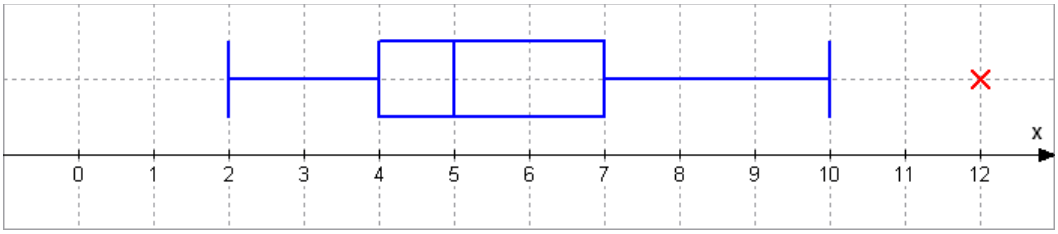
Beschreibende Statistik – Kenngrößen in der Übersicht Ac 2006-2021



„Boxplot“ - Beispiel

Im folgenden wird die Berechnungsweise des **TI 83** (sowie von SPSS, s. unten) verwendet. Diese geht auf eine Festlegung von Moore und McCabe (2002) zurück .

In der Literatur existieren insbesondere für die Berechnung der Quartile Q1 und Q3 (s.u.) noch ganz andere Formeln.
Zum Beispiel rechnen EXCEL und das NRW-Schulportal Learn-line nach einer Methode von John W. Tukey (1983; Begründer der Explorativen Datenanalyse) !

2 Beispiele:

Gegeben sei eine Liste x(i) von n Daten . Eine ungeordnete Liste muss zunächst sortiert werden.

U.a. werden jeweils 5 Kennzahlen ermittelt: xmin, xmax, med, Q1, Q3 (siehe Erläuterung unten).

Beispiel 1 mit n=9: 1 2 | 3 3 3 4 4 | 5 6
xmin=1 Q1=2,5 med=3 Q3=4,5 xmax=6

Beispiel 2 mit n=10: 0 1 2 4 4 | 6 7 8 9 9
xmin=0 Q1=2 med=5 Q3=8 xmax=9

Ersichtlich gibt es Unterschiede in der Behandlung bei geraden und ungeraden n !

1) Formeln zur Berechnung der Kenngrößen (mit dem TI 83):

Die Spannweite R (Range): Differenz zwischen Minimum und Maximum der Liste

range = xmax – xmin Im Beispiel 1: range = 6-1= 5

Bei den folgenden Kenngrößen kann es vorkommen, dass der Listenindex zwischen 2 ganzen Zahlen liegt, z.B. x(6,5) . Es muss dann das arithmetische Mittel verwendet werden .

statt $x(6,5)$ verwende $\frac{x(6)+x(7)}{2}$.

Der Modalwert D:

Der Modalwert D (bzw. x_M) ist das am **häufigsten** vorkommende Element der Liste.
Der Modalwert ist nicht eindeutig , existiert aber immer.
Im extremsten Fall kommen alle Elemente 1-mal vor ; dann ist jedes Element Modalwert !

Der Zentralwert (Median oder med): Der Wert in der Mitte der sortierten Liste

n ungerade: $\text{med} = x\left(\frac{n+1}{2}\right)$

n gerade: $\text{med} = \frac{x\left(\frac{n}{2}\right)+x\left(\frac{n}{2}+1\right)}{2}$

Arithmetischer Mittelwert (mean):

$$\bar{x} = \frac{x(1) + x(2) + \ldots + x(n)}{n}$$

Varianz (empirische):

Die Varianz (bzw. Streuung) V(x) ist als mittlere Abweichung der Quadrate vom Mittelwert definiert:

$$V(x) = \sigma_{n-1}^2 = \frac{(x(1)-\bar{x})^2 + (x(2)-\bar{x})^2 + \ldots + (x(n)-\bar{x})^2}{n-1}$$

Standardabweichung (standard deviation; empirisch):

Die Standardabweichung σ_{n-1} (Sx beim TI 83) ist die Wurzel aus der Varianz:

$$\sigma_{n-1} = \sqrt{\frac{(x(1)-\bar{x})^2 + (x(2)-\bar{x})^2 + \ldots + (x(n)-\bar{x})^2}{n-1}}$$

Achtung:
Bei stochastischen Verteilungen (z.B. Binomialverteilung) verwendet man σ_n statt σ_{n-1} .
Bei diesem wird in der Wurzel durch n statt durch n-1 dividiert !

Daten mit Häufigkeiten:

In vielen Fällen liegen die Daten x_i zusammen mit Besetzungszahlen (Häufigkeiten) vor.
Dann ändern sich die Formeln für Mittelwert und Standardabweichung:

Treten k Merkmalsausprägungen $x_1, x_2, x_3, \ldots, x_k$ mit k Besetzungszahlen (absoluten Häufigkeiten) $H_1, H_2, H_3, \ldots, H_k$ auf, so gelten folgende Formeln :

Relative Häufigkeit: $h_j = \frac{H_j}{n}$, n = Summe der Besetzungszahlen bzw. Umfang der Stichprobe

Arithmetisches Mittel: $\bar{x} = \frac{H_1 x_1 + H_2 x_2 + \ldots + H_k x_k}{n}$

empirische Varianz: $V(x) = \sigma_{n-1}^{\,2} = \frac{H_1 (x_1 - \bar{x})^2 + H_2 (x_2 - \bar{x})^2 + \ldots + H_k (x_k - \bar{x})^2}{n - 1}$

Weitere Kenngrößen

Das erste (untere) Quartil Q1: Q1 gibt den oberen Bereich des ersten Viertels der Liste an.
Andere Definition: Q1 ist der Median der links von **med** liegenden Liste.

n ungerade: $Q1 = x\left(\frac{n+1}{4}\right)$ n gerade: $Q1 = x\left(\frac{n+2}{4}\right)$

Im Beispiel 1: $Q1 = x(2,5) = \frac{2+3}{2} = 2,5$ Im Beispiel 2: $Q1 = x(3) = 2$

Das dritte (obere) Quartil Q3: Q3 gibt den oberen Bereich des dritten Viertels der Liste an.
Andere Definition: Q3 ist der Median der rechts von **med** liegenden Liste:

n ungerade: $Q3 = x\left(\frac{3n+3}{4}\right)$ n gerade: $Q3 = x\left(\frac{3n+2}{4}\right)$

Im Beispiel 1: $= Q3 = x(7,5) = \frac{4+5}{2} = 4,5$ Im Beispiel 2: $Q3 = x(8) = 8$

Der Interquartilsabstand IQR (interquartile range) :

IQR = Q3-Q1 Im Beispiel 1: IQR = 4,5 – 2,5 = 2

Wichtig: Im IQR (zwischen Q1 und Q3) liegt genau die Hälfte aller Daten ! Begründung ?

Ausreißer: Ein Wert, der mehr als das 1,5-fache des IQR von den Quartilen abweicht.
Wie findet man Ausreißer ? Man definiert ein Intervall $[z_u ; z_o] = [Q1 - 1,5 \cdot IQR ; Q3 + 1,5 \cdot IQR]$
Liegt ein Wert der Liste außerhalb dieses Bereichs, so ist er ein Ausreißer.
Im Beispiel 1 ist das Intervall [-0,5 ; 7,5] . Es gibt dort also keine Ausreißer .

Vorteile des Medians gegenüber dem arith.Mittel sowie des IQRs geg. der Standardabweichung:

Median und IQR sind unempfindlich gegenüber Ausreißern und unzuverlässigen Messungen oder Übertragungsfehlern, weil sie **keine Gewichtung** der Daten vornehmen !!

Die Quantile: Der Begriff Quantil ist ein Oberbegriff bzgl. Quartil und Median .

Quantile = Punkte einer nach Rang oder Größe geordneten Datenliste.
Z.B. gibt das 0,35-Quantil die Obergrenze für 35% der unteren geordneten Liste an .

Beispiel für n=8 :

3	4	5	5	6	7	8	9
x(1)	x(2)	x(3)	x(4)	x(5)	x(6)	x(7)	x(8)

Das 0,2-Quantil ist x(2) = 4 . Das 0,5-Quantil(Median) ist x(4,5) = 5,5 .

Markus Paul gibt für das q-Quantil folgende Berechnungen an :

$$\begin{cases} \dfrac{x(n\cdot q)+x(n\cdot q+1)}{2}, & \text{falls } n\cdot q \in \mathbb{N} \\ x(trunc(n\cdot q+1)), & \text{falls } n\cdot q \notin \mathbb{N} \end{cases}$$

Die Perzentile: Spezialfall der Quantile.

Perzentile P unterteilen eine **geordnete Liste in 100 Teile (0 < P < 100)** .
Es sind Punkte P, welche die Obergrenze für die Hundertstel der unteren geordneten Liste angeben.
Z.B. ist das Perzentil P=30 der Wert, für den 30% aller anderen Werte kleiner oder gleich diesem Wert sind.
Eine Unterteilung in Perzentile ist nur für große Listen sinnvoll.

Formeln für das P-Perzentil:

$$\begin{cases} \dfrac{x(n\cdot P/100)+x(n\cdot P/100+1)}{2}, & \text{falls } n\cdot P/100 \in \mathbb{N} \\ x(trunc(n\cdot P/100+1)), & \text{falls } n\cdot P/100 \notin \mathbb{N} \end{cases}$$

2) Grafische Darstellung von Datenreihen:

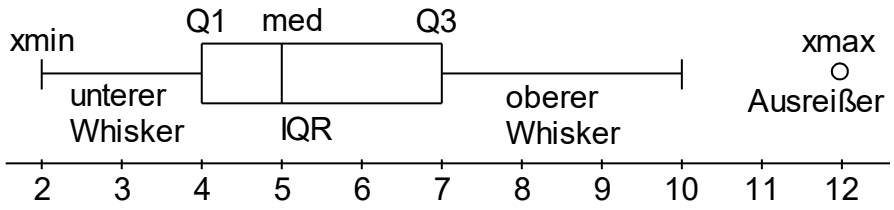
2.A) Boxplot (Box-Whisker-Plot):

Der Boxplot stellt die Kenngrößen mittels einer Box dar.

Willkürliches Beispiel: **2 3 4 4 5 5 6 6 7 10 12**

Es gelten dann: xmin=2 xmax=12 med=5 Q1=4 Q3=7 IQR=3

(„modifizierter Boxplot“)

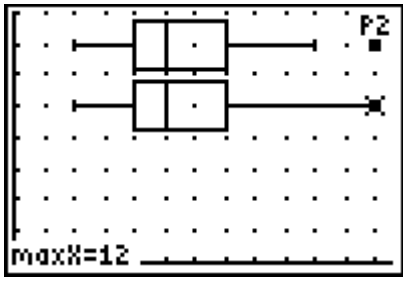


Whisker (Schnurrbarthaare) sind beim **normalen Boxplot** die Verbindungslinien von Q1 zu xmin sowie von Q3 zu xmax.
Beim **modifizierten Boxplot** (siehe Grafik oben) kann es aber vorkommen, dass die Whisker einen der Randpunkte oder gar beide Randpunkte (xmin, xmax) nicht erreichen, weil Ausreißer immer außerhalb des Bereichs der Whisker gezeichnet werden .
Genauer:
In obiger Grafik gilt IQR = 3 und somit 1,5·IQR = 4,5 .
Das für den Ausschluss von Ausreißern zu betrachtende Intervall ist [Q1-1,5·IQR ; 7+1,5·IQR].
Setzt man die entsprechenden Zahlen ein, so erhält man das Intervall [-0,5 ; 11,5].
xmin = 2 liegt innerhalb dieses Intervalls, ist also kein Ausreißer .
xmax = 12 liegt außerhalb dieses Intervalls und ist demnach ein Ausreißer ! Aus diesem Grunde schließt der obere Whisker die Zahl 12 nicht ein, sondern er erstreckt sich von Q3 bis zum letzten Wert, der noch innerhalb des Intervalls [-0,5 ; 11,5] liegt . In diesem Fall ist das die Zahl 10 .

Whisker und IQR sind Bereiche, keine Punkte. Im IQR liegen 50% aller Daten, im Bereich zwischen xmin und Q1 sowie xmax und Q3 liegen nochmals je 25% aller Daten .

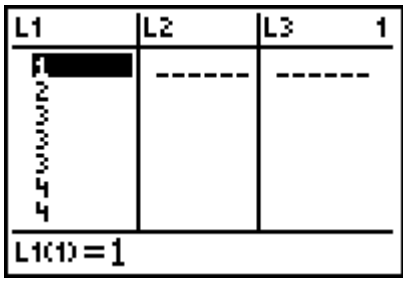
Genauso wie bei den Quartilen gibt es für die Definition der Lage der Whisker in der Literatur verschiedene Möglichkeiten .

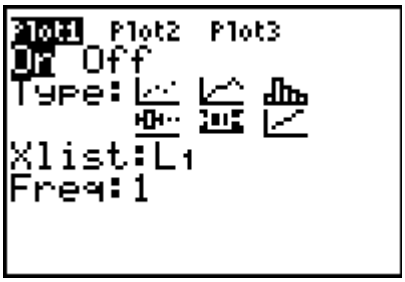
Darstellung mit dem TI83.
Modifizierter Boxplot (mit Ausreißer) und normaler Boxplot im Vergleich:

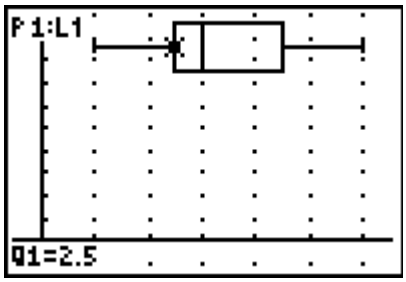


Boxplots mit dem TI 83

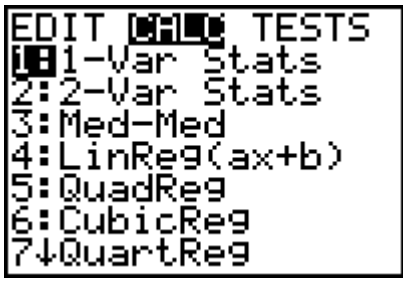
Zunächst nochmal das einführende Beispiel : 1 2 3 3 3 4 4 5 6

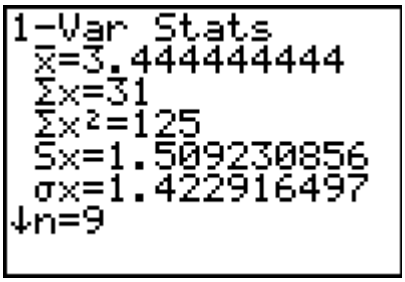


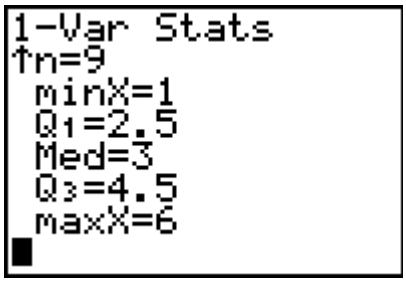




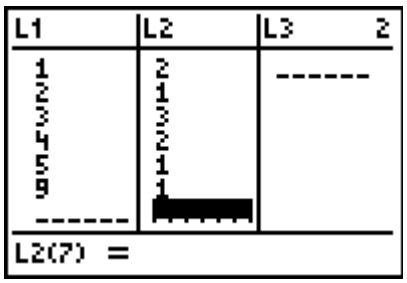
Dies ist ein **normaler Boxplot** (Type 5) . Mögliche Ausreißer würden hier nicht gezeichnet !
Die Kennzahlen können mit **TRACE** abgefragt werden (siehe Bild 3 oben).
Alternativ können sie auch mittels STAT CALC 1-Var Stats ausgegeben werden (Bilder unten).

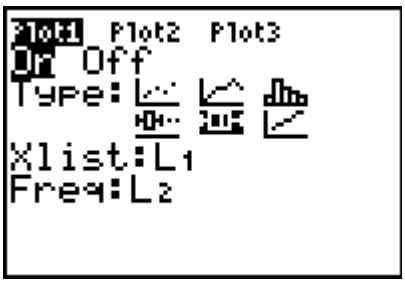


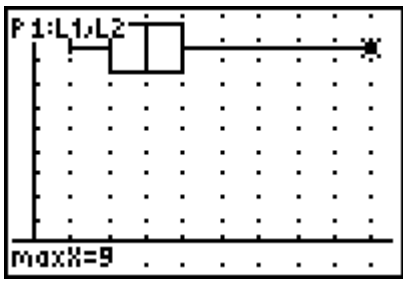




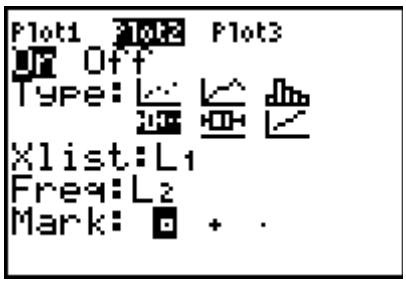
Ein weiteres Beispiel, diesmal werden absolute Häufigkeiten in L_2 mitverwendet . Im Statplot bei Freq den Wert L_2 (statt 1) eintragen ! Zuerst der normale Boxplot:

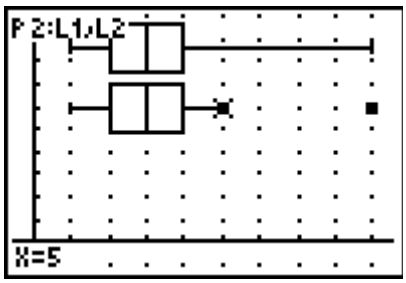






Verwendet man zusätzlich den modifizierten Boxplot (Type 4) , so können Ausreißer angezeigt werden. Hier ist das der Wert 9 . Der Whisker geht dann rechts nur noch bis zu x=5. Dies ist der letze Wert der um den Ausreißer gekürzten Liste.





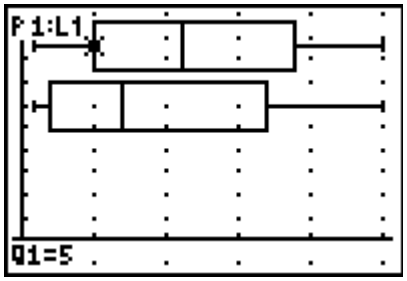
Ohne GTR findet man bei diesem einfachen Beispiel die Kennzahlen ebenfalls mühelos:

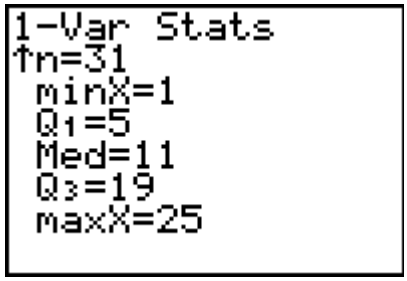
Betrachte die sortierter Liste: 1 1 2 3 3 3 4 4 5 9
Ablesen: xmin=1 xmax=9 med=3 Q1=2 Q3=4 IQR=2 1,5*IQR=3
Das durch 1,5*IQR definierte Intervall zum Ausschluss der Ausreißer ist dann [-1;7] . Da 9 nicht in diesem Bereich liegt ist es ein Ausreißer und somit geht der obere Whisker nur bis 5 .

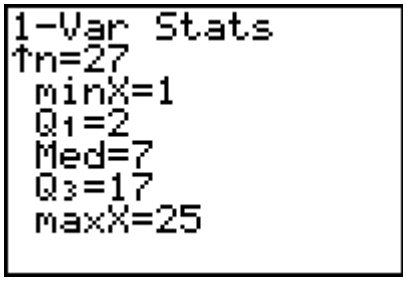
Wozu dienen Boxplots ? Vergleich zweier Datensätze:

Die Schüler der Klassen 9a und 9b geben die Entfernung (in km) ihres Wohnortes zur Schule an:
9a: 1 7 25 3 5 6 12 11 16 1 25 9 1 21 2 6 25 10 18 1 1 20 13 23 18 18 6 8 18 19 25
9b: 7 9 1 3 25 5 12 2 17 1 21 21 6 15 19 18 2 3 1 17 21 1 1 7 14 2 7

Aufgabe: Ordne die Listen und bestimme die Kennzahlen. Vergleiche mit den Ergebnissen des TI 83.







Hinweis: Die zweite Statistik erhält man mittels STAT CALC 1-Var Stats L_2 ENTER .

Schlussfolgerungen aus den Ergebnissen:

- Die 9b wohnt im Schnitt näher an der Schule
- In der 9b hat die **Hälfte** höchstens 7 km Schulweg, in der 9a höchstens 11 km
- Ein Viertel der 9b wohnt höchstens 2 km von der Schule entfernt, in der 9a sind es höchstens 5 km
- Keine Unterschiede gibt es beim kürzesten bzw. längsten Schulweg der beiden Klassen

2.B) Histogramme:

Das sind Rechtecke , deren ***Flächen proportional zur klassenspezifischen Häufigkeit*** *sind .*
Die Breite der Rechtecke (Klassenbreite !) kann variabel sein, was aber der TI83 nicht beherrscht .

Beispiel von oben : Wohnortentfernung von Schülern:
Man gibt am besten für jede Klasse 2 Listen ein, und zwar die jeweilige Entfernung und die dazugehörige Häufigkeit (Anzahl der Schüler mit dieser Entfernung).

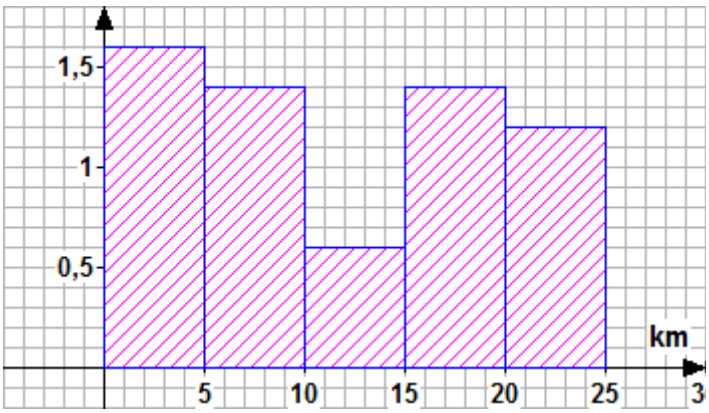
Für die 9a ist das z.B.

Entf / km	Anzahl
1	5
2	1
3	1
4	0
5	1
6	3
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	0
15	0
16	1
17	0
18	4
19	1
20	1
21	1
22	0
23	1
24	0
25	4

Teilt man nun die Entfernungen in Klassen der Breite 5km ein, so erhält man eine neue Liste mit der Klassenbreite 5, bei der die Rechtecksflächen der Anzahl (Häufigkeit) entsprechen. Folglich ergibt sich die jeweilige Rechteckshöhe aus Fläche / Klassenbreite = Anzahl / Klassenbreite !

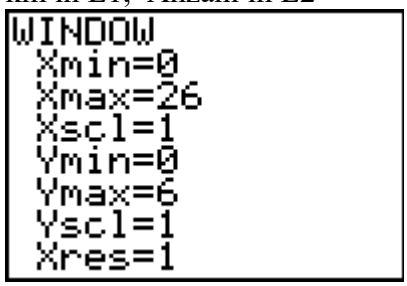
Entfernung in km [> a ; b]	Anzahl	Rechteckshöhe = Anzahl / 5
[0 ; 5]	8	1,6
[5 ; 10]	7	1,4
[10 ; 15]	3	0,6
[15 ; 20]	7	1,4
[20 ; 25]	6	1,2

Korrekt dargestelltes Histogramm (mit KarloPlot)



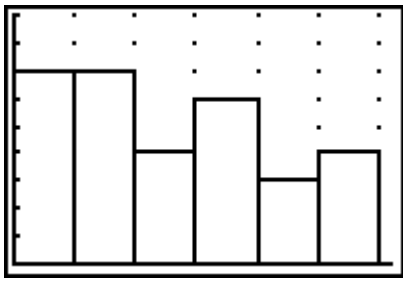
Histogramm mit TI83:

km in L1, Anzahl in L2



Einstellungen wie links:

mit Xscl=5 Xmax=31 Ymax=9



Man erkennt, dass
- die Rechteckshöhen nicht an die Flächen angepasst werden,
- insgesamt 6 Rechtecke statt 5 entstehen .

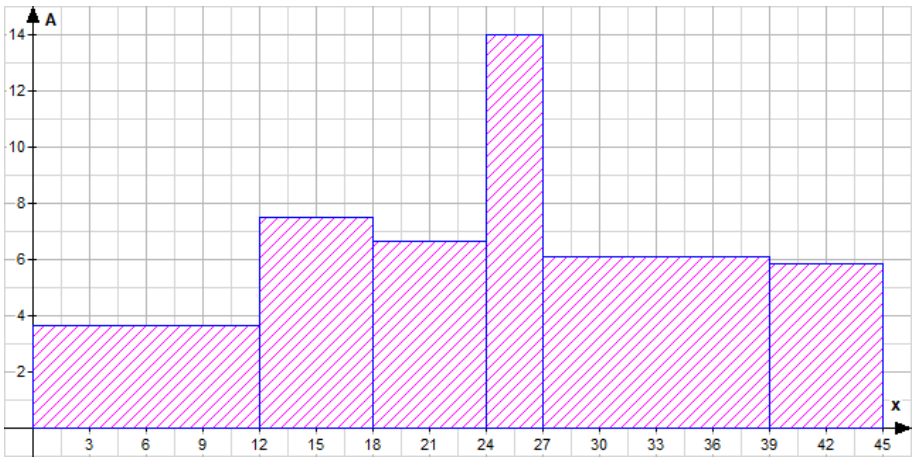
Offensichtlich ist der TI83 nicht für Histogramme geeignet !

Histogramme können völlig verschiedene Klassenbreiten haben !

Daten:

x von;bis	y (abs. Häufigk.)
0;12	44
12;18	45
18;24	40
24;27	42
27;39	73
39;45	35

Die **Rechtecksfläche** entspricht der abs. Häufigk.



2.C) Weitere gebräuchliche grafische Darstellungen

Außer den Histogrammen sind noch gebräuchlich:
Stängel-Blatt-Diagramm,
Stabdiagramm,
Häufigkeitspolygon ,
Kreisdiagramm (Torten-),
Punktdiagramm(Scatter).

Der TI 83 bietet hiervon nur Histogramm, Scatter und Häufigkeitspolygon .

2.D) Speziellere grafische Möglichkeiten:

Normal-Quantil-Plot:

Sind die erhobenen Daten annähernd normalverteilt ?
Um dies zu entscheiden, kann über das Histogramm die Normalverteilungskurve mit entsprechendem Mittelwert und Standardabweichung gelegt werden.
In der explorativen Datenanalyse jedoch verwendet man Normal-Quantil-Plots.
Hierbei werden die Quantile der Häufigkeitsverteilung mit entsprechenden Quantilen der Standard-normalverteilung verglichen.

Liegen die Punkte auf einer Geraden, so spricht das für eine annähernde Normalverteilung .

Der TI 83 bietet hierfür den Plot-Type 6 .

Eine genauere Betrachtung ist nachzulesen bei
Markus Paul (T^3 Europe): „Beschreibende Statistik und explorative Datenanalyse“

3) Anmerkungen zu anderer Software:

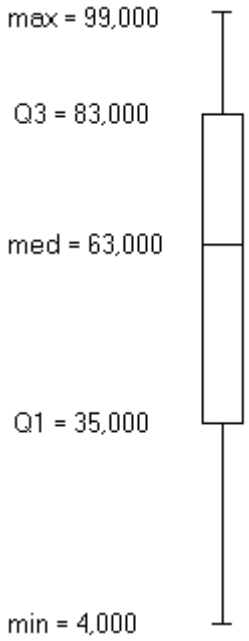
3.1) EXCEL u.a. berechnen nach der Tukey-Methode die Quartile folgendermaßen:

$$Q1 = x\left(\frac{\left[\frac{n+1}{2}\right]+1}{2}\right) \quad \text{und} \quad Q3 = x\left(n+1-\frac{\left[\frac{n+1}{2}\right]+1}{2}\right)$$

[z] ist die sog. Gaußklammerfunktion
(größte ganze Zahl ≤ z)

Bei dieser Methode wird bei ungeradem n der med-Wert in der Teiliste links (bzw. rechts) mitgezählt !
Bei geradem n ist die Methode identisch mit derjenigen des TI 83 .

3.2) Mehrere Softwarepakete zeichnen Boxplots vertikal statt horizontal (siehe Grafik) oder sie bieten beide Darstellungsmöglichkeiten.



3.3) Die Länge der Whisker im modifizierten Boxplot wird sehr unterschiedlich gehandhabt:

a) maximal bis zum 1,5-fachen IQR-Abstand von der Box; falls xmax bzw. xmin kleiner als dieser Abstand ist, dann bis zu xmax bzw. xmin

b) genau bis zum 0,05- bzw. 0,95-Quantil .

c) genau bis zum 0,025- bzw. 0,975-Quantil .

3.4) Gängige Statistik-Software-Pakete (kommerziell) sind:
- Fathom
- Minitab
- S-Plus
- SPSS

Meist erfordert diese Software eine nicht unbeträchtliche Einarbeitungszeit.

Es gibt aber auch freie (oder sehr preisgünstige) Pakete:
- Statistik-Labor(FU Berlin)
- VU-Statistik (Verlag Schroedel)
- GrafStat
- GeóGebra (außer Statistik noch viele weitere Themen)
- Excel bzw. Calc (von OpenOffice)
- usw.