Deskriptivstatistik

a) Univariate Statistik

Weiters zum Thema der statistischen Informationsverdichtung

M a ß z a h l e n

Statistiken
bei Stichproben

Parameter
bei Grundgesamtheiten

Maßzahlen
zur Beschreibung
univariater Verteilungen

Maßzahlen
der zentralen Tendenz
(Lagemaße bzw. Mittelwerte)

Maßzahlen
der Variabilität
(Streuungsmaße)

Deskriptivstatistik

a) Univariate Statistik

Hierzu noch weitere Begriffserläuterungen:

Statistik

Sofern eine Maßzahl eine summarische Aussage über den Informationsgehalt der erhobenen Daten macht, wird diese Maßzahl - in Anlehnung an den englischen Begriff „statistic“ - als Statistik bezeichnet.

Parameter

In Funktionen und Gleichungen eine neben den eigentlichen Variablen auftretende, entweder unbestimmt gelassene oder konstant gehaltene Hilfsgröße.

Deskriptivstatistik

a) Univariate Statistik

MAßZAHLEN DER ZENTRALEN TENDENZ

Generell gilt:
Eine Maßzahl der zentralen Tendenz soll der Kennwert sein, der die gesamte Verteilung am besten repräsentiert.

1. : Modus /Modalwert

Symbol =h

Der Modus ist der Wert, der in einer Verteilung am häufigsten vorkommt (dichtester Wert).

Beispiel A:

5, 6, 6, <u>7, 7, 7</u>, 8, 8, 9, 10

Der Wert 7 kommt hier am häufigsten (= 3mal) vor, also ist: **h = (3 x 7)/3=7**

Beispiel B:

5, 6, 6, <u>7, 7, 7, 8, 8, 8</u>, 9, 10

Hier ist: h = $\left[\left(3\cdot 7\right)+\left(3\cdot 8\right)\right]:6=7{,}5$

(aufgrund von benachbarten

Häufigkeitsmaxima)

Deskriptivstatistik

a) Univariate Statistik

Treten zwei nicht benachbarte Messwerte mit relativen Häufigkeitsmaxima auf, dann werden diese beiden Messwerte als Modalwerte angegeben:

Die Verteilung ist dann bimodal.

B e i s p i e l:

x_i	f_i
4	1
5	2
6	5
7	2
8	1
9	5
10	1
11	1

h = 6 ; h =9

Deskriptivstatistik
a) Univariate Statistik

Bitte beachten:

Da der Modalwert (h) schon auf nominalem Meßniveau genutzt werden kann, ist h demnach auch für alle weiteren Meßniveaus zulässig.

Deskriptivstatistik

a) Univariate Statistik

Maßzahlen der zentralen Tendenz

→ die erst ab ordinalem Skalenniveau berechnet werden können:

2. Median ►**Symbol:** $\tilde{x}$

→ **auch „x-Schlange“**
oder „x-Tilde“ genannt.

Begriff:

Median kommt vom lateinischen

medianus

(= „in der Mitte befindlich“)

Definition:

Der Median ist der Wert, der eine nach ihrer Größe geordnete Reihe von Messwerten halbiert.

Medianbestimmung:

Die Bestimmung des Medians für nicht-klassifizierte Daten ist abhängig von der Anzahl der Fälle:

a) liegt eine <u>ungerade</u> Anzahl von Fällen vor, so ist der Median der Wert des mittleren Falles.

Deskriptivstatistik

a) Univariate Statistik

Allgemeine Formel für die Berechnung von $\tilde{x}$ bei ungerader Anzahl der Fälle:

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

Beispiel: **x_i = 2 ; 3 ; 4 ; 5 ; 9; 16 ; 22 ; n=7**

$$\tilde{x} = x_{\left(\frac{7+1}{2}\right)} = x_{\left(\frac{8}{2}\right)} = x_{[4]}$$

Demnach ist hier der Median ($\tilde{x}$)
der Wert des 4.Falles, also: 5

► denn es liegen gleichviel Fälle
(hier jeweils 3 Fälle)
unterhalb wie oberhalb des 4.Falles

Deskriptivstatistik

a) Univariate Statistik

b) liegt eine gerade Anzahl von Fällen vor, so ist der Median der Wert, der die beiden mittleren Fälle halbiert:

Formel:

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

Beispiel: gegeben sei eine Stichprobe über das monatliche Einkommen (in Euro) von Studenten mit folgenden Werten:

637 697 719 750 898 912 981 1032

Bestimmung des Medians mit Hilfe der zuvor benannten Formel:

$$\tilde{x} = \frac{x_{\left(\frac{8}{2}\right)} + x_{\left(\frac{8}{2}+1\right)}}{2} = \frac{x_4 + x_5}{2} = \frac{750+898}{2} = 824$$

Deskriptivstatistik

a) Univariate Statistik

Medianbestimmung bei klassifizierten Häufigkeiten:

1) Zunächst müssen die kumulierten Häufigkeiten gebildet werden.

2) Dann ermitteln wir n/2 bzw. (n+1)/2. Damit ist die Lage des Medians festgelegt.

3) Wir bestimmen nun aufgrund der kumulierten Häufigkeit diejenige Messwertklasse, in die der Median fällt.

$\longrightarrow$ Dies wird auch der Eingriffsspielraum genannt

4) Wir formulieren die exakten Grenzen des Eingriffsspielraumes.

5) Wir berechnen den Median nach der Formel:

$$\tilde{x} = U + \frac{\left(\frac{N}{2}\right) + FU}{Fm} \times h$$

wobei: U = exakte untere Grenze des Eingriffsspielraumes (=Medianintervall)
N = Anzahl der Fälle
FU = kumulierte Häufigkeit unterhalb des Medianintervalls
Fm = Häufigkeit im Medianintervall
h = Klassenbreite

Deskriptivstatistik

a) Univariate Statistik

Beispiel:

1) Wir bilden die kumulierte Häufigkeit (cum f_i)

4) exakte Grenze 11,5

Klassenintervall	f_i	cum f_i
6 – 8	3	3
9 – 11	10	13
12 – 14	14	27
15 – 17	13	40
18 – 20	11	51

3) Eingriffs-spielraum (Median-intervall)

2) Wir ermitteln die Stellung des Medians:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{51+1}{2}\right)} = x_{(26)}$$

→ **also der 26. Wert**

5) $\tilde{x} =$ Exakte untere Grenze des Eingriffsspielraumes (Medianintervalls) + $\left(\frac{\text{N/2 – cum } f_i \text{ unterhalb des Medianintervalls}}{f_i \text{ im Medianintervall}} \right) * h$

<u>Mit den in die Formel eingesetzten Werten:</u>

$$\tilde{x} = 11{,}5 + \left(\frac{51/2 - 13}{14} \right) * 3$$

$$\tilde{x} = 14{,}18$$

<u>Achtung</u>: der Ausdruck „klein h“ bezeichnet hier die Breite des Klassenintervalls!

Deskriptivstatistik

a) Univariate Statistik

Bitte beachten Sie:

Der Median zeichnet sich durch **Unempfindlichkeit gegenüber Extremwerten** aus.

Befinden sich am oberen und/oder am unteren Ende einer Verteilung extreme Werte (sogenannte „Ausreißer“), so bleiben sie bei der Berechnung des Medians unberücksichtigt!

Deskriptivstatistik

a) Univariate Statistik

Maßzahlen der zentralen Tendenz

→ die erst ab Intervall-/Ratioskalenniveau (= metrischem Skalenniveau) berechnet werden können:

3. Das arithmetische Mittel

► **Symbol:** $\overline{x}$

Definition:

Das arithmetische Mittel (Durchschnittswert) ist definiert als die Summe der Messwerte, geteilt durch ihre Anzahl, als Formel geschrieben:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum_{i=1}^{N} x_i}{n}$$

Liegt mehr als eine Häufigkeit vor gilt:

$$\overline{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + ... + f_k x_k}{n} = \frac{\sum_{i}^{k} f_i \cdot x_i}{n}$$

Deskriptivstatistik

a) Univariate Statistik

Beispiel: x_i = Alter(in Jahren) von Häusern in einem Sanierungsgebiet

$\mathbf{x_i}$	$\mathbf{f_i}$
30	1
40	9
50	7
100	2
400	1
Σ	20

$$\bar{x} = \frac{(1\times 30)+(9\times 40)+(7\times 50)+(2\times 100)+(1\times 400)}{20}$$

$$\bar{x} = \frac{(30)+(360)+(350)+(200)+(400)}{20}$$

$$\bar{x} = \frac{1340}{20} = 67$$

ohne „Ausreißer“:

$$\bar{x} = \frac{(1\times 30)+(9\times 40)+(7\times 50)+(2\times 100)}{19} = \frac{940}{19} = 49{,}47$$

Deskriptivstatistik

a) Univariate Statistik

Bitte beachten Sie:

Das arithmetische Mittel ist **empfindlich gegenüber Extremwerten** (sogenannten „Ausreißern“).

Das arithmetische Mittel ist aber **am informativisten** gegenüber den anderen Lagemaßen, da hier **alle** Messwerte in die Berechnung mit aufgenommen werden.

Deskriptivstatistik
a) Univariate Statistik

Besondere Eigenschaften des arithmetischen Mittels

1) Die Summe der Abweichungen aller Messwerte von ihrem arithmetischen Mittel ist gleich Null.

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right) \quad = 0$$

Beispiel: 5, 6, 7, 8, 9

$$\overline{x} = 7$$

x_i	$x_i - \overline{x}$
5	-2
6	-1
7	0
8	1
9	2
Σ	0

Deskriptivstatistik

a) Univariate Statistik

Besondere Eigenschaften des arithmetischen Mittels

Problem zu 1) , „verursacht“ durch die Idealwissenschaft Mathematik:

Wenn gilt, dass $\sum_{i=1}^{n}(x_i - \bar{x}) \quad = 0$

, dann kann ich an dieser Stelle nicht weiterrechnen, den es gilt auch:

$$0 + 0 = 0$$

und

$$0 \cdot 0 = 0$$

und

$$x \cdot 0 = 0$$

zudem ist der Ausdruck $\frac{x}{0} \longrightarrow$ **nicht definiert;**

dagegen ist festgelegt das gilt: $x^0 = 1$ **,**

und der Ausdruck $\frac{0}{x}$ **ist festgelegt mit:**

$$\frac{0}{x} = 0$$

Deskriptivstatistik

a) Univariate Statistik

Besondere Eigenschaften des arithmetischen Mittels

„Abhilfe“ bietet hier
folgende mathematische Gesetzmäßigkeit:

2) Die Summe der Quadrate der Abweichungen
aller Messwerte von ihrem arithmetischen Mittel
ist kleiner
als die Summe der Quadrate der Abweichungen
aller Messwerte von einem beliebigen Wert der
Verteilung, formelhaft dargestellt:

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \quad < \quad \sum_{i=1}^{n}\left(x_i-x_0\right)^2$$

Mit anderen Worten:
Die Summe der Abweichungsquadrate
ist für das arithmetische Mittel ein Minimum
- aber eben nicht 0 - , formelhaft dargestellt:

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \quad = \min!$$

Deskriptivstatistik
a) Univariate Statistik

Besondere Eigenschaften des arithmetischen Mittels

Beispiel: x_i= 5 , 6 , 7 ,8 , 9 ► $\bar{x} = 7$

Der beliebige Wert sei hier angenommen mit x_0 = 6

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$x_i - x_0$	$(x_i - x_0)^2$
5	-2	4	-1	1
6	-1	1	0	0
7	0	0	1	1
8	1	1	2	4
9	2	4	3	9
$\sum$	**0**	10	5	15

Also ergibt sich: 10 < 15

Deskriptivstatistik

a) Univariate Statistik

Besondere Eigenschaften des arithmetischen Mittels

3) Die Addition (oder: Subtraktion)
einer bestimmten Zahl (=c)
zu allen Einzelwerten (= x_i)
einer Verteilung
vergrößert (bzw. verkleinert)
das arithmetische Mittel
um diese Zahl,
formelhaft dargestellt:

$$(x_i + c) \Rightarrow \bar{x} + c$$

bzw.

$$(x_i - c) \Rightarrow \bar{x} - c$$

Beispiel für : $(x_i + c) \blacktriangleright \bar{x} + c$

Jährliche Reparaturaufwendungen (in €)
für ein Auto bei fünf Befragten:

100 ; 200 ; 300 ; 400 ; 500 ; n=5

$$\bar{x} = \frac{100+200+300+400+500}{5} = \frac{1500}{5} = 300$$

Deskriptivstatistik
a) Univariate Statistik

Besondere Eigenschaften des arithmetischen Mittels

Zusätzliche jährliche Benzinkosten, verursacht durch Preisteigerungen , je Auto: 80 € ; (also: c=80)

$$\bar{x} = \frac{(100+80)+(200+80)+(300+80)+(400+80)+(500+80)}{5} = \frac{1900}{5} = 380$$

Für $\bar{x} + c$ **ergibt sich dann hier für:**

$\bar{x} = 300$; $c = 80$

$\bar{x} + c$ ► **300 + 80 = 380**

Deskriptivstatistik

a) Univariate Statistik

Conclusio:

Die Maßzahlen der zentralen Tendenz sind:

- **Modus** (h): typischer Wert einer Verteilung
- **Median** ($\tilde{x}$): zentraler Wert einer Verteilung
- **Arithmetisches Mittel** ($\bar{x}$): Durchschnittswert

Maßzahlen der zentralen Tendenz und Skalenniveau

		Maßzahlen der zentralen Tendenz		
		Modus h	Median $\tilde{x}$	arithm. Mittel $\bar{x}$
Skalen-niveau	**Nominal**	X		
	Ordinal	X	X	
	Intervall-/Ratio	X	X	X

Die besprochenen Mittelwerte h , $\tilde{x}$, $\bar{X}$ dienen der Kennzeichnung empirischer Verteilungen durch **eine** Maßzahl.

Folgen:
- kleinere oder größere Ungenauigkeiten in Bezug zur gesamten Verteilung;
- Informationsverlust in Bezug zur gesamten Verteilung.

Deskriptivstatistik

a) Univariate Statistik

Maßzahlen der zentralen Tendenz

Übungsaufgabe: Bestimmen Sie den Modus, Median und das arithmetisches Mittel!

x_i	f_i
1	2
3	5
7	3
11	2

► Literaturhinweis zu den Maßzahlen der zentralen Tendenz

KROMREY, Helmut, a.a.O., 2006(11.üb.Aufl.), Kap.8.2.3:435 - 446.