Streuwerte

Streuwerte oder Streuungswerte geben an, wie die Daten um einen **Lage**wert herum (meist um den Mittelwert) **verteilt** oder gestreut sind.

Nehmen wir noch einmal die verschiedenen Werte der Stichprobe der Lernenden aus der Übung 2. Die Werte sind jetzt aufsteigend geordnet und um 3 Werte ergänzt.

158 160 164 168 169 171 173 173 173 173 173 175 176 176 176 177 178 178 179 180 181 183 185 188 191

Den ersten Streuungswert, den wir sofort erkennen können, ist die

Die Spannweite r

(r aus dem englischen range = Bereich).

Die Spannweite r gibt den **gesamten Bereich** an, innerhalb welchem die Daten angeordnet sind.

158 160 164 168 169 171 173 173 173 173 173 175 176 176 176 177 178 178 179 180 181 183 185 188 191

$r = x_{max} - x_{min}$ in unserem Beispiel: $r = 191 - 158 =$ **33**

Unser Beispiel ergibt eine sinnvolle Spannweite. Wir können uns sogar einfach vorstellen, wie alle Lernenden dieser Klasse im Grössenvergleich stehen würden und auch den Grössenunterschied von 33 cm zwischen dem Grössten und dem Kleinsten.

Doch manchmal ergibt die Spannweite nicht die gewünschte Übersicht. Denken Sie nur an die Löhne der Angestellten einer Grossbank aus Übung 8. Was sollen wir mit der Spannweite von r = 39'000 – 48 = 38'052 kFr anfangen?

Wo liegt das Problem? Es liegt beim weit abgelegenen Lohn (=Ausreisser) des Herrn CEO. Auch die nur 48'000 Fr Lohn (eventueller Lohn des Reinigungspersonals, das nicht direkt von der Bank angestellt ist), dürften die Aussage über Banklöhne stark beeinflussen.

Wie bei der Übung 8 schon festgestellt, ist der **Median** stabil gegenüber Ausreissern. Der Median ist auch jener Lagewert, der mit dem Streuungswert „Spannweite“ am meisten zu tun hat.
Wir behalten also den stabilen Median und verknüpfen ihn so mit unseren Werten der Stichprobe, dass die Streuung sinnvoll wird.

Das Quartil (die Quartile)

Quartile sind weitere Kennwerte einer Stichprobe. Man teilt die Reihe der **geordneten** Werte in 4 Quartile ein.
Wichtig sind dabei das untere Quartil $Q_{0,25}$ und das obere Quartil $Q_{0,75}$.
In der Mitte liegt der **Median z**, manchmal auch als $\mathbf{Q_{0,5}}$ bezeichnet. Unterhalb des $Q_{0,25}$ liegen damit 25 % der Werte und oberhalb des $Q_{0,75}$ ebenfalls 25 % der Werte der Stichprobe.
Somit sind 50 % der Werte im Bereich $Q_{0,25}$ bis $Q_{0,75}$ vorhanden.
Dieser Bereich heisst Quartilabstand QA oder IQR = interquartile range.

Hat man genügend Werte bei einer Stichprobe erfasst, so kann man die Spannweite auf den Quartilabstand reduzieren und ist damit sicher, dass die **Ausreisser wegfallen**.

Die Quartile werden folgendermassen bestimmt:
Zuerst wird die Position des Quartils in der geordneten Reihe mit n Werten bestimmt.

$Q_{0,25} = \frac{n}{4}$ $\quad Q_{0,75} = \frac{n}{4} \cdot 3$ Jeweils auf die nächsthöhere Zahl aufrunden.

Dann werden die **Werte** (nicht die Position) eingesetzt.

Beispiel

Die Lohnstichprobe der Übung 8 hatte folgende 10 Stichprobenwerte angegeben:
48 114 153 87 68 145 129 39'000 117 98 Liste ordnen! 48 68 **87** 98 114 117 129 **145** 153 39'000 geordnet

$Q_{0,25} = \frac{10}{4} = \frac{n}{4} = 2{,}5$ *gerundet* $\Rightarrow$ 3. *Position*

$Q_{0,25} = 87$

$Q_{0,75} = \frac{n}{4} \cdot 3 = 7{,}5$ *gerundet* $\Rightarrow$ 8. *Position*

$Q_{0,75} = 145$

Quartilabstand $QA = Q_{0,75} - Q_{0,25} = 145 - 87 = 58$

Beispiel aus der Verkehrsplanung: Die Geschwindigkeit v_{85}

V steht für Geschwindigkeit. Wenn z.B. die Wirkung einer Tempo-30-Zone überprüft werden soll, so ist der sogenannte v_{85}-Wert wichtig als Überprüfungswert. Also nicht die Durchschnittsgeschwindigkeit, weil diese von einem einzelnen Raser verfälscht werden könnte.
Das heisst: v_{85} ist jene Geschwindigkeit, die von **85%** der Fahrzeuge **eingehalten** wird (während 15% sie überschreiten).

Übung 13 (müssen Sie nicht machen, ist zuwenig wichtig)
Bestimmen Sie die Quartile $Q_{0,25}$, $Q_{0,75}$ und den Quartilabstand der geordneten Reihe mit 25 Werten.
158 160 164 168 169 171 173 173 173 173 173 175 176 176 176 177 178 178 179 180 181 183 185 188 191

Übung 14 + 15 gestrichen

Wichtigster Streuwert: die Standardabweichung $s_{\bar{x}_A}$ oder σ_X

Wie aus dem Symbol ersichtlich, bezieht sich der Streuwert **Standardabweichung** auf den Lagewert des **Arithmetischen Mittels**. Definition:
Die Standardabweichung ist die Wurzel aus der mittleren quadratischen Abweichung der Daten von ihrem arithmetischen Mittel. Als Formel geschrieben heisst das:

$s_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	$\sum_{i=1}^{n}$	Bedeutet Summe, die sich auf jeden Wert (Index i von 1 bis n) bezieht.
	$(x_i - \bar{x})^2$	Quadrierte Differenz zwischen Wert x_i und arithmetischen Mittelwert $\bar{x}_A$
	n-1	Anzahl der Stichprobenwerte –1
Ausgeschrieben:	mit 3 bis n Werten:	$s_{\bar{x}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \text{K} + (x_n - \bar{x})^2}{n-1}}$

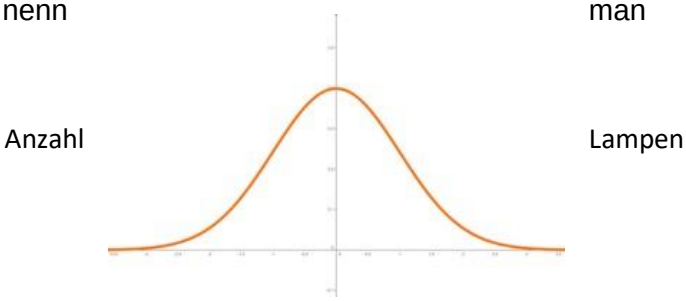
Oft wird auch nur durch n statt durch n-1 geteilt. Der Unterschied ist meist minimal, Details:

Summe der Werte	Σ_X
Anzahl Werte	n
Arithmetisches Mittel	$\bar{x}$ oder μ
Standardabweichung der Stichprobe	s_{Xn-1} oder σ_{Xn-1}
Standardabweichung der Grundgesamtheit	s_{Xn} oder σ_{Xn}

Übung 16
Berechnen Sie Mittelwert und Standardabweichung aus den 4 Noten **2, 6, 5, 3**.
Vergleichen Sie wenn möglich die Funtions des Taschenrechners mit einer Berechnung von Hand gemäss der Definitionsformel. Nötige Funktionen: Taste Σ+ für jeden Wert, dann Taste σ_{Xn-1}
Mittelwert: Σ+ = 4, σ_{Xn-1} = 1.825…
Standardabweichung: $\frac{(x1 - x|)^2 + (x2 - x|)^2 + (x3 - x|)^2 + (x4 - x|)^2}{N - 1} \rightarrow \frac{(2-4)^2 + (6-4)^2 + (5-4)^2 + (3-4)^2}{4-1} \rightarrow$
$\frac{(-2)^2 + (2)^2 + (1)^2 + (-1)^2}{3} \rightarrow \frac{4+4+1+1}{3} \rightarrow$ 10 / 3 → x = √3.333.., y = 1.825…

Die Normalverteilung

Die Lebensdauer in Stunden einer Stichprobe von Glühlampen wird erfasst.
Sobald genügend Werte einer Stichprobe vorhanden sind, bemerkt man eine bestimmte HäufigkeitsVerteilung der Werte um das arithmetische Mittel herum, nach links und rechts abnehmend. Die Form der Verteilfunktion ist eine Kurve, die in der Mathematik als **«Gauss'sche Glockenkurve»** bekannt ist. Dies nenn man „Normalverteilung.



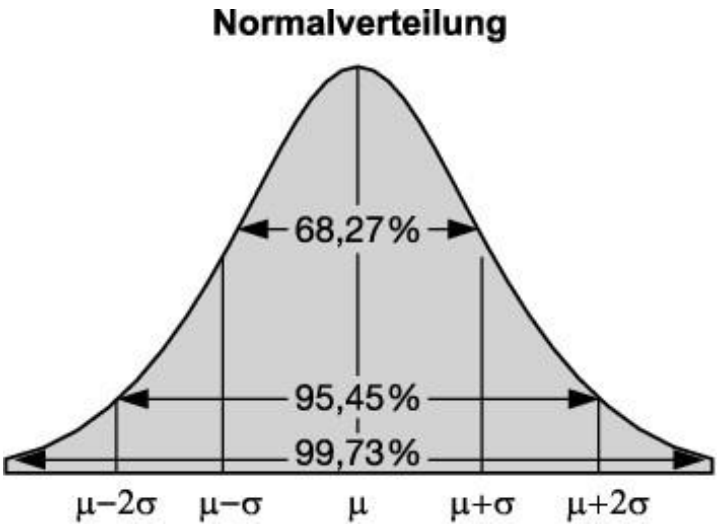
Eigenschaften dieser glockenförmigen Verteilung:

- Die Verteilung ist spiegelsymmetrisch um den Mittelwert μ
- Das Maximum der Verteilung liegt an der Stelle x = μ Mittelwert
- Die „Breite" der Glocke wird durch die Standardabweichung σ bestimmt.

Die Standard-Normalverteilung

Sie ist das wichtigste Verteilmodell in der Statistik.
Obige Glockenkurve stimmt für die Brenndauer von Glühlampen. Würden wir die Körpergrösse unserer Lernenden darstellen, so erhielten wir ebenfalls eine glockenförmige Kurve, aber das arithmetische Mittel wäre natürlich nicht 1000 Stunden, sondern vielleicht 176 cm und die Standardabweichung z.B. 7 cm. Genauso wären die Werte für Batteriespannungen oder Toleranzen in der Uhrmacherkunst ganz andere, obschon die Kurvenform die gleiche bleiben würde. Eben die Normalverteilung.
Die **Standard-**Normalverteilung belässt die Glockenform und normiert die Achsen (macht sie unabhängig von den Einheiten).
Mit der Standardnormalverteilung lassen sich eine Vielzahl natur-, wirtschafts- und ingenieurwissenschaftlicher Zusammenhänge, wie beispielsweise die Zufallsvariablen Körpergrösse, Gewicht, Messfehler, Zeiten bis hin zur Grösse von Sternen entweder exakt oder zumindest in guter Näherung beschreiben.
Nach der Normierung ist

- auf der x-Achse (Abszisse) die Anzahl der Standardabweichungen σ , die vom arithmetischen Mittelwert μ abweichen.
- Der Scheitelpunkt in der Mitte mit μ = 0 (Scheitelpunkt der Kurve)

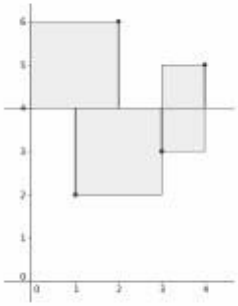


Da die Gauss'sche Glockenkurve symmetrisch ist, liegen 50 % der Werte links vom arithmetischen Mittelwert μ und 50 % sind rechts davon.

Bei einer guten Normalverteilung liegen 68 % der Werte innerhalb der einfachen Standardabweichung σ oder $S_{\bar{x}_A}$. Man spricht von der $^2/_3$ – Regel oder 1σ – Regel.

Aus einem andern Statistik-Kurs (M. Kriener Statistik):

Bei einer Liste von Zahlen $\{x_1, x_2, x_3, x_4\} = \{2, 6, 3, 5\}$ (das könnten zum Beispiel Ihre letzten Mathematiknoten sein) sind wir nicht nur am Mittelwert interessiert (der ist hier $\mu = 4$), sondern auch an der Streuung der Daten - hier hat man den Eindruck, dass die Noten ziemlich hin- und herspringen. Bei einem anderen Schüler sind die Noten (4.5, 3.5, 4, 4). Der gleiche Mittelwert, aber viel weniger Streuung. Wie können wir das Ausmass der Streuung mit einer Zahl erfassen? Einfach den Durchschnitt der **Abweichungen** $x_k - e\mu$ berechnen funktioniert nicht, denn $\Sigma(x_k - \mu) = 0$ (das ist die Robin-Hood-Gleichung). Wir möchten ausserdem, dass grosse Abweichungen stärker "zählen" sollen als kleine.



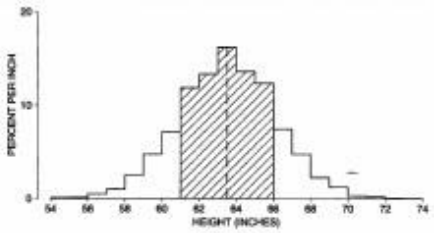
Wir haben gesehen, dass der Mittelwert μ die Summe der quadratischen Abweichungen $\sum(x_k - \mu)^2$ minimiert. Wir nennen den Mittelwert dieser Summe die **Varianz** und die Wurzel der Varianz die **Standardabweichung** einer Liste von Daten:

$$\sigma := \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2} \qquad (3)$$

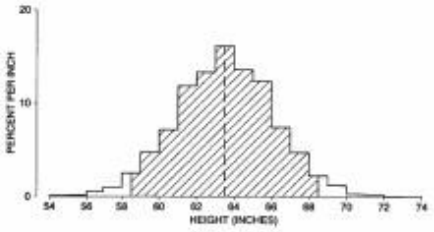
Beispiel: Für die Liste $(x_1, x_2, x_3, x_4) = (2, 6, 3, 5)$ von oben (Noten mit dem Mittelwert $\mu = 4$) ist die Standardabweichung

$$\sigma = \sqrt{\frac{2^2 + 2^2 + 1^2 + 1^2}{4}} = \sqrt{2.5} \approx 1.58$$

Was bedeutet diese Zahl? Ganz grob kann man sagen, dass die Standardabweichung "Normalität" misst - alles innerhalb *einer* Standardabweichung σ entfernt vom Durchschnitt ist noch "normal".



Beispiel: Die beiden Histogramme rechts zeigen zweimal die Grössenverteilung amerikanischer Frauen. Die gestrichelte Linie ist der Mittelwert, und der schattierte Bereich im oberen Histogramm ist der Bereich, der maximal eine Standardabweichung σ von μ entfernt ist - das sind etwa zwei Drittel aller Frauen. Der schattierte Bereich im unteren Histogramm ist der Bereich, der maximal zwei Standardabweichungen 2σ von μ entfernt ist - das sind schon 95% der Frauen.



Es gibt eine Daumenregel, die bei sehr vielen Daten gut funktioniert, nämlich bei **normalverteilten** Daten. Eigenschaften sind dann normalverteilt, wenn sie das Resultat von vielen voneinander unabhängigen Einflüssen sind. Körpergrösse und Intelligenz sind zum Beispiel normalverteilt. Die Daumenregel geht so:

- etwa 68% der Daten liegen innerhalb einer Standardabweichung vom Mittelwert - d.h. im Intervall $[\mu - \sigma; \mu + \sigma]$ liegen etwa 68% der gesamten Daten. Das entspricht der schattierten Fläche im oberen Histogramm.
- etwa 95% liegen innerhalb von zwei Standardabweichungen vom Mittelwert - d.h. im Intervall $[\mu - 2\sigma; \mu + 2\sigma]$ liegen etwa 95% der gesamten Daten. Das entspricht der schattierten Fläche im unteren Histogramm.
- etwa 99.7% einer Population liegen innerhalb von drei Standardabweichungen vom Mittelwert. Hier müsste man praktisch die gesamte Fläche schattieren.

Übung . Berechnen sie die Standardabweichung der folgenden Listen:

a) {4, 5, 4.2, 4.8}

b) {10, 8, 13, 11, 12, 9, 7}

A) Standartabweichung = 0.476…

B) Standartabweichung = 2.160…