Kennwerte eindimensionaler Häufigkeitsverteilungen
– Einführung –

Statistische Kennwerte von Verteilungen sind numerische Maße mit der Funktion, zusammenfassend einen Eindruck von 1) dem „Schwerpunkt", 2) der Variabilität und 3) der Form einer Merkmalsverteilung zu geben.

Man unterteilt statistische Kennwerte dazu in Maße der:

1) Zentralen Tendenz
Leitfrage: Welcher Wert kennzeichnet die Lage des Zentrums einer Merkmalsverteilung am besten?

2) Dispersion (Variabilität)
Leitfrage: Wie kann das Ausmaß an Unterschiedlichkeit (Variabilität) in den Messwerten gekennzeichnet werden?

3) Verteilungsform
Leitfrage: Welche Werte kennzeichnen die Merkmalsverteilung hinsichtlich Symmetrie und Schmalheit/Breite?

Maße der Variabilität (1)
– Einführung I –

Beispiel: Sechs Verteilungen unterschiedlicher Variabilität

1	***2***	***3***	*4*	*5*	*6*
7.0	**7.2**	**40.2**	7.0	10.0	97.8
7.0	**7.1**	**40.1**	7.0	10.0	88.5
7.0	**7.1**	**40.1**	6.0	9.0	83.4
7.0	**7.1**	**40.1**	5.0	7.0	76.2
7.0	**7.0**	**40.0**	4.0	5.0	69.9
			4.0	4.0	67.3
			4.0	3.0	58.4
			3.0	2.0	44.7
				0.0	
				0.0	

Z2 < Z3

V2 = V3

Maße der Variabilität (2)
– Einführung II –

Maße der Variabilität (Dispersion, Streuung) spielen neben den Maßen der zentralen Tendenz eine wichtige Rolle in den Sozialwissenschaften.

Drei Gründe:

1) **Beschreibende Funktion:**
Während die Maße der zentralen Tendenz angeben, welcher Wert eine Häufigkeitsverteilung am besten repräsentiert, informieren die Maße der Variabilität (Dispersion, Streuung) über die Unterschiedlichkeit bzw. Homogenität der Messwerte einer Häufigkeitsverteilung.

2) **(Relativierende) Funktion bei Vergleichen:**
Um die Abweichung eines einzelnen Messwertes von einem Mittelwert bzw. Normwert oder aber Unterschiede in zwei Mittelwerten sinnvoll interpretieren zu können, muss man Informationen über die Variabilität der untersuchten Merkmale hinzuziehen.

3) **Suche nach „Ursachen":**
Die Suche nach Erklärungen („Ursachen") von Merkmalsausprägungen wie z.B. Intelligenz, Kontrollbedürfnis, deviantes Verhalten erfolgt in den Sozialwissenschaften über die statistische Erklärung von Unterschiedlichkeit.

8(41)

Maße der Variabilität (3)
– Einführung III –

Variabilität spielt in den Sozialwissenschaften eine besondere Rolle:

Wären alle Menschen gleich, könnte das Verhalten einer ganzen Population anhand des Verhaltens eines einzigen Individuums vorhergesagt werden und weitere Forschung wäre überflüssig; alles, was für eine x-beliebige Person gelten würde, würde für alle anderen Personen gleichermaßen gelten.

Dies ist aber nicht der Fall: Die entscheidende Aufgabe sozialwissenschaftlicher Forschung ist es, Variation zu erklären: Warum sind manche Jugendliche gewalttätiger als andere, warum verhalten sich Menschen in unterschiedlichen Situationen unterschiedlich etc.

Abgesehen von der relativierenden Funktion von Maßen der Variabilität (oder Dispersion) beim Vergleich eines einzelnen Wertes mit dem Normwert, ist die statistische Erklärung von Unterschiedlichkeit eine zentrale Funktion der Sozialwissenschaften.

9(41)

Maße der Variabilität (4)
– Einführung IV –

Beispiel:

Stellen Sie sich vor, sie hätten in einem Leistungstest einen Punktwert von 75 erzielt.

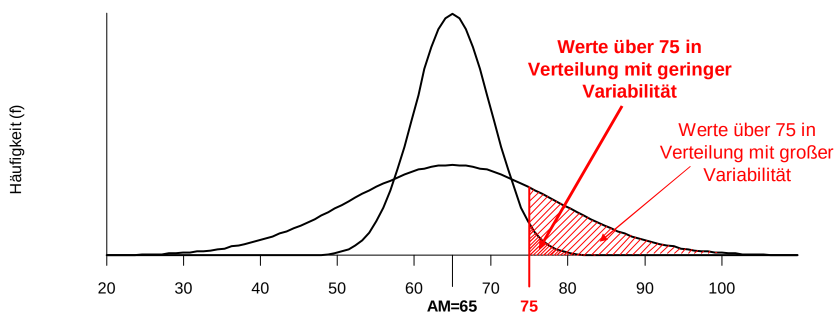
Die mittlere Punktzahl bei diesem Test liege bei 65.

Die minimal mögliche Punktzahl sei 0.

Die maximal mögliche Punktzahl sei 100.

Maße der Variabilität (5)
– Einführung V –

Beispiel: Häufigkeitspolygone zweier Verteilungen mit gleichem Mittelwert und unterschiedlicher Variabilität



Maße der Variabilität (6)

– Maße für mind. nominalskalierte Merkmale I –

1) ***Variationsratio VR*** *(= variation ratio)*

Ausmaß, in dem sich Fälle auf den typischen Wert konzentrieren:

$$VR = 1 - \left(\frac{N_{Modalkategorie}}{N_{total}} \right)$$

Nachteile:

Nicht geeignet für multimodale Verteilungen,

abhängig von der Anzahl der Kategorien

15(41)

Maße der Variabilität (7)

– Maße für mind. nominalskalierte Merkmale II –

2) ***Index qualitativer Variation IQV*** *(= index of qualitative variation)*

Prozentualer Anteil der beobachteten Variation an der maximal möglichen Variation (gegeben die Anzahl der Fälle und Kategorien):

$$IQV = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_{beob_i} \cdot N_{beob_j}}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} N_{erw_i} \cdot N_{erw_j}}$$

N_{erw} entspricht dabei der Anzahl der Fälle, die wir in einer Kategorie erwarten würden, wenn alle Fälle gleichmäßig über alle Kategorien verteilt wären

$$N_{erw} = \frac{N_{total}}{\text{Anzahl Kategorien}}$$

Eine einfachere und äquivalente Formel lautet:

$$IQV = \frac{k \cdot (N^2 - \sum f^2)}{N^2 \cdot (k-1)}$$

16(41)

Maße der Variabilität (8)
– Maße für mind. nominalskalierte Merkmale III –

Beispiel: Hinrichtungsmethoden in den USA 1977 – 2000

Kategorie	Häufigkeit (N)
Tödliche Injektion	518
Elektrokution	149
Tödliches Gas	11
Erhängung	3
Erschießung	2
total	683

$$IQV = \frac{k \cdot (N^2 - \sum f^2)}{N^2 \cdot (k-1)}$$

$$IQV = \frac{5 \cdot (683^2 - (518^2 + 149^2 + 11^2 + 3^2 + 2^2))}{683^2 \cdot (5-1)}$$

$$= 0.471$$

Maße der Variabilität (9)
– Maße für mind. ordinalskalierte Merkmale I –

(1) ***Spannbreite*** (= Spannweite, Variationsweite, *range*)
Differenz zwischen größtem und kleinsten Wert:

$$\text{Range} = x_{max} - x_{min}$$

Nachteile:
Sehr anfällig für Ausreißerwerte,
stark abhängig von Stichprobengröße,
kein Bezug zu einem Maß der zentralen Tendenz (Modus, Median, arithmetisches Mittel)

Maße der Variabilität (10)
– Maße für mind. ordinalskalierte Merkmale II –

Obwohl dies Verfahren intuitiv einleuchtet, ist diese Kennziffer häufig irreführend, da die extremen Werte einer Verteilung häufig „exzentrisch“ und untypisch für den Rest der Verteilung sind.

Beispiel:

Verteilung A	10	10	9	7	5	4	3	2	0	0
Verteilung B	10	6	5	5	5	5	5	5	4	0

Im o.g. Beispiel ist die Spannweite der Verteilungen A und B jeweils 10 obwohl in Verteilung B die Werte 0 und 10 sehr viel extremere Werte darstellen als in Verteilung A.

Maße der Variabilität (11)
– Maße für mind. ordinalskalierte Merkmale III –

(2) ***Interquartilsabstand Q_{diff}*** *(= inter-quartile range IQR)*
Differenz zwischen dem dritten und ersten Quartil (bzw. zwischen dem 75. Centil und 25. Centil) der Verteilung:

$$Q_{diff} = Q_3 - Q_1 = C_{75} - C_{25}$$

Bereich, in dem 50 % aller Werte liegen; wenig anfällig für Ausreißerwerte

Davon abgeleitet:
Mittlerer Quartilsabstand QD (= mittlere Quartilsdistanz, *semi-interquartile range*):

$$QD = (Q_3 - Q_1)/2$$

Maße der Variabilität (12)

– Maße für mind. intervallskalierte Merkmale I –

(1) ***AD-Streuung*** (= mittlere absolute Abweichung, *average deviation*, Abweichungsdurchschnitt; Fechners *e*)

Durchschnittliche Abweichung der Messwerte vom arithmetischen Mittel der Verteilung unter Verwendung des Absolutbetrages der Abweichungen:

$$AD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Bei gruppierten Daten gilt entsprechend (mit f_i = Häufigkeit der *i*-ten Klassenmitte und x_i = Kategorienmitte der Klasse *i*)

$$AD = \frac{\sum_{i=1}^{n} f_i \, |x_i - \bar{x}|}{n}$$

Maße der Variabilität (13)

– Maße für mind. intervallskalierte Merkmale II –

Würde statt des arithmetischen Mittels der Median gewählt, wäre das Maß der AD-Streuung minimal (d.h. unter Verwendung des arithmetischen Mittels ist das Maß des Abweichungsdurchschnitts *nicht* der kleinstmögliche Wert).

Aufgrund dieser Minimumseigenschaft und der Benutzung der Absolutbeträge hat sich diese Kennziffer der Variabilität für die Verwendung in weiteren statistischen Analysen nicht so gut bewährt wie die im folgenden dargestellte Kennziffern der Varianz bzw. der Standardabweichung.

Maße der Variabilität (14)
- Maße für mind. intervallskalierte Merkmale III –

(2) ***Varianz s²*** (= *variance*)

Durchschnittliche quadrierte Abweichung der Messwerte vom arithmetischen Mittel:

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

Davon abgeleitet:

Standardabweichung s (= StdDev; *standard deviation*; Streuung)

Positive Quadratwurzel aus der Varianz:

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

24(41)

Maße der Variabilität (15)
– Maße für mind. intervallskalierte Merkmale IV –

Die durchschnittliche quadrierte Abweichung von einer Konstanten ist minimal, wenn für diese Konstante das arithmetische Mittel gewählt wird.

Aufgrund dieser Eigenschaften ist die Varianz bzw. die Standard-abweichung eine zentrale Kennziffer der Statistik, die auch in anderen statistischen Verfahren eine besondere Bedeutung hat.

25(41)

Maße der Variabilität (16)

– Maße für mind. intervallskalierte Merkmale V –

Achtung: Die hier dargestellten Kennziffern der Varianz und der Standardabweichung dienen dazu, die Variabilität einer Stichprobe (oder einer kompletten Population) zu *beschreiben*.

Würde diese Kennziffer dazu benutzt werden, aufgrund der Daten einer Stichprobe die Variabilität der Werte in der Population zu *schätzen*, würde die Variabilität der Populationswerte systematisch unterschätzt (verzerrt) werden.

Deshalb wird in fast allen Statistikprogrammpaketen eine andere Formel für die Varianz bzw. Standardabweichung benutzt, in der diese Verzerrung dadurch korrigiert wird, dass im Nenner (n – 1) an Stelle von n benutzt wird (dabei wird die geschätzte Populationsvarianz bzw. Standardabweichung durch den griechischen Buchstaben Sigma mit einem *accent circonflex* – häufig „Sigma Dach" ausgesprochen – gekennzeichnet):

$$\text{geschätzte Populationsstandardabweichung} = \hat{\sigma}_x = \sqrt{\hat{\sigma}_x^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

26(41)

Maße der Variabilität (17)

– Maße für mind. intervallskalierte Merkmale VI –

Da die meisten Statistikprogrammpakete nur die korrigierten Populationsschätzer berechnen, muss – falls die unkorrigierte Varianz oder Standardabweichung berechnet werden sollte – die Korrektur wie folgt rückgängig gemacht werden:

$$s_x^2 = \frac{n-1}{n} \cdot \hat{\sigma}_x^2$$

bzw.

$$s_x = \sqrt{\frac{n-1}{n}} \cdot \hat{\sigma}_x$$

27(41)

Maße der Variabilität (18)
– Effekte von Transformationen –

Effekte von Transformationen auf arithmetisches Mittel, Standardabweichung und Varianz

Operation	Effekt auf arithmetisches Mittel	Effekt auf Standardabweichung	Effekt auf Varianz
Addition einer Konstanten *k* zu jedem Wert	neuer Mittelwert = alter Mittelwert **+ *k***	keiner	keiner
Subtraktion einer Konstanten *k* von jedem Wert	neuer Mittelwert = alter Mittelwert **− *k***	keiner	keiner
Multiplikation jedes Wertes mit einer Konstanten *k*	neuer Mittelwert = alter Mittelwert · ***k***	neue s = alte s · **|*k*|**	neue s^2 = alte $s^2 \cdot k^2$
Division jedes Wertes durch eine Konstante *k*	neuer Mittelwert = alter Mittelwert **/ *k***	neue s = alte s **/ |*k*|**	neue s^2 = alte s^2 / k^2

41(41)