deskriptive Statistik

Silke Bott

Sommersemester 2023

Einleitung

Unter Statistik verstehen wir einerseits die Gesamtdarstellung einer konkret vorliegenden Datenmenge, etwa die Statistik der Verkehrstoten in einem Jahr oder die Statistik der gemeldeten Einwohner einer Gemeinde, in der eine gesamt vorliegende Datenmenge in Zahlen, Tabellen und Graphiken zusammengefasst wird. Andererseits ist *Statistik* aber auch die Gesamtheit aller Methode, die für die Gewinnung und Verarbeitung empirischer Informationen relevant sind.

Einleitung

Die statistische Methodenlehre beschäftigt sich dementsprechend mit den folgenden drei Grundfragen:

- Datenbeschreibung (*Deskription*).
- Datenerhebung (*Exploration*).
- Schlussfolgerung (*Induktion*).

Grundbegriffe der Datenerhebung

Definition (Grundbegriffe)

- Ein *Merkmalsträger* oder eine *statistische Einheit* ist ein Objekt, an dem interessierende Größen erfasst werden.
- Die *Grundgesamtheit* oder *statistische Masse* ist die Menge aller für die Fragestellung relevanten Merkmalsträger.
- Eine *Stichprobe* ist die tatsächlich untersuchte Teilmenge der Grundgesamtheit.
- Ein *Merkmal* oder eine *Variable* ist die interessierende Größe.
- Eine *Merkmalsausprägung* ist der konkrete Wert eines Merkmals bei einem konkreten Merkmalsträger.

Grundbegriffe der Datenerhebung

Übung

Beschäftigungsverhältnis ist

- ein Merkmalsträger?
- ein Merkmal?
- eine Merkmalsausprägung?

Lösung:

Beschäftigungsverhältnis ist eine Merkmal.

Grundbegriffe der Datenerhebung

Wir unterscheiden zwischen *endlichen Grundgesamtheiten* Ω und *unendlichen Grundgesamtheiten* Ω. Bei unendlichen Grundgesamtheiten unterscheiden wir ferner zwischen abzählbar unendlichen (wie etwa $\Omega = \mathbb{N}$) oder überabzählbar unendlichen (wie etwa $\Omega = \mathbb{R}$).
Ferner unterscheiden wir Grundgesamtheiten auch nach zwei Grundtypen:

- Bestandsmassen: Grundgesamtheit zu einem fixierten Zeitpunkt.
- Bewegungsmassen: Grundgesamtheit über eine Zeitperiode.

Grundbegriffe der Datenerhebung

Definition (Skalentypen)

- Ein Merkmal heißt *nominalskaliert*, wenn die Merkmalsausprägungen Namen oder Kategorien sind, die den Merkmalen zugeordnet werden können, und diese Ausprägungen zwar unterschieden werden können aber keine natürliche Rangfolge aufweisen.
- Ein Merkmal heißt *ordinalskaliert*, wenn die Merkmalsausprägungen geordnet und in eine natürliche Reihenfolge gebracht werden können, ohne dass aber die Abstände zwischen den Merkmalsausprägungen interpretiert werden können.
- Ein Merkmal heißt *kardinalskaliert* oder *metrisch skaliert*, wenn die Merkmalsausprägungen in eine Rangordnung gebracht werden können, und zusätzlich noch bestimmt werden kann, in welchem Ausmaß sich je zwei verschiedene Merkmalsausprägungen unterscheiden.

Grundbegriffe der Datenerhebung

Übung

Bei einer Befragung werden folgende Merkmale erhoben:

1. Geschlecht der befragten Person.
2. (Höchster) Schulabschluss der befragten Person.
3. Durchschnittliches monatliches Nettoeinkommen der befragten Person (in €).
4. Körpergröße der befragten Person.
5. Zufriedenheit der der befragten Person mit ihrer gegenwärtigen Situation auf einer Skala von -3 (sehr unzufrieden) bis 3 (sehr zufrieden).

Bestimmen Sie den Merkmalstyp der erhobenen Merkmale.

Grundbegriffe der Datenerhebung

Lösung:

Es gilt:

1. Geschlecht der befragten Person: nominalskaliert.
2. (Höchster) Schulabschluss der befragten Person: ordinalskaliert.
3. Durchschnittliches monatliches Nettoeinkommen der befragten Person (in €): metrisch skaliert.
4. Körpergröße der befragten Person: metrisch skaliert.
5. Zufriedenheit der der befragten Person mit ihrer gegenwärtigen Situation auf einer Skala von -3 (sehr unzufrieden) bis 3 (sehr zufrieden): ordinalskaliert.

Grundbegriffe der Datenerhebung

Bei Kardinalskalen unterscheiden wir noch nach *Intervallskalen*, bei denen die Abstände zwischen zwei Ausprägungen verglichen werden können und *Verhältnisskalen* bei denen zusätzlich noch ein natürlicher Nullpunkt dazukommt und damit auch die Quotienten von Merkmalsausprägungen sinnvoll interpretiert werden können.
Das Merkmal „Geburtszeitpunkt“ ist intervallskaliert, aber nicht verhältnisskaliert, das Merkmal „Größe“ ist verhältnisskaliert.

Grundbegriffe der Datenerhebung

Bei Merkmalen unterscheiden wir *diskrete Merkmale*, bei denen endlich viele oder abzählbar unendlich viele Merkmalsausprägungen möglich sind, und *stetige Merkmale*, die sich dadurch auszeichnen, dass alle Zahlen eines Intervalls als Merkmalsausprägungen angenommen werden können.

Hier sind jedoch auch Zwischenstufen und Mischformen möglich!

Ein Merkmal heißt *kategorisiert* oder *klassiert* oder *gruppiert*, wenn die eigentlichen Merkmalsausprägungen zu Klassen oder Gruppen zusammengefasst werden.

Grundbegriffe der Datenerhebung

Bei einer *einfachen Stichprobe* werden Teilmengen der Grundgesamtheit so erhoben, dass jede dieser Teilmengen dieselbe Wahrscheinlichkeit besitzt, gezogen zu werden. Notwendig ist dazu, dass die Grundgesamtheit nummeriert ist und (zumindest theoretisch) als Liste vorliegt. Zur Realisierung wird dann jede Nummer auf ein Los geschrieben und in eine Urne gesteckt. Dann wird die Stichprobe blind aus dieser Urne gezogen.

Grundbegriffe der Datenerhebung

In der Praxis werden Stichproben häufig systematisch gezogen.
Eine andere Variante ist, die Grundgesamtheit zunächst in disjunkte Teilmengen zu zerlegen und dann aus jeder Teilmenge eine Zufallsstichprobe zu ziehen (*geschichtete Zufallsstichprobe* oder auch *Klumpenstichprobe*, falls die Grundgesamtheit in natürlicher Weise in disjunkte Teile zerfällt).
Auch mehrstufige Auswahlverfahren für Stichproben sind möglich.

Eindimensionale deskriptive Statistik

Definition

Univariater Daten oder *eindimensionaler Daten* gehen aus der Beobachtung eines einzigen Merkmals hervor, dessen Ausprägungen wir in geeigneter Weise (in der Regel durch reelle Zahlen) darstellen.

Wir untersuchen ein Merkmal X. Dazu machen wir eine Erhebung von n Merkmalsträgern dieses Merkmals und beobachten an diesen die Merkmalsausprägungen $x_1, \ldots, x_n \in \mathbb{R}$ (*Roh–* oder *Primärdaten*, *Urliste*).

Beispiel

In einer Vorlesung wird das Klausurnoten der Vorlesungsteilnehmer erhoben (in ganzen Notenstufen). Die Erhebung ergibt die Urliste

$$(2, 5, 3, 4, 1, 1, 2, 2, 1, 2, 2, 2, 4, 4, 5)$$

Häufigkeiten

Die Merkmalswerte der Urliste seien $a_1, \ldots, a_k$, wobei wir annehmen, dass $a_1 < a_2 < \ldots < a_k$.

Definition (Häufigkeiten)

- Die *absolute Häufigkeit* $h_l = h(a_l)$ der Ausprägung a_l ist die Anzahl der i mit $x_i = a_l$,

$$h(a_l) = |\{i \in \{1, \ldots, n\} | \, x_i = a_l\}|$$

- Die *relative Häufigkeit* $f(a_l)$ von a_l ist $f(a_l) = \frac{h(a_l)}{n}$.
- $(h(a_1), \ldots, h(a_k))$ heißt *absolute Häufigkeitsverteilung* der Stichprobe.
- $(f(a_1), \ldots, f(a_k))$ heißt *relative Häufigkeitsverteilung* der Stichprobe.

Häufigkeiten

Beispiel

Bei der Notenerhebung treten die Merkmalsausprägungen $a_1 = 1, a_2 = 2, a_3 = 3, a_4 = 4$ und $a_5 = 5$ auf, und zwar mit den folgenden Häufigkeiten

$$h(1) = 3,\ h(2) = 6,\ h(3) = 1,\ h(4) = 3,\ h(5) = 2$$

Die Häufigkeitsverteilung ist

$$H = (3, 6, 1, 3, 2)$$

Die relative Häufigkeitsverteilung ist

$$F = \left(\frac{3}{15}, \frac{6}{15}, \frac{1}{15}, \frac{3}{15}, \frac{2}{15}\right)$$

Häufigkeiten

Definition (kumulierte Häufigkeiten)

- $H(x) = \sum_{a_l \leq x} h(a_l)$ heißt *absolute kumulierte Häufigkeit* der Stichprobe zum Wert x.
- $F(x) = \sum_{a_l \leq x} f(a_l)$ heißt *relative kumulierte Häufigkeit* der Stichprobe zum Wert x.

Häufigkeiten

Übung

Bestimmen Sie die kumulierten und relativen kumulierten Häufigkeiten im Notenbeispiel mit Häufigkeitsverteilung

$$H = (3, 6, 1, 3, 2)$$

Skizzieren Sie den Graphen der relativen kumulierten Häufigkeit.

Lösung:

Es gilt

	$x < 1$	$1 \leq x < 2$	$2 \leq x < 3$	$3 \leq x < 4$	$4 \leq x < 5$	$x \geq 5$
$H(x)$	0	3	9	10	13	15
$F(x)$	0	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{10}{15}$	$\frac{13}{15}$	1

Häufigkeiten

Bemerkung

Es gilt $F(x) = \frac{H(x)}{n}$.

Bemerkung

Bei $F(x)$ spricht man auch von der *empirischen Verteilungsfunktion*.

Häufigkeiten

Wir betrachten zunächst ein (mindestens) nominal skaliertes Merkmal X.

Definition

Der *Modus* x_{mod} der Stichprobe ist der Merkmalswert mit der größten Häufigkeit,

$$x_{mod} = a_l \iff h(a_l) \geq h(a_j) \quad \forall j \neq l$$

Bemerkung

Der Modus muss nicht eindeutig sein, es können mehrere Werte mit der gleichen (größten) Häufigkeit auftreten.

Beispiel

Im Notenbeispiel ist $x_{\text{mod}} = 2$.

Median und Quantile

Wir betrachten nun ein mindestens ordinal skaliertes Merkmal X, und wir nehmen an, dass schon die x_i angeordnet sind, also $x_1 \leq x_2 \leq \ldots \leq x_n$.

Definition

Der *Median* x_{med} ist definiert wir folgt

- Falls n ungerade ist, so ist

$$x_{med} = x_{\frac{n+1}{2}}$$

- Falls n gerade ist, so ist

$$x_{med} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$$

Median und Quantile

Übung

Bestimmen Sie den Median im Notenbeispiel mit Häufigkeitsverteilung

$$H = (3, 6, 1, 3, 2)$$

Lösung:

Es ist $x_{\text{med}} = 2$.

Median und Quantile

Definition

Für eine Zahl p mit $0 < p < 1$ ist ein *p–Quantil* x_p der Stichprobe eine Merkmalsausprägung x_p, für die gilt

$$\frac{|\{i \in \{1, \ldots, n\} | \, x_i \leq x_p\}|}{n} \geq p \quad \text{und} \quad \frac{|\{i \in \{1, \ldots, n\} | \, x_i \geq x_p\}|}{n} \geq 1-p$$

Bemerkung

Das p–Quantil ist so gewählt, dass mindesten $p \cdot n$ der Merkmalswerte kleiner oder gleich x_p sind und mindestens $(1-p) \cdot n$ der Merkmalswerte größer oder gleich x_p sind.
Ein p–Quantil existiert immer. Wenn $n \cdot p \notin \mathbb{Z}$, so ist es eindeutig bestimmt. Falls $n \cdot p \in \mathbb{Z}$, so ist jede Zahl $x_p \in [x_{np}, x_{np+1}]$ ein p–Quantil. Es ist üblich, in diesem Fall $x_p = \frac{1}{2} \cdot (x_{np} + x_{np+1})$ zu wählen.

Median und Quantile

Bemerkung

Spezielle p-Quantile sind die Quartile,

$x_{0.25}$ = 25 %–Quantil = unteres Quartil.

$x_{0.75}$ = 75 %–Quantil = oberes Quartil.

Das „mittlere Quartil“ einer Stichprobe ist der Median der Stichprobe.

Bemerkung

Die *Fünf–Punkte–Zusammenfassung* einer Verteilung besteht aus den Werten

$$x_{min},\ x_{0.25},\ x_{med},\ x_{0.75},\ x_{max}$$

Lageparameter

Wir betrachten nun ein metrisch skaliertes Merkmal X.

Definition

Das *arithmetisches Mittel* $\overline{x}$ der Stichprobe ist definiert als

$$\overline{x} = \frac{1}{n} \cdot (x_1 + \cdots + x_n) = \frac{1}{n} \cdot \sum_{l=1}^{n} x_l$$

Bemerkung

Das arithmetische Mittel berechnet sich aus den relativen Häufigkeiten als

$$\overline{x} = a_1 \cdot f(a_1) + \cdots + a_k \cdot f(a_k) = \sum_{l=1}^{n} a_l \cdot f(a_l)$$

Lageparameter

Beispiel

Wir betrachten das Merkmal *Körpergröße* von fünf Personen

k	1	2	3	4	5
x_k	176	182	190	182	174

Dann ist

$$\overline{x} = \frac{1}{5} \cdot (176 + 182 + 190 + 182 + 174) = 180.8$$

Lageparameter

Übung

Wir betrachten das Merkmal *Körpergewicht* von sechs Personen

k	1	2	3	4	5	6
x_k	84	72	69	80	94	81

Bestimmen Sie das arithmetische Mittel dieser Merkmalsverteilung.

Lösung:

Es ist

$$\overline{x} = \frac{1}{6} \cdot (84 + 72 + 69 + 80 + 94 + 81) = 80.0$$

Lageparameter

Bemerkung

Das arithmetische Mittel reagiert relativ empfindlich auf Ausreißer.

Bemerkung

Das arithmetische Mittel ist der Wert, von dem alle Merkmalswerte die geringste mittlere quadratische Abweichung haben, also dasjenige μ, für das gilt

$$\sum_{i=1}^{n}(x_i - \mu)^2 \quad \text{ist minimal}$$

Lageparameter

Definition (Streuungsparameter)

Die *empirische Varianz der Stichprobe* (oder *empirische Stichprobenvarianz* $\widetilde{s}^2$ ist erklärt als

$$\widetilde{s}^2 = \frac{1}{n} \cdot \left((x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2 \right) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Die *Standardabweichung* $\widetilde{s}$ der Stichprobe ist erklärt als

$$\widetilde{s} = \sqrt{\widetilde{s}^2} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Bemerkung

Die Varianz ist das Mittel der quadratischen Abweichungen der Merkmalswerte vom arithmetischen Mittel.

Lageparameter

Definition

Die *korrigierte empirische Varianz der Stichprobe* s^2 ist definiert als

$$s^2 = \frac{1}{n-1} \cdot \left((x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2 \right) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Bemerkung

Die korrigiere empirische Varianz der Stichprobe unterscheidet sich von der empirischen Varianz der Stichprobe im Normierungsfaktor. Eine Begründung für diesen anderen (zunächst seltsam erscheinenden) Normierungsfaktor werden wir in der induktiven Statistik geben.

Lageparameter

Beispiel

Wir betrachten wieder das Merkmal *Körpergröße* von fünf Personen

k	1	2	3	4	5
x_k	176	182	190	182	174

von dem wir schon wissen, das $\overline{x} = 180.8$.
Dann ist

$$\begin{aligned} \widetilde{s}^2 &= \frac{1}{5} \cdot (((176 - 180.8)^2 + (182 - 180.8)^2 + (190 - 180.8)^2 \\ &\quad + (182 - 180.8)^2 + (174 - 180.8)^2) \\ &= \frac{1}{5} \cdot 156.8 \\ &= 31.36 \end{aligned}$$

und $s^2 = \frac{1}{4} \cdot 156.8 = 39.2$.

Lageparameter

Übung

Wir betrachten das Merkmal *Körpergewicht* von sechs Personen

k	1	2	3	4	5	6
x_k	84	72	69	80	94	81

Bestimmen Sie die empirische Varianz dieser Merkmalsverteilung.

Lösung:

Es ist

$$\begin{aligned} \widetilde{s}^2 &= \frac{1}{6} \cdot (((84-80)^2 + (72-80)^2 + (69-80)^2 \\ &\quad + (80-80)^2 + (94-80)^2 + (81-80)^2) \\ &= \frac{1}{6} \cdot 398 \end{aligned}$$

und $s^2 = \frac{1}{5} \cdot 398 = 79.6$.

Lageparameter

Satz (Verschiebungssatz)

Für jedes $c \in \mathbb{R}$ gilt

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - c)^2 = \widetilde{s}^2 + (\overline{x} - c)^2$$

Speziell gilt also

$$\widetilde{s}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

Satz (Transformationssatz)

Ist $Y = a \cdot X + b$ (also ist $y_i = a \cdot x_i + b$ für alle i), so gilt

$$\overline{y} = a \cdot \overline{x} + b, \qquad \widetilde{s}_y = |a| \cdot \widetilde{s}_x$$

Lorenzkurve

Betrachte Merkmal X mit Merkmalsausprägungen $x_1, \ldots, x_n$, wobei

1. $x_i \geq 0$ für alle i.
2. $x_1 \leq x_2 \leq \ldots \leq x_n$.

In den Wirtschafts– und Sozialwissenschaften oft gesucht (z.B. bei den Einkommensverteilungen):
Ein Maß, wie stark die x_i voneinander abweichen.

Lorenzkurve

Für $k = 1, \ldots, n$ setze

$$u_k = \frac{k}{n}$$

$$v_k = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{n} x_i}$$

und für $k = 0$:

$$u_0 = 0, \quad v_0 = 0$$

Definition

Der Polygonzug, der die Punkte

$$(0, 0) = (u_0, v_0), (u_1, v_1), \ldots, (u_{n-1}, v_{n-1}), (u_n, v_n) = (1, 1)$$

miteinander verbindet, heißt **Lorenzkurve** zu den Merkmalsausprägungen $x_1, \ldots, x_n$.

Lorenzkurve

Beispiel

Betrachte vier Unternehmen mit den folgenden Umsätzen:

Unternehmen	1	2	3	4
Umsatz	25	25	25	25

Es gilt

k	1	2	3	4
u_k	0.25	0.50	0.75	1.00
v_k	0.25	0.50	0.75	1.00

Lornezkurve

Beispiel

Betrachte vier Unternehmen mit den folgenden Umsätzen:

Unternehmen	1	2	3	4
Umsatz	5	5	5	85

Es gilt

k	1	2	3	4
u_k	0.25	0.50	0.75	1.00
v_k	0.05	0.10	0.15	1.00

Gini–Koeffizient

Bemerkung

Die Fläche zwischen der Diagonale d und der Lorenzkurve l ist ein Maß für die Ungleichverteilung der gegebenen Daten.

Übung

Fünf Unternehmen haben die folgenden Umsätze:

Unternehmen	1	2	3	4	5
Umsatz	50	20	30	60	40

Bestimmen Sie die Lorenzkurve in dieser Situation.

Gini–Koeffizient

Lösung:

Zunächst sortieren wir die Umsätze der Größe aufsteigend und erhalten folgende sortierte Liste:

Unternehmen	2	3	5	1	4
Umsatz	20	30	40	50	60

Daraus ergibt sich

k	1	2	3	4	5
u_k	0.20	0.40	0.60	0.80	1.00
v_k	0.10	0.25	0.45	0.70	1.00

Gini–Koeffizient

Definition

Die Größe

$$G = \frac{\text{Fläche zwischen Diagonale und Lorenzkurve}}{\text{Fläche zwischen Diagonale und } x\text{–Achse}}$$

heißt **Gini–Koeffizient** zu den Merkmalsausprägungen $x_1, \ldots, x_n$.
Für $n \geq 2$ ist der **normierte Gini–Koeffizient** G^* definiert als

$$G^* = \frac{n}{n-1} \cdot G$$

Bemerkung

$$G = 2 \cdot \text{ Fläche zwischen Diagonale und Lorenzkurve}$$

Gini–Koeffizient

Beispiel

Im ersten Beispiel ist offensichtlich $G = 0$, $G^* = 0$.

Beispiel

Im zweiten Beispiel ist die Fläche **unterhalb** der Lorenzkurve gleich

$$F = \int_0^1 l(x)\, dx = 0.20$$

Damit gilt

$$G = 2 \cdot \left(\frac{1}{2} - 0.20 \right) = 0.60$$

und

$$G^* = \frac{4}{3} \cdot 0.60 = 0.80$$

Gini–Koeffizient

Übung

Berechnen Sie G und G^* im dritten Beispiel.

Lösung:

Die Fläche **unterhalb** der Lorenzkurve ist in diesem Fall

$$\begin{aligned} F &= \tfrac{1}{2} \cdot 0.2 \cdot 0.10 + \tfrac{1}{2} \cdot 0.2 \cdot (0.25 + 0.10) + \tfrac{1}{2} \cdot 0.2 \cdot (0.45 + 0.25) \\ &\quad + \tfrac{1}{2} \cdot 0.2 \cdot (0.70 + 0.45) + \tfrac{1}{2} \cdot 0.2 \cdot (1.00 + 0.70) \\ &= 0.40 \end{aligned}$$

Damit ist

$$G = 2 \cdot (0.50 - 0.40) = 0.20, \qquad G^* = 0.20 \cdot \frac{5}{4} = 0.25$$

Gini–Koeffizient

Der Ginikoeffizient kann unmittelbar aus den Daten $(x_1, \ldots, x_n)$ berechnet werden:

Satz

Es gilt

$$G = \frac{2 \cdot \sum_{k=1}^{n} k \cdot x_k}{n \cdot \sum_{k=1}^{n} x_k} - \frac{n+1}{n}$$

Satz

Genau dann gilt $G^* = 0$, wenn die Merkmalsausprägungen gleichverteilt sind, dh. wenn $x_1 = x_2 = \cdots = x_n$.
Genau dann gilt $G^* = 1$, wenn $x_1 = x_2 = \cdots x_{n-1} = 0, x_n \neq 0$ (also wenn maximale Konzentration vorliegt).

Gini–Koeffizient

Der Ginikoeffizient wird häufig herangezogen, um Vermögens– oder Einkommensverteilungen zu untersuchen.

Beispiel

Die Gini–Koeffizienten in einigen Ländern für das Vermögen V und das Einkommen E

Land	Gini(V/2000)	Gini(V/2016)	Gini(E)
Russland	0.699	0.923	0.377
Indien	0.669	0.876	0.368
Italien	0.609	0.687	0.354
USA	0.801	0.862	0.415
Deutschland	0.667	0.789	0.291
UK	0.697	0.732	0.332