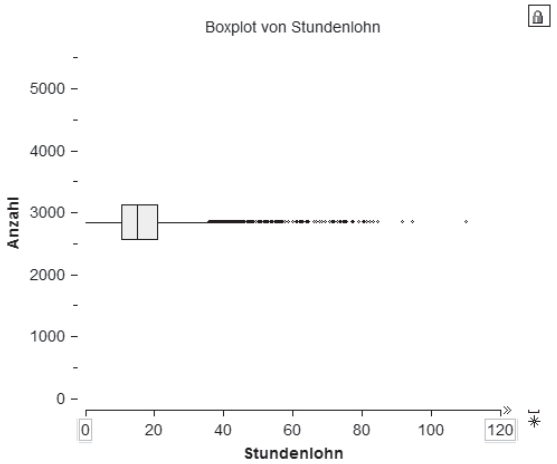


Abb. 1b: Boxplot in TinkerPlots

Dem Boxplot (Abb. 1b) kann man entnehmen, dass die mittleren 50 % zwischen ca. 11,70 € und 22,70 € und ca. 25 % der befragten Arbeitnehmerinnen und Arbeitnehmer bis zu 11,70 € verdienen. Eine weitere Möglichkeit ist die Erstellung[11] eines Histogramms mit frei wählbarer Klassenbreite (Abb. 2). Wir können auch in Abbildung 2 den Hauptgipfel aus dem Punktdiagramm (Abb. 1a) identifizieren. Gibt es vielleicht noch weitere Auffälligkeiten in dieser Verteilung? Eine Verfeinerung der Klassenbreite (Abb. 3) kann hier weitere Einsichten bringen. Die Klassenbreite $b = 2$ € bringt den bereits entdeckten Nebengipfel besser zum Vorschein.

Wir wenden uns nun der Frage nach den Unterschieden zwischen Männern und Frauen zu. Die Software TinkerPlots hält für uns die Möglichkeit bereit, nach dem Merkmal Geschlecht zu unterscheiden, indem wir dieses z. B. auf die Hochachse ziehen.

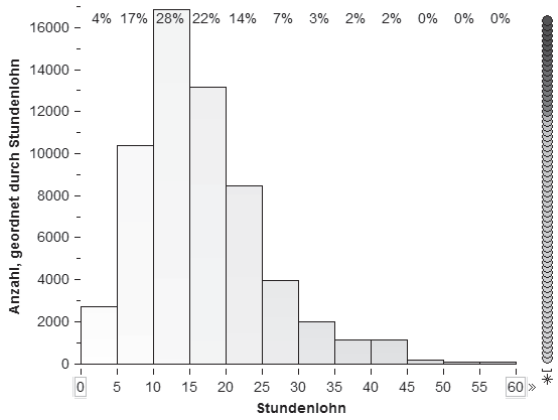


Abb. 2: Histogramm in TinkerPlots, Klassenbreite $b = 5$ €

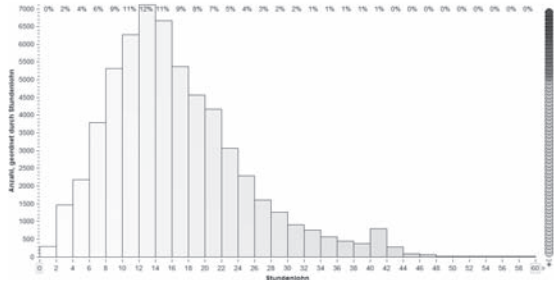


Abb. 3: Histogramm in TinkerPlots, Klassenbreite $b = 2$ €

Die Abbildungen 4 bzw. 5 zeigen uns die beiden Verteilungen des Merkmals „Stundenlohn“ getrennt nach Geschlecht. Als erstes bleibt festzuhalten, dass die Anzahl der Befragten in beiden Gruppen unterschiedlich ist, es sind 26178 Arbeitnehmerinnen und 33258 Arbeitnehmer befragt worden. Wenn einem

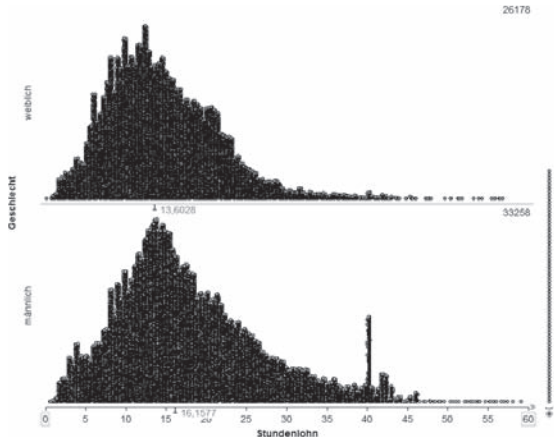


Abb. 4: Verteilungen des Merkmals „Stundenlohn“ getrennt nach dem Merkmal „Geschlecht“ (mit jeweils eingeblendeten Median)

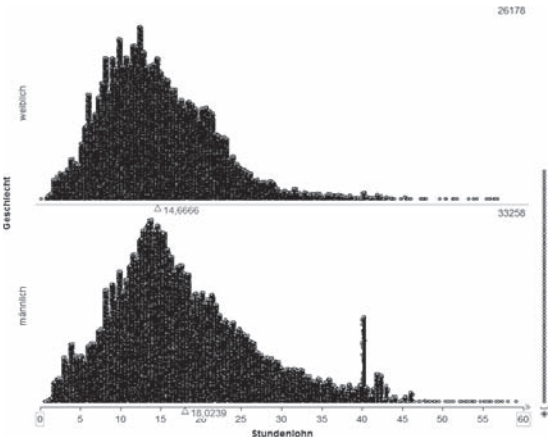


Abb. 5: Verteilungen des Merkmals „Stundenlohn“ getrennt nach dem Merkmal „Geschlecht“ (mit jeweils eingeblendeten arithmetischen Mittelwerten)

die Verteilungen noch nicht so ganz vertraut sind, kann man zunächst Unterschiede zwischen den Mittelwerten anschauen. Die Software TinkerPlots hält hier voreingestellte Funktionen bereit, die das arithmetische Mittel (siehe Abb. 5 – blaues Dreieck) wie auch den Median (siehe Abb. 4 – rotes, umgedrehtes „T“) eines quantitativen Merkmals direkt berechnen.

Man erkennt sehr schön, dass hier die Mediane die „Zentren“ der beiden Verteilungen viel besser als die arithmetischen Mittelwerte wiedergeben. Will man sich für einen Mittelwert bei rechtsschiefen Verteilungen entscheiden, dann wird i. A. der Median empfohlen. Die Nutzung des Medians wird auch in der offiziellen Einkommensstatistik bevorzugt. Betrachten wir die Unterschiede im Median, so sehen wir, dass die Arbeitnehmer im Median ca. 2,55 € mehr verdienen als die Arbeitnehmerinnen. Wir sehen aber, dass die Arbeitnehmer ca. 3,36 € im Durchschnitt pro Stunde mehr verdienen als die Arbeitnehmerinnen. Wie erklärt sich dieser Unterschied? Der Nebengipfel und die relativ vielen Werte am rechten Rand haben eine vergleichsweise größere Auswirkung auf den männlichen Durchschnitt im Vergleich zum Durchschnitt bei den Frauen. Dass beide jeweils größer sind als der Median ergibt sich aus der Schiefe der Verteilungen. Dabei ergeben sich Durchschnittswerte, die leicht von denen unseres Zeitungsartikels abweichen. Diese sind vermutlich durch unseren oben beschriebenen Datenreinigungsprozess (Eliminierung der großen Ausreißer) zu erklären.

Betrachten wir einmal andere Darstellungsformen, z. B. ein Histogramm mit der Klassenbreite $b = 2$ € und betrachten die relativen Häufigkeiten der einzelnen Klassen – wie in Abbildung 6 zu sehen ist. Hier werden unsere bereits bezüglich der Form und der Gipfel in den einzelnen Verteilungen getätigten

Vermutungen aus einer weiteren Perspektive verdeutlicht. Es bleibt anzumerken, dass TinkerPlots die relativen Häufigkeiten (in Form der Höhe der jeweiligen Rechtecke) nicht angemessen umsetzt.

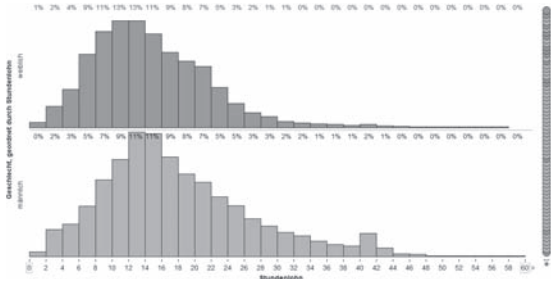


Abb. 6: Verteilungen des Merkmals „Stundenlohn“ getrennt nach dem Merkmal „Geschlecht“

Wir wollen nun die Verteilungen des Merkmals „Stundenlohns“ hinsichtlich des Merkmals „Geschlecht“ anhand ihrer Streumaße vergleichen: Betrachtet man die Spannweite der beiden Verteilungen (z. B. in Abb. 4) so lässt sich dort kaum ein Unterschied feststellen. Allerdings ist die Spannweite (da sie nur vom Maximal- und Minimalwert der jeweiligen Verteilung abhängt) kein wirklich aussagekräftiges Streuungsmaß. Die Streuungs-Aussagen lassen sich präzisieren, wenn man Boxplots als Darstellung (Abb. 7) nimmt. Diese lassen sich in TinkerPlots als vorgefertigte Darstellung per Knopfdruck erzeugen.[12]

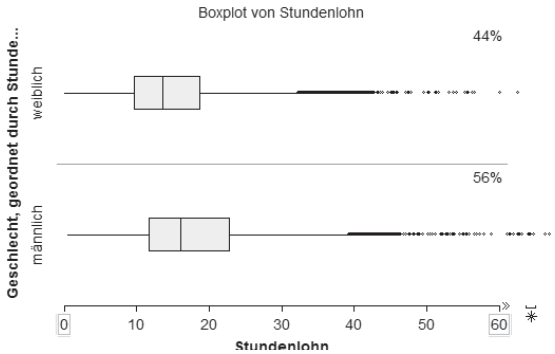


Abb. 7: Boxplots in TinkerPlots

Betrachten wir den Interquartilsabstand: 11,05 € bei der Verteilung des Merkmals „Stundenlohn“ bei den männlichen Arbeitnehmern vs. 9,10 € bei der Verteilung des Merkmals „Stundenlohn“ bei den weiblichen Arbeitnehmern. Wir sehen ein typisches Phänomen: die Streuung wächst mit dem Mittelwert. Bilden wir die Quotienten aus Interquartilabstand und Median, so bekommen wir bei den Männern 0,68 und bei den Frauen 0,67. Das heißt, dass die „relative Streuung“ sogar in etwa gleich groß ist. In der Statistik werden auch andere Koeffizienten (z. B. Standardabweichung durch das arithmetische Mittel) ermittelt, die hier aber nicht so angemessen sind.

Anhand der Boxplots (Abb. 6) können wir nun auch die Art der Verschiebung der Verteilung weiter präzisieren. Die Box liegt bei den Männern weiter rechts als bei den Frauen. Die konstante relative Streuung weist auf eine multiplikative Verschiebung hin. Dieses wird bestätigt, wenn wir die Quotienten der einzelnen Kennwerte, die zwischen 1,19 und 1,22 schwanken (siehe Tab. 1), betrachten, die also praktisch konstant sind. Männer verdienen ca. 20 % mehr als Frauen. Diese Aussage gilt nicht nur für die Mediane, sondern für die ganze Verteilung. Der Quotient der arithmetischen Mittelwerte liegt auch bei diesem Zahlenwert, nämlich bei 1,228.

	Q1	Median	Q3	aMittel
Stundenlohn (m)	11,69	16,18	22,75	18,02
Stundenlohn (w)	9,62	13,60	18,72	14,67
Quotient	**1,22**	**1,19**	**1,21**	**1,22**

Tab. 1: Tabelle „multiplikative Verschiebung“

Mit den Boxplots kann man nun quantil-basierte Vergleiche durchführen (z. B. die ersten Quartile oder die Mediane der beiden Teilgruppen vergleichen). Schülerinnen und Schüler nutzen aber häufig auch den Median der einen Verteilung als Vergleichslinie. Während 50 % der Frauen höchstens 13,60 € Stundenlohn haben, entnehmen wir dem Boxplot, dass es bei den Männern ein Anteil zwischen 25 % und 50 % ist. TinkerPlots hat nun den großen Vorteil, dass man die 13,60 € als Vergleichslinie (Einteiler[13]) einzeichnen kann und die relativen Anteil links davon ermitteln lassen kann, so dass man nicht auf Schätzungen angewiesen ist. Man erkennt in Abbildung 8, dass bei den Männern nur 35 % unterhalb des Medians der Frauen verdienen. So kann man beispielsweise formulieren, dass ca. 50 % der weiblichen Arbeitnehmer weniger als 13,60 € pro Stunde verdienen und bei den männlichen Arbeitnehmern dieser Anteil nur 35 % beträgt. Dass der Median bei den Frauen genau 50 % abteilt, liegt an dem großen Datensatz. Bei kleineren Datensätzen, zumal, wenn sie Bindungen[14] im Median aufweisen, können das auch mehr als 50 % sein.

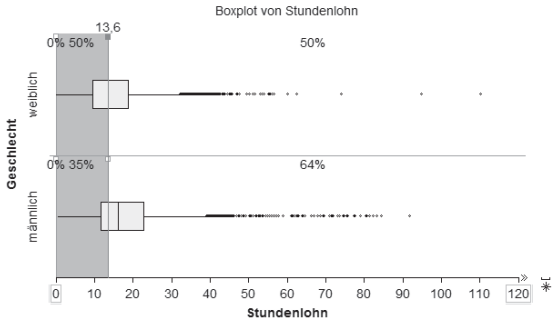


Abb. 8: Boxplots und Divider (Einteiler) in TinkerPlots

Wie man weiterhin anhand der relativen Häufigkeiten im Intervall [13,6 €; 120 €] in beiden Gruppen ablesen kann, verdienen 64 % der Männer 13,60 € pro Stunde oder mehr, bei den Frauen sind das 50 %.

Wie sieht es mit „Vielverdienern“ in den beiden Gruppen aus? Gibt es bei den männlichen Befragten mehr „Vielverdiener“ als bei den weiblichen Befragten? Es stellt sich die Frage: Wie wollen wir einen Vielverdiener festlegen? Wir definieren, das ist natürlich willkürlich: Eine Person, die 30 € oder mehr pro Stunde verdient, ist ein Vielverdiener, analog definieren wir, dass eine Person, die weniger als 8 € brutto pro Stunde verdient als Wenigverdiener gilt. In Abbildung 9 sehen wir, dass der Anteil von solchen Vielverdienern bei den Männern bei 11 %, bei den Frauen aber nur bei 3 % liegt. Wir nennen dies in Anlehnung an Biehler (2001, S. 110) h-basierte Vergleiche (h steht für Häufigkeiten). Es lassen sich auch Fragen in die „umgekehrte“ Richtung stellen: „Wie hoch ist der Stundenlohn bei den oberen 10 % der männlichen Arbeitnehmer mindestens? Hier vergleichen wir an Hand des 90 %-Quantils. Wir sprechen auch von q-basierten Vergleichen. Beide Vergleichsansätze lassen sich mit den Einteilern der Software schön realisieren.

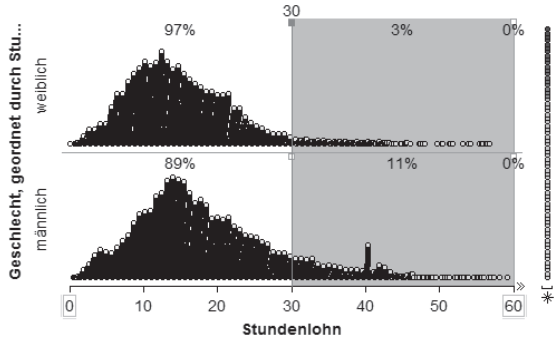


Abb. 9: Unterschiede zwischen den Anteilen an „Vielverdienern“ (h-basierter Vergleich)

Anhand von Abbildung 10 wird Folgendes deutlich: 16 % der Frauen sind Wenigverdiener, bei den Männern sind das gerade einmal 10 %. Im Sinne q-basierter Vergleiche können wir fragen: Wie hoch ist der Stundenlohn bei den oberen 10 % der männlichen Arbeitnehmer mindestens? Wie man der Abbildung 11 entnehmen kann, verdienen die Arbeitnehmer in der Gruppe der oberen 10 % mindestens 30,60 € pro Stunde, bei den weiblichen Arbeitnehmern sind das gerade mal 23 € pro Stunde.

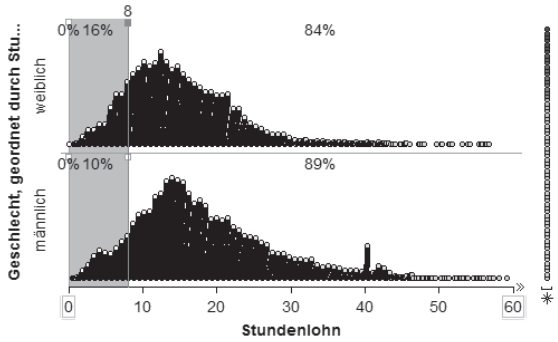


Abb. 10: Unterschiede zwischen den Anteilen an „Wenigverdienern (h-basierter Vergleich)

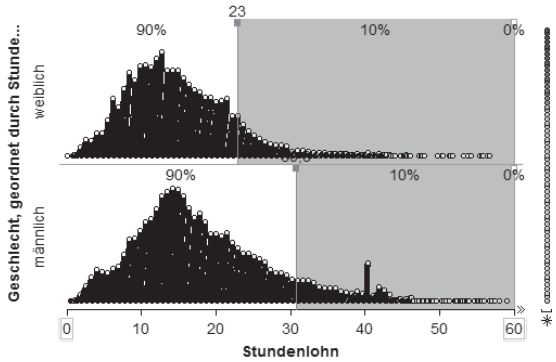


Abb. 11: q-basierter Vergleich in TinkerPlots

Analog können wir auch untersuchen, wie viel die unteren 10 % bei den Arbeitnehmerinnen und Arbeitnehmern jeweils höchstens verdienen. TinkerPlots erlaubt hier anschaulich mit Einteilern beliebige Quantile zu finden und die Bereiche zu markieren. In anderen Programmen geht das oft nur über ein „Quartil-Kommando“, wobei die genaue mathematische Definition der Quartile nicht so ganz einfach ist. Bei kleineren Datensätzen kann man oft nicht jeden Prozentsatz abtrennen (z. B. 13 % bei 40). Das sieht man dann aber direkt beim Arbeiten mit TinkerPlots. Natürlich liefert der Boxplot standardmäßige Vergleichsmöglichkeiten an Hand der Quartile und dem Median. Je nach Kontext können aber andere Quantile vergleichsrelevant sein. Wir haben aufgezeigt, wie sich mit TinkerPlots ein solcher Vergleich anschaulich auf der Basis von Verteilungen durchführen lässt. Es könnten sicherlich noch viele weitere Unterschiede insbesondere auch Unterschiede in Hinblick der verschiedenen Berufsgruppen herausgearbeitet werden – diese findet man z. B. bei Engel (2014). Er thematisiert zwei verschiedene Arten, das Gehaltsgefälle zwischen Männern und Frauen zu unterscheiden: den „raw gender pay gap“ und den „adjusted gender pay gap“. Beim „raw gender pay gap“ werden Unterschiede bezüglich anderer Variablen (wie Ausbildung, Berufserfahrung, etc.) für den Vergleich der Löhne außen vor gelassen. Hingegen werden all diese Faktoren beim „adjusted gender pay gap“ berücksichtigt. In der Folge ist dieser dann auch geringer, allerdings immer noch bei einem Unterschied von 8 % (Engel 2014). Wir wollen unseren „Wanderweg“ an diesem Beispiel nun beenden.

Berufsgruppe	Arbeitnehmer	Arbeitnehmerinnen
Angestellter	11764 (35%)	8479 (32%)
Teilzeitbeschäftigter_ mehr_18h/Woche	1602 (5%)	6964 (27%)
Facharbeiter	8062 (24%)	1099 (4%)
Arbeiter	5779 (17%)	2029 (8%)
Teilzeitbeschäftigter_ Weniger_18h/Woche	1705 (5%)	3299 (13%)
Beamter_Vollzeit	1972 (6%)	1631 (6%)
Azubi	1529 (5%)	1134 (4%)
Beamter_Teilzeit	365 (1%)	1505 (6%)
Meister	504 (2%)	27 (0%)
Heimarbeiter	30 (0%)	25 (0%)

Abb. 12. Verteilung des Merkmals „Geschlecht“ auf die Berufsgruppen (aufsteigend nach Anteil der Berufsgruppe an der Gesamtheit der befragten Arbeitnehmerinnen und Arbeitnehmer)

Im zweiten Teil lassen wir uns nun vom Kontext leiten und nehmen eine „Wunderer“-Haltung ein. Wir fragen („wundern“) uns: wie kann man in den Daten Erklärungsansätze für die festgestellten Unterschiede finden? Oder: wie kann man die in den Abbildungen 4–6 identifizierbaren Gipfel und Häufungen von Teilgruppen erklären? Man könnte vermuten, dass männliche Arbeitnehmer tendenziell „höher gestellte“ Berufe als weibliche Arbeitnehmer haben und deshalb der Stundenlohn höher ist und nicht etwa weil nicht gleicher Lohn für gleiche Arbeit gezahlt würde.

Es sind zwei Phänomene denkbar: In allen Berufsgruppen verdienen Männer im Schnitt mehr als Frauen. In den einzelnen Berufsgruppen gibt es keine Verdienstunterschiede, der unterschiedliche Medianverdienst in der Gesamtgruppe kommt deshalb zustande, weil die Frauen eher in schlechter bezahlten Berufen arbeiten. Es könnte sogar sein, dass die Frauen in jeder Berufsgruppe mehr verdienen als die Männer und trotzdem in der Gesamtgruppe die Männer vorne sind. Dann hätte man ein Beispiel für das Vorliegen von Simpsons Paradoxon (vgl. z. B. Jahnke 1993). Offen ist wie man „Berufsgruppe“ fassen könnte. Wir müssen einen Kompromiss machen und schauen, welche Informationen dazu in unserem Datensatz verfügbar sind, denn wir können keine neuen Daten erhaben. In Abbildung 12 zeigen wir die Verteilung auf die Ausprägungen des im Datensatz vorhandenen Merkmals „Berufsgruppe“.

Sowohl bei den männlichen als auch bei den weiblichen Arbeitnehmern stellt die Berufsgruppe der Angestellten den größten Anteil da: Knapp ein Drittel der männlichen und knapp ein Drittel der weiblichen befragten Arbeitnehmer gehören dieser Berufsgruppe an. Während bei den männlichen Arbeitnehmern die Berufsgruppen „Facharbeiter“ und „Arbeiter“ ca. 41 % ausmachen, beträgt der Anteil bei den weiblichen Beschäftigten gerade einmal 12 %. Eine weitere Diskrepanz wird deutlich, wenn wir die Anteile der „Teilzeit-Arbeiter“ unter den weiblichen und männlichen Arbeitnehmern betrachten: Zirka 40 % der weiblichen Arbeitnehmer gehen einer Teilzeitbeschäftigung (weniger_18h/Woche und mehr_18h/Woche)[15] nach, der Anteil der Teilzeitbeschäftigten bei den männlichen Arbeitnehmern beträgt gerade einmal ca. 10 %. Im Folgenden nehmen wir die Verteilungen des Merkmals Stundenlohn zunächst (Abb. 13) getrennt nach Berufsgruppen ins Visier, ohne nach Geschlecht zu unterscheiden. Um die Komplexität der Darstellung zu minimieren, beschränken wir uns hier nur auf die Boxplots der einzelnen Verteilungen und ordnen sie in TinkerPlots dem Median nach aufsteigend an (Abb. 13).

Im Median verdienen die Beamten, die in Teilzeit arbeiten, am meisten, gefolgt von den Beamten (Vollzeit), den Meistern und Angestellten. Den mit

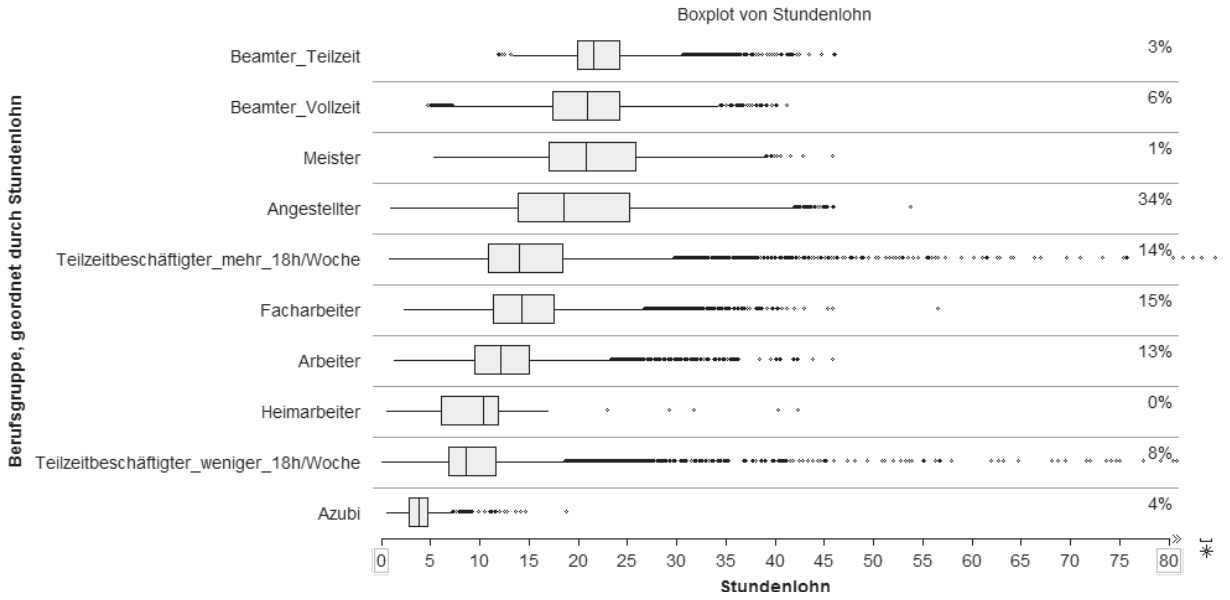
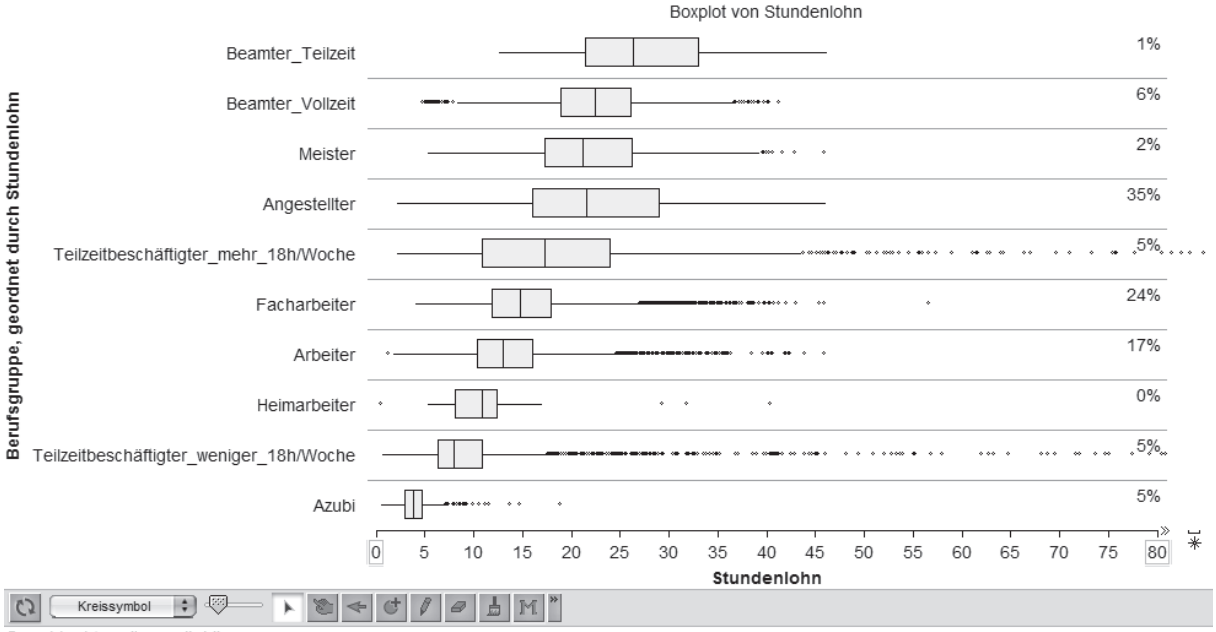


Abb. 13: Verteilungen des Stundenlohns getrennt nach „Stellung im Beruf“

Abstand geringsten Wert im Median haben die Verteilungen des Stundenlohns der Auszubildenden. Man könnte nun hier drei Gruppen von „Stundenlöhnen“ unterscheiden: Beamte (Vollzeit & Teilzeit), Angestellte und Meister, die einen vergleichsweise „hohen“ Stundenlohn haben, Facharbeiter, Teilzeitbeschäftigte_mehr_18h/Woche und Arbeiter deren Stundenlohn man eher als „mittel“ einstufen könnte und Teilzeitbeschäftigte_weniger_18h/Woche und Auszubildende, die eher einen „geringen“ Stundenlohn haben. Dabei bleibt zu bemerken, dass die Gruppe der Angestellten, die deutlich größte Gruppe mit einem Anteil von 34 % an den befragten Arbeitnehmerinnen und Arbeitnehmern darstellt und somit einen großen Einfluss auf die Gesamtverteilung des Merkmals Stundenlohn hat. Dieses ist gleichzeitig auch die Gruppe mit der größten Streuung. Gründe für die Heterogenität dieser Berufsgruppe können vielfältig sein: Prämien, Zugehörigkeit zur Firma, Tariflohn, etc. Betrachten wir nun die Verteilungen des Stundenlohns männlicher und weiblicher Arbeitnehmer getrennt nach „Berufsgruppe“. TinkerPlots bietet hier eine „Filterfunktion“, die es ermöglicht nur bestimmte Teilmengen im Graph anzuzeigen.

Abbildung 14 ist für einen Männer-Frauen – Vergleich nicht so gut geeignet. Dazu wäre es besser,



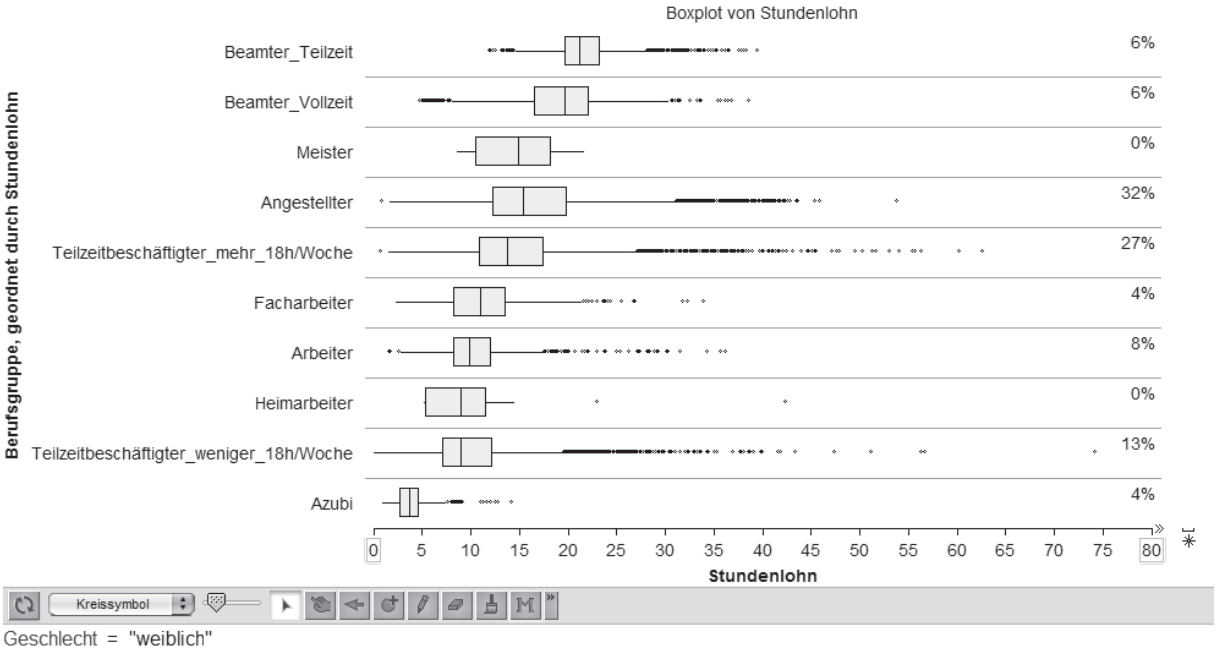


Abb. 14: Verteilungen des Stundenlohns männlicher (oben) und weiblicher (unten) Arbeitnehmer getrennt nach „Berufsgruppe“

es stünden die Männer-Frauen Boxplots in einer Berufsgruppe direkt untereinander. Das aber ist in Tinkerplots nicht realisierbar. Der erste Eindruck zeigt, dass in fast allen Gruppen höhere Medianeinkommen (pro Stunde) der Männer zu finden sind. Das höhere Medianeinkommen der Männer in der Gesamtverteilung erklärt sich vor allem durch diese Unterschiede in fast jeder Gruppe.

Eine weitere interessante Beobachtung lässt sich anhand der unterschiedlichen Streuungen in den einzelnen Gruppen machen. Die Gruppe der Teilzeitbeschäftigten, die weniger als 18 Stunden pro Woche arbeiten, scheint sehr homogen zu sein – hier gibt es keine großen Unterschiede, was z. B. daran liegen kann, dass viele dieser Personen einem 400 €-Job nachgehen. Ebenso ist das Verhalten bei den Auszubildenden sehr homogen, welches auf ähnliche Gründe (festgelegtes Gehalt in den einzelnen Lehrjahren) zurückzuführen ist. Generell lassen sich hier die Verteilungen sehr schön anhand der „Zentrums-Streuungs" – Interpretation (Biehler 2001, S. 108), die die Boxplots bieten, vergleichen.

Bei den weiblichen Arbeitnehmern (Abb. 14 unten) ist es auch die Berufsgruppe der Beamten, die im Median am meisten verdienen. Im Gegensatz zu den männlichen Arbeitnehmern ist diese Gruppe deutlich homogener, was durch die schmale Box und Antennen verdeutlicht wird. Weiterhin gehören Meister (ebenfalls sehr geringer Anteil), Angestellte und Teilzeitbeschäftigte zu den Berufsgruppen mit vergleichsweise höheren Stundenlöhnen bei den weiblichen Arbeitnehmern.

Graphik 13 und 14 sind noch nicht günstig eingerichtet, um Verdienstunterschiede in den Berufsgruppen untersuchen zu können. Es hilft eine „Datenreduktion" auf die Differenz (Abb. 15a) und den Quotienten (Abb. 15b) der Mediane der jeweiligen Berufsgruppen.

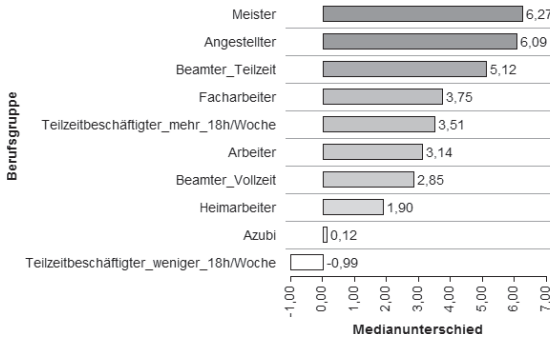


Abb. 15a: Differenz der Mediane der Stundenlöhne der männlichen und weiblichen Arbeitnehmer aufgeschlüsselt nach Berufsgruppen

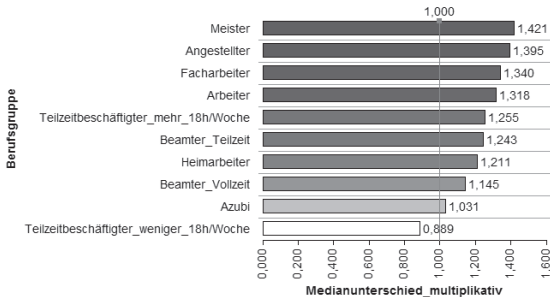


Abb. 15b: Quotient der Mediane der Stundenlöhne der männlichen und weiblichen Arbeitnehmer aufgeschlüsselt nach Berufsgruppen

Die Abbildungen 15a und 15b zeigen bemerkenswerte Unterschiede. Der multiplikative Vergleich passt zu der Zeitungsmeldung und differenziert diese aus. Wir schauen zwei Gruppen aufgrund ihrer großen Anteile an der Gesamtarbeitnehmerschaft genauer an: die Angestellten (Abb. 16) und die Teilzeitbeschäftigten (Abb. 17).

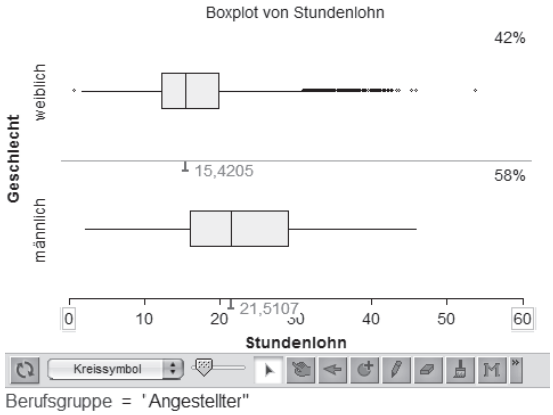


Abb. 16: Verteilungen des Bruttomonatsgehaltes weiblicher (oben) und männlicher (unten) Angestellter

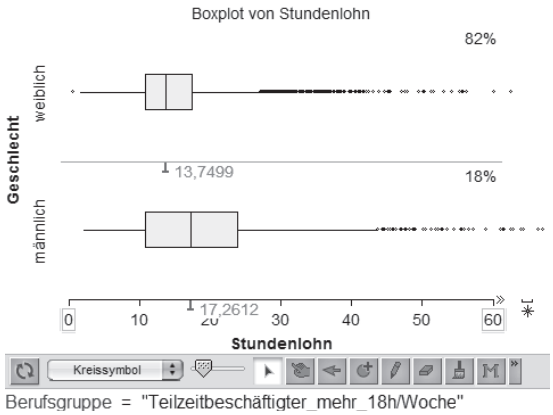


Abb. 17: Verteilungen des Bruttomonatsgehaltes weiblicher (oben) und männlicher (unten) Teilzeitbeschäftigter

Der Unterschied ist in der Gruppe der Angestellten am größten und für die Gesamtverteilung von besonderer Bedeutung: zum einen verdienen die männli-

chen Angestellten im Median 6,09 € pro Stunde mehr (bzw. ca. das 1,4fache) als die weiblichen, zum anderen macht diese Teilgruppe einen großen Anteil an der Gesamtgruppe der gefragten männlichen (35 %) und weiblichen (32 %) Arbeitnehmer aus (siehe Abb. 12), so dass hiervon die Gesamtverteilung stark beeinflusst wird. Auch hier lassen sich Unterschiede, betrachtet man die Tabelle 2, approximativ als eine multiplikative Verschiebung beschreiben.

	Q1	Median	Q3	aMittel
Stundenlohn (m)	16,05	21,51	29,03	23,12
Stundenlohn (w)	12,31	15,42	19,86	16,68
Quotient	1,30	1,39	1,46	1,38

Tab. 2: Tabelle „multiplikative Verschiebung“ der Verteilungen des Merkmals „Stundenlohn“ bei den Angestellten

Einen weiteren möglichen Grund für den geschlechterspezifischen Gehaltsunterschied stellen die Anteile der Teilzeitbeschäftigten, deren Gehaltsklasse niedrig ist, dar. Während von den männlichen Arbeitnehmern gerade einmal 5 % einer Teilzeitbeschäftigung im Umfang von 18 oder mehr Stunden pro Woche nachgehen, sind es bei den weiblichen Arbeitnehmern 27 %.

Bei den Teilzeitbeschäftigten, die unter 18 Stunden pro Woche arbeiten, ergibt sich ein ähnliches Bild: 13 % der weiblichen Arbeitnehmer stehen dort 5 % der männlichen Arbeitnehmer entgegen. Dieses bewirkt auch die lokale Häufung der niedrigen Gehälter in der Gesamtverteilung des Merkmals „Stundenlohns“ bei den weiblichen Arbeitnehmern.

5 Fazit

In diesem Artikel haben wir exemplarisch einen differenzierten Verteilungsvergleich durchgeführt und demonstriert, wie man diesen mit TinkerPlots umsetzen kann. Dabei haben wir zwei Vorgehensweisen unterschieden: Zum einen haben wir einen Wanderweg (entlang der Fragestellung „Inwiefern unterscheiden sich die Arbeitnehmerinnen und Arbeitnehmer hinsichtlich ihres Stundenlohns?“) durch einen Verteilungsvergleich mit TinkerPlots eingeschlagen und zum anderen haben wir uns vom Kontext leiten lassen und weiterführende Untersuchungen in den Daten angestrebt sowie Ansätze gesucht, die das Ungleichgewicht des Merkmals Stundenlohn von weiblichen und männlichen Arbeitnehmern erklären. Weitere interessante Merkmale im Datensatz wie „Ausbildung“, „Region“, usw. laden zu weiterführenden Explorationen mit der Software ein.

Anmerkungen

1 http://www.mathematik.uni-dortmund.de/ak-stoch/stellung.html (aufgerufen am 27.9.2013)

2 http://www.forschungsdatenzentrum.de/campus-file.asp (aufgerufen am 26.4.2014)

3 Dieser (aus einer nicht-repräsentativen Umfrage entstandene) Datensatz enthält über 50 Variablen zum Freizeitverhalten und Medienkonsum von 538 Schülerinnen und Schülern aus Nordrhein-Westfalen aus dem Jahr 2000.

4 An dieser Stelle könnte man auch auf die Feinheiten bei der Generierung von statistischen Fragestellungen/Hypothesen eingehen. Dieses soll an dieser Stelle allerdings nicht weiter ausgeführt werden und wird ausführlich in Biehler (2001, S. 98) beschrieben.

5 Man kann sich zum Beispiel vorstellen beim Computer-Verhalten zwischen Wenig- und Vielnutzern zu unterscheiden, diese geeignet zu definieren und dann zu vergleichen. Für eine differenziertere Darstellung dieser Vergleichsmöglichkeit („h-basiert“) siehe auch Biehler (2001, S. 110).

6 Hier kann man sich beispielsweise fragen wie viel die oberen q % einer Verteilung mindestens verdienen. Für eine differenziertere Darstellung dieser Vergleichsmöglichkeit („q-basiert“) siehe auch Biehler (2001, S. 110).

7 Informationen zur Arbeit mit TinkerPlots im deutschsprachigen Raum finden sich hier: http://lama.uni-paderborn.de/personen/rolf-biehler/projekte/tinkerplots.html (aufgerufen am: 19.3.2014)

8 Zitat (Hessisches Statistisches Landesamt, Hans-Peter Hafner, 17.11.2009)

9 http://www.handelsblatt.com/politik/deutschland/aktuelle-statistik-frauen-liegen-beim-gehalt-deutlich-zurueck/3449220.html (aufgerufen am 30.4.2014)

10 Die Grenzen für die Ausreißer sind dabei wie folgt definiert: $f_u = Q_1 - 1{,}5 \cdot (Q_3 - Q_1)$ sowie $f_o = Q_3 + 1{,}5 \cdot (Q_3 - Q_1)$.

11 Es bleibt hier nochmals anzumerken, dass es in TinkerPlots – bis auf wenige Ausnahmen (wie z. B. Boxplots) – keine vorgefertigten und „auf Knopfdruck fertigen“ Darstellungen gibt, sondern – wie in Abschnitt 3 angedeutet – diese sukzessive durch wiederholte Operationen erstellt werden müssen. Auf diesen Prozess wollen wir hier allerdings nicht näher eingehen. Eine ausführliche Erläuterung findet sich in Biehler & Frischemeier (2013).

12 Ein Vorteil der Software bei der Darstellung von Boxplots ist, dass auch die Datenpunkte unterhalb des Boxplots sichtbar bleiben können. Dieses kann helfen die Boxplots zu lesen und zu vergleichen sowie Bezüge herzustellen und Fehlinterpretationen zu vermeiden. Andererseits kann es für einen Vergleich auch sinnvoll sein, die Daten auszublenden.

13 Einteiler bieten in TinkerPlots die Möglichkeit absolute und relative Häufigkeiten in freiwählbaren Intervallen zu bestimmen.

14 Wenn in einem Datensatz Werte mehrfach vorkommen spricht man von „ties“ (dt. „Bindungen“)

15 Es bleibt hier zu bemerken, dass die Einteilung auf die Berufsgruppen eine Partition auf die Gesamtmenge darstellt. Definition (Variablenliste): Teilzeitbeschäftigte sind Arbeitnehmer/innen, deren Arbeitszeit aufgrund eines Arbeitsvertrages unter der betrieblichen Arbeitszeit liegt.

Danksagung

Wir danken Joachim Engel für die vielen hilfreichen Anmerkungen und Kommentare bei der Entstehung und Überarbeitung dieses Artikels.

Zusatzmaterial

Die Daten sind in verschiedenen Formaten auf der Website erhältlich (Tinkerplots, Fathom, EXCEL).

Literatur

Biehler, R. (2001). Statistische Kompetenz von Schülerinnen und Schülern – Konzepte und Ergebnisse empirischer Studien am Beispiel des Vergleichens von statistischen Verteilungen. In: Borovcnik, M., Engel, J., Wickmann, D. (Hrsg.) *Anregungen zum Stochastikunterricht.* Hildesheim: Franzbecker 2001, 97–114.

Biehler, R., Kombrink, K., & Schweynoch, S. (2003). MUFFINS – Statistik mit komplexen Datensätzen – Freizeitgestaltung und Mediennutzung von Jugendlichen. *Stochastik in der Schule, 23*(1), 11–25.

Biehler, R. (2007a). Arbeitsumgebungen zur Entwicklung von Datenkompetenz ab Klasse 1 – Das Potential der Software TinkerPlots. In: *Beiträge zum Mathematikunterricht 2007.* Hildesheim: Franzbecker.

Biehler, R. (2007b). TINKERPLOTS: Eine Software zur Förderung der Datenkompetenz in Primar- und früher Sekundarstufe. *Stochastik in der Schule, 27*(3), 34–42.

Biehler, R. (2007c). Denken in Verteilungen – Vergleichen von Verteilungen. *Der Mathematikunterricht, 53*(3), 3–11.

Biehler, R., Ben-Zvi, D., Bakker, A. & Makar, K. (2013). Technology for Enhancing Statistical Reasoning at the School Level. In: K. Clements, A. Bishop, C. Keitel, J. Kilpatrick & F. Leung (Hrsg). *Third International Handbook of Mathematics Education.* New York: Springer 2013.

Biehler, R.; Frischemeier, D. & Podworny, S. (2013). TinkerPlots 2.0 – von realen Handlungen über Computersimulationen zum stochastischen Denken. In G. Greefrath, F. Käpnick, & M. Stein: *Beiträge zum Mathematikunterricht 2013,* WTM Verlag, Münster, 144–147.

Biehler, R. & Frischemeier, D. (2013). Spielerisches Erlernen von Datenanalyse – Von Datenkarten und lebendiger Statistik zur Software TinkerPlots. Ein Workshop im Rahmen einer Lehrerfortbildung für die Primarstufe. *Stochastik in der Schule, 33*(3), 2–9.

Engel, J. (2014). Open data, civil society and monitoring progress: challenges for statistics education. *Submitted for Proceedings of ICOTS-9.*

Frischemeier, D. (2014). Comparing groups by using TinkerPlots as part of a data analysis task – Tertiary students' strategies and difficulties. In: K. Makar, B. de Sousa & R. Gould (Eds.), Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Frischemeier, D. & Biehler, R. (2011). Spielerisches Erlernen von Datenanalyse mit der Software TinkerPlots – Ergebnisse einer Pilotstudie. In R. Haug & L. Holzapfel (Hrsg.), *Beiträge zum Mathematikunterricht.* Münster: WTM, 275–278.

Jahnke, T. (1993). Das Simpsonsche Paradoxon verstehen – ein Beitrag des Mathematikunterrichts zur Allgemeinbildung. *Journal für Mathematik-Didaktik, 14*(3/4), 221–242.

Garfield, J. & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice.* Berlin: Springer.

Konold, C. & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W.G. Martin & D. E. Schifter (Eds.). A research companionn to principles and standards for school mathematics Reston, VA: NCTM, 193–215.

Konold, C. & Miller, C. (2011). TinkerPlots TM Version 2 [computer software]. Emeryville, CA: Key Curriculum Press.

Krüger, K. (2012). Haushaltsnettoeinkommen – ein Vorschlag zur Nutzung der GENESIS-Online Datenbank im Unterricht. *Stochastik in der Schule, 32*(3), 8–14.

Makar, K. & Confrey, J. (2014). Wondering, Wandering or Unwavering? Learners' Statistical Investigations with Fathom. In: Wassong, T.; Frischemeier, D.; Fischer, P. R.; Hochmuth, R.; Bender, P. (Eds.): *Mit Werkzeugen Mathematik und Stochastik lernen – Using Tools for Learning Mathematics and Statistics.* Wiesbaden: Springer Spektrum.

Anschrift der Verfasser

Rolf Biehler
Universität Paderborn
Institut für Mathematik
Warburger Straße 100
33098 Paderborn
biehler@math.upb.de

Daniel Frischemeier
Universität Paderborn
Institut für Mathematik
Warburger Straße 100
33098 Paderborn
dafr@math.upb.de