Statistik Kurs SS 2004

3.Tag

Grundlegende statistische Maße

Mittelwert (mean) $\bar{x}$
Durchschnitt aller Werte

Varianz (variance) s^2
Durchschnittliche quadrierte Abweichung aller Werte vom Mittelwert

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

>> Die Varianz ist ein Maß für die Variabilität der Werte
(anders formuliert: Die Varianz ist ein Maß dafür, wie die einzelnen Daten um den Mittelwert verteilt sind.)

Standardabweichung (standard deviation, SD, Streuung) s
Quadratwurzel der Varianz

$$s = \sqrt{s^2}$$

Schreibweisen & Erläuterungen:

Die **Standardabweichung** s(*x*) oder sdv(*x*) ist die Wurzel aus der Varianz und wird oft als mittlerer quadratischer Fehler der Einzelwerte bezeichnet:

$$\mathbf{s}(x) = \mathbf{sdv}(x) = \sqrt{\mathbf{var}(x)} = \sqrt{s^2(x)}$$

Notation

Maß	Stichprobe	Grundgesamtheit
Mittelwert	$\bar{x}$	μ
Standardabweichung	s	σ
Varianz	s^2	σ^2
Umfang	n	N

μ heißt auch Erwartungswert.

(Arithmetisches) Mittel

- Maß für die Lage der Verteilung
  - *für metrisch skalierte Daten*
- „Durchschnittlicher Wert“ der Daten
- empfindlich gegen „Ausreißer“

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Bsp.:
ermittelte Daten: 1,2,3

~> $\bar{x}$ = (1+2+3) / 3 = **2**

Bestimmung des *gewichteten Mittels* anhand von z.B. folgendem Daten, die in einer fiktiven Klasse ermittelt wurden:

Anzahl der Brüder	0	1	2	3	4
Häufigkeit des Auftretens	8	4	3	2	0

⇨ n = 8 + 4 + 3 + 2 = 17 (Anzahl Schüler)

⇨ $\bar{x}$ = 1/17 (0*8 + 1*4 +2*3 +3*2 +4*0) = 0,94

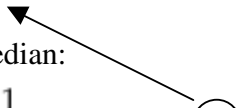
Median

- Maß für die Lage der Verteilung
  - *‚für ordinal skalierte Daten*
- Mittlere Wert der geordneten Daten
  - ungerade n: Mittlerer Wert
  - gerade n: Durchschnitt der beiden mittleren Werte

  => Position des Median = $\frac{n+1}{2}$
- unempfindlich (robust) gegen „Ausreißer“

~> Vorgehen:

Rohdaten: 3 14 25 65 34 4 12 => n: ungerade (Anzahl der Messwerte)
sortiert: 3 4 12 14 25 34 65
Position: 1 2 3 4 5 6 7



Bestimmung Median:

Median = $\frac{n+1}{2}$ = (7+1) : 2 = 4

- **14**

- Bei geradem n anschließende Ermittlung des Median-Wertes durch Bestimmung des Mittelwertes der beiden mittleren Werte der Rangheihenfolge:

  Daten: 3 5 8 9
  Sortiert: 1 2 3 4
  => (5 + 8) : 2 = **6,5**

Modus

- Maß für die Lage der Verteilung
  - ‚ebenso *für nominal skalierte Daten*
- häufigster Wert der Meßreihe
  - möglich sind kein, einer oder mehrere Modi
- Verwendung bei qualitativen und quantitativen Daten möglich
- unempfindlich (robust) gegen „Ausreißer“

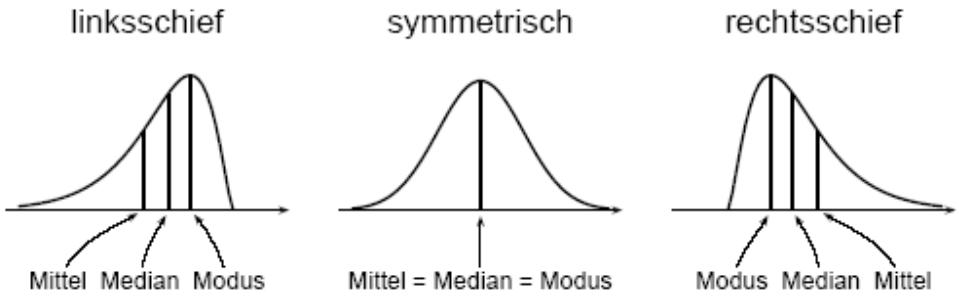
	Rohdaten					
Kein Modus:	4	8	10	11	16	17
Ein Modus:	4	67	77	77	88	91
Mehrere Modi:	6	6	7	8	12	12

=> All dies sind Möglichkeiten die Lage der Verteilung zu beschreiben.

Zusammenfassung:

Begriff	Formel	Beschreibung	Skalenniveau
Mittel (arithm.)	$\bar{x} = \frac{\sum x_i}{n}$	„Durchschnitt“	metrisch
Median	Position $= \frac{n+1}{2}$	mittlerer Wert bei geordneten Daten	ordinalsk.
Modus		häufigster Wert	nominalsk.

Formen der Verteilung:



Noch interessant hierbei weitere Formen des Mittelwertes:

’Getrimmte’ Mittelwert

Im Gegensatz zu dem Median ist der Mittelwert – insbesondere bei kleinen Stichproben – von extremen Werten stark abhängig.

Bsp (1): Reaktionszeiten in ms: (200 300 400 500 600)

$\bar{x}$ = 400

Md(x) = 400

Bsp (2): Reaktionszeiten in ms mit extrem großem Wert. (200 300 400 500 8000)

$\bar{x}$ = 1880

Md(x) = 400

Bei der Berechnung des 'getrimmten Mittelwerts' werden die extremen Werte nach oben und unten weggelassen (abgeschntten).
Wenn z.B. 20 % der Werte je von oben und unten abgeschnitten werden (also insgesamt 40 %) und von dem Rest der Daten der Mittelwert berechnet wird, dann heißt der entsprechende Mittelwert $x_{0.2}$.

Bsp.:
Bei dem obigen Beispiel (2) die 'Extremwerte' 200 und 8000 weggelassen, so ergibt sich für $x_{0.2} = 400$
Im Extremfall $x_{0.4} = 400$ ist der getrimmte Mittelwert gleich dem Median.

Das arithmetische Mittel ist zwar der am häufigsten berechnete Mittelwert, aber nicht der einzige:

ebenso noch: **Median**, **Median**, **geometrische Mittel**

<u>Streuungsmaße</u>

- Unter der **Spannweite** versteht man die Differenz vom größtem und kleinstem Wert
- **Quartile** (werden wir als Grafik anschaulich kennen lernen)
- **mittlere Abweichung**
- **Varianz oder mittlere quadratische Abweichung**

Die Wurzel aus der Varianz nennt man **Standardabweichung s**.

- Die Varianz lässt sich einfacher auch folgendermaßen berechnen: $s^2 = 1/n \, \Sigma \, x_i^2 - \overline{x}^2$

>> Die Standardabweichung ist das am häufigsten verwendete Streuungsmaß

Statistische Kennwerte

Perzentilwerte

Quartile	Zeigt das 1., 2. und 3. Quartil an. 1. Quartil (Q1): ist derjenige Punkt der Messwertskala, unterhalb dessen 25% der Messwerte liegen 2. Quartil (Q2): ist derjenige Punkt der Messwertskala, unterhalb dessen 50% der Messwerte liegen. * bezeichnet man auch als Median * 3. Quartil: ist derjenige Punkt der Messwertskala, unterhalb dessen 75% der Messwerte liegen. * liegen die Daten nur in Form einer Rangordnung vor, wird der mittlere Quartilabstand als Streuungsmaß benutzt ~> Definiert als Q = (Q3-Q1): 2.
Perzentile	z.B. die Werte 25, 50 und 75, so erhält man Quartile. Beliebige Perzentilwerte, z.B. 39 und 63. Im ersten Fall (39) wird der Wert der gewählten Variablen angezeigt, unterhalb dessen 39% der Werte liegen, im zweiten Fall (63) ist es der Wert, unterhalb dessen sich 63% der Werte befinden (z.B. mit SPSS)

Streuungsmaße:

Standardabweichung	Standardabweichung ist ein Maß für die Streuung der Messwerte - sie ist die Quadratwurzel aus der Varianz * trägt man die Standardabweichung zu beiden Seiten des Mittelwertes auf, so liegen bei normalverteilten Werten ca. 67% der Werte in diesem Intervall *
Varianz	Varianz ist das Quadrat der Standardabweichung und somit ebenfalls ein Maß für die Streuung der Messwerte * Berechnet aus der Summe der Abweichungsquadrate aller Messwerte von ihrem arithmetischen Mittel, dividiert durch die um 1 verminderte Anzahl der Werte
Spannweite	Spannweite ist die Differenz zwischen dem größten (Maximum) und dem kleinsten Wert (Minimum) der Messreihe
Minimum	kleinste ermittelte Wert

Maximum:	größte ermittelte Wert
Standardfehler	handelt sich hierbei um den Standardfehler des Mittelwertes. * trägt man den Standardfehler zu beiden Seiten des Mittelwertes auf, liegt mit etwa 67%-iger Wahrscheinlichkeit der Mittelwert der Grundgesamtheit in diesem Intervall * * Berechnung: Standardabweichung dividiert durch die Quadratwurzel des Stichprobenumfangs

Lagemaße

Mittelwert	Mittelwert ist der arithmetische Mittel der Messwerte und berechnet sich so aus der Summe der Messwerte geteilt durch ihre Anzahl N Summe von 6 ermittelten Messwerten: 600, so folgt: Mittelwert x = 600:6 = 100
Median	Median ist derjenige Punkt der Messreihe unterhalb und oberhalb dessen jeweils die Hälfte der Messwerte liegen * z.B. 2 5 8 5 8 6 3 2 2 1 7 >> so schreibt man diese zunächst der Größe nach sortiert auf: 1 2 2 2 3 5 5 6 7 8 8 Median: 5 Es liegen insgesamt n = 11 Messwerte vor, so dass der 6te Messwert der Median ist. So liegen nämlich 5 Messwerte unterhalb und 5 Messwerte oberhalb dieses Wertes. - bei ungeradem n ist der Median also ein tatsächlich auftretender Messwert. - bei geradem n ist der Median das arithmetische Mittel der mittleren benachbarten Messwerte. * z.B. 3 3 4 5 6 7 7 8 8 9 >> so Median: (6 + 7): 2 = 6.5
Modalwert	Modalwert ist der am häufigsten auftretende Wert in den Stichproben. * haben mehrere Werte dieselbe maximale Häufigkeit, so mehrere Modi
Summe	Summe aller Werte

Ergänzung für mathematisches Verstääöndnis:

Mathematische Definition der Standardabweichung

$$\sigma := \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

σ: Standardabweichung
μ: Erwartungswert
N: Umfang der Grundgesamtheit
x_i: Merkmalsausprägungen am *i*-ten Element der Grundgesamtheit

$$\mu := \frac{1}{N}\sum_{i=1}^{N} x_i$$

Die Standardabweichung σ der Grundgesamtheit kann aus einer Stichprobe auf verschiedene Weise geschätzt werden:

Schätzung der Standardabweichung aus einer Stichprobe

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

=> Formel ist eine Schätzung, die durch keine besondere Eigenschaft ausgezeichnet ist.

S: Schätzung für die Standardabweichung
X_i: Merkmalsausprägung am *i*-ten Element der Stichprobe
$\bar{X}$: arithmetische Mittelwert der Stichprobe (~> damit erwartungstreue Schätzung für den Erwartungswert).
n: Stichprobenumfang

Bsp.:
Stichprobe mit 5 ermittelten Werten: 3, 4, 5, 6, 7

~> Schätzung für die Standardabweichung ergibt:

⇨ Summe der quadrierten Abweichungen vom Stichprobenmittelwert
Wert: 10
~> um eine *erwartungstreue Schätzung* für die Varianz zu erhalten, wird diese Summe durch die Zahl der Messwerte n weniger 1 geteilt, wie oben in Formel ersichtlich

Hieraus ergibt die Schätzung für die Varianz 10/4 = **2,5**
Daraus nimmt man die Quadratwurzel:

$\sqrt{2,5} \approx 1,581$ Ergibt die gebräuchliche Schätzung für Standardabweichung: 1,581

Begriff der „Freiheitsgrade“ ☺

Freiheitsgrade (***df***, degrees of freedom) sind die Zahl der Elemente einer Stichprobe, die frei variieren können.

(In einer (**theoretischen**) Population entspricht die Zahl der **Freiheitsgrade** einfach der Zahl der Elemente: ***df*** = *n*)

Wenn man weiß, dass die Elemente bestimmte Randbedingungen erfüllen müssen, gehen **Freiheitsgrade** verloren: ein Freiheitsgrad pro einzuhaltende Randbedingung (Verminderung der anzunehmenden Zustände, Analogie etwa ein 6-seitiger Würfel mit ***df*** =6 hin zu einem 5-seitigen mit ***df*** =5)
Man könnte auch sagen, Anzahl der freien möglichen Zustände.

Beispiel: In der Formel für die empirische Varianz steht (*n-1*) statt *n*, weil die Varianz ja relativ zum Mittelwert berechnet werden muss – einer der Messwerte ist also durch die anderen genau festgelegt, wenn alle zusammen den Mittelwert ergeben sollen.

Viele Wahrscheinlichkeitsverteilungen hängen von der Zahl der **Freiheitsgrade** ab!

8.6.04
Rouven Saleika
☺