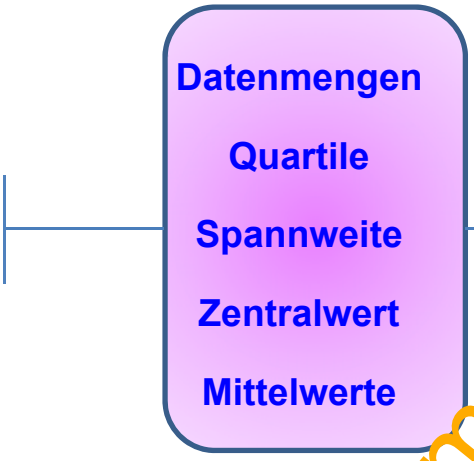
Statistik

Boxplot

Es geht um



Stand 28. Juli 2020

Datei-Nr. 31030

Friedrich W. Buckel

Internetbibliothek für Schulmathematik

www.mathe-cd.de

Vorwort

Der Boxplot ist ein sehr einfaches Werkzeug um die Verteilung der Werte eines Datensatzes einschätzen zu können. Die meisten Daten werden weggelassen und einige wichtige Kenngrößen werden berechnet. Daraus gestaltet man den Boxplot.

Da es verschiedene Methoden hierzu gibt, habe ich mich entschlossen, auch zwei verschiedene Methoden zu zeigen. Dass dabei leicht abweichende Ergebnisse entstehen, ist dem Statistiker egal. Ihm kommt es nur auf eine großzügige Erkenntnis über die Verteilung der Daten an.
Es gibt also verschiedene Boxplots, die sich aber gleichermaßen zur Auswertung eignen.

Inhalt

1 Der einfache Boxplot

Beispiel 1: Pflanzenwachstum.

Klaus hat die Pflanzensamen in Erde gelegt und will ihr Wachstum untersuchen.
Nach fünf Wochen misst er die Höhe der 19 aufgegangenen Pflanzen. Er ordnet die Daten gleich und erhält folgendes Ergebnis:

6,2 6,7 7,0 7,4 7,8 8,2 8,4 8,9 9,1 9,5 9,5 9,6 9,8 9,8 9,9 10,0 10,0 10,4 11,0

Die Statistik hat eine Methode entwickelt, die eine Übersicht über die Verteilung der Daten liefert.

1. Schritt: Bestimmung des Zentralwerts = Median

Überlegung: Wenn eine ungerade Anzahl von Daten gegeben ist, dann gibt es einen eindeutigen Zentralwert. Wir haben 19 Daten. Die Mitte berechnet man so:

$\boxed{m = \frac{n-1}{2} + 1}$, Dabei ist n die Anzahl der Daten und m die Nummer des Medians.

Hier: $m = \frac{19-1}{2} + 1 = 10$. Der 10. Wert ist also der Median: $\bar{x}_{med} = x_{10} = \boxed{9,5}$

6,2 6,7 7,0 7,4 7,8 8,2 8,4 8,9 9,1 **9,5** 9,5 9,6 9,8 9,8 9,9 10,0 10,0 10,4 11,0

9 Werte 9 Werte

Mit dem Median haben wir den Datensatz halbiert.

2. Schritt: Bestimmung der Mediane der beiden Datenhälften.

Die untere (bzw. linke) Hälfte besteht aus einer ungeraden Anzahl von Daten. Also gibt es einen eindeutigen Zentralwert oder Median. Die beiden Hälften sind nun jeweils etwa ein Viertel des ganzen Datensatzes. Daher nennt man die Marke bzw. den Wert, der diese Viertel erzeugt das **1. Quartil**. Es ist hier der Wert $Q1 = \bar{x}_{0,25} = \boxed{7,8}$.

Dasselbe gilt entsprechend für die obere (bzw. rechte) Hälfte des Datensatzes. Die Marke, die diese Hälfte halbiert und damit wieder zwei Viertel erzeugt, heißt 3. Quartil: $Q3 = \bar{x}_{0,75} = \boxed{9,9}$.

Q1 $\bar{x}_{med}$ Q3

6,2 6,7 7,0 7,4 **7,8** 8,2 8,4 8,9 9,1 **9,5** 9,5 9,6 9,8 9,8 **9,9** 10,0 10,0 10,4 11,0

Übrigens nennt man den Median aus ersichtlichen Gründen auch das 2. Quartil. Diese 3 Quartile teilen den Datensatz in vier „gleich große“ Bereiche auf.

Diese Aussage ist oft nur näherungsweise richtig, denn eine ungerade Anzahl von Daten kann man weder halbieren noch vierteln. Daher gibt es einige Varianten zur Festlegung der Quartile.

Daher schauen wir uns die Zahlen nochmals genauer an:

Beginnen wir beim Median. Er ist die Marke, welche die Mitte des Datensatzes kennzeichnet und ist in unserem Beispiel der Zentralwert 9,5. Man liest oft, dass links und rechts vom Median 50% der Daten liegen. Schauen wir nach: Links und rechts liegen jeweils 9 der 19 Daten. Das sind keine 50%! Wenn man aber den Median mit einschließt, dann enthalten der linke Bereich $\{x_{min}\,;\ldots;\bar{x}_{med}\}$ und der rechte Bereich $\{\bar{x}_{med}\,;\ldots;x_{max}\}$ ungefähr aber mindestens 50% aller Daten, nämlich 10 der 19 Werte, also etwa 52,6% .

Nun beleuchten wir das 1. Quartil Q1 = 7,8. Er ist der Zentralwert von 9 Werten. Links vom Median. Links von Q1, also in $\{6{,}2\,;\,6{,}7\,;\,7{,}0\,;\,7{,}4\}$, liegen 4 der 9 Werte, also wieder weniger als die Hälfte, und von der ganzen Datenmenge weniger als ein Viertel. Daher ist es üblich, den Q1-Wert mit dazu zu nehmen: In $\{6{,}2\,;\,6{,}7\,;\,7{,}0\,;\,7{,}4\,;\,\mathbf{7{,}8}\}$ liegen 5 der 9 Werte, also ungefähr aber mindestens 50%.

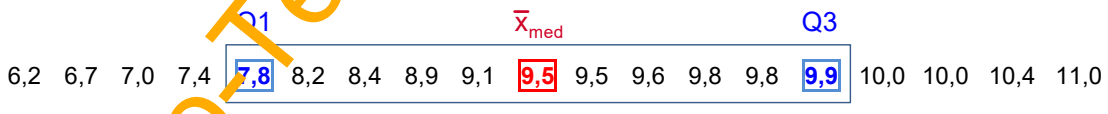
Dasselbe gilt für die Daten rechts vom 1. Quartil, also in $\{\mathbf{7{,}8}\,;\ 8{,}2\,;\,8{,}2\,;\,8{,}9\,;\,9{,}1\}$.

Nun untersuchen wir den Bereich vom 1. bis 3. Quartil. In $\{Q1\,;\ldots;Q3\}$ liegen 11 Werte. das ist deutlich mehr als die Hälfte von 19. Links bzw. rechts vom Median bis zur Quartilsmarke, also in $\{Q1\,;\ldots;\mathbf{9{,}5}\}$ und in $\{\mathbf{9{,}5}\,;\ldots;Q3\}$, liegen jeweils 6 der 11 Werte, das entspricht 54,5 %. Also sind wir wieder bei ungefähr aber mindestens 50%.

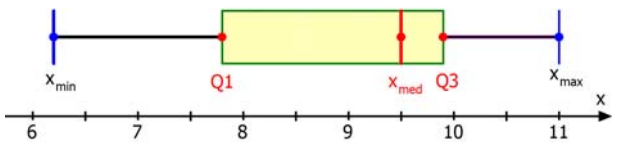
Für statistische Zwecke ist dieses „ungefähr" kein Problem. Die Bedeutung der Statistik liegt darin, Datensätze so zu vereinfachen, dass die wesentlichen Fakten deutlich werden wobei viele Details verschwinden. Da kommt es auf einen Wert mehr oder weniger nicht an, vor allem dann, wenn man große Datensätze untersuchen soll.

3. Schritt: Erstellung des Boxplots.

Im 2. Schritt haben wir das 1. und 3. Quartil bestimmt:

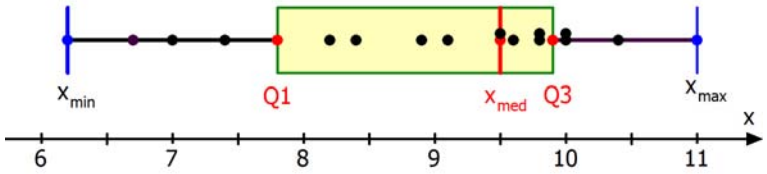


Um diese beiden Quartile legt man nun ein Rechteck (die Box) und zeichnet es maßstäblich.:



Das ist der fertige Boxplot. Man lässt alle Daten außer den fünf Kenngrößen weg.
So entstehen vier Bereiche $\{x_{min}\,;\ldots;Q1\}$, $\{Q1\,;\ldots;\bar{x}_{med}\}$, $\{\bar{x}_{med}\,;\ldots;Q3\}$ und $\{Q3\,;\ldots;x_{max}\}$, in denen jeweils ungefähr aber mindestens 50% aller Daten liegen.

Man sieht hier schnell, dass der kleinste Wert vielleicht fragwürdig ist, und dass etwa ein Viertel der Werte im Bereich $\{\overline{x}_{med}; \ldots; Q3\}$ eng zusammenliegen. Diese Pflanzen wachsen offenbar nur noch langsam. Ich trage nun in diesen Boxplot alle Werte ein, was nicht üblich ist, nur um zu zeigen, wie sie verteilt sind:



Für die Übersichtlichkeit der Daten reichen also wirklich der kleinste und größte Wert, der Median sowie die beiden Quartile Q1 und Q3.

Für das Erstellen eines Boxplots muss man also lediglich diese Werte eintragen, die Box zeichnen und die beiden „Antennen“ (sie heißen auch Whisker) links und rechts anbringen.

Beispiel 2: Pflanzenwachstum.

Hier wird es etwas schwieriger, weil teilweise gerade Anzahlen von Daten auftreten

Klaus hat die Samen zweier Sorten Pflanzen in Erde gelegt und will ihr Wachstum untersuchen. Nach fünf Wochen stellt er folgendes fest (ich ordne die Werte gleich der Größe nach):

Von der Sorte 1 sind 9 junge Pflanzen aufgegangen. Er misst ihre Größen und schreibt sie auf:

6,2 6,7 8,4 8,9 9,4 9,8 10,1 10,3 11 (cm)

Von der Sorte 2 haben es 8 geschafft:

6,8 7 7,4 7,6 8 8,2 9,6 10 (cm)

Erstelle jeweils einen Boxplot.

1. Schritt: Bestimmung des Median

Bei der **Sorte 1** hat Klaus eine ungerade Anzahl von Daten. Daher gibt es ist einen Zentralwert, der genau in der Mitte der Datenreihe steht. Es ist der Wert 9,4 (cm). Links von 9,4 stehen 4 (kleinere) Werte, rechts davon 4 (größere) Werte:

6,2 7,1 8,4 8,9 **9,4** 9,8 10,1 10,3 11 (cm)

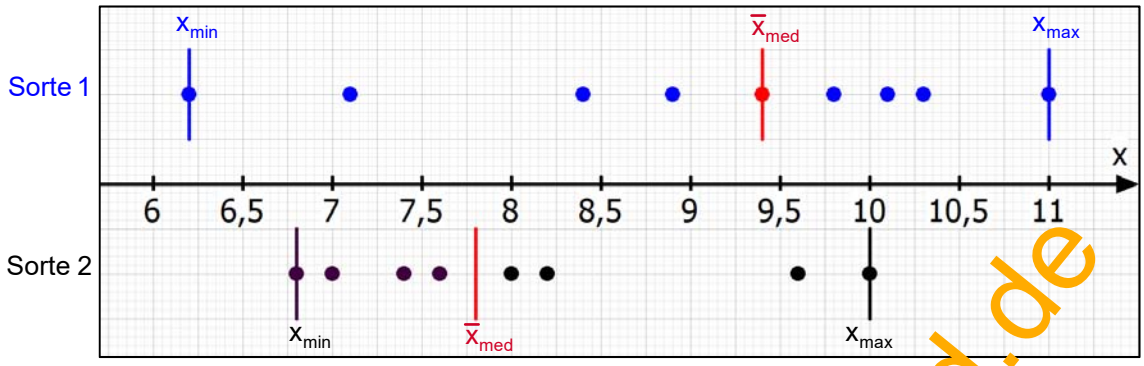
$\overline{x}_{med} = 9,4$

Bei der **Sorte 2** liegen 8 Messwerte vor. Bei einer geraden Anzahl von Daten gibt keinen mittleren Wert. Dieser müsste zwischen 7,6 und 8 liegen. Im Grunde könnte man jeden Wert zwischen diesen Zahlen nehmen. Man verwendet aber meistens deren arithmetisches Mittel:

6,8 7 7,4 7,6 **7,8** 8 8,2 9,6 10 (cm)

$\overline{x}_{med} = \frac{7,6+8}{2} = 7,8$

Nun stelle ich die Daten maßstäblich besser dar, indem ich sie entlang einer x-Achse abtrage.



2. Schritt: Bestimmung der Quartile Q1 und Q3.

Bleiben wir zuerst bei der Sorte 1.

Das geht für die linke (also untere Hälfte) so: Der dort gesuchte neue Median (er heißt unteres Quartil oder 1. Quartil oder Q1) soll wieder die jeweilige Hälfte halbieren. Wir haben jedoch in jeder Hälfte genau 4 Daten, also eine gerade Anzahl. Daher ist die Mitte keine der Daten.

6,2 7,1 | 8,4 8,9 | 9,4 | 9,8 10,1 | 10,3 11 (cm)

Q1? $\overline{x}_{med}$ Q3?

In der Statistik gibt es dazu leider verschiedene Varianten. Ich zeige zwei unterschiedliche Methoden. (von denen der Leser vielleicht schon eine kennt.)

Die erste Methode verwendet die Mittelwerte der beiden mittigen Werte:

$Q1 = \overline{x}_{0,25} = \frac{7,1+8,4}{2} = 7,75$ und $Q3 = \overline{x}_{0,75} = \frac{10,1+10,3}{2} = 10,2$

In diesem Fall sind diese Quartile keine vorhandenen Werte.

Die zweite Methode verlangt, dass ungefähr aber mindestens 25% **aller** Messwerte kleiner oder gleich dem unteren Quartil sein sollen (also folgt logischerweise, dass ungefähr aber mindestens 75% **aller** Messwerte sollen größer oder gleich dem untere Quartil sein sollen). Dazu die Rechnung:

Bei der Sorte 1 gibt es 9 Werte. 25% davon sind $0,25 \cdot 9 = 2,25$, also muss man das untere Quartil bei der drittkleinsten Zahl ansetzen, also bei 8,4: $\overline{x}_{0,25} = 8,4$

6,2 7,1 8,4 8,9 9,4 9,8 10,1 10,3 11 (cm)

Jetzt sind 3 von 9 Werten kleiner oder gleich **8,4** also ein Drittel, d.h. etwa 33%.
Das ist **mindestens ein Viertel**.

Und 7 von 9 Werten sind größer oder gleich 8,4 also etwa 78 %.
Das sind **mindestens drei Viertel**.

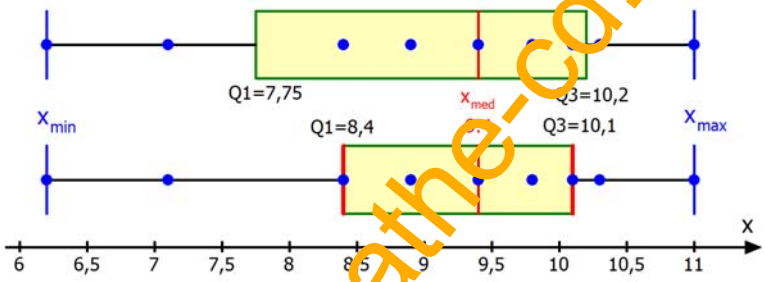
Analog dazu bestimmt man das obere Quartil $\boxed{\bar{x}_{0,75} = 10,1}$ (es heißt auch 3. Quartil oder Q3),

denn dann sind ungefähr aber mindestens 25 % aller Werte (3 Werte von 9) größer
und ungefähr aber mindestens 75% kleiner.

Die Boxplots beider Methode unterscheiden sich in der Lage von Q1 und Q3.
Das sieht hier nach viel aus. Der Unterschied liegt links und rechts nur in einem Wert, der bei der Mittelwert-Methode für die Quartile in der Box liegt, bei der Prozent-Methode auf dem Rand der Box , wo er dann zu beiden Bereichen dazugerechnet wird.

Methode 1 mit Mittelwerten:

Methode 2 mit Prozent-Methode:



Bei großen Datensätzen ist kaum ein Unterschied zu sehen!
Der Boxplot soll ja effektiv sein und die Lage von Datenmengen sichtbar machen. Da spielen 2 Punkte keine Rolle.

Auswertung der Sorte 2

Hier gibt es 8 Werte: 6,8 7 7,4 7,6 | 8 8,2 9,6 10 (cm)

Da es keinen Zentralwert gibt, wird der Median als Mittelwert der beiden mittigen Daten festgelegt:

$$\overline{x}_{med} = \frac{7{,}6+8}{2} = 7{,}8$$

Im Bereich links davon gibt es 4 Werte (rechts ebenfalls). Auch dort gibt es keinen Zentralwert.
Zur Festlegung der Quartile Q1 und Q3 verwendet …

die erste Methode die Mittelwerte der beiden mittleren Werte:

$Q1 = \overline{x}_{0,25} = \frac{7+7{,}4}{2} = 7{,}2$ und $Q3 = \overline{x}_{0,75} = \frac{8{,}2+9.6}{2} = 8{,}9$

In diesem Fall sind diese Quartile keine vorhandenen Werte.

die zweite Methode die Bedingung, dass ungefähr aber mindestens 25% **aller** Messwerte kleiner oder gleich dem unteren Quartil sein sollen. 25% von 8 sind 2 Werte.

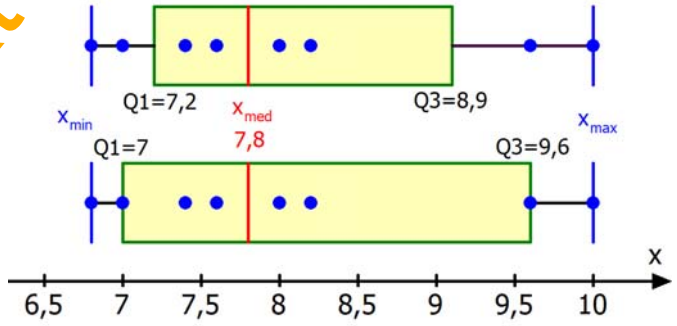
Für $Q1 = \overline{x}_{0,25} = 7$ gilt, dass 2 von 8 Werten kleiner oder gleich groß sind, das sind hier genau 25%, und dass 6 von 8 Werten größer oder gleich groß sind, das sind 75%.

Für $Q3 = \overline{x}_{0,75} = 9{,}6$ gilt, dass 2 von 8 Werten größer oder gleich groß sind, das sind hier genau 25%, und dass 6 von 8 Werten kleiner oder gleich groß sind, das sind 75%.

Nun zeige ich die beiden Boxplots für die Sorte 2

Methode 1 mit Mittelwerten:

Methode 2 mit Prozent-Methode:



Auch hier liegt der Unterschied nur darin, das zwei Werte auf dem Rand der Box liegen oder außerhalb. Bei großen Datenmengen ist das völlig unerheblich.

Achtung: Die Punkte lässt man bei einem Boxplot weg, da man ja vereinfacht.
Ein Boxplot besteht also nur aus der Box, dem Median und den beiden Antennen (bzw. Whiskers).

Zusammenfassung:

Welchen Sinn hat ein Boxplot?

Er gibt eine grobe Übersicht über die Verteilung (Streuung) der Messwerte.
Der Sinn der Vereinfachung liegt darin, dass damit die Verteilung übersichtlicher wird.
Er besteht also nur aus der Box mit dem Median und den beiden Antennen (Whiskers).
Die einzelnen Punkte (Daten, Werte) lässt man einfach weg.

Wie interpretiert man einen Boxplot

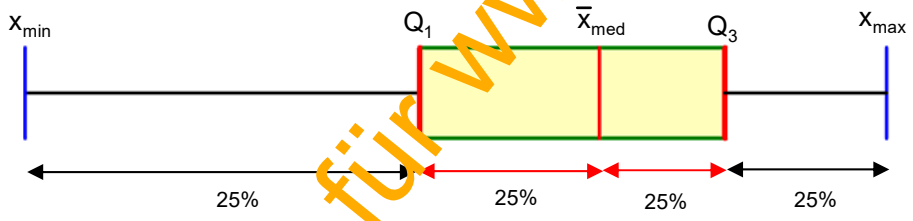
1. Zunächst geben die äußeren Striche die **Spannweite** an.
   Im Intervall vom kleinsten bis zum größten Wert liegen alle Messwerte.
   Die Spannweite ist die Länge dieses Intervalls und wird als Differenz berechnet:

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

2. In der **Box** (das Rechteck vom 25%-Quartil bis zum 75%-Quartil) liegen
   ungefähr aber mindestens 50% der Daten.

3. Im Bereich vom Minimum und dem 1. Quartil Q1 liegen **ungefähr aber mindestens 25%** der Werte. Im Bereich von Q3 bis zum Maximum ebenso.

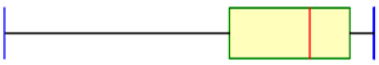
Damit teilt der Boxplot die Lage der Daten in 4 etwa inhaltsgleiche Intervalle, die je etwa 25 % der Messwerte enthalten.



Mit Hilfe eines solchen Boxplots kann man rasch einige Aussagen über die Daten machen.
Hier etwa diese:

Weil der Median nicht in der Mitte der Spannweite liegt, sondern nach rechts (oben) verschoben ist, kann man erkennen, dass diese Pflanzen anfangs schneller wachsen. Im oberen Bereich liegen ihre Größen enger beisammen, dort ist das Wachstum gebremst. Vermutung: Wären die Messwerte einige Zeit später erfasst worden, wäre der Median noch weiter rechts und 50% der größeren Pflanzen würden sich in einem engeren Bereich rechts vom Median befinden. Denn irgendwann sind die Pflanzen ausgewachsen und das zeigt sich dann in einem Boxplot von etwa der Art:

Am linken Ende gibt es offenbar eine Pflanze, die nicht richtig gewachsen ist. Man wird prüfen, ob sie auszuschließen ist, weil sie kein normales Wachstum aufweist. Man bezeichnet sie als „Ausreißer“. Wir lernen in Abschnitt 3, wie man eine Grenze ziehen kann, so dass man Ausreißer sinnvoll erfassen kann. Das ist aber lediglich eine sinnvolle Festlegung, die der Erfahrung gut entspricht.



2 Boxplot schnell erstellen.

Beispiel 3: Körpergröße von Schülern einer Schulklasse.

Die folgende Tabelle zeigt die Köpergrößen von 24 Schülerinnen und Schülern.

x_i gibt die Merkmalsausprägung (Größe) an, n_i dessen (absolute) Häufigkeit. Der Index i läuft dabei von 1 bis 16, denn es gibt 16 verschiedene Werte: Die Merkmalsausprägung x_2 = 155 (cm) kommt mit der Häufigkeit n_2=1 (also einmal) vor, während x_7 = 164 (cm) genau dreimal vorkommt.

x_i	154	155	156	157	158	161	164	165	167	169	170	171	172	173	180	183
n_i	1	1	1	1	1	1	3	2	1	2	2	2	1	2	2	1

Zur **Erstellung eines Boxplots** müssen wir beachten, dass es mehrere gleich große Werte gibt. Wir rechnen also mit 24 Werten, nicht mit 16!

1. Berechnung des Medians:

Da die Anzahl der Messwerte mit 24 eine gerade Zahl ist liegt der Median zwischen dem 12. Wert (167) und dem 13. Wert (169). Der Übersicht wegen schreibe ich ein Teilstück der Anordnung auf, wobei ich nun alle durchnummeriere:

$$\ldots,\ x_7 = x_8 = x_9 = \boxed{164},\ x_{10} = x_{11} = \boxed{165},\ x_{12} = \boxed{167},\ x_{13} = x_{14} = \boxed{169},\ldots..$$

Bei 24 Daten liegt die Mitte zwischen dem 12. und dem 13. Wert, also zwischen 167 und 169. Der Median ist also $\overline{x}_{med} = \boxed{168}$. *Dieser Wert ist nur eine Marke. Er muss nicht im Datensatz vorkommen*. Kommt er jedoch als Wert vor, heißt er auch Zentralwert.

2. Berechnung der Quartile Q1 und Q3.

1. Mittelwert-Methode: Q1 ist dann die Mitte des Bereichs links vom Median, also *der Mittelwert von x_6 und x_7, also* ***162,5***. Q3 ist dann die Mitte des Bereichs rechts vom Median also *der Mittelwert von x_{18} und x_{19}, also* ***171,5***.

2. Die 25%-Methode: Zunächst berechne ich, wieviel 25% der 24 Werte sind: $\frac{1}{4} \cdot 24 = 6$.

usw. auf der CD