Ich wollte einige Dinge klären, die mir noch unklar sind:

1. Unterschied Transformation/Kodierung: Kann man sagen, dass Transformation die Verwandlung von Merkmalen ist während die Kodierung die Verschlüsselung ist?

2. Wenn man mal rel. Häufigkeiten zu berechnen hat, muss man die IMMER auf 3 Nachkommastellen runden oder ist das unerheblich?

Antwort:

1. Bei der Kodierung werden den Merkmalsausprägungen Zahlen zugeordnet, welche die entsprechende Ausprägung repräsentieren. Da es nicht möglich ist, mit Zeichenketten zu rechen, kann zum Beispiel { "JA", "NEIN"} als {1, 0} kodiert werden (Vgl. Folie 1.32 und S. 14-15 im Buch "Deskriptive Statistik"von Toutenburg et al.).

   Bei einer Transformation werden die Ausprägungen eines Merkmals mit Hilfe einer Zuordnungsvorschrift auf neue Ausprägungen des gleichen oder eines anderen Merkmals übertragen (Vgl. Folie 1.33).

   Auf Folie 1.34 findet man zwei Beispiele für Transformationen von metrischen Merkmalen: die Temperaturumrechnung von °F in °C und die Umrechnung von $ in €.

   Ein Beispiel für eine Transformation eines nominalskalierten Merkmales findet man auf S. 17 vom Buch "Deskriptive Statistik".

2. Die relativen Häufigkeiten geben den Anteil der Untersuchungseinheiten in der Erhebung an, die die Ausprägung a_j besitzen.

   $$f_j = f(a_j) = \frac{n_j}{n}, \quad j = 1, \ldots, k$$

   Bei der Darstellung dieses Anteils als Dezimalzahl werden in der Vorlesung und in den Übungen normalerweise drei Nachkommastellen verwendet.

Könnten Sie mir bitte noch einmal die Bedeutung der Eigenschaft „translationsäquivariant“ erklären? Wenn möglich auch mit einem Beispiel.

Antwort:

Die Eigenschaft der Translationsäquivarianz wurde im Zusammenhang mit den Lagemaßen eingeführt und besprochen. Der Modus, der Median und das arithmetische Mittel sind translationsäquivariant. Für eine Lineartransformation der Daten, d.h., eine Transformation der Form $y_i = a + bx_i$ mit a,b beliebige reelle Zahlen, soll gelten

$$L(y_1, \ldots, y_n) = a + b \cdot L(x_1, \ldots, x_n),$$

wobei mit $L(\cdot)$ der jeweilige Lageparameter bezeichnet wird. (Vgl. Toutenburg et a. S. 43) Das bedeutet, dass z.B. das arithmetische Mittel der linear transformierten Werte gleich der linearen Transformation des arithmetischen Mittels der ursprünglichen Werte ist.

Wir betrachten die Merkmale:

X: Temperatur in Celsius

Y: Temperatur in Fahrenheit

Es gilt: $y_i = a + bx_i$, mit $a = 32$ und $b = 1{,}8$

Nun sei folgender Datensatz gegeben:

°C	°F
35	95,0
37	98,6
37	98,6
40	104,0

Modus, Median und arithmetisches Mittel ergeben sich zu:

$$\bar{x}_M = 37 \qquad \bar{y}_M = 98{,}6$$

$$\tilde{x}_{0,5} = 37 \qquad \tilde{y}_{0,5} = 98{,}6$$

$$\bar{x} = 37{,}25 = \frac{35 + 37 + 37 + 40}{4} \qquad \bar{y} = 99{,}05 = \frac{95 + 98{,}6 + 98{,}6 + 104}{4}$$

Aufgrund der Translationsäquivarianz gilt:

$$\bar{y}_M = 98{,}6 = 32 + 1{,}8 \cdot 37 \quad = 32 + 1{,}8 \cdot \bar{x}_M$$

$$\tilde{y}_{0,5} = 98{,}6 = 32 + 1{,}8 \cdot 37 \quad = 32 + 1{,}8 \cdot \tilde{x}_{0,5}$$

$$\bar{y} \quad = 99{,}05 = 32 + 1{,}8 \cdot 37{,}25 = 32 + 1{,}8 \cdot \bar{x}$$

$\Rightarrow$ Es ist also egal, ob ich die Daten transformiere und aus den transformierten Daten den Mittelwert (Median, Modus) berechne oder ob ich den Mittelwert (Median, Modus) der ursprünglichen Daten transformiere. Ich erhalte jeweils dasselbe Ergebnis.

Auch die Quantile sind translationsäquivariant (Vgl. Folie 3.12). Es gilt also:

$$\tilde{y}_\alpha = 32 + 1{,}8 \cdot \tilde{x}_\alpha$$

Herr Professor Heumann hat in der Vorlesung das Bsp. mit dem Median gebracht, wenn $n = 4$, also gerade ist. Die Werte waren 3,4,5 und 8. Wieso ist der Median 4,5? Ich wende die Formel an und erhalte ein anderes Ergebnis:

$$\begin{aligned} \tilde{x}_{0,5} &= \frac{1}{2}(x_{(2)} + x_{(3)}) \\ \tilde{x}_{0,5} &= \frac{1}{2}(x_{(5)}) \\ \tilde{x}_{0,5} &= 2{,}5 \end{aligned}$$

Antwort:

Falls n gerade ist, wird der Median mit Hilfe der folgenden Formel (FS S. 2) berechnet:

$$\tilde{x}_{0,5} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$$

Wir betrachten die geordneten Daten:

$$x_{(1)} = 3,\ x_{(2)} = 4,\ x_{(3)} = 5,\ x_{(4)} = 8$$

Durch Anwendung der Formel erhalten wir den Median:

$$\begin{aligned} \tilde{x}_{0,5} &= \frac{1}{2}(x_{(2)} + x_{(3)}) \\ \tilde{x}_{0,5} &= \frac{1}{2}(4 + 5) \\ \tilde{x}_{0,5} &= 4{,}5 \end{aligned}$$

$x_{(2)}$ bezeichnet die zweite Beobachtung in der geordneten Stichprobe und $x_{(3)}$ die dritte (Siehe FS S. 1). Es gilt also $x_{(2)} = 4$ und $x_{(3)} = 5$. Die Zahlen in den runden Klammern bezeichnen die Positionen der Werte in der geordneten Stichprobe, sind also Indizes. Um den Median zu berechnen, muss man die entsprechenden Werte einsetzen.

Wieso ist auf der Folie 18 bei A der „obere Whisker“ bei 9 (kein Ausreißer) und bei B wird die 12 als Ausreißer gesehen? Ich muss ehrlich zugeben, dass ich A genauso gezeichnet hätte aber die 12 ist in meinen Augen eigentlich eher ein Extremwert. Wo ist der Unterschied zwischen Extremwert und Ausreißer?

Antwort:

Auf Folie 3.16 wird zwischen Ausreißer und Extremwert unterschieden:

- Ausreißer sind Werte, die mehr als 1,5 mal dem Quartilsabstand von einem der beiden Quartile entfernt liegen.
- Extremwerte sind Werte, die mehr als 3 mal dem Quartilsabstand von einem der beiden Quartile entfernt liegen.

Für das Beispiel auf Folie 3.18 ergibt sich:

$$d_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25} = 7 - 5 = 2$$

$$\Rightarrow 1{,}5 \cdot 2 = 3 \text{ und } 3 \cdot 2 = 6$$

Nun muss also geprüft werden, ob 12 größer als 7 + 3 ($\rightarrow$ Ausreißer) oder größer als 7 + 6 ($\rightarrow$ Extremwert) ist!

Somit gilt:

$$7 + 3 = 10 < 12 < 13 = 7 + 6$$

Die 12 ist also ein Ausreißer, kein Extremwert.

Ist das Merkmal „Alter“ intervallskaliert? Welche Regel gibt es, um zu sagen, ob man die Merkmalsausprägungen ins Verhältnis setzen darf bzw. ob sie einen natürlichen Nullpunkt haben (also ob sie verhältnis- oder intervallskaliert sind)?

Antwort:

Das Merkmal „Alter“ ist ein verhältnisskaliertes Merkmal. Das Alter kann in Tagen, Monaten oder Jahren gemessen werden. Der natürliche Nullpunkt ist die Geburt der Person (Siehe auch Übungsblatt 1, Aufgabe 3d). Quotienten (Verhältnisse) sind interpretierbar: „Person A ist doppelt so alt wie Person B“ ist eine sinnvolle Aussage.

Auch die Körpergröße ist ein verhältnisskaliertes Merkmal. Diese besitzt einen natürlichen Nullpunkt, nämlich 0 cm. (Vorlesung vom 07.10. Videoaufzeichnung zum Thema „Elementare Begriffe“ 63:20-64:44).

Ein gutes Beispiel für den Unterschied zwischen Intervall- und Verhältnisskala ist das Merkmal „Temperatur“. Wenn die Temperatur in Kelvin gemessen wird, handelt es sich um ein verhältnisskaliertes Merkmal, denn der Nullpunkt dieser Skala ist ein natürlicher Nullpunkt. Wird die Temperatur dagegen in Grad Celsius gemessen, handelt es sich um ein intervallskaliertes Merkmal, denn der Nullpunkt auf dieser Skala ist kein natürlicher, sondern ein willkürlich festgelegter Nullpunkt. Hier können Verhältnisse nicht interpretiert werden sondern lediglich Differenzen, 4°C ist z.B. nicht doppelt so warm wie 2°C sondern lediglich um 2°C wärmer.

Skript Kap.1 S.10: Für die Nominalskala sind nur eindeutige Transformationen zulässig. Was bedeutet hier eindeutig?

Antwort:

Bei dieser Art von Transformation wird jeder Merkmalsausprägung eines Merkmals eine beliebige Zahl zugeordnet. Wir nehmen z.B. an, das Merkmal „Geschlecht“ sei wie folgt kodiert: weiblich = 0, männlich = 1.

Eindeutige Transformationen wären:

- weiblich = 3, männlich = 4.
- weiblich = 20, männlich = 5.

Das heisst also, dass auch wenn die Zahlenzuordnung willkürlich ist, diese dennoch eine eindeutige Unterscheidung der Merkmalsausprägungen ermöglicht (Videoaufzeichung vom 14.10. 7:00-7:44).

Ein weiteres Beispiel zu diesem Thema findet man im Buch von Toutenburg auf Seite 17.

Kap.2 S.2: Was bedeutet „Indikatorfunktion“ genau und wie wende ich die Formel an?

Antwort:

Die Funktion $I_T : X \to \{0,1\}$ für eine Teilmenge $T \subseteq X$ definiert durch

$$I_T(x) = \begin{cases} 1 & \text{falls } x \in T \\ 0 & \text{sonst} \end{cases}$$

heisst Indikatorfunktion.

Anwendungsbeispiele

- Beispiel 1: Sei die Menge $X = \{A, B, C, D, E\}$, dann ergibt sich für die Teilmenge $T = \{B, D, E\}$

$$I_T(A) = 0$$
$$I_T(B) = 1$$
$$I_T(C) = 0$$
$$I_T(D) = 1$$
$$I_T(E) = 1$$

- Beispiel 2: Sei die Menge $X = \{A, B, C, D, E\}$, dann ergibt sich für die Teilmenge $W = \{B\}$:

$$I_W(A) = 0$$
$$I_W(B) = 1$$
$$I_W(C) = 0$$
$$I_W(D) = 0$$
$$I_W(E) = 0$$

Ein Beispiel für die Anwendung der Indikatorfunktion bei der Bestimmung von absoluten Häufigkeiten wurde in der Vorlesung besprochen (Siehe Folie 2.7 bzw. Antwort auf Frage 9).

Kap.2 S.3: Wie ist die Tabelle auszufüllen / zu verstehen bzw. warum sind die Zwischensummen immer Null (müsste doch eigentlich für blau, zwei für gelb und eins für grün ergeben?!)

Antwort:

Auf Seite 3 findet sich die Ausgangstabelle. Bevor man mit der Bestimmung der absoluten Häufigkeiten beginnt, sind alle Zwischensummen gleich Null.

Für dieses Beispiel wurden drei Indikatorfunktionen definiert, nämlich:

$$I_{\{\text{blau}\}}(x_i) = \begin{cases} 1 & \text{falls } x_i \in \{\text{blau}\} \\ 0 & \text{sonst} \end{cases}$$

$$I_{\{\text{gelb}\}}(x_i) = \begin{cases} 1 & \text{falls } x_i \in \{\text{gelb}\} \\ 0 & \text{sonst} \end{cases}$$

$$I_{\{\text{grün}\}}(x_i) = \begin{cases} 1 & \text{falls } x_i \in \{\text{grün}\} \\ 0 & \text{sonst} \end{cases}$$

Nun betrachten wir die Beobachtung $x_1 =$ blau. Es gilt:

$I_{\{\text{blau}\}}(x_1) = 1$

$I_{\{\text{gelb}\}}(x_1) = 0$

$I_{\{\text{grün}\}}(x_1) = 0$

$\Rightarrow$

Ausprägung	blau	gelb	grün
Zwischensumme	1	0	0

Die nächste Beobachtung ist $x_2 =$ gelb. Es gilt:

$I_{\{\text{blau}\}}(x_2) = 0$

$I_{\{\text{gelb}\}}(x_2) = 1$

$I_{\{\text{grün}\}}(x_2) = 0$

$\Rightarrow$

Ausprägung	blau	gelb	grün
Zwischensumme	1	1	0

Mit Hilfe der Indikatorfunktionen werden die absoluten Häufigkeiten der Merkmalsausprägungen ermittelt.

Wie kann e_{j-1} die untere Klassengrenze sein? Dann wäre ja jede Klasse nur eine Einheit breit?

Antwort:

Wir betrachten klassierte Daten (bei stetigen Merkmalen), z.B.:

$$| \underbrace{x_{(1)}, x_{(2)}, x_{(3)}}_{K_1} | \underbrace{x_{(4)}, x_{(5)}}_{K_2} | x_{(6)}, \ \ldots \ | \underbrace{\ldots, x_{(n)}}_{K_k} |$$

k	Anzahl der Klassen
K_j	j-te Klasse
e_{j-1}	untere Grenze der j-ten Klasse
e_j	obere Grenze der j-ten Klasse
$d_j = e_j - e_{j-1}$	Klassenbreite der j-ten Klasse

Für $j = 1$ ist $e_{j-1} = e_{1-1} = e_0$ die untere Grenze der ersten Klasse.

Für $j = 2$ ist $e_{j-1} = e_{2-1} = e_1$ die untere Grenze der zweiten Klasse.

Die Klassenbreite der ersten Klasse ist $d_1 = e_1 - e_0$.

Die Klassenbreite der zweiten Klasse ist $d_2 = e_2 - e_1$.

Beispiel aus Übungsblatt 2, Aufgabe 2:

$e_0 = 1$

$e_1 = 2{,}0$

$e_2 = 2{,}4$

$d_1 = e_1 - e_0 = 2{,}0 - 1{,}0 = 1$

$d_2 = e_2 - e_1 = 2{,}4 - 2{,}0 = 0{,}4$

Kap.2 S.8: Warum bildet man in der Formel für die relativen Häufigkeiten in Histogrammen nicht das Produkt aus Breite mal Höhe sondern das Kreuzprodukt?

Antwort:

Das Symbol „$\times$“ auf Seite 8 bezeichnet nicht das Kreuzprodukt, sondern das Produkt zweier reeller Zahlen.

Was sollen die beiden Diagrammbeispiele auf S.8 zeigen? Ist die Aussage, dass man die Gleichverteilung nicht erkennen kann, was impliziert, dass es ein schlecht gemachtes Diagramm ist?

Antwort:

Das Aussehen von Histogrammen hängt wesentlich davon ab, wie man die Anzahl der Klassen und die Klassenbreite wählt. Die Beispiele auf Seite 8 sollen diesen Sachverhalt veranschaulichen.
Im linken Diagramm wurden für die Darstellung der Daten zehn Klassen mit einer Klassenbreite von jeweils 0,1 gewählt.
Für das rechte Diagramm wurden drei Klassen gewählt. Bei der ersten und zweiten Klasse beträgt die Klassenbreite 0,3. Die dritte Klasse hat eine Breite von 0,4.
Der optische Eindruck der Datenverteilung ist bei beiden Diagrammen unterschiedlich, obwohl ihnen dieselben Daten zugrunde liegen.
Man kann die Gleichverteilung nicht erkennen, weil nur 100 Zufallszahlen gezogen wurden (Videoaufzeichnung vom 14.10. 50:00 - 50:35).

Kap.2 S.11: Was ist ein Polygonzug?

Antwort:

Ein Polygonzug ist die Vereinigung der Verbindungsstrecken einer Folge von Punkten. Es handelt sich also um eine stückweise lineare Funktion. In der Abbildung auf Seite 12 werden die Punkte $(e_j; F(e_j))$ durch einen Polygonzug verbunden (vgl. Videoaufzeichnung vom 21.10. 18:40 -19:05).

Frage:

Warum wird bei der Erstellung empirischer Verteilungsfunktionen stetiger Merkmale zwischen Originaldaten und klassierten Daten unterschieden?

Antwort:
Je nachdem, ob Originaldaten oder klassierte Daten vorliegen, geht man bei der Bestimmung der empirischen Verteilungsfunktion anders vor (Folie 2.33).

Bei Originaldaten wird zu jeder beobachteten Merkmalsausprägung x_i der Wert $F(x_i)$ gemäß der Formel auf Seite 1 der Formelsammlung berechnet. Die Wertepaare $(x_i; F(x_i))$ werden dann durch einen Polygonzug verbunden.

Bei klassierten Daten wird innerhalb der Klassen eine Gleichverteilung der Merkmalsausprägungen angenommen. Die empirische Verteilungsfunktion ist damit innerhalb Klasse eine Diagonale, die die Punkte $(e_{j-1}; F(e_{j-1}))$ und $(e_j; F(e_j))$ verbindet.

Im Beispiel 2.2.2 (Toutenburg S. 27-28) werden anhand eines Datensatzes beide Fälle (Originaldaten und klassierte Daten behandelt).

Bemerkungen:

- Bei stetigen Merkmalen ist die Anzahl der beobachteten Merkmalsausprägungen oft sehr groß, manchmal sogar gleich der Anzahl der Beobachtungen. Die daraus entstehende Häufigkeitsverteilung besitzt daher nur geringe Aussagekraft. Um eine interpretierbare Verteilung zu erhalten, werden mehrere Merkmalsausprägungen zu einer Klasse zusammengefasst (vgl. Toutenburg S. 23).
- Es kann auch der Fall eintreten, dass die Originalwerte eines stetigen Merkmals nicht mehr vorliegen und man daher nur die klassierten Daten zur Verfügung hat.

Frage:

Ist e_0 die niedrigste Klasse und e_k die höchste?

Antwort:

e_0 ist die untere Grenze der ersten Klasse (siehe Antwort auf Frage 10).

e_k ist die obere Grenze der letzten Klasse (Folie 2.9. und Formelsammlung S. 1).

Frage:

Kap.2 S.12: Warum ist im Diagramm die Linie ab ca. 4000 € gestrichelt?

Antwort:

Die Linie ist ab 4500 gestrichelt, weil die letzte Klasse $[4500; \infty[$ offen ist. Das heisst, man kann nicht genau sagen, wann die empirische Verteilungsfunktion den Wert Eins erreicht.
Die gestrichelte Linie deutet die Steigung der Gerade im letzten Abschnitt lediglich an. Um diese genau zu bestimmen, müßte man eine obere Grenze festlegen und die Funktion bis zu dieser Grenze auf Eins ansteigen lassen.
(Videoaufzeichnung vom 21.10. 23:30-24:00)

Frage:

Welche Formel liegt der Berechnung der Häufigkeit des 4000 Euro Einkommens zugrunde?

Antwort:

Die Formel für die Berechnung der empirischen Verteilungsfunktion von klassierten Merkmalen findet man im Abschnitt 2.33 vom Skript (S. 11).

Für $x = 4000 \in [2600; 4500[= [e_{4-1}; e_4[$ ergibt sich:

$$\begin{aligned} F(4000) &= F(e_{4-1}) + \frac{f_4}{d_4}(4000 - e_{4-1}) \\ &= F(e_3) + \frac{f_4}{d_4}(4000 - e_3) \\ &= F(2600) + \frac{0.189}{1900}(4000 - 2600) \end{aligned}$$

mit $d_4 = 4500 - 2600 = 1900$.

Zu den Rechenregeln für stetige Merkmale: warum ist f(x) gleich Null?

Antwort:

Die relativen Häufigkeiten $f(x)$ sind bei stetigen Merkmalen gleich Null, da die empirische Verteilungsfunktion ein Polygonzug ist (vgl. Antworten auf Frage 13 und Frage 14). Es handelt sich um eine stetige Funktion, denn sie besitzt keine Sprungstellen. Es ist daher irrelevant, ob man bei der Bildung von Differenzen ein einzelner Punkt im betrachteten Intervall enthalten ist oder nicht (Vorlesungsaufzeichnung vom 21.10. 27:40 - 29:07).

Somit ist z.B. $H(x < d) = H(x <= d)$.

Frage:

Kap.3 S.3: Wie funktioniert der Strahlensatz bzw. die lineare Interpolation, wenn man einen Punkt innerhalb einer Klasse berechnen möchte?

Antwort:

Die Formel auf Seite 3 (Abschnitt 3.9) bezieht sich auf die Berechung des Medians bei klassierten Daten (siehe auch Antwort auf Frage 20). Ein Beispiel für die Anwendung dieser Formel findet sich im Buch von Toutenburg, S. 48-49.

In der Vorlesung wurde ein Beispiel für die Berechnung der Verteilungsfunktion an einem Punkt innerhalb einer Klasse ausführlich besprochen (vgl. Videoaufzeichnung vom 21.10. 24:15 - 27:25 und Antwort auf Frage 17).

Frage:

Wie berechnet man den Median innerhalb von Klassen m.H. des Dreisatzes (Längen verhalten sich wie...)?

Antwort:

Die Berechnung des Medians bei klassierten Daten wurde in der Vorlesung anhand eines konkreten Beispiels mit Abbildungen und ausführlichen Erklärungen vorgeführt (vgl. Videoaufzeichnung vom 21.10. 48:36 -54:11).

Warum nimmt man bei der Berechnung von Quantilen (wenn $n\alpha$ ganzzahlig) nicht $x_{(n\alpha)}$, sondern die Formel $\frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)})$?

Antwort:

Wenn $n\alpha$ ganzzahlig ist, dann gilt die Forderung $F(\tilde{x}_\alpha) = \alpha$ für alle Zahlen im Intervall zwischen $x_{(n\alpha)}$ und $x_{(n\alpha+1)}$. Da wir uns für eine dieser Zahlen entscheiden müssen, wählen wir den Mittelwert der beiden Intervallgrenzen (vgl. Toutenburg S. 49)
Offensichtlich wird dieses Problem beim 50%-Quantil, also dem Median. Hier gilt $\alpha = 0{,}5$, und $n\alpha$ ist genau dann ganzzahlig, wenn n eine gerade Zahl ist. Sei beispielsweise $n = 10$, dann wäre der Median das arithmetische Mittel aus $x_{(5)}$ und $x_{(6)}$. Dies ist auch intuitiv einleuchtend, da es bei 10 Werten ja offensichtlich keinen Wert gibt, der eindeutig in der Mitte liegt. Die Werte $x_{(5)}$ und $x_{(6)}$ liegen gleichermaßen in der Mitte der Daten.

Was bedeutet die (35) im Ergebnis von Kendalls (Skript Kap. 6, Abschnitt 4.3) bzw. die (26) im Ergebnis von Kendalls/Stuarts (Skript Kap. 6, Abschnitt 4.4)?

Antwort:

Kendalls τ_b beträgt 0.230535, wenn man das Ergebnis auf sechs Nachkommastellen rundet.
Analog beträgt Kendalls/Stuarts τ_c: 0,203726.

Formelsammlung S.3 Formel für Varianz: Warum wendet man hier nicht die zweite binomische Formel an, um den zweiten Term zu erhalten?

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Antwort:

Man erhält den zweiten Term gerade durch die Anwendung der zweiten binomische Formel:

$$\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2
\end{aligned}$$

Kapitel 4, Seite 7: Ich habe das Beispiel mit der Währungssumme nicht ganz verstanden: Warum multipliziert man €2 mit 100 und nicht wie in der Umrechnung mit 10?

Antwort:

Die Frage bezieht sich auf ein Beispiel, das in der Vorlesung besprochen wurde (Videoaufzeichung vom 04.11.2014, 21:50 - 23:25).

Wir betrachten die Merkmale:

X: Preis in Euro

Y: Preis in Rand

1 € $\hat{=}$ 10 Rand

Es gilt also : $y_i = a + bx_i$, mit $a = 0$ und $b = 10$.

Die Varianz von Y (lineare Transformation von X) ergibt sich zu:

$$\begin{aligned} s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - (a + b\bar{x}))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(b(x_i - \bar{x}))^2 \\ &= \frac{b^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= b^2 s_x^2 \end{aligned}$$

$\Rightarrow s^2_{\text{Rand}} = 10^2 s^2_{€} = 100 s^2_{€}$

Kapitel 5 Seite 3: Was ist ein quadratischer Graph mit Kantenlänge 1?

Antwort:

Ein quadratischer Graph mit Kantenlänge 1 ist ein Quadrat mit der Seitenlänge 1.

Im kartesischen Koordinaten System wird das Quadrat mit den Eckpunkten (0,0), (1,0), (1,1), (0,1) als *Einheitsquadrat des 1. Quadranten* bezeichnet.

Bei der Lorenzkurve ist die Koordinate (u_0,v_0) immer (0,0) und die Koordinate (u_n,v_n) immer (1,1). Die Lorenzkurve ist also eine Funktion im Einheitsquadrat des 1. Quadranten. Deswegen ist im Skript von einer Abbildung auf einen quadratischen Graphen mit Kantenlänge 1 die Rede (siehe auch Folie 5.9).

Video vom 04.11.14, 40:25 – 40:47: Warum ist es anders bzw nicht zu vergleichen, wenn die Graphen sich schneiden? Wie soll man das dann interpretieren?

Antwort:

Wenn sich zwei Lorenzkurven schneiden, kann man nicht eindeutig sagen, welche Kurve die Verteilung mit der größeren Konzentration aufweist. Das ist nur dann möglich, wenn eine Lorenzkurve an jedem Punkt, (d.h. die gesamte Kurve) unterhalb einer anderen liegt.
Bei sich schneidenden Lorenzkurven ist diese Beziehung nur abschnittsweise erfüllt, so dass ein rein graphischer Vergleich der Konzentration in den entsprechenden Verteilungen nicht möglich ist. Daher wird eine Maßzahl wie z.B. der Gini-Koeffizient benötigt.

Video vom 18.11.2014, 22:00 : Warum ist beim Beispiel D=0 und K=0? Man hat doch noch die ausgefüllten Felder, die man zusammenrechnen kann?! Und fängt man bei Diskordant immer links oben an oder?

Antwort:

- Beim ersten Beispiel ($\gamma = 1$) ist $D = 0$ und $K > 0$. Wenn nur die Felder auf der Diagonalen besetzt sind, dann gibt es keine diskordanten Paare. Wenn jede Beobachtung (x_i,y_i) mit jeder anderen verglichen wird, findet man kein Paar, von dem man sagen kann, dass die Beobachtung mit größerem x das kleinere y aufweist.
- Beim zweiten Beispiel ($\gamma = -1$) ist $K = 0$ und $D > 0$. Zu den Beobachtungen (x_i,y_i), deren Häufigkeiten in den besetzten Feldern eingetragen sind, gibt es keine konkordanten Paare, das heißt keine Paare, bei denen sowohl die x-Koordinate, als auch die y-Koordinate größer bzw. kleiner ist.
- Beim Zählen der diskordanten Paare ist es ratsam und praktischer, aber nicht zwingend links oben anzufangen.

Ich habe eine Frage bezüglich der Quantile. Wir haben in einem Beispiel in der Vorlesung von folgenden Daten: 4, 5, 5, 6, 6, 6, 7, 7, 9 das untere Quantil $x_{0,25}$ wie folgt ausgerechnet: Zuerst haben wir $n \cdot 0{,}25$, also $9 \cdot 0{,}25$ berechnet und 2,25 herausbekommen.
Daraus haben wir geschlossen, dass $x_{(3)}$ das untere Quantil beschreibt. Ich frage mich jedoch warum wir aufgerundet haben, schließlich ist $2{,}25 < 2{,}5$, weswegen ich auf 2,0 abrundet hätte und somit auf $x_{(2)}$ gekommen wäre. Im oben genannten Beispiel macht dies zwar keinen Unterschied, da $x2 = x3 = 5$, jedoch macht es bei anderen Beispielen doch einen Unterschied.
Meine Frage ist daher, ob wir immer aufrunden bei Quantilen sobald unser Wert größer als eine natürlich Zahl ist, oder ob die allgemeinen Auf- und Abrundregeln hier auch gelten?

Antwort:

Das α-Quantil ist wie folgt definiert (FS S.3):

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{falls } n\alpha \text{ keine ganze Zahl ist,} \\ & k \text{ ist dann die kleinste,} \\ & \text{ganze Zahl} > n\alpha, \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{falls } n\alpha \text{ ganzzahlig ist.} \end{cases}$$

Für $n\alpha = 9 \cdot 0{,}25 = 2{,}25$ ist $x_{0.25} = x_{(3)}$, denn $k = 3$ ist die kleinste ganze Zahl größer als $2{,}25 = n\alpha$.

Ich habe eine Frage zu der Interpretation des Odd- Ratio (Blatt 6, Aufgabe 1c): Woher weiß ich in der Aufgabe dass der OR aussagt dass die Chance auf Kurzzeitarbeitslosigkeit 1,77 mal näher ist als bei den Frauen? Wie würde der OR aussehen wenn ich das Verhältnis der Chance auf Langzeitarbeitslosigkeit ausrechne? Geht das überhaupt?

Antwort:

Die richtige Interpretation lautet: Die Chance auf Kurzzeitarbeitslosigkeit ist bei den Männern 1,77 mal höher als bei den Frauen.

Der Odds-Ratio kann als Verhältnis von Chancen gesehen werden:

$$OR = \frac{o_{\mathrm{M}}}{o_{\mathrm{F}}} = \frac{\text{„Chance auf kurzzeitige Arbeitslosigkeit bei Männern“}}{\text{„Chance auf kurzzeitige Arbeitslosigkeit bei Frauen“}} \approx \frac{2{,}41}{1{,}36} \approx 1{,}77$$

Betrachtet man die Ausprägung „Langfristige Arbeitslosigkeit“, so kann man folgendes Chancenverhältnis bilden:

$$\begin{aligned} OR^* &= \frac{o_{\mathrm{M}}}{o_{\mathrm{F}}} = \frac{\text{„Chance auf langfristige Arbeitslosigkeit bei Männern“}}{\text{„Chance auf langfristige Arbeitslosigkeit bei Frauen“}} \\ &\approx \frac{167/403}{175/239} \approx \frac{0{,}414}{0{,}735} \approx 0{,}56 \end{aligned}$$

Dies entspricht einer Vertauschung der Spalten der gegebenen (2×2)-Kontingenztafel:

		Arbeitslosigkeit		
		Langzeit-	Kurzzeit-	
Geschlecht	männlich	167	403	570
	weiblich	175	238	413
		641	342	983

$$OR = \frac{a \cdot d}{b \cdot c} = \frac{167 \cdot 238}{403 \cdot 175} \approx 0{,}56$$

Werden Zeilen oder Spalten vertauscht, so ändert sich der Odds-Ratio auf den Kehrwert!

Hier: $\mathrm{OR}^* = \frac{1}{\mathrm{OR}} = \frac{1}{1{,}77} \approx 0{,}56$

Wenn nach der empirischen Verteilungsfunktion gefragt wird, hat diese dann bei diskreten Merkmalen immer das Aussehen einer gezackten Linie? Ich weiß nämlich nicht wie ich die Form mit ausschließlich horizontalen Linien, die an den Enden kleine Kreise haben (meist das linke Ende ausgemalt und das rechte leer) einordnen soll bzw. wann man diese benötigt.

Antwort:

Die empirische Verteilungsfunktion bei diskreten Merkmalen ist eine Treppenfunktion, die an den Ausprägungen $a_1, \ldots, a_k$ um die entsprechende relative Häufigkeit nach oben springt. Der obere Wert an den Sprungstellen, d.h. die Treppenkante, ist der zugehörige Funktionswert. Die Funktion ist somit rechtsstetig. Diese Eigenschaft wird bei der Darstellung mit den horizontalen Linen und den Kreisen deutlich (Vgl. Toutenburg, S. 27 und Aufgabe 2, Übungsblatt 2). Oft werden aber die Stufen durch vertikale Linien verbunden, was zu der anderen Darstellungsform führt (Vgl. Folie 2.31).