

7. Streuungsmaße

Nur durch das Extreme hat die Welt ihren Wert, nur durch das Durchschnittliche ihren Bestand

Paul Valery

7.1 Varianz und Standardabweichung

Die Varianz ist der wichtigste Streuungsparameter. Analog zum arithmetischen Mittel gilt: Anwendung immer dann, wenn kardinalskalierte Daten vorliegen und keine der in den Unterabschnitten ab Kap. 7.2 aufgeführten Gründe für die Wahl anderer Streumaße sprechen.

Sei eine Urliste kardinalskalierter Merkmalsausprägungen x_i mit arithmetischem Mittel \bar{x} gegeben. Dann ist die Varianz das arithmetische Mittel der quadrierten Abweichungen von \bar{x} :

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

Aus einer Häufigkeitstabelle bzw. aus klassierten Daten berechnet man s_x^2 analog wie beim Mittelwert:

$$s_x^2 = \frac{1}{n} \sum_{j=1}^m (x_j - \bar{x})^2 h_j \quad \text{aus der Häufigkeitstabelle}$$

$$s_x^2 \approx \frac{1}{n} \sum_{k=1}^K (x_k^* - \bar{x})^2 h_k \quad \text{aus klassierten Daten}$$

7.2. Eigenschaften der Varianz

- Wie beim arithmetischen Mittel gilt: Geeignet zur Beschreibung von eingipfligen und nicht zu unsymmetrischen Verteilungen.
- Es ist empfindlich gegenüber Ausreißern. *Beispiel:* Für die Güte des Verkehrsflusses ist die Varianz der Geschwindigkeiten eine wichtige Messgröße. Messungen mit Mess-Schleifen ergaben folgende Werte (in km/h): 98 107 102 104 99 99 91 100 0 100. Offensichtlich ist bei einer Messung der Detektor ausgefallen (n.B.: warum ist hier diese Folgerung sogar zwingend?). Errechnen Sie, wie dieser Ausreißer die gemessene Varianz verfälscht!
- Arithmetisches Mittel und Varianz “gehören” insofern “zusammen”, dass der Mittelwert die Quadratsumme der Abweichungen minimiert und diese dann die mit n multiplizierte Varianz darstellt:

$$F = \sum_{i=1}^n (x_i - c)^2 = \min = ns_x^2, \quad \text{falls } c = \bar{x}$$

7.2(b) Mathematische Eigenschaften der Varianz

- Zur leichteren Berechnung “per Hand” nutzt der **Verschiebungssatz**:

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

Für Häufigkeitstabellen und klassierte Daten ersetzt man wie üblich: $\sum_{i=1}^n \Rightarrow \sum_{j=1}^m h_j$ für Häufigkeitstabellen, sowie $\sum_{i=1}^n \Rightarrow \sum_{k=1}^K h_k$ und $x_i \Rightarrow x_k^*$ für klassierte Daten.

Übungsbeispiel: Gegeben ist die geordnete Urliste vom vorherigen Abschnitt: 2 4 5 5 6 6 6 7 8 8 9 9 10 13. Ermitteln Sie die Varianz

- (i) direkt aus der Urliste,
 - (ii) aus der dazugehörigen Häufigkeitstabelle,
 - (iii) nach Klassierung der Daten ($d = 3, x_1^u = 0.5$)
- Hängt die Größe Y linear von X ab, dann kann man die Varianz von Y direkt aus der für X berechnen:

$$Y = a + bX \quad \Rightarrow \quad s_y^2 = b^2 s_x^2$$

Insbesondere hängt s_y nicht von a ab!

Beispiel: Die gemessene Geschwindigkeitsvarianz beträgt 5 $(\text{m/s})^2$. Wie groß ist sie in $(\text{km/h})^2$ ausgedrückt?

7.2(b) Mathematische Eigenschaften der Varianz II

Bisweilen liegen zwei unabhängige Datensätze x_i , $i = 1, \dots, n_x$ und y_i , $i = 1, \dots, n_y$ desselben Merkmals mit gegebenen Mittelwerten \bar{x} , \bar{y} und gegebenen Varianzen s_x^2 , s_y^2 vor. Dann gilt für den gesamten aus x_i und y_i bestehenden Datensatz z_i (Umfang $n = n_x + n_y$):

$$\begin{aligned}\bar{z} &= f_x \bar{x} + f_y \bar{y}, \\ s_z^2 &= f_x [s_x^2 + (\bar{x} - \bar{z})^2] + f_y [s_y^2 + (\bar{y} - \bar{z})^2]\end{aligned}$$

mit den relativen Gewichten (relativen Häufigkeiten) $f_x = n_x/n$ und $f_y = n_y/n$.

Beispiel: Auf der rechten Spur einer zweispurigen Autobahn misst man einen Verkehrsfluss von 1000 Fahrzeugen/h, eine mittlere Geschwindigkeit von 90 km/h und eine Standardabweichung $\sqrt{s^2}$ von 10 km/h. Auf der linken Spur (2000 Fahrzeuge/h) misst man ein Mittel von 120 km/h und eine Standardabweichung von 20 km/h. Wie groß sind Mittel und Standardabweichung für den Gesamtverkehr?

7.2(c). Standardabweichung

Da alle Einheiten der Merkmalsausprägungen bei der Varianz quadriert werden, ist sie nicht anschaulich. Ein Maß für die tatsächliche Streubreite gibt dagegen die

$$\text{Standardabweichung} \quad V_x = s_x = \sqrt{s_x^2}$$

- Für beliebige Verteilungen gilt: Höchstens ein Anteil p der Werte liegt außerhalb des Intervalls $[\bar{x} - s/\sqrt{p}, \bar{x} + s/\sqrt{p}]$. (manche Bücher enthalten schärfere Formulierungen, die jedoch i.A. falsch sind)

Beispiel: Höchstens 1 % der Werte können mehr als 10 Standardabweichungen vom Mittelwert entfernt sein.

- Sind die Merkmalsausprägungen annähernd gaußverteilt (mit Varianz σ^2), gilt: 68% der Werte sind innerhalb $[\bar{x} \pm \sigma]$, 95% innerhalb $[\bar{x} \pm 2\sigma]$, 99.7% innerhalb $[\bar{x} \pm 3\sigma]$.

Beispiel: Ein Hersteller von “Radarfallen” gibt für seine Geräte eine Genauigkeit von ± 2 km/h an. Wenn nichts weiter gesagt wird, ist damit die Standardabweichung gemeint. Bei Annahme einer Gaußverteilung haben damit 32% der Messungen einen Messfehler > 2 km/h! Deshalb gibt man häufig eine “3- σ -Grenze” an, innerhalb der 99% aller Messungen liegen.

7.2(d) Variationskoeffizient

Der

$$\text{Variationskoeffizient } V_x = \frac{s_x}{\bar{x}}$$

ist ein Maß für die relative Streuung und wird vor allem bei Streuungs-*Vergleichen* angewandt.

- Im Gegensatz zu s_x ist V_x eineitenlos. Als Faustregel gilt: ist $V_x > 50\%$, dann ist das arithmetische Mittel kein guter Repräsentant der Einzelwerte mehr.
- Er beschreibt direkt die tatsächliche Situation und hängt nicht von willkürlichen Maßeinheiten ab.

Beispiel 1: Auf einer deutschen Autobahn wurden Geschwindigkeiten von (100 ± 10) km/h gemessen, auf einer amerikanischen (55 ± 6) mph. Fließt anhand dieser Daten der Verkehr auf der amerikanischen Autobahn ruhiger?

Beispiel 2: Die auf ein Jahr bezogene relative Schwankung V_{DAX} des DAX (auf beliebiger Börsen-Website VDAX eingeben) ist deutlich kleiner als die des TECDAX, obwohl es sich mit den absoluten Schwankungen genau umgekehrt verhält.

- Er ist nur für positive und mindestens verhältnisskalierte Merkmalsausprägungen sinnvoll.

Gegenbeispiel: Im letzten Monat war der Mittelwert der (nur intervallskalierten) Celsius-Temperatur $\bar{\theta} = 10^\circ\text{C}$ und die Standardabweichung $s_\theta = 7^\circ\text{C}$. Einen Variationskoeffizient $V_\theta = 0.7$ zu berechnen wäre sinnlos. Dies wird spätestens für den Februar mit $\bar{\theta} = 0^\circ\text{C}$ klar ...

7.2(e) Standardisierung

Will man verschiedene Merkmale mit unterschiedlichen Einheiten in einer Untersuchung vergleichen (multivariate Analyse, Kap. 18-21) oder tabellierte Verteilungen bei der Berechnung von Wahrscheinlichkeiten nutzen, ist eine Standardisierung der Merkmalswerte nötig:

Die lineare Transformation $X \Rightarrow Y = a + bX$ heißt **Standardisierung** des kardinalskalierten Merkmals X , wenn für die Variable Y gilt: $\bar{y} = 0$, $s_y^2 = 1$.

Aufgabe: Ermitteln Sie a und b !

7.3 Mittlere absolute Abweichung

Anstelle der Summe der quadrierten Abweichungen kann man auch die Summe der Beträge der Abweichungen (“lineare Streuung” bzw. “mean absolute deviation”, MAD als Streuungsmaß definieren:

$$s_{\text{MAD}} = \frac{1}{n} \sum_{i=1}^n |x_i - x_{0.5}|$$

- Eine Anwendung ist dann sinnvoll, wenn die MAD unmittelbare Bedeutung hat wie bei Wegezeiten.
- Der Median $x_{0.5}$ und die MAD “gehören” insofern “zusammen”, dass der Median die Betragssumme der Abweichungen minimiert und diese dann die mit n multiplizierte mittlere absolute Abweichung darstellt:

$$F = \sum_{i=1}^n |x_i - c| = \min = \frac{s_{\text{MAD}}}{n}, \quad \text{falls } c = x_{0.5}.$$

Daher wird s_{MAD} üblicherweise bezüglich des Medians definiert.

- Die MAD ist empfindlich gegenüber Ausreißern, aber weniger ausgeprägt als bei der Standardabweichung.

Beispiel: Berechnen Sie s_{MAD} bei der schon bekannten Geschwindigkeitsreihe 98 107 102 104 99 99 91 100 0 100 mit und ohne Ausreißer.

7.4 Interquartilsabstand und Spannweite

Die Spannweite ist die Differenz zwischen dem größten und dem kleinsten Merkmalswert:

$$R = \max(x_i) - \min(x_i) \quad \text{bzw.} \quad R = x_K^o - x_1^u$$

Der Interquartilsabstand ist der Abstand zwischen dem dritten und dem ersten Quartil,

$$s_{IQ} = x_{0.75} - x_{0.25}$$

- Während die Spannweite extrem anfällig gegenüber Ausreißern ist, ist der Interquartilsabstand als einziges der behandelten Streuungsmaß unempfindlich dagegen (Test an der obigen Geschwindigkeitsreihe!).
- Anwendung:
 - bei ordinalskalierten Daten, bei denen man weder s_x^2 noch s_{MAD} berechnen kann,
 - wenn Unempfindlichkeit gegenüber Ausreißern wichtig ist (s_{IQ}).
- Visualisierung im “Box-and-Whisker Plot”
- Für beliebige Verteilungen gilt: $s_{MAD} \leq s_x \leq R$

8. Maßzahlen für die Form der Verteilung

Für **multimodale** Häufigkeitsverteilungen sind dies vor allem Lage (d.h. Modus x_D) und Höhe aller "Gipfel".

Für **unimodale** Verteilungen gibt es zwei Kategorien von Kennzahlen:

- Asymmetriemaße, z.B. "Schiefe" Γ ,
- Wölbungsmaße, z.B. "Exzess" (Kurtosis) K .¹

Zur Definition von Schiefe und Exzess werden die zentralen Momente benötigt:

Das N -te zentrale Moment von n kardinalskalierten Merkmalswerten ist gegeben durch

$$M_N = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^N \text{ bzw. } M_N = \sum_{i=1}^m (x_i - \bar{x})^N f_i.$$

Damit definiert man diese Formmaße:

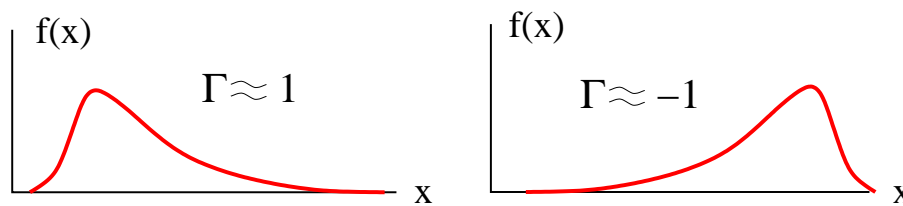
$$\text{Schiefe } \Gamma = \frac{M_3}{s_x^3}, \quad \text{Exzess } K = \frac{M_4}{s_x^4} - 3.$$

Man spricht von einer *rechtsschiefen* bzw. *linkssteilen* Verteilung, falls $\Gamma > 0$. Die Varianz ist übrigens gleich M_2 .

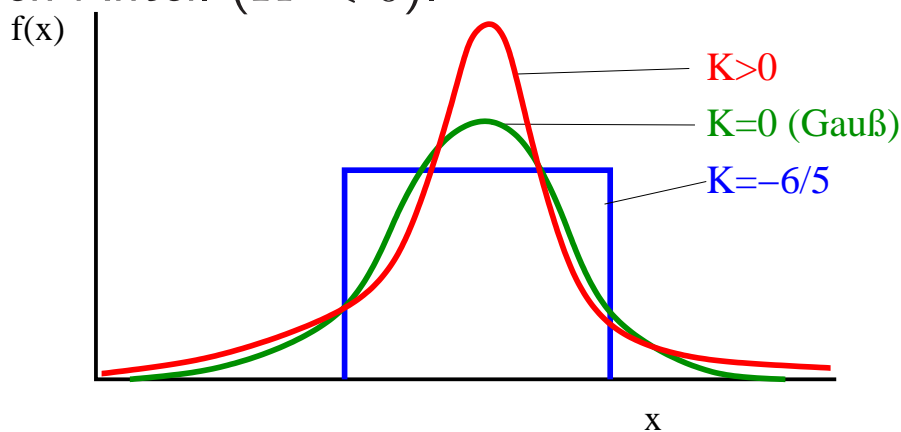
¹ Wegen Verwechslungsgefahr mit der Standardabweichung bzw. dem Zeichen für Erwartungswert werden Schiefe und Exzess nicht mit S bzw. E bezeichnet.

8.2 Veranschaulichung von Schiefe und Exzess

Die Schiefe gibt an, ob die Werte der Verteilung vom Modus aus links ($\Gamma > 0$, linkssteil bzw. rechtsschief) oder rechts ($\Gamma < 0$, rechtssteil bzw. linksschief) schneller abfallen; das "lange Ende" der Verteilung ist jeweils auf der anderen Seite:



Der Exzess, auch Kurtosis genannt, gibt an, ob es, im Vergleich zur Gaußverteilung, einen höheren Anteil von "Ausreißern" (d.h. mehrere Standardabweichungen vom Mittel abweichende Werte) gibt ($K > 0$) oder einen geringeren Anteil ($K < 0$).



Neben vielen Ausreißern gibt es bei Verteilungen mit positivem Exzess auch viele Werte nahe dem Mittel.

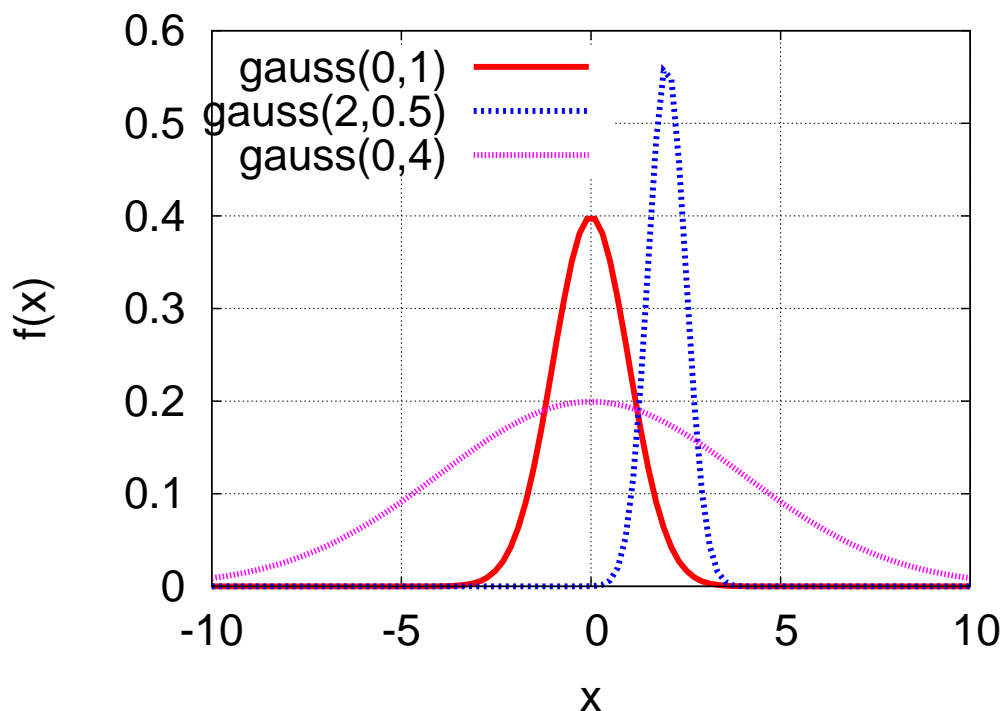
Wichtig: Die mittlere *Größe* der Abweichungen wird durch s gemessen; hier geht es rein um die *Form* mit der Gaußverteilung als Referenz: Die -3 in der Formel für den Exzess kommt daher, dass bei der Gaußverteilung $M_4/s_x^4 = 3$ gilt.

8.3 Eigenschaften von Schiefe und Exzess

- Im Gegensatz zu \bar{x} und s_x sind sowohl Γ als auch K relative Größen und somit einheitenlos.
- Beide Größen hängen weder von Verschiebungen noch von Skalierungen der Verteilung ab. Genauer:

$$Y = a + bX \Rightarrow \Gamma_y = \Gamma_x, \quad K_y = K_x.$$

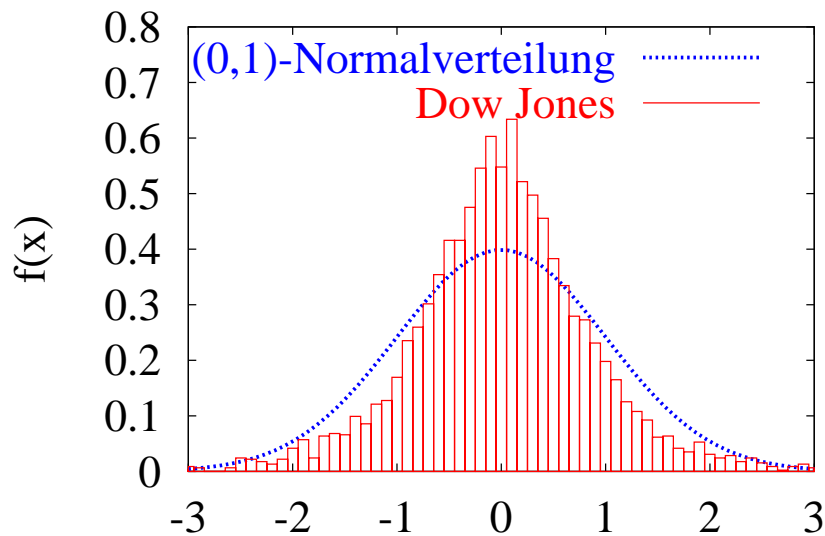
Aufgabe: Zeigen Sie dies.



Alle drei Kurven sind Gaußverteilungen und haben damit dieselbe Schiefe $\Gamma = 0$ und denselben Exzess $K = 0$.

8.3(b) Beispiel

Betrachtet wird die Häufigkeitsverteilung der prozentualen täglichen Wertänderungen des Dow Jones:



auf (0,1) normierte taegliche Wertaenderung X

Es gibt häufiger große negative Veränderungen (Crashes) als große positive (Boom-Phasen); damit fällt die Verteilung der Wertänderungen auf der rechten Seite schneller ab und die Schiefe ist negativ.

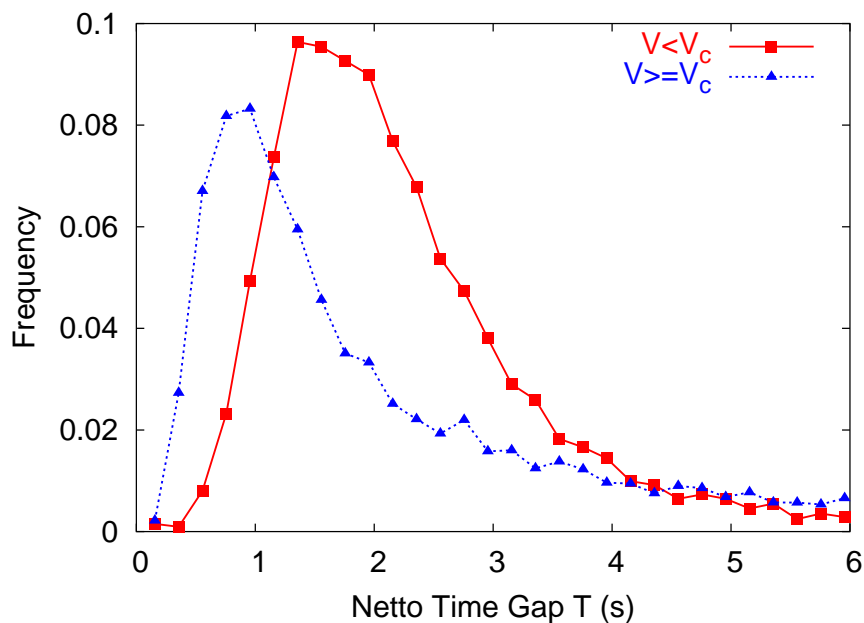
Ferner gibt es an vielen Tagen kaum Veränderungen. Viel häufiger als bei der Gaußverteilung sind sie aber mehrere Standardabweichungen groß. Der Exzess ist also positiv.

In gewisser Hinsicht stellt der Exzess ein *Konzentrationsmaß* dar, darf aber nicht mit den Konzentrationsmaßen des nächsten Kapitels verwechselt werden.

Beispielsaufgabe zu Kap. 6.6(b), Kap. 7 und 8

Abstandsverteilungen im Autobahnverkehr

Auf einem Autobahnabschnitt wurde der zeitliche Netto-Abstand von aufeinander folgenden Fahrzeugen gemessen (echte Messdaten!). Dabei wurden in zwei Messreihen die relativen Klassenhäufigkeiten einmal für freien Verkehr ($V \geq 60$ km/h) und einmal für gestauten Verkehr ($V < 60$ km/h) bestimmt:



Im nicht dargestellten Bereich ($x \geq 6$ s) liegen 20% der Daten des freien Verkehrs mit einer mittleren Folgezeit von 15s, und ebenso 5% der Daten des gestauten Verkehrs mit einer mittleren Folgezeit von 10s. Im dargestellten Bereich ($x < 6$ s) sind die Messwerte in Klassen gleicher (aber nicht mehr bekannter) Breite eingeteilt.

Obwohl die Klassenbreiten nicht bekannt sind, macht sich ein Verkehrsanalyst an die Auswertung dieser Daten. Er teilt die Daten in neue Klassen ein und weist ihnen zunächst die Fläche A_k unter der Kurve zu:

Klasse	1	2	3	4	5	6
Folgezeit x in s	0-0.5	0.5-1	1-1.5	1.5-2	2-3	3-6
Fläche A_k^{frei} in s/100	1.8	3.9	3.1	2.1	2.6	3.6
Fläche A_k^{stau} in s/100	0.2	1.3	3.8	4.4	5.4	4.4

Beispielsaufgabe (Fortsetzung)



1. Sind die dargestellten Kurven empirische Dichtefunktionen? Wenn nicht, berechnen Sie diese. Berechnen Sie auch die relativen Klassenhäufigkeiten! Bestimmen Sie auch die relativen Häufigkeitsdichten f_k^D !
Lösungshinweise: Welche Bedingung muss für die gesamte Fläche unter den Kurven gelten? Schließen Sie daraus auf die relativen Häufigkeiten f_k der neuen Klassen. Den nicht dargestellten Bereich ($x \geq 6$ s) können Sie in Klasse 7 aufnehmen.
2. Berechnen Sie für gestauten Verkehr den wahrscheinlichsten Wert (Modus) der Folgezeit, den Median und das arithmetische Mittel. Nehmen Sie ggf. als Klassenobergrenze der größten Klasse eine Folgezeit von 15 s an sowie als Untergrenze der niedrigsten Klasse 0 s. Für welche der obigen Lagemaße benötigt man diese zusätzlichen, dem Sachverhalt angemessenen Ad-Hoc Informationen? Kann man allein aus diesen Lagemaßen schließen, dass es sich um eine linkssteile Verteilung handeln muss?
3. In der Fahrschule lernt man "Sicherheitsabstand mindestens gleich halber Tacho". Wieviel Prozent der Leute halten sich im freien und im gestauten Verkehr daran? Im Prinzip kann man bei Unterschreiten eines zeitlichen Abstandes von 0.9 s bestraft werden. Für welchen Anteil würde dies zutreffen?
4. Berechnen Sie für freien Verkehr das arithmetische Mittel des zeitlichen Abstandes (und wundern Sie sich über das Ergebnis!) Was lernt man daraus über das arithmetische Mittel?
5. Berechnen Sie alle Streumaße für die Verteilung der Folgezeiten im gestautem Verkehr: Varianz, Standardabweichung, Variationskoeffizient, Mittlere absolute Abweichung, Interquartilsabstand und Spannweite
6. Berechnen Sie Schiefe und Exzess für die Verteilung der Folgezeiten im gestautem Verkehr.
7. Wie ändern sich der Zahlenwert von Mittelwert, Varianz, Standardabweichung, Schiefe und Exzess, wenn die Folgezeit nicht in s, sondern in ms gemessen wird, also alle Zahlenwerte der Folgezeiten mit 1000 multipliziert werden?
8. Gerade als der Verkehrsanalyt die obigen Ergebnisse präsentieren will, erfährt er, dass das Messgerät für die Messung der Folgezeit systematisch alle Zeiten um 0.2 s zu niedrig gemessen hat. Kann er ad-hoc (ohne Rechnung) dennoch die richtigen Ergebnisse für Mittelwert, Varianz, Standardabweichung, Schiefe und Exzess präsentieren?