Beschreibende Statistik und explorative Datenanalyse

mit dem TI-83 und der Stats/List-Applikation für TI-92plus/TI-89

Markus Paul

e-mail: markus.paul@utanet.at

Ein Unterrichtsbehelf zum Einsatz moderner Technologien im Mathematikunterricht

T³ Österreich / ACDCA am PI-Niederösterreich, Hollabrunn

Vorwort

Bisher haben wir uns wenig im Mathematikunterricht mit „schmutzigen Daten“ herumgeschlagen. Statistik ist ein ungeliebtes Stoffgebiet der Mathematiklehrer.

Woran liegt das?

Ein wesentlicher Grund liegt darin, dass die statistischen Methoden meistens sehr rechenintensiv sind. Bisher war es praktisch nicht möglich, im Klassenzimmer große Datenmengen in einem vertretbaren Zeitrahmen grafisch, tabellarisch und rechnerisch auszuwerten.

Nun stehen uns aber mit den neuen Texas-Rechnern im Klassenzimmer Tools zur Verfügung, die die wichtigsten statistischen Konzepte integriert haben und mit denen große Datensätze (fast) wie am PC verarbeitet werden können. Nun ist die Zeit reif, dass wir im Mathematikunterricht in die Niederungen der Empirie hinabsteigen. Wir haben nun die Möglichkeit, Echtdaten im Klassenzimmer zu erheben und auszuwerten und damit einen spannenden Praxisbezug herzustellen, der bisher undenkbar war. Das fördert die Akzeptanz des Fachs Mathematik, wenn die Schüler sehen, wozu die mathematischen Modelle gut sind und wo sie angewendet werden.

Mit den TI-Rechnern kann somit im Mathematikunterricht ein Hauch von professioneller Statistikanalyse Einzug halten. Schüler, die später im Rahmen eines sozial-, naturwissenschaftlichen oder medizinischen Studiums etwa mit SPSS (Statistical Packages for Social Sciences) Daten auszuwerten haben, werden es dem Mathematikunterricht danken. Sie werden den Umstieg relativ leicht bewältigen. Wir sollten uns immer dessen bewusst sein, dass die Statistik jenes Teilgebiet der Mathematik ist, mit dem unsere Maturanten am ehesten in einem Studium konfrontiert werden.

Markus Paul, im April 2002

Inhalt

0. Was ist „explorative Datenanalyse“?

„Exploratory data analysis is detective work.“

Mit dieser Feststellung beginnt der amerikanische Statistiker John W. Tukey (Princeton University und Bell Telephone Laboratories, 1915-2000) sein Buch „Exploratory Data Analysis“ [1], in dem er neue Verfahren zur Visualisierung von Datenmaterial vorstellt, untern anderen stem-and-leaf-displays (Stängel- und Blatt-Diagramme) und box-and-whisker-plots (Kastendiagramme). In den Daten verstecken sich Informationen, die der Statistiker durch trickreiche Darstellungen entlocken kann. Er befindet sich in derselben Situation wie der Detektiv, der unter den Verdächtigen ein Geheimnis aufdecken will. Explorative Datenanalyse steht für Tukey am Anfang jeder statistischen Tätigkeit:

„Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step.”

Moderne Statistikbücher haben die Konzepte von John Tukey übernommen und die Exploration als eine der drei Grundaufgaben der Statistik integriert: „Beschreiben (Deskription), Suchen (Exploration) und Schließen (Induktion). So widmet sich die beschreibende oder deskriptive Statistik der Beschreibung und Darstellung von Daten. Die explorative Statistik befasst sich mit dem Auffinden von Strukturen, Fragestellungen und Hypothesen, während die induktive Statistik Methoden bereit stellt, um statistische Schlüsse mittels stochastischer Modelle ziehen zu können.“ So verstehen Ludwig Fahrmeir, Rita Künstler, Iris Pigeot und Gerhard Tutz in ihrem Lehrbuch „Statistik. Der Weg zur Datenanalyse“ [2] ihr Fachgebiet.
Die Abgrenzung zwischen deskriptiver Statistik und explorativer Datenanalyse ist unscharf. So viel kann gesagt werden: „Die explorative Datenanalyse geht weiter als die deskriptive Statistik. Sie verwendet zwar ebenfalls keine Stochastik, also auf Wahrscheinlichkeitstheorie basierende Verfahren, aber einige ihrer Methoden sind durchaus von der induktiven Statistik beeinflusst. Über die Darstellung von Daten hinaus ist sie konzipiert zur Suche nach Strukturen und Besonderheiten in den Daten und kann so oft zu neuen Fragestellungen und Hypothesen in den jeweiligen Anwendungen führen. Sie wird daher typischerweise eingesetzt, wenn die Fragestellung nicht genau definiert ist oder auch die Wahl eines geeigneten statistischen Modells unklar ist.“ ([2], S.12)

Fragestellungen: (vgl. [3], S.3f.)

deskriptive Statistik	explorative Datenanalyse EDA
Wie kann man eine Verteilung eines Merkmals beschreiben?	Was ist an einer Verteilung eines Merkmals bemerkenswert oder ungewöhnlich? ⇓ mit induktiver Statistik explorative Vermutungen in signifikante Aussagen überführen

Bisher haben wir uns wenig im Mathematikunterricht mit „schmutzigen Daten“ herumgeschlagen. Einen wesentlichen Grund sehe ich darin, dass die Methoden der explorativen Datenanalyse meistens sehr computerintensiv sind. Bisher war es praktisch nicht möglich, im Klassenzimmer große Datenmengen in einem vertretbaren Zeitrahmen grafisch, tabellarisch und rechnerisch auszuwerten.

Die neuen Texas-Rechnern aber haben nun die wichtigsten Konzepte der explorativen Datenanalyse integriert und wir können nun im Klassenzimmer große Datenmengen (fast) wie am PC verarbeiten.

Der Einsatz der TI-Rechner ermöglicht im Mathematikunterricht einen Paradigmenwechsel: Die deskriptive Statistik wird ergänzt durch die explorative Datenanalyse.

Um den TI-92plus bzw TI-89 mit Statistik nachzurüsten, können Sie gratis die Applikation „Statistics with List Editor – Stats/List Editor“ von der TI-homepage

```
www://education.ti.com/product/tech/92p/apps/apps.html
www://education.ti.com/product/tech/89p/apps/apps.html
```

herunterladen und über das Graph-Link-Kabel auf Ihren Rechner übertragen.

Weiters steht Ihnen mit der PDF-Datei statsle.pdf (Statistics with List Editor Application for the TI-89/TI-92Plus) ein sehr detailliertes Handbuch dieser Applikation zur Verfügung (196 Seiten!).

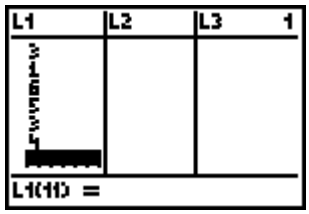
Durch das Konzept der Flash-Technologie ist es möglich geworden, Funktionalitäten, die bisher auf spezielle Rechner beschränkt waren, auf andere Modelle zu übertragen. Nun können insbesondere Statistik- und Finanzmathematik-Menüs, über die bisher nur der TI-83 verfügte, auch für die großen Brüdern TI-92plus und TI-89 nachgerüstet werden. Damit eröffnen sich für TI-92plus und TI-89 völlig neue Anwendungsgebiete der Mathematik. Der Vollständigkeit halber sei noch erwähnt, dass ein Großteil der Aufgaben am TI-92 auch über den DATA-MATRIX-Editor behandelt werden kann.

Welche Plattform schlussendlich im Unterricht verwendet wird, hängt von der Schwerpunktsetzung des Lehrers/ der Lehrerin ab und nicht zuletzt auch von der finanziellen Schmerzgrenze der Eltern.
Ich selbst verwende im Unterricht in der HAK den TI-83, weil er vom Preis vertretbar ist und weil das Verständnis meiner Schüler dort endet, wo die Überlegenheit des TI-92plus beginnt.

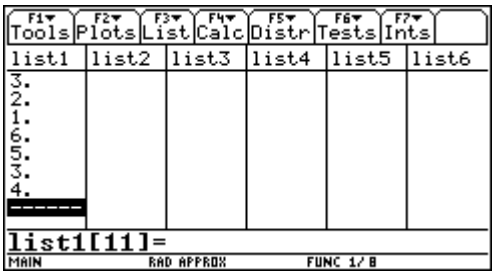
1. Lage- und Streuungsmaße

Statistik betreibt man nicht mit 10 Daten, aber man muss die Konzepte mit kleinen übersichtlichen Datensätzen lernen.

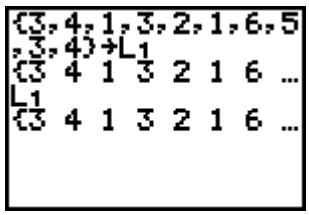
Beispiel: In einem Wohnblock wird die Haushaltsgröße (Anzahl der im Haushalt lebenden Personen) erhoben. Es ergeben sich folgende Werte:
3, 4, 1, 3, 2, 1, 6, 5, 3, 4.
Bestimmen Sie die wichtigsten statistischen Kennzahlen und stellen Sie die Daten grafisch dar.



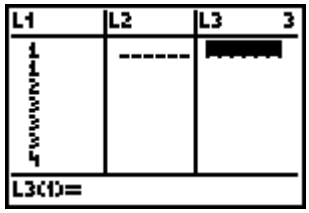
Statistische Daten können in ***Listen*** (mathematisch sind das endliche Folgen) gespeichert werden. Am einfachsten können auf dem TI-83 die Werte über den Statistik-Editor [**STAT] > EDIT > 1:Edit** in einer der vordefinierten Listen L1, L2, ..., L6 eingegeben werden. Geben Sie in der Liste L1 die Daten der Reihe nach ein.
(Sollten auf Ihrem Rechner diese Listen nicht vorhanden sein, so können Sie diese mit **[STAT] > 5:SetUpEditor** erzeugen.)
Sie können mit **DEL** einen falschen Wert löschen, mit **INS** können Sie einen Wert einfügen.
So löschen Sie die gesamten Inhalte einer Liste: Stellen Sie den Cursor in den Kopf der Liste und drücken Sie **CLEAR** und **ENTER.** Die Inhalte der Liste werden gelöscht, die Liste selbst bleibt im Editor. (ACHTUNG: Mit **DEL** löschen sie die Liste aus dem Editor! Die Liste ist aber immer noch im Speicher und kann über **[LIST]** > **NAMES** wieder aktiviert werden. Über das Menü **[MEM] > 2:Mem Mgmt/Del > 4:List** kann eine Liste (unwiderruflich) aus dem Speicher gelöscht werden.)



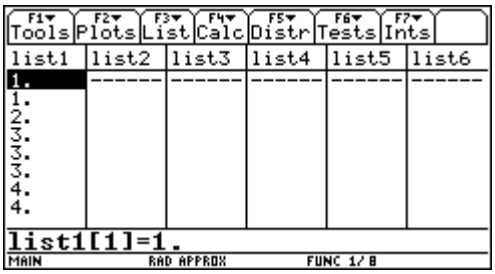
Auf dem TI-92plus müssen Sie über **[APPS] > 1:FlashApps > Stats/List Editor** die Statistik-Applikation starten. Sie können nun einen Folder (Ordner) wählen bzw. anlegen. Wählen Sie für den Anfang den main-Ordner. Nun gelangen Sie in den Stats/List Editor, dort können Sie in der **list1** die Daten der Reihe nach eingeben.



Sie können die Liste auch im Hauptbildschirm eingeben, indem Sie die Listenwerte innerhalb von geschwungenen Klammern { und } eingeben und in der Liste **L1** mit **STO** speichern.
Sie können dann im Hauptbildschirm mit **L1 [ENTER]** die Liste wieder ausgeben lassen oder Sie wechseln in den Statistik-Editor **[STAT] > EDIT > 1:Edit**, wo Sie die Liste tabellarisch als Spalte vorfinden.



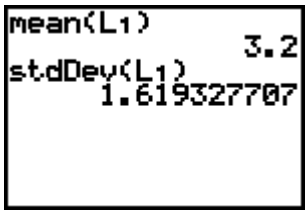
Sie können nun als ersten Schritt die Liste aufsteigend sortieren lassen, indem Sie im Hauptbildschirm im Menü **[LIST] > OPS > 1:SortA** aufrufen und die Liste L1 einfügen: **SortA(L1)**
Betätigen Sie die ENTER-Taste, sagt der Rechner DONE und im Statistik-Editor finden Sie die sortierte Liste vor.



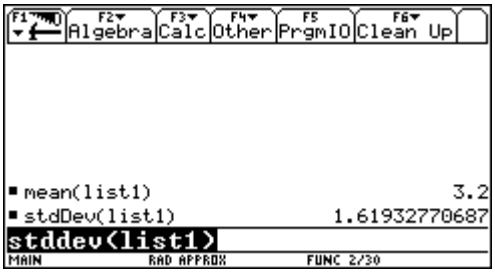
Auf dem TI-92plus finden Sie diese Funktion im Stats/List Editor im Menü **[F3] (List) > 2:Ops > 1:Sort List**.

Über das Menü [**LIST] > MATH** können die wichtigsten Kennzahlen berechnet werden:

1:min(Liste)	Minimaler Wert der Liste
2:max(Liste)	Maximaler Wert der Liste
3:mean(Liste)	Arithmet. Mittel der Liste
4:median(Liste)	Median der Liste
5:sum(liste)	Summe der Listenelemente
6:prod(Liste)	Produkt der Listenelemente
7:stdDev(Liste)	Standardabweichung der Liste (n-1-Gewichtung!)
8:variance(Liste)	Varianz der Liste (n-1-Gewichtung!)

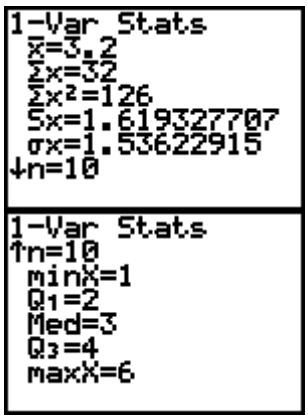


Auf dem TI-92plus finden Sie diese Funktionen im Menü **[F3] (List) > 3: Math**.
Die Funktionen können auch im Hauptbildschirm über **[CATALOG]** aufgerufen werden. Die Listen finden Sie im Menü **[VAR-LINK]** im entsprechenden Ordner, etwa im main-Ordner. Sie können die Funktionen und die Listennamen aber auch einfach eintippen, mit **mean(list1)** können Sie z.B. den Mittelwert der vorliegenden Daten berechnen, mit **stddev(list1)** die Standardabweichung.

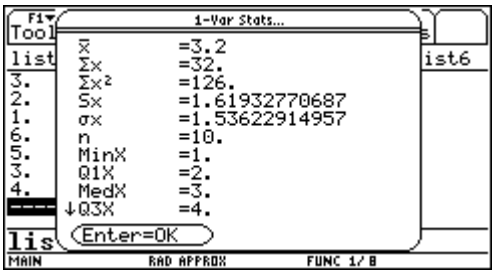


Einfacher erhalten Sie im Rechenmenü **[STAT] > CALC > 1:1-Var Stats** mit **1-Var Stats L1** die wichtigsten Kennzahlen, unter anderem:

arithmetisches Mittel	$\bar{x}$ = 3,2
Standardabweichung einer Stichprobe	Sx = 1,619 (n-1-Gewichtung!)
Standardabweichung der Gesamtpopulation	σx = 1,536 (n-Gewichtung!)
minimaler Wert	MinX = 1
erstes Quartil	Q1 =2
Median	Med = 3
drittes Quartil	Q3 = 4
maximaler Wert	MaxX = 6



Auf dem TI-92plus im Stats/List Editor über das Menü **[F4] (Calc) > 1:1-Var Stats**. Alle Outputs werden in Variablen gespeichert, die Sie überall abrufen können, $\bar{x}$ etwa unter dem Variablennamen **x_bar**, ∑x unter **sumx** (s.Handbuch statsle.pdf, download von Apps-Seite auf education.ti.com). Die Variablen sind in dem Ordner **statvars** gespeichert. Um die Variablen im Hauptbildschirm abzurufen, müssen Sie diesen Ordner als Pfad angeben: **statvars\x_bar** (entweder eintippen oder über **[VAR-LINK]** Ordner **STATVARS** abrufen



Hier begegnen wir einem grundlegenden Konzept von John Tukey, bei dem die Daten durch 25%-Quantile charakterisiert werden, der ***5-number-summary*** (Fünf-Punkte-Zusammenfassung):

Minimum, erstes Quartil, Median, drittes Quartil, Maximum

Um den Median zu bestimmen, untersuchen wir die sortierten Daten:

1 1 2 3 3 | 3 4 4 5 6
↑
Med

Jener Wert, der die Daten in zwei Hälften teilt, ist der Median, hier: $\text{Med} = \frac{3+3}{2} = 3$

Nun kann wieder von jeder Hälfte der Median bestimmt werden, wir erhalten das untere Quartil Q1 und das obere Quartil Q3:

1 1 [2] 3 3 | 3 4 [4] 5 6
Q1 Q3

Was ist, wenn wir eine ungerade Anzahl von Werten haben? Wir löschen von der sortierten Liste den ersten Wert 1 (im Statistik-Editor mit Cursor auf Listenelement **L1(1)** und mit **[DEL]** löschen) und haben nun in Liste L1 folgende 9 Werte:

1 2 3 3 [3] 4 4 5 6

Als Median erhalten wir $x_5 = 3$. Sollen wir nun für die Ermittlung der Quartile den Median in die beiden Hälften aufnehmen oder sollen wir ihn draußen vor der Tür lassen?
Hier scheiden sich die Geister:
Tukey „faltet“ die Daten und nimmt die „Angelpunkte“ (die Quartile heißen bei ihm „hinges“) in die 5-number-summary:

1 [3] 6
2 3 4 5
[3] [4]
Q1 Q3

Den Median nimmt Tukey in beide Hälften auf und errechnet Q1 = 3 und Q3 = 4.

So rechnet auch EXCEL (Funktion QUARTILE).

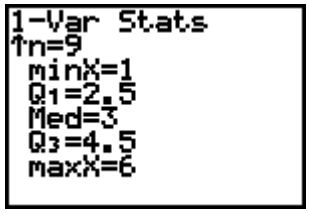
	A	B	C	D	E
1	1				
2	2				
3	3				
4	3				
5	3				
6	4				
7	4				
8	5				
9	6				
10					
11	Min =	1	=QUARTILE($A$1:$A$9;0)		
12	Q1 =	3	=QUARTILE($A$1:$A$9;1)		
13	Med =	3	=QUARTILE($A$1:$A$9;2)		
14	Q3 =	4	=QUARTILE($A$1:$A$9;3)		
15	Max =	6	=QUARTILE($A$1:$A$9;4)		

Das hat den Vorteil, dass bei 5 Werten diese mit der 5-number-summary übereinstimmen:

Min		Med		Max
1		3		5
	2		4	
	Q1		Q3	

Die TI-Rechner aber lassen den Median weg und errechnen

1 2 | 3 3 **3** 4 4 | 5 6
Q1=2,5 Q3 = 4,5



SPSS liefert dasselbe Ergebnis wie die TI-Rechner:

Statistiken

HAUSHALT

N	Gültig	9
	Fehlend	1
Perzentile	25	2,50
	50	3,00
	75	4,50

Sie sehen, die Definition der Quartile ist nicht eindeutig! Tatsächlich erhalten wir mit verschiedenen Technologien unterschiedliche Ergebnisse: Unter Umständen liefern die TI-Rechner andere Ergebnisse als EXCEL, EXCEL andere als SPSS, willkommen im Sumpf der Statistik.

Man muss sich auf eine mathematische Definition einigen, etwa auf folgende:

Das *p*-Quantil berechnet sich für

np nicht ganzzahlig: $\overline{x}_p = x_{[np+1]}$, wobei [ ] größte ganze Zahl ≤ dem Klammerausdruck ist (Gauß-Klammer); *p* der Anteil und *n* die Anzahl der Elemente

np ganzzahlig: $\overline{x}_p = \frac{x_{np} + x_{np+1}}{2}$

(vgl. [4], S.23)

Danach wäre $\overline{x}_{0,25} = x_{[0,25 \cdot 9+1]} = x_3 = 3 = \text{Q1}$ und $\overline{x}_{0,75} = x_{[0,75 \cdot 9+1]} = x_7 = 4 = \text{Q3}$

Lakonisch heißt es im Statistik-Buch [2] dazu: „Statistische Programmpakete benützen zum Teil unterschiedliche Definitionsvarianten, durch die sich abweichende Quantilswerte ergeben können." ([2], S.63)

Also: Es lassen sich sinnvoll nur Klassen angeben, in denen die Quartile liegen. Wie dann innerhalb dieser Klassen das Quartil berechnet wird, hängt vom statistischen Modell bzw. von der verwendeten Software ab. (Schließlich ist auch die Festlegung des Medians als Mittelwert der beiden mittleren Werte eine willkürliche Definition, eigentlich besitzt jeder Wert zwischen den beiden mittleren Werten die Eigenschaft, dass er das Datenmaterial in zwei gleiche Hälften teilt.)
Das einfachste, aber wenig aussagekräftige Streuungsmaß ist die Spannweite

$$RANGE = maxX - minX$$

Mit Hilfe der Quartile kann neben der Standardabweichung und der Spannweite als weiteres Streuungsmaß der ***Interquartilabstand*** IQR (interquartile range, bei Tukey „H-spread = difference between values of hinges") definiert werden:

$$IQR = Q3 - Q1$$

hier: IQR = 4 – 2 = 2
Im Intervall [Q1; Q3] liegen 50% der Werte, 25% der Werte liegen jeweils unterhalb und oberhalb dieses Intervalls.

Eine besondere Eigenschaft des Interquartilabstands ist seine Resistenz oder Robustheit gegenüber „Ausreißern"! (Was ein „Ausreißer" eigentlich ist, muss erst noch definiert werden!) Der IQR hat gegenüber der Standardabweichung bei den Streuungsmaßen also jenen Vorteil, den der Median gegenüber dem arithmetischen Mittel bei den Zentralmaßen hat: Unempfindlichkeit gegenüber Ausreißern, damit auch Unabhängigkeit von schlechten oder wenig verlässlichen Messungen. (Ein Übertragungsfehler kann sich fatal auf das arithmetische Mittel auswirken, der Median bleibt davon unbeeindruckt.)

Was ist nun ein „Ausreißer"? Ein Wert, der „weit von der Masse der Werte entfernt" liegt. Aber wie weit muss ein Wert von der Masse der Werte abweichen, dass wir ihn als „Ausreißer" identifizieren? Mit Hilfe des IQR hat Tukey ein Kriterium für „Ausreißer" angegeben: ein Wert, der um mehr als das 1,5fache des IQR („Step = 1.5 times H-spread")von den entsprechenden Quartilen abweicht. Er definiert den ***inneren Zaun*** („inner fences" are 1 step outside hinges):

$$[z_u; z_o] \text{ mit } z_u = Q1 - 1{,}5*IQR \text{ und } z_o = Q3 + 1{,}5*IQR$$

x_i heißt ***Ausreißer (outside value*** oder ***outlier)***, wenn $x_i \notin [z_u; z_o]$.

In unserem Beispiel:
$z_u = Q1 - 1{,}5*IQR = 2 - 1{,}5*2 = -1$; $z_o = Q3 + 1{,}5*IQR = 4 + 1{,}5*2 = 7$
Alle Werte liegen innerhalb des Zauns [-1; 7], in unserem Datensatz gibt es keinen Ausreißer.

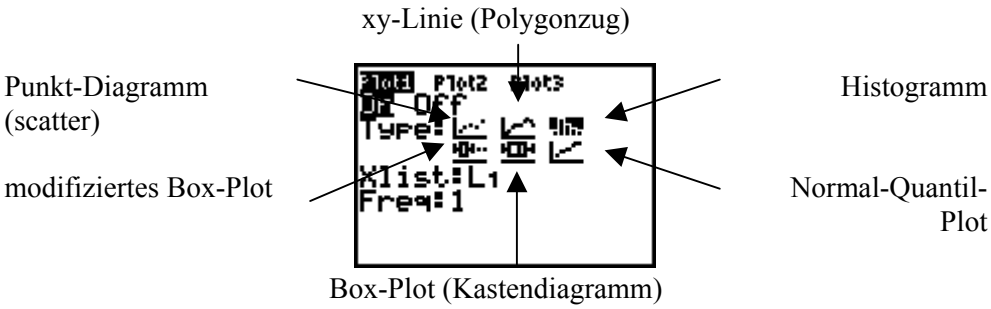
Warum nimmt Tukey gerade die 1,5fache IQR als Abweichungsmaß? Auf diese Frage soll Tukey geantwortet haben: „because 1 is too small and 2 is too large."

"Where do we stand?", fragt John Tukey an dieser Stelle in seinem Buch und meint:
„We have not looked at our results until we have displayed them effectively." ([1], S.56)

Also nun zu den grafischen Darstellungen von Datenmaterial!

2. Grafische Darstellungen

Der TI-83 bietet im Menü [**STAT PLOT**] sechs verschiedene Darstellungsformen für Listen an:

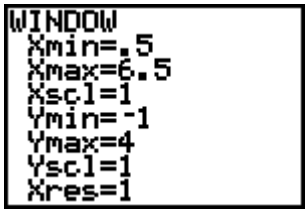


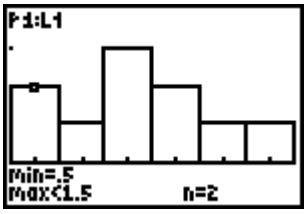
Auf dem TI-92plus finden Sie diese Grafiken im Stats/List Editor im Menü **[F2] (Plots) > 1:Plot Setup > [F1] Define**:

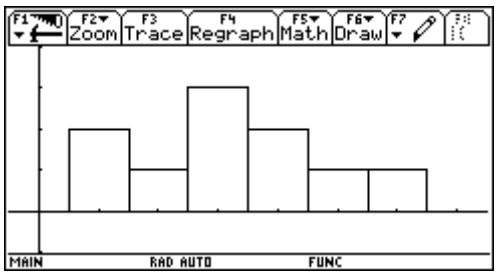
2.1. Histogramm

Für die grafische Darstellung der Daten müssen Sie den Statistik-Plot **[STAT PLOT] > 1:Plot1** aufrufen. (Deaktivieren Sie die eingegebenen Funktionen Y1 bis Y0.) Wählen Sie bei „Type“ das Histogramm . Als „Xlist“ geben Sie L1 ein.

Wählen Sie ein geeignetes Bildschirmfenster im [**WINDOW**]-Menü:
Xmin = 0,5 (kleiner als minX=1), Xmax = 6,5 (größer als maxX=6), Xscl=1 Klassenbreite!; Ymin=-1 (Platz lassen unter der Abszissenachse für die Beschriftung); Ymax=4 (größer als die höchste Klassenhäufigkeit)

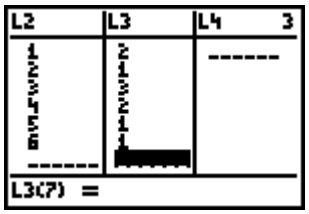


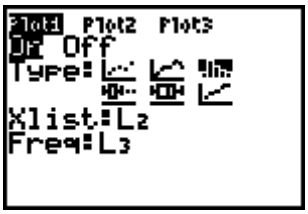




Hier wurde Xmin = -0.5 und Xmax = 7.5 gewählt.
Als Hist. Bucket Width bleibt die Voreinstellung 1.

Mit [**GRAPH**] erhalten Sie ein Histogramm, das Sie mit [**TRACE**] abtasten können. So können Sie die Häufigkeitsverteilung erstellen, die Sie in den Listen L2 (Werte) und L3 (Häufigkeiten) speichern können:

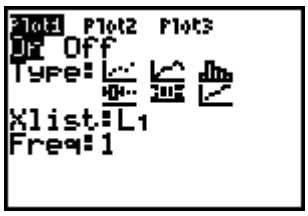


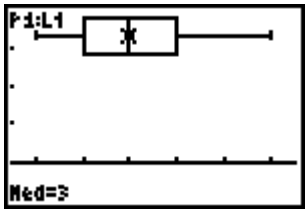


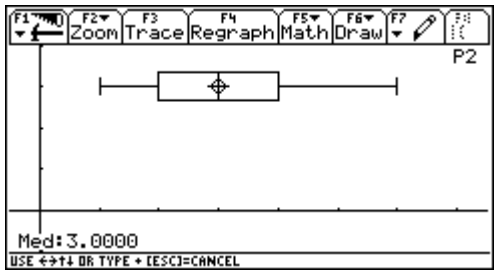
Sie erhalten wiederum den Statistik-Plot, indem Sie bei „Xlist“ L2 und bei „Freq“ L3 (Frequence = Häufigkeit) eingeben;
für die 1-Variablen-Statistik müssen Sie im Menü **[STAT] > CALC >1:1-Var-Stats** auswählen und die Liste L2 mit der Häufigkeit L3 angeben: **1-Var Stats L2,L3**

2.2. Box-Plot

Wenn Sie im Statistik-Plot bei „Type“ das Box-Plot-Diagramm wählen, wird die Fünf-Punkte-Zusammenfassung **minX, Q1, Med, Q3, maxX** in einem Box-and-Whisker-Plot dargestellt:



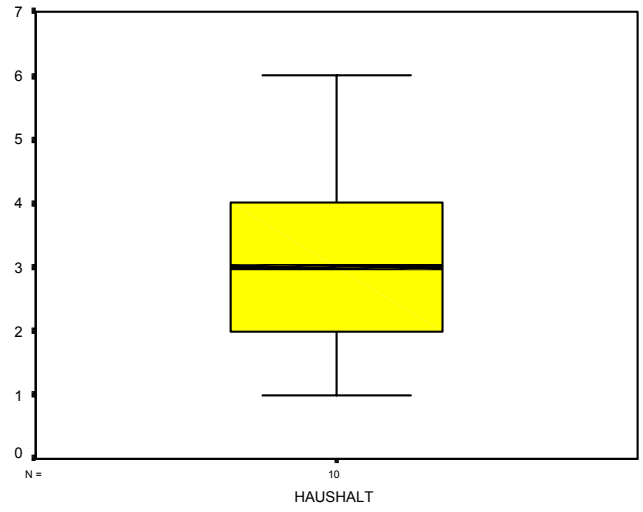




Aufbau des Box-Plots:
1. Die Box erstreckt sich über das Quartilintervall [Q1; Q3], die Länge der Box ist die IQR;
2. Der Median wird in der Box durch einen senkrechten Strich markiert;
3. Zwei Linien (Whiskers = Schnurrhaare der Katze) außerhalb der Box gehen bis minX und maxX.

Diese grafische Darstellung der Daten eignet sich sehr gut zum Vergleich verschiedener Verteilungen. Es lässt sich schnell ein Eindruck darüber gewinnen, ob die Beobachtungen annähernd symmetrisch verteilt sind und ob Ausreißer in dem Datensatz auftreten.

SPSS gibt die Box-Plots standardmäßig so aus:



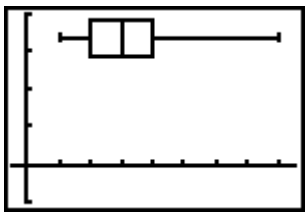
In EXCEL gibt es den Box-Plot nicht als Standard-Diagramm, ein Box-Plot kann in EXCEL nur höchst trickreich erstellt werden. Hier sind die TI-Rechner moderner als EXCEL!

2.3. Modifiziertes Box-Plot

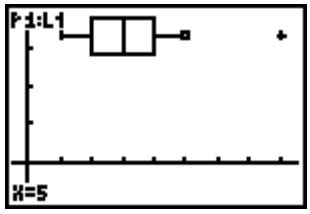
Mit dem modifizierten Box-Plot können „Ausreißer“ identifiziert werden. Die Whiskers außerhalb der Box werden nur dann bis zu minX und maxX gezogen, wenn minX und maxX innerhalb des Zauns $[z_u; z_o]$ liegen. Ansonsten gehen die Whiskers nur bis zum kleinsten bzw. größten Wert innerhalb des Zauns und die außerhalb liegenden Werte werden als „Ausreißer“ individuell markiert.

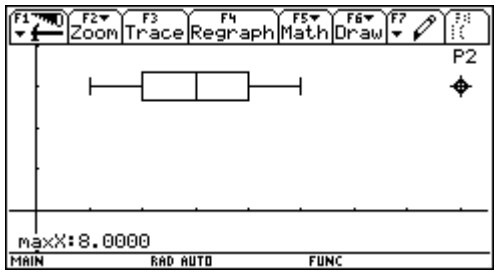
In unserem Beispiel: $[z_u; z_o] = [-1; 7]$
Wir ersetzen den Wert 6 durch 8 (in der Liste L1 überschreiben). Damit haben wir einen Ausreißer produziert. Beim normalen Box-Plot erhalten wir einen langen Whisker, beim modifizierten Box-Plot wird der Ausreißer individuell markiert, die Box reicht nur bis zum Wert 5:

Box-Plot:



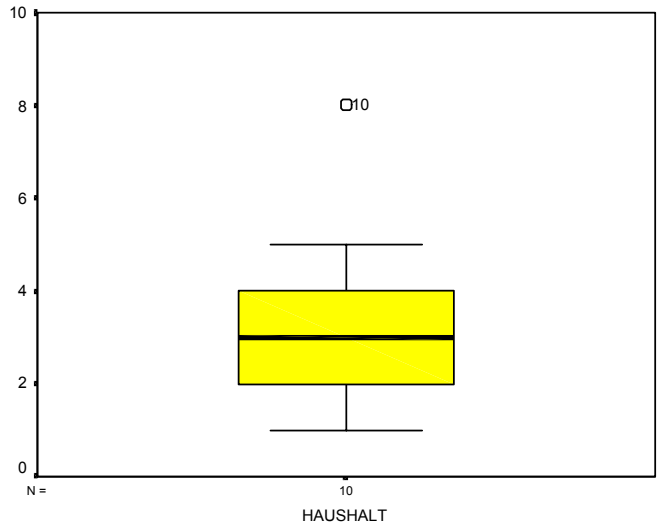
modifiziertes Box-Plot:





Natürlich sind die [WINDOWS]-Werte anzupassen!

Mit SPSS:



2.4. Normal-Quantil-Plot

Eine wesentliche Grundfrage der Statistik lautet: Sind die erhobenen Daten annähernd normalverteilt. Um dies zu entscheiden, kann über das Histogramm die Normalverteilungskurve mit entsprechendem Mittelwert und Standardabweichung gelegt werden.
Die explorative Datenanalyse hat für die Untersuchung dieser Fragestellung Normal-Quantil-Plots entwickelt, bei denen die Quantile der Häufigkeitsverteilung mit entsprechenden Quantilen der Standardnormalverteilung verglichen werden. Dazu fasst man die geordneten Werte $x_{(1)}, ..., x_{(n)}$ als Quantile der Häufigkeitsverteilung auf und trägt sie gegen entsprechende Quantile der Standardnormalverteilung ab.

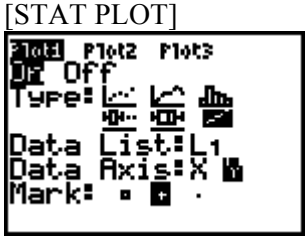
In unserem Beispiel: n = 10 (ursprüngliche Liste!), dann ist $x_{(1)}$ das 1/10 = 0,1-Quantil, $x_{(2)}$ das 2/10 = 0,2-Quantil usw. Allerdings hat es sich als günstig erwiesen, statt dieser einfachen Quantile die korrigierten Quantile (i – 0,5)/n aufzutragen. Durch diese Stetigkeitskorrektur wird die Approximation der empirischen Verteilung durch eine Normalverteilung verbessert.
Für diese Quantile werden dann die (i – 0,5)/n-Quantile z_i der Standardnormalverteilung berechnet.
Der Normal-Quantil-Plot besteht aus den Punkten $(z_i|x_{(1)}), ..., (z_n|x_{(n)})$ im z-x-Koordinatensystem.

i	1	2	3	4	5	6	7	8	9	10
sortierte Werte x_i	1	1	2	3	3	3	4	4	5	6
Quantile	1/10	2/10	3/10	4/10	5/10	6/10	7/10	8/10	9/10	10/10
korrigierte Quantile	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0,95
z-Werte z_i	-1,645	-1,036	-0,674	-0,385	-0,126	0,126	0,385	0,674	1,036	1,645

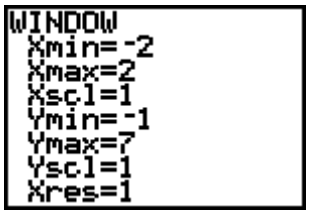
$z_1 = -1{,}645$ ergibt sich aus $\Phi^{-1}(0{,}05)$ usw.

Man sieht: Der Rechenaufwand ist beträchtlich! Wir übergeben die Berechnung dem TI-83:

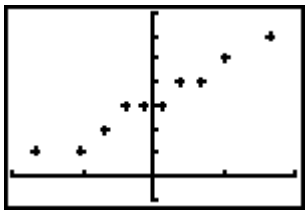
[STAT PLOT]



[WINDOW]



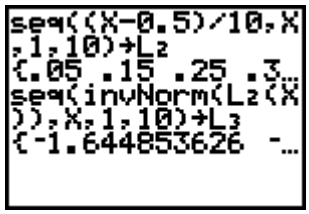
[GRAPH]



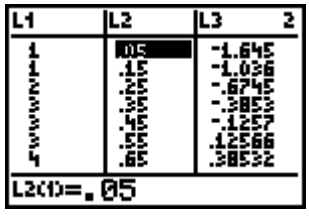
Für Spezialisten: Mit Listen können die Berechnungen nachvollzogen werden:
In der Liste L2 geben wir die korrigierten Quantile 0,05, 0,15,..., 0,95 ein. Das können wir elegant mit dem Sequence-Befehl **seq** aus dem Menü [LIST] > OPS > 5:seq erledigen:

seq(*Ausdruck*, *Variable*, *Anfangswert*, *Endwert* [,*Schrittweite*])

liefert eine Liste des ausgewerteten *Ausdrucks* in Abhängigkeit einer *Variable* für alle Werte vom *Anfangswert* bis zum *Endwert* (erhöht um die *Schrittweite*, deren Voreinstellung 1 ist).

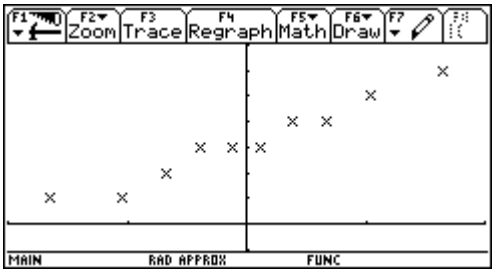


seq((X-0,5)/10, X, 1, 10) → L2
erzeugt die Liste {.05, .15, …, .95} und speichert diese in L2



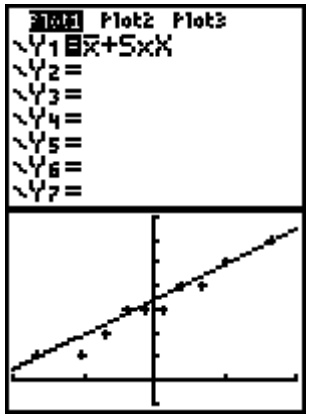
seq(invNorm(L2(X)), X, 1, 10) → L3
erzeugt die Liste {-1.645, -1.036, …, 1.645} und speichert diese in L3.
(die Funktion **invNorm** finden Sie im Menü **[DISTR] > 3:invNorm**, sie berechnet Quantile der Normalverteilung)

Auf dem TI-92plus finden Sie den Normal-Quantil-Plot im Stats/List Editor in einem eigenen Menü:
[F2] (Plots) > 2:Norm Prob Plot.
Der Rechner erzeugt automatisch eine Liste **zscores**, in der die oben berechneten z-Werte gespeichert und im Stats/List Editor angezeigt werden. (Falls im Stats/List Editor im Menü **[F1] (Tools) > 9:Formats** der Schalter **Results->Editor** auf **YES** gesetzt ist)



Mit den NQ-Plots lassen sich Schiefe und Wölbung im Vergleich zu Normalverteilungen und andere Besonderheiten der Daten erkennen. Ist die Variable X annähernd normalverteilt mit dem Mittelwert $\mu = \bar{x}$ und der Standardabweichung $\sigma = s_{n-1}$, so ist die standardisierte Variable $Z = (X - \mu)/\sigma$ annähernd standardnormalverteilt und die Punkte $(z_i|x_{(i)})$ des NQ-Plots liegen in etwa auf der Geraden $x = \mu + \sigma \cdot z$. Abweichungen von dieser Geraden indizieren dann Schiefe, Wölbung oder andere Besonderheiten wie Ausreißer oder Bimodalität.

Wir können in unserem Beispiel die Normalverteilungshypothese prüfen, indem wir diese Gerade im Y-Editor eingeben. Dazu verwenden wir die Statistik-Variablen $\bar{x}$ und **Sx** aus dem Menü **[VARS] > 5:Statistics**.
Sie sehen: Die Werte gruppieren sich recht eng um diese Gerade. Tatsächlich kann durch einen Normalverteilungstest, etwa den Kolmogorow-Smirnow-Test, die Hypothese, dieser Datensatz sei normalverteilt, nicht abgelehnt werden. Wer hätte das aus der Betrachtung des Histogramms behaupten wollen?



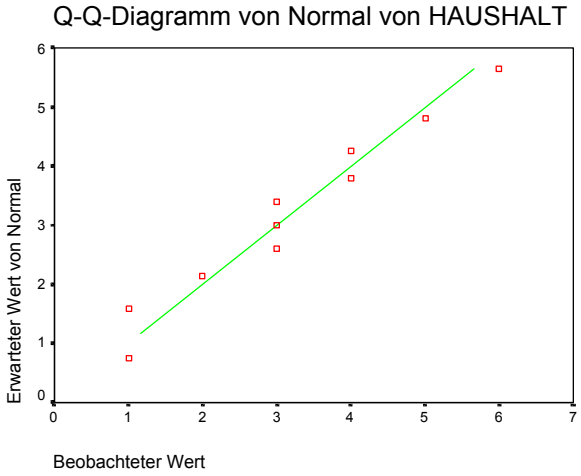
mit SPSS durchgeführte explorative Datenanalyse:

Tests auf Normalverteilung

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistik	df	Signifikanz	Statistik	df	Signifikanz
HAUSHALT	,151	10	,200*	,952	10	,666

*. Dies ist eine untere Grenze der echten Signifikanz.

a. Signifikanzkorrektur nach Lilliefors

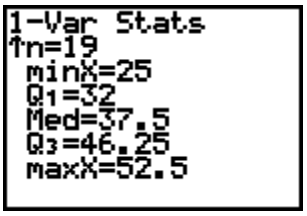


Anregungen für Aufgaben und kleine Projekte:

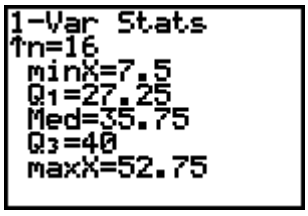
Ohne großen Aufwand können Echtdaten der Schüler/innen erhoben und in der Klasse ausgewertet werden. Die Schüler/innen erleben so in Ansätzen, wie empirische Sozialforschung betrieben wird. Die Schüler/innen machen das mit großer Begeisterung, da sie sich in dem Datenmaterial wieder finden können. Die Analyse nach einfachsten Faktoren (Geschlecht, Schulklassen) provoziert heftige Diskussionen. So drängt sich sehr rasch die Frage auf, ob Unterschiede „signifikant“ sind, oder ob Unterschiede aus der Zufälligkeit der Stichprobenerhebung zu erklären sind. Den Schüler/innen wird sehr früh schon bewusst, dass die statistische Analyse bei der Deskription und Exploration nicht stehen bleiben darf. Aber wie sagte Tukey: Exploration kann nicht alles sein, aber es ist der erste Schritt.

Ü1. Auswertung der Punkte-Ergebnisse des Känguru-Tests 2002 der Kategorie Junior aller zweiten Jahrgänge der BHAK Schwaz:
statistische Kennzahlen:

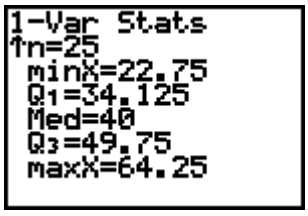
2AK:



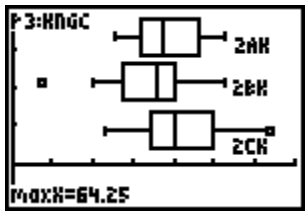
2BK:



2CK:



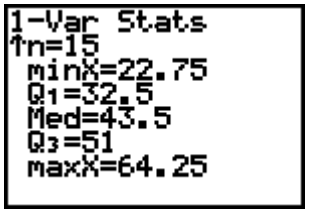
Boxplots:



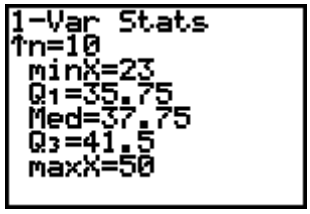
Klassische Fragen der explorativen Datenanalyse:
Gibt es Ausreißer?
Gibt es auffallende Unterschiede zwischen den Klassen? Welche Klasse hat „am besten“ abgeschnitten?

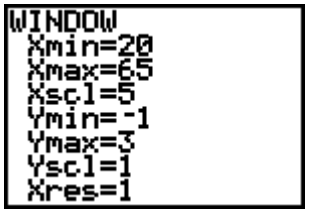
Ergebnis nach Geschlecht in der 2CK:

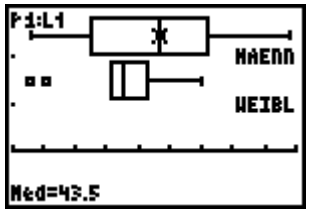
männlich:



weiblich:







Haben die Buben besser abgeschnitten als die Mädchen?
Gibt es Ausreißer?
Wie streuen die Punkte bei Mädchen und wie bei den Buben?

Ü2. Wir erheben die Körpergröße der Schüler/innen aller Klassen eines Jahrgangs und stellen die Verteilung der Körpergröße a) nach dem Faktor Schulklasse, b) nach dem Faktor Geschlecht in Box-Plots dar.
Mögliche Fragestellungen: Gibt es Ausreißer in dieser Population? (Besonders große oder besonders kleine Schüler/innen) Gibt es auffallende Unterschiede zwischen den Schulklassen? (Vermutlich wird die Verteilung keine auffallenden Unterschiede liefern) Gibt es auffallende Unterschiede zwischen den Geschlechtern? (Vermutlich werden die Buben auffallend größer sein als die Mädchen)

Ü3. Wir werten die Punkte-Ergebnisse einer Mathematik-Schularbeit nach dem Faktor Geschlecht aus und erstellen zwei Box-Plots der Punkteverteilung für Mädchen und für Buben? Analoge Fragestellungen wie oben. (Ausreißer, auffallende Unterschiede zwischen den Geschlechtern)

Ü4. Wie jedes Jahr soll auch diesmal die 8. Klasse mit den besten Sportleistungen belohnt werden. Zur Entscheidungshilfe werden die Ergebnisse im Weitsprung herangezogen. Zwei Stunden vor der Siegerehrung hängen die Ergebnisse des diesjährigen Sportfestes aus. Vanessa, Alex und Basti schauen sich die Weitsprung-Tabellen der 8. Klassen an:

Klasse 8a: (LW8A)
4.32 4.19 4.30 4.44 4.31 4.09 4.13 4.72 4.40 4.20 4.15 4.35 4.20 4.56 4.31 4.27 4.32 4.16 4.25 4.33

Klasse 8b: (LW8B)
4.12 4.20 4.42 4.38 4.50 4.53 4.03 4.19 4.14 4.07 4.41 4.36 4.53 4.32 4.19 4.21 4.50 4.01 4.12 4.27 4.42 4.39

Klasse 8c: (LW8C)
4.11 4.54 4.66 4.22 4.20 4.40 4.12 4.65 4.08 4.24 4.12 4.57 4.13 4.21 4.07 4.12 4.62 4.57 4.11 4.08 4.15 4.52 4.57 4.01

Klasse 8d: (LW8D)
4.32 4.21 4.28 4.34 4.46 4.30 4.30 4.32 4.27 4.22 4.33 4.26 4.36 4.10 4.39 4.23 4.34 4.37 4.21 4.37 4.22 4.21 4.30

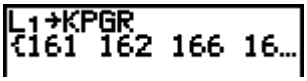
Die drei diskutieren, welche Klasse wohl den alljährlichen Sonderpreis für die besten Leistungen bekommen wird:
Alex: „Ich fände es gut, wenn die Klasse mit dem besten Springer den Preis bekommt!“
Vanessa: „Quatsch! Das wäre doch ungerecht, wegen einem guten Sportler die ganze Klasse zu belohnen. Man sollte den Preis der Klasse mit den meisten guten Springern geben.“
Basti: „Und wenn in dieser Klasse auch die ganzen schlechten Springer sind? Ich meine, der Preis sollte an die Klasse mit den ausgeglichensten Ergebnissen gehen.“
Wie denkst du darüber?
Die folgenden Fragen können bei der Entscheidung helfen:
1) Welche Klasse ist die „beste“?
2) In welcher Klasse ist eine Leistung von 4.40 „am meisten wert“, d.h., in welcher Klasse gehört man mit dieser Sprungweite zu den besseren Sportlern?
3) Welche Klasse ist die „ausgeglichenste“?
4) Welche Klasse hat die „stärkste Spitze“?
(aus: [5])

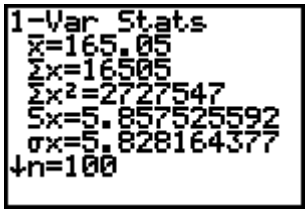
3. Klassenbildung

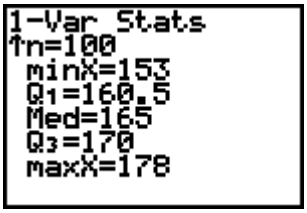
Beispiel: Erwin Kreyszig arbeitet in seinem Buch „Statistische Methoden und ihre Anwendungen“ [6] mit Daten, die er vom statistischen Amt des Magistrats Graz bekommen hat, der Größe von 100 achtzehnjährigen Mittelschülerinnen: (LKPGR)

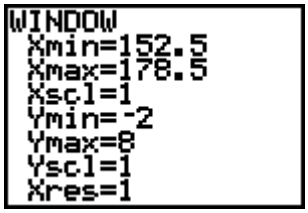
161	162	166	161	171	159	160	174	165	163
161	178	157	156	160	172	167	162	164	156
177	162	167	168	157	164	176	166	171	169
171	155	170	158	171	167	161	172	169	161
160	164	162	170	168	165	173	159	173	166
170	154	165	162	174	158	156	165	160	165
172	167	173	166	164	168	175	158	163	169
171	166	159	162	159	171	163	158	167	168
163	153	172	170	158	164	162	175	165	169
170	155	169	159	163	159	166	157	166	175

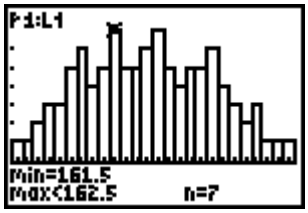
Diese Daten geben wir in **L1** ein und speichern sie mit **L1 [STO>]** in der Liste **KPGR**. Diese kann nun über **[LIST] > NAMES** aufgerufen werden. Statistische Daten mit Grafik:

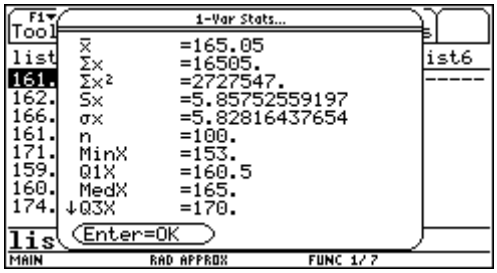


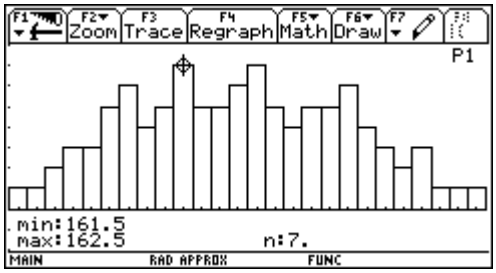












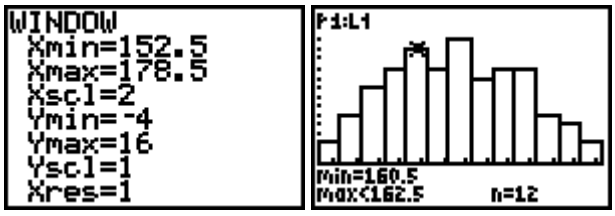
Hier ist es sinnvoll, mehrere Werte in Klassen zusammenzufassen.
Für die Anzahl *m* der zu bildenden Klassen bei einer Stichprobe vom Umfang *n* gelten folgende *Faustregeln*: (vgl. [4], S.17)

Stichprobenumfang:	Klassenanzahl:
$n \leq 30$:	$m = 5$
$30 < n < 400$	$m \approx \sqrt{n}$
$n \geq 400$	$m = 20$

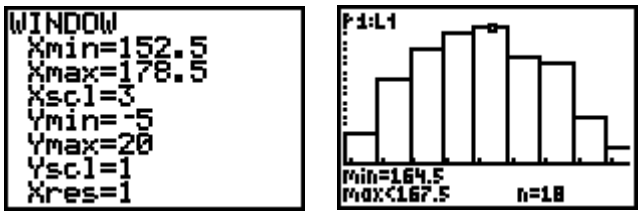
Für unser Datenmaterial empfehlen sich also ca. $m = \sqrt{100} = 10$ Klassen.
Die Spannweite beträgt RANGE = 178 – 153 = 25.
Die Klassenbreite beträgt also 25/10 = 2,5; sinnvoll ist also die Klassenbreite 3.

Mit dem TI-83 können nun leicht Klassen verschiedener Breite gebildet werden. Dazu müssen wir nur im Menü [WINDOW] die Skalierung der x-Achse **Xscl** verändern.

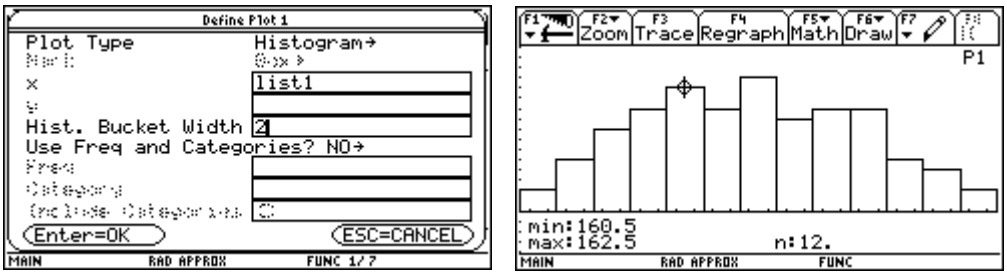
Für Xscl=2 erhalten wir 13 Klassen:



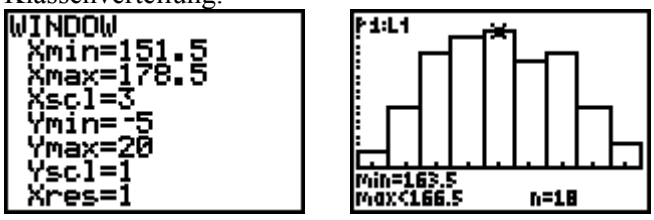
Für Xscl=3 erhalten wir 9 Klassen:

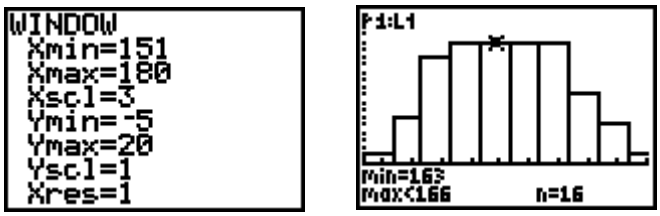


Auf dem TI-92plus können Sie die Klassenbreite mit **Hist. Bucket Width** einstellen (im Stats/List Editor Menü **[F2] (Plots) > 1:Plot Setup > [F1] Define)**

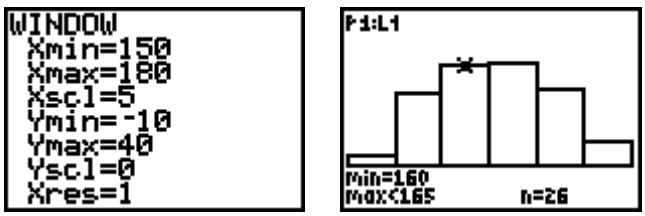


Sie können auch die Klassenuntergrenze Xmin beliebig verändern und erhalten jeweils eine andere Klassenverteilung.

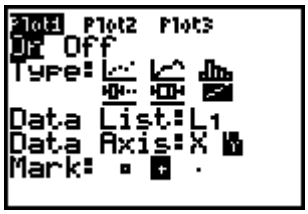


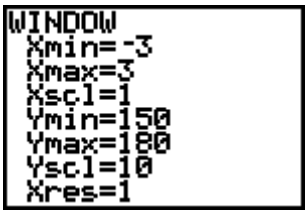


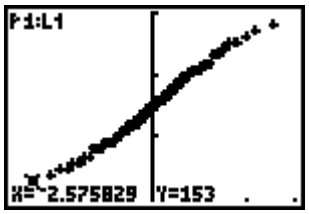
Für Xscl=5 erhalten wir 6 Klassen:



Hier zeigt sich die Normalverteilung der Körpergröße recht deutlich. Das können wir noch durch einen NQ-Plot überprüfen:



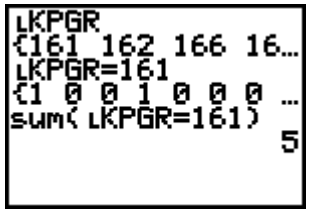




Häufigkeitsverteilung

Aus dem Histogramm können wir mit **[TRACE]** die absoluten Häufigkeiten der einzelnen Körpergrößen ablesen. Wie können wir diese Häufigkeiten in einer Liste berechnen lassen?

Der TI-83 verfügt über eine seltsame Funktion, die der TI-92 in dieser Form nicht hat: Er erzeugt bei logischen Abfragen mit Listen eine neue Liste mit den Wahrheitswerten 0 für falsch bzw. 1 für wahr. Beim TI-92 werden die Wahrheitswerte true, bzw. false ausgegeben, daher muss man etwas anders an die Sache herangehen, wie im Anschluss demonstriert wird.



Z.B. erzeugt die logische Abfrage **ʟKPGR=161** eine neue Liste mit derselben Dimension wie **ʟKPGR**, in der statt der Werte 161 eine 1 steht und 0 sonst. (Diesen Hinweis verdanke ich Walter Heinzle.) Die Summe dieser Liste

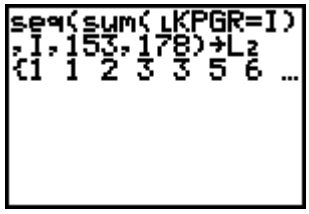
sum(ʟKPGR=161) ([LIST] > MATH > 5:sum)

liefert die absolute Häufigkeit 5 des Werts 161 in der Liste **ʟKPGR**.
Mit Hilfe des *Sequence*-Befehls aus dem Menü **[LIST]**

seq(*Ausdruck, Variable, Anfang, Ende* [,*Schrittweite*])

der eine Liste des ausgewerteten *Ausdrucks* in Abhängigkeit einer *Variable* für alle Werte vom *Anfangswert* zum *Endwert* (erhöht um die *Schrittweite*) liefert, kann nun eine Liste der absoluten Häufigkeiten erzeugt werden:

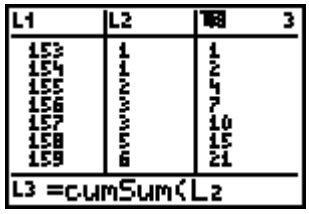
seq(sum(ʟKPGR=I),I,153,178)



Wir erzeugen nun eine Liste der 26 verschiedenen Körpergrößen 153 bis 178:

seq(X, X, 153, 178) → L1

Mit **cumSum(L2**) in der Liste L3 (aus **[LIST] > OPS > 6:cumSum**) kann eine zusätzliche Liste mit den kumulierten Häufigkeiten erzeugt werden.

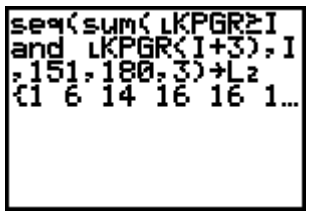


Nun erzeugen wir die Häufigkeitsverteilung der Klasseneinteilung mit der Breite 3 und dem Bereich 151 bis 180, indem wir die Schrittweite auf 3 erhöhen (Histogramm s. oben):

seq(sum(ʟKPGR ≥I and ʟKPGR <I+3), I, 151, 180, 3) → L2

einige Details zur Eingabe der Relationszeichen und logischen Verknüpfungen:

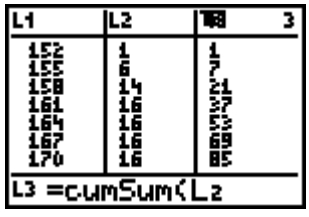
≥ : [TEST] > 4: ≥
and: [TEST] > LOGIC > 1:and
< : [TEST] > 5: <



Überprüfen Sie in der Grafik mit [TRACE] die Häufigkeiten!
Mit

seq(X, X, 152, 179, 3)

erzeugen wir die Liste der Klassenmitten.

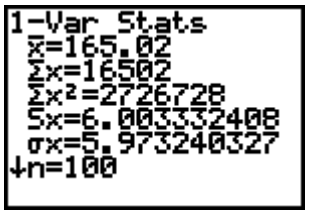


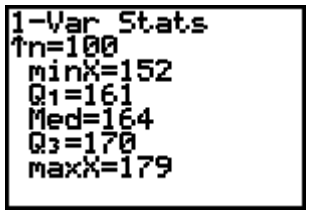
Mit

1-Var Stats L1,L2

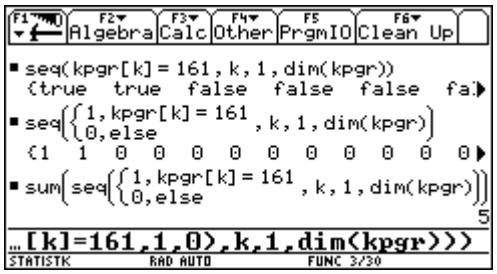
kann die Klasseneinteilung statistisch ausgewertet werden.

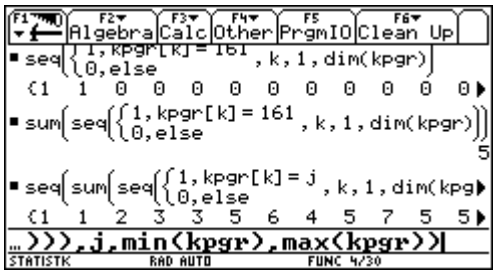
Beachten Sie, dass der Mittelwert der Klassen–einteilung 165,02 nur ganz wenig vom Mittelwert der Urliste 165,05 abweicht.





Nun zu Behandlung am TI-92





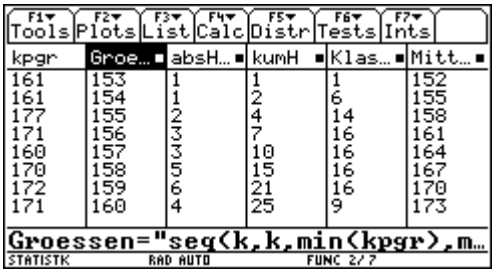
seq(kpgr[k]=161,1,dim(kpgr)) → {true, true, false, ...} – hat die Größe den Wert 161?

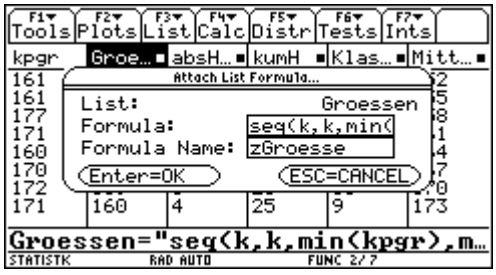
seq(when(kpgr[k]=161,1,0),k,1,dim(kpgr)) → {1, 1, 0, ...} – 1 für true und 0 für false

sum(seq(when(kpgr[k]=161,1,0),k,1,dim(kpgr)) → 5 – 5 mal tritt die Größe 161 auf

seq(sum(seq(when(kpgr[k]=j,1,0),k,1,dim(kpgr))),j,min(kpgr),max(kpgr)) → {1, 1, 2, ...} – die geordnetenabsoluten Häufigkeiten **abshaeuf**

Damit steht uns das Rüstzeug für den Aufbau der Listen im Listeneditor zur Verfügung. Die listen-erzeugenden Formeln werden direkt im Editor eingegeben. (Sie können mit Cut and Paste auch aus demHomescreen übertragen werden.





In der zweiten Spalte wird eine Liste der möglichen Größen, beginnend mit dem kleinsten Wert der Liste **kpgr** und endend mit dem größten erzeugt. Die Formeln werden über F3 > 4:Attach List Formula in die Liste eingefügt - **groessen**.

seq(k,k,min(kpgr),max(kpgr)) -→ {153, 154, ..., 178}

Daneben platzieren wir die Liste der absoluten Häufigkeiten **abshaeuf** , welche nun geordnet auftreten.

Die kumulierten Häufigkeiten **kumH** ergeben sich dann mit **cumSum(abshaeuf))**

Für die Klasseneinteilung erzeugen wir ähnlich wie beim TI-83 die Liste **Klasse3** über die Formel

seq(sum(seq(when(kpgr[k] ≥ j and kpgr[k] < j+3,1,0),k,1,dim(kpgr))),j,151,180,3)

und über **seq(k,k,151,179,3)** die Liste **Mitten** der Klassenmitten.

4. Histogramm mit Normalverteilung

Im Programmpaket SPSS kann automatisch in einem Histogramm diejenige Normalverteilungskurve dargestellt werden, die sich am besten an die Daten anpasst ($\mu = \bar{x}$, σ = s). Mit den TI-Rechnern können wir diese Funktionalität eines professionellen Programmpakets recht elegant selbst basteln. Die Schüler bekommen so schon in der deskriptiven Statistik einen guten visuellen Eindruck, ob empirische Daten normalverteilt sind oder nicht.

Wir führen die eindimensionale Statistik-Analyse **1-Var Stats** der Liste **KPGR** durch, sie liefert folgende Werte:

Mittelwert $\bar{x}$=165,05;

Standardabweichung σx=5,828.

Diese Kennzahlen sind nun in den Variablen $\bar{x}$ und σx gespeichert, die Sie aus im Menü **[VARS] > 5:Statistics > XY** aufrufen können

Wir definieren im [Y=]-Editor

Y1=normalpdf(X, $\bar{x}$, σx)*Xscl*dim(ʟKPGR)

Details:

normalpdf: Dichtefunktion der Normalverteilung (**normal p**robability **d**ensity **f**unction) im Menü [DISTR] > 1:normalpdf;

$\bar{x}$ und **σx:** Variablenmenü [VARS] > 5:Statistics,

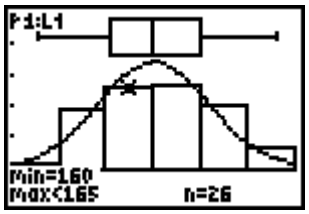
Xscl: Skalierung auf der x-Achse in [VARS] > 1:Window > 3:Xscl;

dim: Anzahl der Elemente der Liste in [LIST] > OPS > 3:dim.

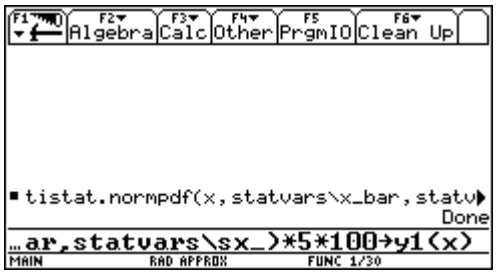
Aktivieren Sie den Plot1 mit dem Histogramm und den Plot2 mit dem BoxPlot.

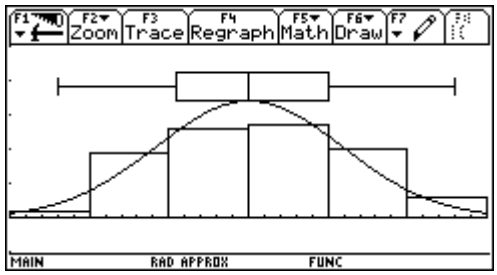
Für die [WINDOW]-Einstellungen x∈[150;180] mit Xscl=5 und y∈[-10;50] erhalten Sie folgende Verteilung:

Klasse	Klassen-mitte	empir. Wert	theoret. Wert
150 - <155	152,5	2	3
155 - <160	157,5	19	15
160 - <165	162,5	26	31
165 - <170	167,5	27	31
170 - <175	172,5	20	15
175 - <180	177,5	6	4



Auf dem TI-92plus: Normalverteilungskurve in y1(x) speichern

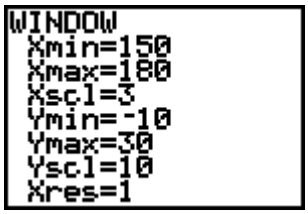


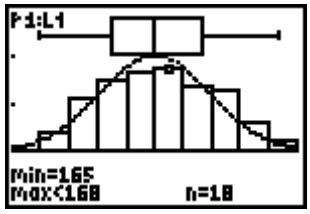


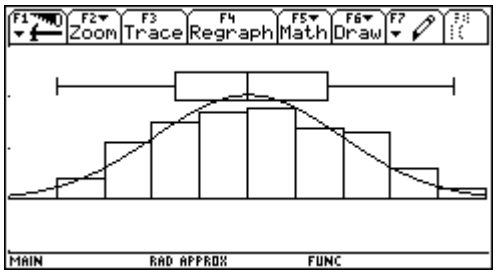
normpdf im Menü [CATALOG] > [F3] (Flash Apps);
x_bar und **sx_** im Menü [VAR-LINK] > STATVARS

Sie können nun die Klassenbreite Xscl im [WINDOW]-Menü beliebig verändern (und die Einstellungen für die y-Achse entsprechend anpassen), Sie erhalten nun immer zur Häufigkeitsverteilung die Normalverteilungskurve gezeichnet.

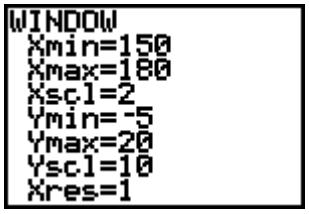
Eine Klasseneinteilung mit Klassenbreite Xscl = 3:

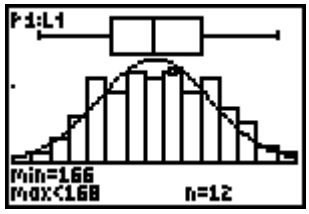


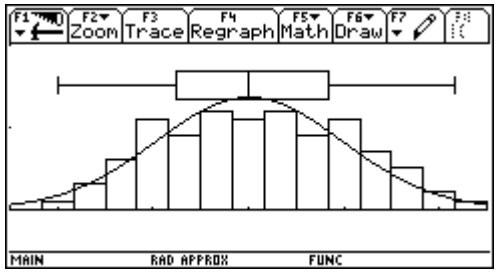




Klasseneinteilung mit Klassenbreite Xscl = 2:

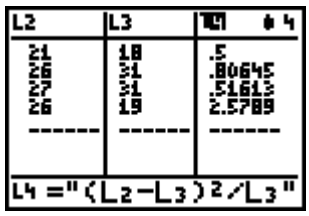






5. CHI-Quadrat-Test auf Normalverteilung [(1)]

Um die Anpassung zu testen, führen wir einen CHI-Quadrat-Test durch.
Dazu fassen wir die Daten so in Klassen zusammen, dass in jeder Klasse mindestens fünf Werte stehen. In unserem Beispiel raffen wir die beiden ersten und die beiden letzten Klassen und erhalten dadurch nur noch 4 Klassen, deren empirische Werte wir in L2 und deren theoretischen Wert in L3 speichern.



Klasse	Klassen-grenzen	Klassen-mitte	empir. Wert O_i	theoret. Wert E_i
1	150 - <160	155	21	15
2	160 - <165	162,5	26	31
3	165 - <170	167,5	27	31
4	170 - <180	175	26	19

(1) Es kann hier nicht auf die Grundlagen des χ^2-Tests eingegangen werden. Ich verweise auf die einschlägige Literatur.

Nun berechnen wir den Chi-Quadrat Testwert

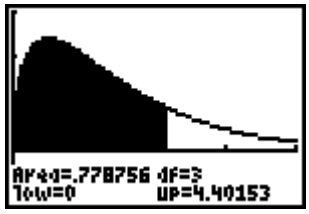
$$c = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(21-15)^2}{15} + \frac{(26-31)^2}{31} + \frac{(27-31)^2}{31} + \frac{(26-19)^2}{19} = 4{,}4$$

Elegant berechnen wir diesen Chi-Quadrat-Testwert in L4 mit (L2-L3)²/L3.

Die Summe ergibt c=4,4.
Für 3 Freiheitsgrade ergibt die Wahrscheinlichkeit P(X≤c) = **χ²cdf(0,C,3)** = 0,778.. < 0,95, d.h. die Nullhypothese kann nicht verworfen werden, d.h. man kann annehmen, dass die Daten normalverteilt sind. (χ²cdf finden Sie im Menü [DISTR] > 7: χ²cdf.)

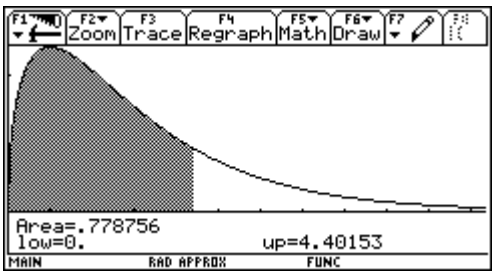
ACHTUNG: Der TI-83 rechnet hier anders als die klassische Teststatistik! In der klassischen Teststatistik wird immer ein kritischer Wert berechnet, ab dem die Nullhypothese verworfen wird. Für die Chi-Quadrat-Verteilung mit df = 3 Freiheitsgraden und dem Signifikanzniveau 95% ergibt sich 7,815.
Unser Testwert c = 4,4 ist kleiner als diese kritische Grenze, es besteht also keine Veranlassung, die Nullhypothese, die Daten seien normalverteilt, abzulehnen.
Mit dem TI-83 können Sie aber die Wahrscheinlichkeit P(X≤4,4) = 0,778 berechnen! Diese Wahrscheinlichkeit enthält mehr Information als der kritische Wert.
Auch SPSS berechnet die Wahrscheinlichkeit, nämlich die Irrtumswahrscheinlichkeit 1- P(X≤c), die sogenannte „Signifikanz".

Mit **Shadeχ²(0,C,3)** (Menü [DISTR] > DRAW > 3:Shadeχ²) kann die χ²-Verteilung auch grafisch dargestellt und der gesuchte Wahrscheinlichkeitsbereich schattiert werden. Dazu muss aber vorher eine geeignete [WINDOW]-Einstellung gefunden werden, hier: x∈[0;8] mit Xscl=2 und y∈[-0,1;0,3].



Auf dem TI-92plus im Stats/List Editor im Menü [F5] (Distr) > 1:Shade > 3:Shade Chi-square

An diesem Beispiel sieht man die besondere Stärke der TI-Rechner, die in den Visualisierungsmöglichkeiten statistischer und wahrscheinlichkeitstheoretischer Verteilungen liegt.



6. Regressionsrechnung

Der TI-83 bietet im Menü **[STAT] > CALC** eine Fülle von Regressionskurven an mit der Syntax

FktReg [Xliste, Yliste, reggl]

3: Med-Med	Zentralwertlinie
4: LinReg (ax+b)	lineare Regression y=ax+b (minimale Fehlerquadratsumme)
5: QuadReg	quadratische Regression $y=ax^2+bx+c$
6: CubicReg	kubische Regression $y=ax^3+bx^2+cx+d$
7: QuartReg	Regression vierten Grades $y=ax^4+bx^3+cx^2+dx+e$
8: LinReg(a+bx)	lineare Regression y=a+bx
9: LnReg	logarithmische Regression $y=a+b\cdot\ln x$ (minimale Fehlerquadrate für linearisierte Gleichung
0: ExpReg	exponentielle Regression $y=a\cdot b^x$ (minimale Fehlerquadrate für linearisierte Gleichung $\ln y=\ln a+x\cdot\ln b$)
A: PwrReg	Potenzfunktionregression $y=a\cdot x^b$ (minimale Fehlerquadrate für linearisierte Gleichung $\ln y=\ln a+b\cdot\ln x$)
B: Logistic	logistische Regression $y=c/(1+a\cdot e^{-bx})$ (iterativ minimale Fehlerquadrate)
C: SinReg	Sinusregression $y=a\cdot\sin(bx+c)$ (iterativ minimale Fehlerquadrate)

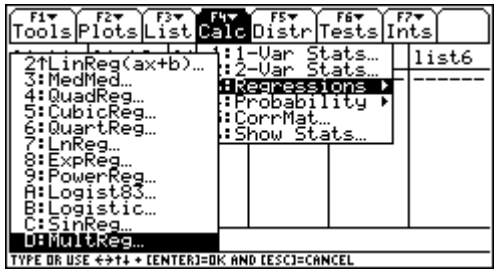
Die Listen **Xliste** und **Yliste** enthalten die Daten der unabhängigen x-Variable und der abhängigen y-Variable.
Bei der Option **reggl** übernimmt der TI-83 den Funktionsterm derRegressionskurve in eine der Funktionen **Y1** bis **Y0** (eingeben über das Menü **[VARS] > Y-VARS > Function**)

Weiters wird automatisch die Liste der Residuen $y_i - \hat{y}_i$ berechnet und in der Liste **LRESID** gespeichert. Sie kann im Menü **[LIST] > NAMES** abgerufen werden.
Ist der Diagnose-Modus aktiviert (**[CATALOG] > DiagnosticOn**), so werden automatisch auch noch das Bestimmtheitsmaß r^2 und der Korrelationskoeffizient r berechnet.

Über das Menü **[STAT PLOT]** können drei Statistik-Plots definiert werden. Für Regressionsdiagramme eignet sich bestens das Streudiagramm. Mit **[ZOOM] > 9:ZoomStat** sucht der TI-83 automatisch eine geeignete Fenstereinstellung, um die x-Liste und die y-Liste grafisch darzustellen. (Feineinstellungen über das Menü [WINDOW]).

Auf dem TI-92plus finden Sie die Regressionsrechnung im Stats/List Editor im Menü
[F4] (Calc) > 3:Regressions. Die Syntax lautet

FktReg [Xliste, Yliste, StoreEqnto, Freq, CategoryList, IncludeCategories]



Zusätzlich zu den Regressionskurven des TI-83 wird auch noch multilineare Regression für bis zu 10 unabhängige Variablen angeboten. Die Liste der Residuen $y_i - \hat{y}_i$ (= Abweichungen der empirischen von den theoretischen Werten) wird automatisch berechnet und in der Liste **resid** gespeichert. Diese Liste wird am Ende des List-Editors eingefügt, wenn im Stats/List Editor im Menü **[F1] (Tools) > 9:Formats** der Schalter für **Results->Editor** auf **YES** gesetzt ist.
Über das Menü **[F2] (Plots) > 1:Plot Setup** können neun Statistik-Plots definiert werden ([F1] Define). Mit **[F2] (Plots) > 1:Plot Setup > [F5] ZoomData** sucht der Rechner eine geeignete Fenstereinstellung, um die x-Liste und die y-Liste grafisch darzustellen.

Somit stehen mit den TI-Rechnern die wichtigsten Kennzahlen der zweidimensionalen Statistik auf Knopfdruck zur Verfügung.

Dadurch wird es möglich, Problemstellungen im Unterricht zu behandeln, an deren numerische Bewältigung bisher im Klassenzimmer nicht zu denken war. Das Hauptaugenmerk verlagert sich dabei vom technischen Rechnen auf die Interpretation der Modelle. Nun endlich kann der Statistikunterricht anwendungsorientiert und praxisbezogen gestaltet werden. Der Zugang zu komplexen Problemstellungen ist zudem experimentell und spielerisch, verschiedene mathematische Modelle können durchgespielt und auf ihre Brauchbarkeit überprüft werden.

6.1. Lineare Regression

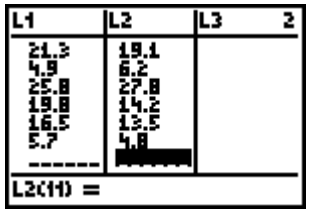
6.1.1. Methode der kleinsten Quadrate

Beispiel: Die folgende Tabelle enthält den Anteil, den die Exporte des betreffenden Landes in die Bundesrepublik an den gesamten Exporten des Landes des Jahres 1990 in Prozent ausmachen sowie den entsprechenden Anteil der Importe aus Westdeutschland an den Gesamtimporten des jeweiligen Landes in Prozent: (Jahresgutachten 1991/92 des Sachverständigenrates zur Begutachtung der gesamtwirtschaftlichen Entwicklung)

Land	**Import-anteil**	**Export-anteil**
Belgien/Luxemburg	24,0	21,3
Dänemark	22,8	17,9
Frankreich	19,0	17,4
Großbritannien	15,9	12,7
Italien	21,3	19,1
Japan	4,9	6,2
Niederlande	25,8	27,8
Schweden	19,8	14,2
Spanien	16,5	13,5
USA	5,7	4,8

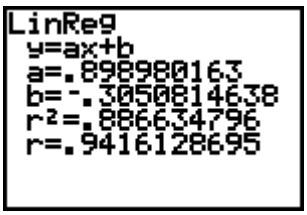
a) Ermitteln Sie die Regressionsgerade mit der unabhängigen Variablen Importanteil und der abhängigen Variablen Exportanteil. Interpretieren Sie die Steigung der Regressionsgeraden! Ermitteln und interpretieren Sie den Korrelationskoeffizienten und das Bestimmtheitsmaß! Stellen Sie die Punkte und die Regressionsgerade in einem Streudiagramm dar.
b) Für die Schweiz liegt der Importanteil bei 33,8. Welcher Anteil der schweizerischen Exporte in die Bundesrepublik an den Gesamtexporten der Schweiz wäre nach der geschätzten Regressionsgeraden zu erwarten?
c) Ermitteln Sie die Residuen und die minimale Fehlerquadratsumme und stellen Sie die Residuen grafisch dar. Kommentieren Sie die Grafik.
d) Der tatsächliche Wert für den Exportanteil der Schweiz beträgt 22,1. Schätzen Sie das Modell erneut unter Einbeziehung der schweizerischen Daten. Stellen Sie beide Regressionsgeraden in einem Streudiagramm dar. Kommentieren Sie dieses Diagramm. (nach [7], S.117)

a) Wir geben den Importanteil im Statistik-Editor **[STAT] > EDIT > 1:Edit** in Liste **L1** ein und den Exportanteil in Liste **L2**.



Die lineare Regressionskurve findet man im Rechenmenü des Statistik-Menüs **[STAT] > CALC > 4:LinReg(ax+b)**. Mit

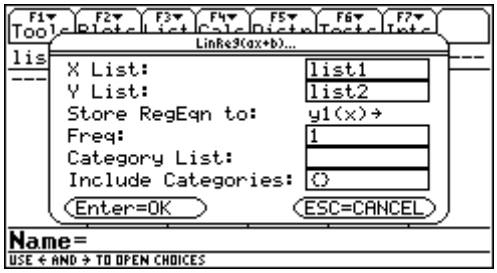
LinReg(ax+b) L1,L2,Y1

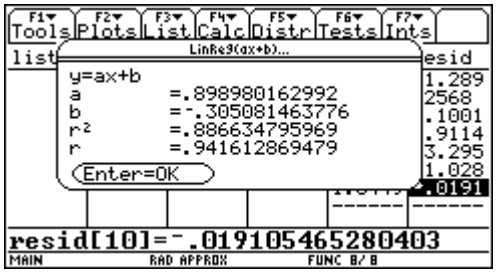


wird die Regressionsgerade berechnet und in der Funktion **Y1** gespeichert. Ist der Diagnose-Modus aktiviert (**[CATALOG] > DiagnosticOn**), so werden automatisch auch noch das Bestimmtheitsmaß r² und die Korrelation r berechnet.
Wir erhalten die Regressionsgerade

$$\hat{y} = 0{,}899 \cdot x - 0{,}305$$

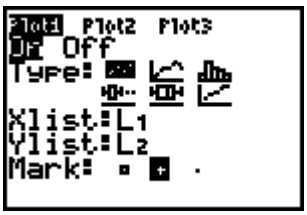
Auf dem TI-92plus finden Sie die lineare Regression im Stats/List-Editor im Menü **[F4] (Calc) > 3:Regressions > 2:LinReg(ax+b)**. Sie erhalten dann eine Dialogbox, in der Sie die entsprechenden Daten eingeben, schließlich eine output-box mit den berechneten Werten. Diese Werte (Steigung a, Achsenabschnitt b, Bestimmtheitsmaß r² und Korrelationskoeffizient r) können Sie über das Menü **[VAR-LINK]** im Ordner **STATVARS** abrufen (Variablen a, b, rsq, r).



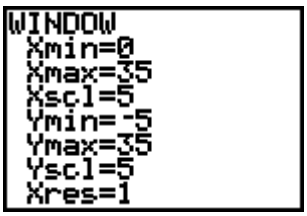


Die Steigung a = 0,899 bedeutet, dass bei den 10 untersuchten Ländern durchschnittlich ein zusätzlicher Prozentpunkt Importanteil nach Westdeutschland mit ca. 0,9 Prozentpunkten zusätzlichem Exportanteil aus der Bundesrepublik verbunden war. Da im vorliegenden Fall der geschätzte Achsenabschnitt b nahe bei null liegt, kann die Steigung a näherungsweise als durchschnittliches Verhältnis von Exportanteil und Importanteil interpretiert werden: Bei den betrachteten Ländern ist der Exportanteil nach Westdeutschland durchschnittlich um $1 - 0{,}899 \approx 0{,}1 = 10\%$ geringer als der entsprechende Importanteil.

Der Korrelationskoeffizient r = 0,94 zeigt einen starken positiven linearen Zusammenhang an.
Das Bestimmtheitsmaß r² = 0,88 besagt, dass sich 88% der Streuung der Exportanteile durch die Regression erklären lassen, nur 12% der Gesamtstreuung sind durch die Regressionsgerade nicht erklärt. (Siehe die anschließende Erläuterung zum Bestimmtsheitsmaß.)



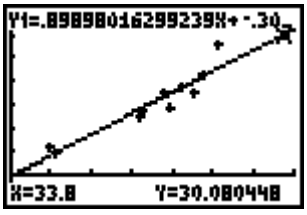
Für die grafische Darstellung wählen wir im Menü **[STAT PLOT] > Plot1** bei Typ das Streudiagramm und als Xlist: L1, als YList: L2, für eine günstige [WINDOW]-Einstellung wählen wir für x den Bereich [0; 35] und für y den Bereich [0; 35]. Mit [GRAPH] wird das Streudiagramm dargestellt und automatisch die in Y1 gespeicherte Regressionsgerade gezeichnet.



b) Um den Exportanteil der Schweiz zu schätzen, müssen wir nur den Funktionswert der Regressionsgeraden an der Stelle 33,8 berechnen:

$$y(33{,}8) = 0{,}899 \cdot 33{,}8 - 0{,}305 = 30{,}08$$

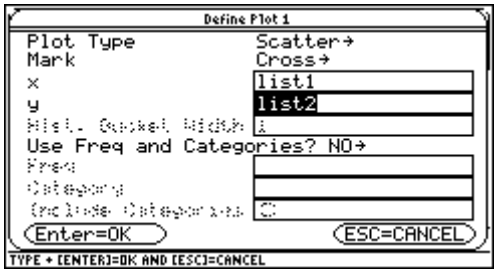
Es gibt mehrere Möglichkeiten, diesen Wert vom TI-83 berechnen zu lassen:
1. in der Grafik: mit [TRACE] den Wert abrufen;
2. im Hauptbildschirm: mit Y1(33,8) (Y1 aus [VARS] > Y-VARS > Function)
3. im [TABLE]-Menü: den Wert abfragen (in [TBLSET] bei Indpnt auf Ask umstellen)

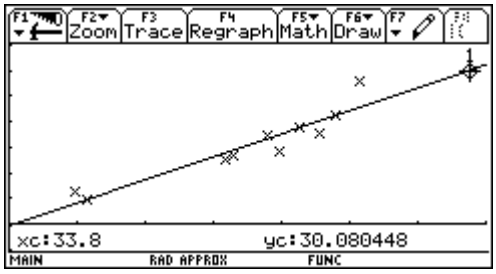


Das Bestimmtheitsmaß ist definiert durch $R^2 =$

$$\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{erklärte Abweichungsquadratsumme von } \bar{y}}{\text{Gesamtabweichungsquadratsumme von } \bar{y}}$$

Daraus ergibt sich die prozentuelle Interpretation: R^2 gibt den Anteil der erklärten Abweichungen von der Gesamtabweichung an.
(Zitat Fahrmeir u.a.: "Das Bestimmtheitsmaß gibt den Anteil der Gesamtstreuung der y_i an, der durch die Regression von Y auf X erklärt wird." (S.159)



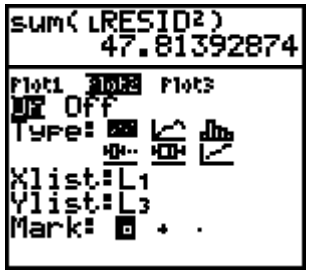


c) Automatisch hat der TI-83 die Liste der Residuen $y - \hat{y}$ berechnet und in der Liste **LRESID** gespeichert. Wir übernehmen diese Liste im Statistik-Editor in die Liste **L3**: Stellen Sie sich in den Kopf der Liste **L3** und rufen Sie im Menü **[LIST] > NAMES** die Liste **LRESID** ab.

Der Exportanteil von Belgien/Luxemburg liegt 0,03 Prozent über der Regressionsgeraden, der Exportanteil von Dänemark liegt 2,29 Prozent unter der Regressionsgeraden usw.
Die Fehlerquadratsumme ermitteln wir im Hauptbildschirm mit

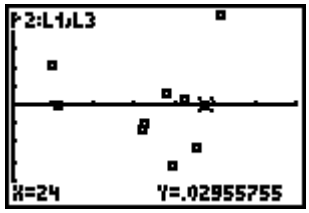
sum(LRESID²)

(**sum** findet man in [LIST] > MATH > 5:sum)



Für die grafische Darstellung definieren wir in **Plot2** als Xlist: L1 und als YList: L3. Eine günstige [WINDOW]-Einstellung für die y-Liste L3 ist der Bereich [-5; 5]. (Mit **1-Var Stats** können min = -3,29 und max = 4,91 ermittelt werden.)

Die Residuen schwanken unsystematisch um die x-Achse und sind nahe bei null. Dies deutet auf eine gute Modellanpassung hin.



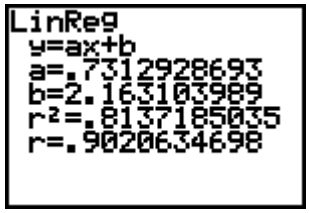
d) Wir fügen am Ende der Liste **L1** den *Importwert der Schweiz 33,8* ein und am Ende der Liste **L2** den *Exportwert der Schweiz 22,1*.
Am Streudiagramm sieht man schon die große Abweichung des Exportwerts der Schweiz von der Regressionsgeraden, sie beträgt

22,1 – 30,08 = -7,98

Diese Abweichung ist fast doppelt so groß wie das größte Residuum.

Wir ermitteln wieder die lineare Regressionskurve und speichern diese in **Y2**:

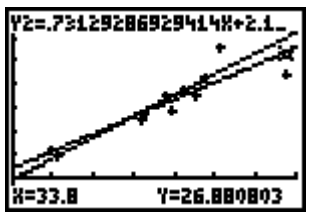
LinReg(ax+b) L1,L2,Y2



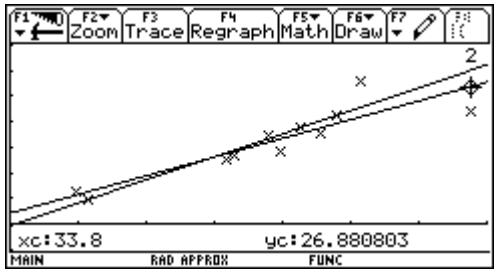
Wir erhalten die Regressionsgerade

$\hat{y} = 0{,}731 \cdot x - 2{,}163$

Die neue Steigung weicht erheblich von der alten ab.
Der Korrelationskoeffizienten r = 0,9 ist schwächer als der alte.



Die Regressionsgerade wird näher zum Punkt (33,8|22,1) hin verschoben.



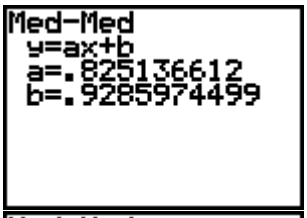
Der Grund für diese starke Abweichung liegt darin, dass die Methode der kleinsten Fehlerquadrate äußerst empfindlich auf Extremwerte reagiert. Da der Datensatz für die Schweiz beim Importanteil weit vom Mittelwert abweicht, beeinflusst dieser Datensatz den Verlauf der Regressionsgerade entscheidend.

Die Aufgabe verdeutlicht zweierlei: ([7], S.144)

1. Einzelne Beobachtungen, die weit von der Punktwolke der übrigen Beobachtungen entfernt liegen, haben (insbesondere wenn sie stark vom Zentrum der Verteilung abweichen) einen starken Einfluss auf die Parameter der Schätzung nach der Methode der kleinsten Quadrate und können die Anpassung des Modells wesentlich verschlechtern. Solche Punkte bezeichnet man als „Ausreißer". Beim Vorliegen derartiger „Ausreißer" sollte man vor Beginn der Analyse überlegen, ob es eventuell gerechtfertigt ist, diese Punkte bei der Analyse unberücksichtigt zu lassen bzw. ob die Anwendung der Methode der kleinsten Quadrate auf den vollständigen Datensatz tatsächlich sinnvoll ist.

2. (Deskriptive) Modelle, die einen Datensatz gut beschreiben, müssen nicht notwendigerweise auch gute Prognosen liefern. Die Güte des Modells (d.h. die Anpassungs- wie die Prognosegüte) kann erst im Rahmen der schließenden Statistik beurteilt werden.

6.1.2. Zentralwertlinie Med-Med:

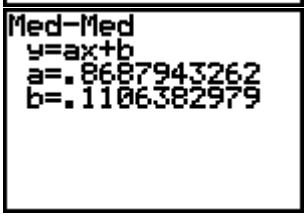


Im vorliegenden Beispiel empfiehlt sich, statt der üblichen Regression (kleinste Quadrate) eine Zentralwert-Regression zu verwenden. Der TI-83 hat diese unter **[STAT] > CALC > 3:Med-Med** implementiert. Mit

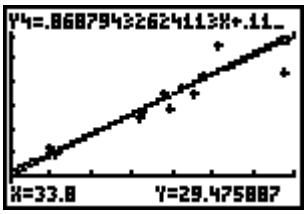
Med-Med L1,L2,Y3 (Werte ohne Schweiz)

sowie

Med-Med L1,L2,Y4 (Werte mit Schweiz)



ermittelt der TI-83 diese Regressionskurve und speichert sie in den Funktionen **Y3** und **Y4**.

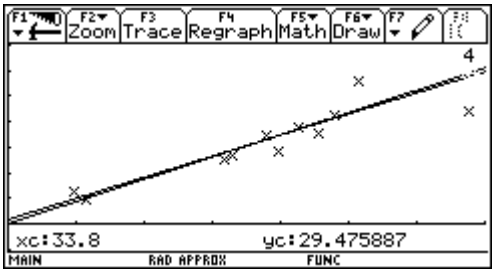


In unserem Beispiel erhalten wir für die Werte *ohne* Schweiz die Zentralwertlinie

$y = 0{,}825 \cdot x + 0{,}929$

für die Werte *mit* Schweiz die Zentralwertlinie

$y = 0{,}869 \cdot x + 0{,}11$



In der Grafik sehen wir, dass die beiden Zentralwertlinien nicht annähernd so stark scheren wie die Kleinstquadrate-Geraden.
Diese Regressionskurve ist wesentlich robuster gegen Ausreißer. (Ähnlich dem Median, der robuster gegenüber Ausreißern ist als das arithmetische Mittel oder ähnlich dem Interquartilabstand, der robuster ist gegenüber Ausreißern als die Standard-abweichung.)

Wie wird die Zentralwertlinie (Daten ohne Schweiz) berechnet?

Die Datenpunkte werden aufsteigend nach der x-Koordinate (Importanteil) sortiert und in drei Gruppen a, b, c eingeteilt (Gruppe a und c müssen die gleiche Anzahl haben, in unserem Beispiel enthalten sie je drei Punkte, die mittlere Gruppe b enthält 4 Punkte).
Nun werden in jeder Gruppe von den x- und den y-Koordinaten die Mediane berechnet, man erhält die drei Median-Punkte A(5,7|6,2), B(19,4|15,8), C(24|21,3).
Die Koordinaten x_1, y_1, x_2, y_2, x_3, y_3 dieser drei Median-Punkte können aus dem Variablenmenü **[VARS] > 5:Statistics > PTS** abgerufen werden.

	Land	Im	Ex	Gruppe	Medianpunkte (ohne Schweiz)
1	B/L	4,9	6,2	a	A(5,7|6,2)
2	DK	5,7	4,8		
3	F	15,9	12,7		
4	GB	16,5	13,5	b	B(19,4|15,8)
5	I	19	17,4		
6	J	19,8	14,2		
7	NL	21,3	19,1		
8	S	22,8	17,9	c	C(24|21,3)
9	E	24	21,3		
10	USA	25,8	27,8		

Durch die Punkte A und C wird eine Gerade g_1: $y = k_1 \cdot x + d_1$ gelegt:

$$k_1 = \frac{y_3 - y_1}{x_3 - x_1} = \frac{21{,}3 - 6{,}2}{24 - 5{,}7} = 0{,}825 \; ; \; d_1 = y_1 - k \cdot x_1 = 6{,}2 - 0{,}825 \cdot 5{,}7 = 1{,}497$$

Parallel dazu legt man eine Gerade g_2: $y = k_2 \cdot x + d_2$ durch B:

$k_2 = k_1$; $d_2 = y_2 - k \cdot x_2$ $15{,}8 - 0{,}825 \cdot 19{,}4 = -0{,}208$

Die Zentralwertlinie ist parallel zu diesen beiden Geraden; der Achsenabschnitt ist das gewichtete arithmetische Mittel der Achsenabschnitte der beiden Geraden - die Gerade durch A und C wird mit 2, die Gerade durch B wird mit 1 gewichtet:

$d = (2 \cdot 1{,}497 + (-0{,}208))/3 = 0{,}929$

Zentralwertlinie (ohne Schweiz): $y = 0{,}825 \cdot x + 0{,}929$.

Nimmt man die Schweizer Daten als 11. Datenpunkt dazu, muss folgendermaßen gruppiert werden:
Gruppe a 4 Punkte, b 3 Punkte, c 4 Punkte.

Probieren Sie's aus!

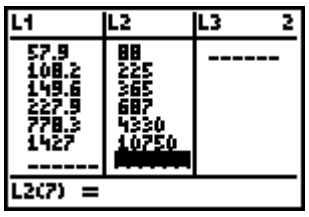
(Das ist eine schöne Übung und Wiederholung von Zwei-Punkt- und Punkt-Richtungsform der Geraden, und kann in diesem Zusammenhang schon in der II.HAK als Anwendung der linearen Funktion vorgestellt werden.)

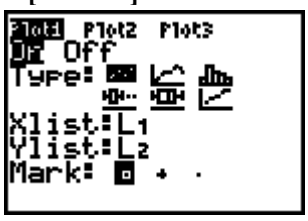
6.2. Nichtlineare Regression:

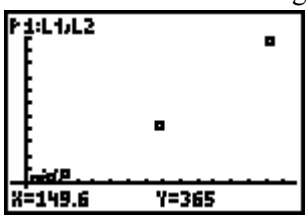
Beispiel: Drittes Keplersches Gesetz.
Die Planeten bewegen sich in elliptischen Bahnen um die Sonne.
Offensichtlich hat der Abstand zur Sonne einen Einfluss auf die Umlaufzeit des Planeten. Wir können annehmen, dass Planeten, die weiter von der Sonne entfernt sind, eine längere Umlaufzeit haben. Aber ist diese Beziehung linear, quadratisch, exponentiell? Oder gibt es eine andere Beziehung? Untersuchen Sie mit Hilfe der Regressionsrechnung den funktionalen Zusammenhang zwischen dem Abstand des Planeten zur Sonne (genauer: der Länge der großen Halbachse der elliptischen Bahn) und der Umlaufzeit. (nach [8])
Die folgende Tabelle gibt den Abstand in Millionen km und die Umlaufzeit in Tagen an:

Planet	**Abstand**	**Umlaufzeit**
Merkur	57,9	88
Venus	108,2	225
Erde	149,6	365
Mars	227,9	687
Jupiter	778,3	4330
Saturn	1427	10750

Wir geben den Abstand im Statistik-Editor **[STAT] > EDIT > 1:Edit** in Liste **L1** ein und die Umlaufzeit in der Liste **L2**. Im Statistik-Plot **[STAT PLOT] > 1:Plot1** wählen wir das Streudiagramm und geben **L1** in Xlist und **L2** in Ylist ein. Mit **[ZOOM] > 9:ZoomStat** erhalten wir das Streudiagramm.

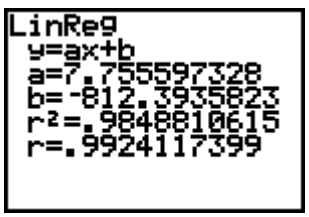


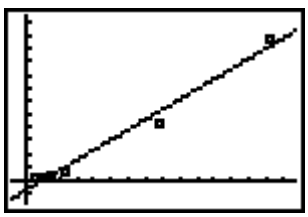


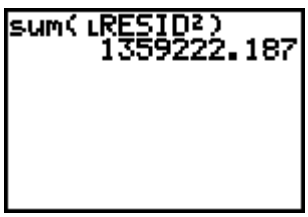


Nun lassen wir den TI-83 verschiedene Regressionskurven berechnen. Als Kriterium für die Güte der Anpassung vergleichen wir den Regressionskoeffizienten und die Fehlerquadratsumme (Summe der quadrierten Residuen) $\sum (y_i - \hat{y}_i)^2$.

Die lineare Regressionskurve findet man im Rechenmenü der Statistik **[STAT] > CALC > 4:LinReg(ax+b)**. Mit **LinReg(ax+b) L1,L2,Y1** wird die Regressionsgerade berechnet und in der Funktion **Y1** gespeichert. Ein Blick auf das Streudiagramm zeigt, dass die Anpassung besser sein könnte. Mit **sum(ʟRESID²)** (**sum** finden Sie in [LIST] > MATH > 5:sum; **ʟRESID**, die Liste der Residuen $y - \hat{y}$, wird automatisch mit der Regressionskurve berechnet, man findet die Liste in [LIST] > NAMES) erhalten wir die Fehlerquadratsumme:







(Falls der Regressionskoeffizient nicht angezeigt wird, müssen Sie den Diagnosemodus in [CATALOG] > DiagnosticOn einschalten!)

Analog ermitteln wir die exponentielle Regressionskurve (**[STAT] > CALC > 0:ExpReg),** speichern sie in Y2 und ermitteln die Fehlerquadratsumme:

ExpReg L1, L2, Y2

Schließlich speichern wir die Potenzfunktion-Regressionskurve (**[STAT] > CALC > A:PwrReg**) in Y3 und die Fehlerquadratsumme:

PwrReg L1, L2, Y3

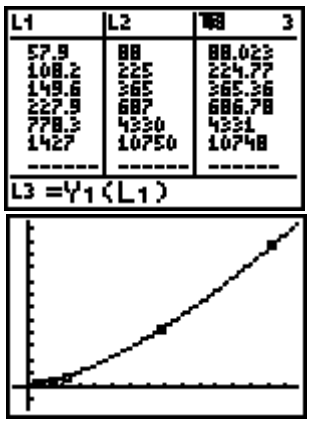
Wir erhalten folgende Ergebnisse:

Regressionstyp	Funktionsvorschrift	Korrelations-koeffizient	Fehlerquadrat-summe
linear (LinReg)	$y = 7{,}7556x - 812{,}3936$	0,99241	1 359 222
exponentiell (ExpReg)	$y = 183{,}56 \cdot 1{,}003^x$	0,9398	42 056 308
potenziell (PwrReg)	$y = 0{,}2003 \cdot x^{1{,}4994}$	0,99999993	3,966

(ACHTUNG: Der Korrelationskoeffizient gilt nur für das linearisierte Modell, siehe unten!)

Die Potenzfunktion-Regression liefert mit Abstand das beste Ergebnis.

Der funktionale Zusammenhang ist gegeben durch $y = 0{,}2 \cdot x^{1{,}5}$,
also gilt $y^2 = c \cdot x^3$. Dies ist bekannt als



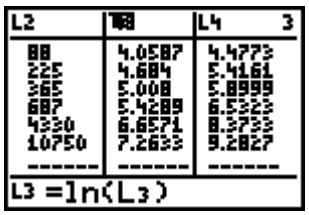
> ***Drittes Keplersches Gesetz***:
> Die Quadrate der Umlaufzeiten der Planeten verhalten sich zueinander wie die dritten Potenzen der großen Halbachsen ihrer Bahnellipsen.

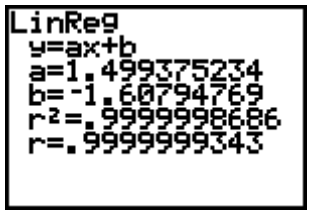
Hinweis zum Korrelationskoeffizienten:
Der Korrelationskoeffizient ist nur für die lineare Regression definiert! Bei einer Potenzfunktion-Regression erhält man diese Maßzahl, indem der Funktionsterm $y = a \cdot x^b$ durch Logarithmieren linearisiert wird:

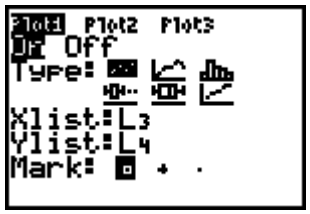
$\ln y = \ln a + b \cdot \ln x$

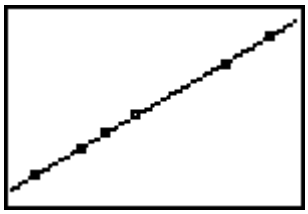
Für die Werte $\ln x$ und $\ln y$ erhalten wir dadurch einen linearen Zusammenhang $y^* = a^* \cdot x + b^*$. Durch Entlogarithmieren erhalten wir $a = \exp(b^*)$ und $b = a^*$

Das können wir mit dem TI-83 nachvollziehen:
Wir definieren die Listen L3 = ln(L1) und L4 = ln(L2), die lineare Regression liefert die linearisierten Parameter a* und b* sowie den Korrelationskoeffizienten.









Aufgabe: Schätzen Sie mit Hilfe der Potenzfunktionregression die Umlaufzeiten für die Planeten Uranus ($2868 \cdot 10^6$ km), Neptun ($4496 \cdot 10^6$ km) und Pluto ($5946 \cdot 10^6$ km)!

(Die wirklichen Werte sind 30664, 60145 und 90739 Tage)

Literatur:

[1] John W. Tukey: Exploratory Data Analysis. Addison und Wesley, 1977.

[2] Ludwig Fahrmeir, Rita Künstler, Iris Pigeot, Gerhard Tutz: Statistik. Der Weg zur Datenanalyse. Springer, 1997.

[3] Wolfgang Polasek: Explorative Datenanalyse. Einführung in die deskriptive Statistik. Springer, 1994

[4] Thomas Sauerbier, Werner Voß: Kleine Formelsammlung STATISTIK. Leipzig: Hanser 2000.

[5] Benno Grabinger, Günter Schmidt: Stochastik mit dem TI-92. Hannover: Schroedel 2001.)

[6] Erwin Kreyszig: Statistische Methoden und ihre Anwendungen. Göttingen: Vandenhoeck & Ruprecht ^{7}1979.

[7] Martin Missong: Aufgabensammlung zur deskriptiven Statistik. Mit ausführlichen Lösungen und Erläuterungen.. München, Wien: Oldenbourg 1998.

[8] Bruce MacMillan: Exploring Planetary Motion. http://www.ti.com/calc/docs/act/mac2.htm (27.11.99)