Beschreibende Statistik und der Übergang zur beurteilenden Statistik

Guido Herweyers
KHBO Campus Oostende
K.U.Leuven

1. Vorwort

Der Einsatz des Voyage 200 erleichert die Verarbeitung von Daten und erlaubt es, eine engere Verbindung zwischen beschreibender und beurteilender Statistik herzustellen.

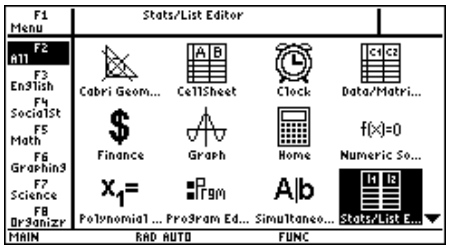
Die Normalverteilung ist die wichtigste stetige Verteilung. In Belgien wird i.a. diese Verteilung erst nach der Wahrscheinlichkeitsrechnung und zwar als Approximation der Binomialverteilung eingeführt.

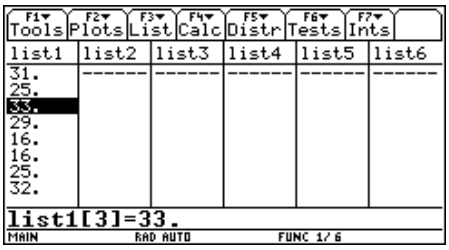
Heute wählen wir die folgende Vorgangsweise:

Das Datenmaterial wird durch ein Histogramm mit geeigneter Klassenbreite dargestellt. Das "Ansehen der Daten" auf diese Weise gibt rasche Information über Form, Streuung und Ausreißer der Verteilung. Die Fläche unter dem Histogramm wird durch Anpassung der Einheit auf der vertikalen Achse zur Maßzahl für die relative Häufigkeit. Als mathematisches Modell wird dann eine Kurve gezeichnet, die möglichst gut über das Histogramm gelegt werden kann. Für ein glockenförmiges Histogramm wird eine Normalverteilung als mathematisches Modell gewählt.

Als zweites Beispiel ist die Einführung des Begriffes Erwartungswert $E(X)$ einer Zufallsvariablen X vorgesehen. Den Erwartungswert erhalten wir als Stabilisierung des Mittelwerts von reellen Zahlen, die man durch oftmalige Ausführung eines Zufallsexperiments erhält. Die Simulation von Zufallsexperimenten mit dem Voyage 200 ist hier sehr behilflich.

Die "Statistics with List Editor" Applikation des Voyage 200, mit dem die Beispiele im folgenden Text ausgearbeitet werden, ist sehr nützlich.





Einzelheiten über diese Applikation werden hier nicht erklärt. Die Abbildungen sind selbsterklärend und die Anleitung zu dieser Applikation (194 Seiten) finden Sie im pdf-Format auf http://education.ti.com/us/product/apps/89/statsle.html.

2. Daten beobachten

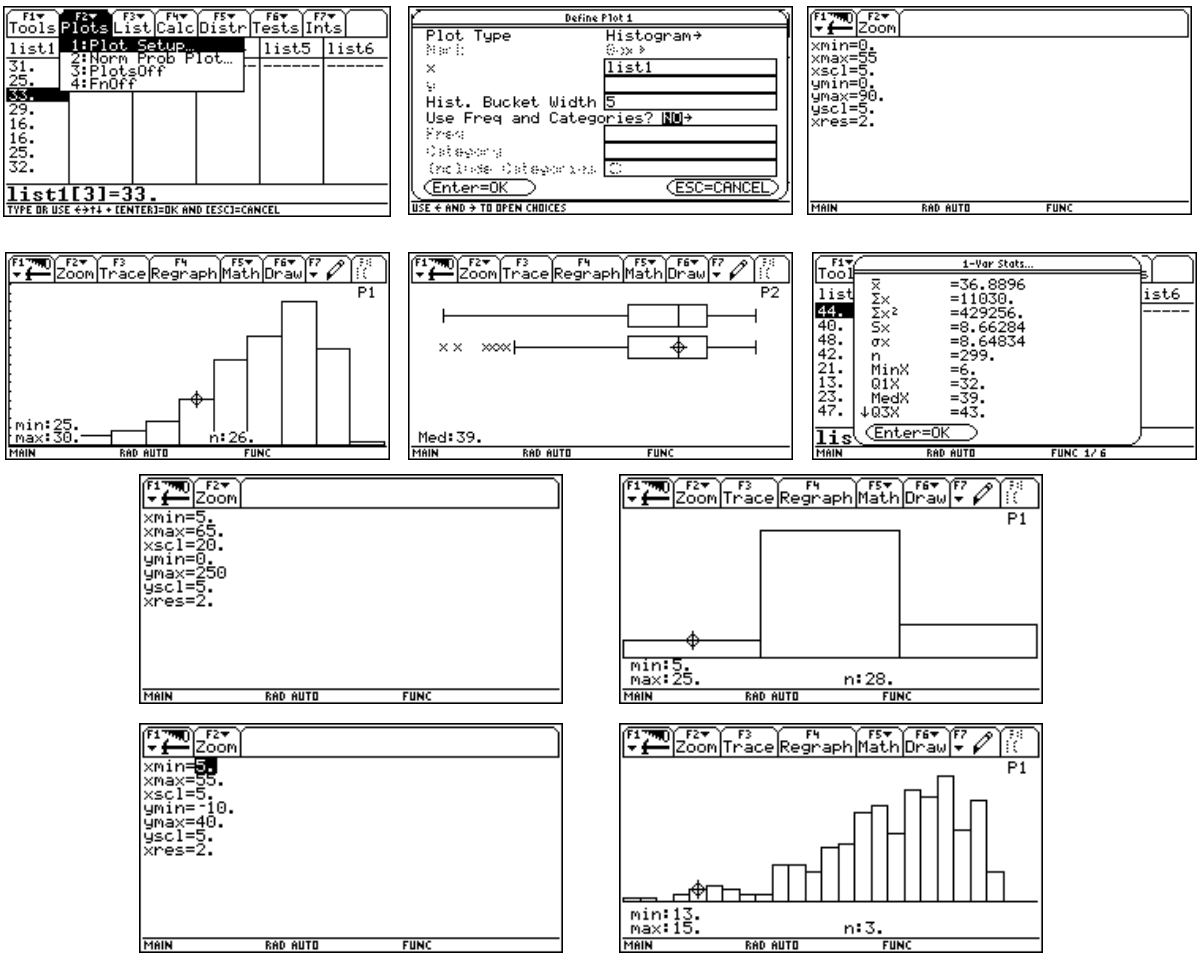
2.1 Eine schiefe Verteilung: Baywatch

Während der Ferienzeit werden die Strände an der belgischen Küste durch Rettungskräfte bewacht. Viel Studenten fühlen sich von diesem abenteuerlichen Urlaubsjob angezogen. Die Ausbildung für diese "Retter am Meer" ist aber nicht zu unterschätzen; neben einem umfassenden theoretischen Lehrgang muss man auch eine Anzahl von schweren Schwimmprüfungen ablegen.
Die Ausbildung wird jährlich organisiert. Zuerst müssen die Studenten bei der theoretischen Prüfung durchkommen. Der Kurs besteht aus sieben Kapiteln. Um die theoretische Prüfung zu bestehen, muss man mindestens 50% der möglichen Punkte für jedes Teilkapitel erreichen.

Zur Ausbildung in 2002-2003 meldeten sich insgesamt 299 Teilnehmer. In Kapitel 1 konnten 50 Punkte erreicht werden. Hier folgen die Ergebnisse der Teilnehmer:

44	46	42	45	45	39	47	37	46	32
40	29	34	44	34	40	41	35	38	36
48	36	40	47	42	33	43	40	33	43
42	44	38	35	24	35	46	45	12	23
21	29	36	27	37	39	29	47	45	42
13	37	43	36	43	42	40	26	42	38
23	32	40	23	42	31	43	34	30	31
47	36	39	45	37	47	40	44	34	49
32	45	44	37	33	49	28	14	47	33
33	48	46	32	42	44	30	25	14	18
40	48	41	41	48	19	45	43	33	40
33	44	44	41	35	25	26	50	41	39
42	47	38	35	38	6	48	40	34	28
32	48	24	32	36	33	44	35	23	34
40	46	43	26	41	41	27	29	8	42
36	33	46	43	42	35	39	44	50	32
23	49	42	46	33	43	42	43	40	37
38	43	18	26	33	44	39	28	49	45
29	41	36	16	42	49	43	22	25	37
18	48	39	31	35	30	35	30	47	47
29	35	35	39	26	20	37	33	40	39
48	27	44	48	46	36	31	28	36	33
42	48	24	25	48	34	39	32	48	36
39	32	38	46	31	29	39	43	42	25
39	30	42	40	33	33	41	16	12	43
44	16	30	38	47	42	47	44	37	48
44	35	47	36	43	46	34	32	24	37
38	49	36	33	40	29	42	44	43	47
48	43	30	43	33	47	23	42	15	36
40	39	44	31	28	45	39	35	45	

Aufgabe der beschreibenden Statistik ist es, diese Daten zu veranschaulichen, zum Beispiel durch ein Histogramm oder ein Boxplot (Kastendiagramm).

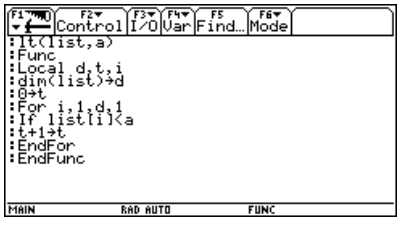


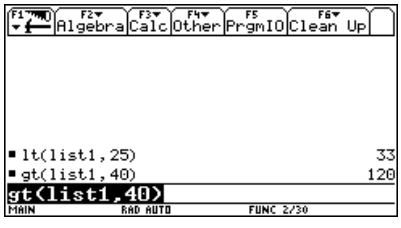
Die Verteilung ist links schief; es gibt mehr Studenten mit guten als mit schlechten Resultaten. Das Histogramm mit der Klassenbreite 20 gibt zu wenig Information und vermittelt den falschen Eindruck einer symmetrischen Verteilung. Hingegen gibt das Histogramm mit Klassenbreite 2 gibt zuviel Detailinformation.

Es ist bequem über Funktionen zu verfügen, die zählen wieviele Elemente in einem bestimmten Intervall liegen. Solche Funktionen kann man leicht programmieren und sie bleiben im Speicher erhalten, solange man sie nicht wieder löscht.

Intervall	Funktionsname (Parameter zwischen Klammern)
$x < a$	`lt(liste,a)`
$x \le a$	`lteq(liste,a)`
$a \le x < b$	`geqlt(liste,a,b)`

usw.





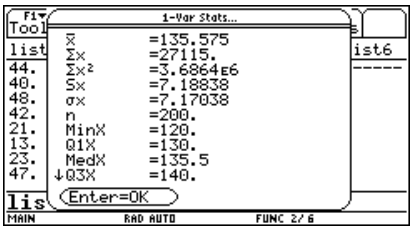
Nur 33 von 299 Studenten haben ein Resultat < 25 und 120 Studenten haben mehr als 40 von 50 möglichen Punkten!

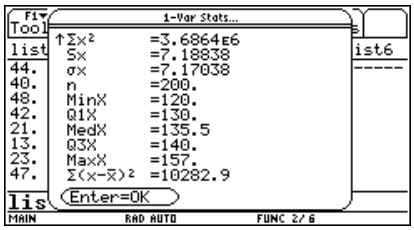
2.2 Eine glockenförmige Verteilung

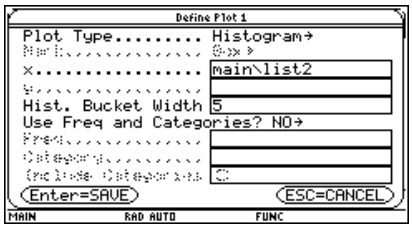
Die folgende Tabelle zeigt die Größen (in cm) von 200 10-jährigen Kinder in Flandern. Die Daten stehen in der Liste "list2"

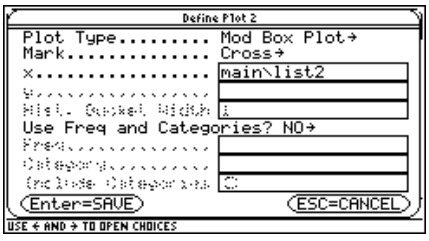
120	123	151	142	137	128	133	142	136	137
144	126	135	142	135	134	140	149	137	129
128	140	129	137	141	137	135	129	133	139
132	125	124	132	129	139	132	145	140	138
137	133	137	138	131	137	131	127	134	134
150	140	144	137	133	139	130	141	136	124
130	135	124	122	136	132	133	133	142	127
142	130	135	125	136	132	153	145	131	131
134	145	139	132	136	143	138	141	141	141
136	148	128	137	134	138	130	145	135	141
131	143	146	132	127	129	133	142	157	133
139	128	123	140	140	152	136	125	130	153
130	126	129	157	144	142	128	138	142	135
141	139	132	135	145	134	140	136	138	143
122	141	122	132	136	129	138	130	129	135
134	141	133	128	121	131	137	140	133	135
138	132	140	145	128	140	134	128	146	132
131	142	133	137	126	128	129	124	137	127
139	141	157	146	128	136	130	141	129	143
137	143	139	141	121	131	128	133	136	146

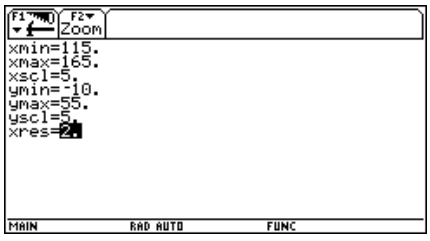
Wir suchen die statistischen Kenngrößen und zeichnen ein Histogramm.

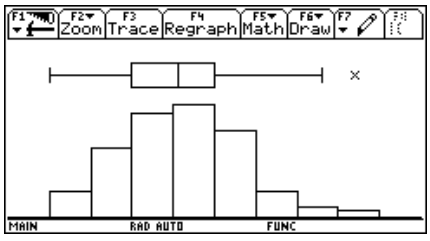












Die Verteilung der Daten ist nahezu symmetrisch und nimmt eine Glockenform an. Neben der "Fünf-Zahlen-Zusammenfassung" wird eine derartige solche Verteilung oft zusammengefasst durch den *Stichprobenmittelwert* $\bar{x}$ und die *Stichprobenstandardabweichung* s (wir fassen die Daten auf als eine Stichprobe aus der Population aller 10-jährigen Kinder in Flandern). Die Stichprobenstandarabweichung ist ein Maß für die Streuung der Daten rund den Mittelwert.

Für eine Liste von n Daten $x_1, x_2, x_3, \ldots, x_n$ werden diese Größen folgendermaßen berechnet:

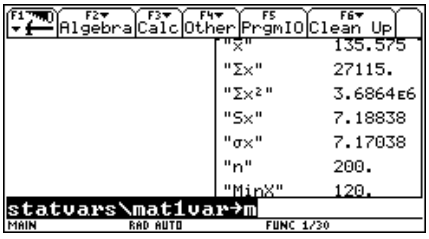
$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

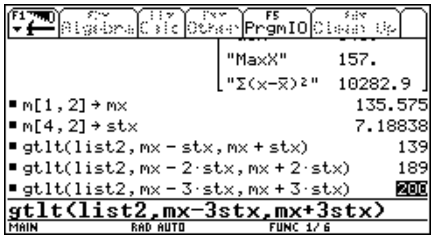
$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Für unseren Daten ist $n = 200$, $\bar{x} = 135.575$ und $s = 7.19$

Der Prozentsatz der Daten gelegen zwischen

- $\bar{x} - s$ und $\bar{x} + s$ oder 128.4 und 142.8 ist 139/200 = 69.5 %,
- $\bar{x} - 2s$ und $\bar{x} + 2s$ oder 121.2 und 150.0 ist 189/200 = 94.5 %,
- $\bar{x} - 3s$ und $\bar{x} + 3s$ oder 114 und 157.1 ist 100%





Für eine glockenförmige Verteilung gilt nun die folgende *Faustregel*:

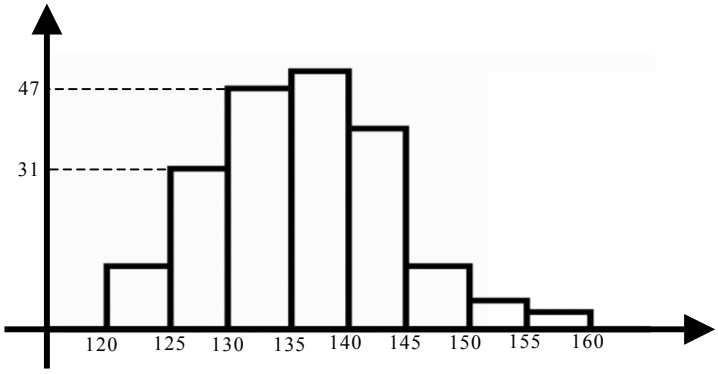
- Ungefähr 68 % der Daten befinden sich innerhalb Mittelwert ± einer Standardabweichung.
- Ungefähr 95 % der Daten befinden sich innerhalb Mittelwert ± zwei Standardabweichungen.
- Nahezu alle Daten befinden sich innerhalb Mittelwert ± drei Standardabweichungen.

Dank dieser Faustregel liefern die Zahlen $\bar{x}$ und s viel Information über eine glockenförmige Verteilung. Im Paragraph 4.2 werden wir sehen, woher die Prozentsätze der Faustregel herkommen.

Von vielen Daten weiß man, dass sie einer glockenförmigen Verteilung gehorchen, z.B. die Größe der Erwachsenen, der Intelligenzquotient, der exakte Inhalt von Cola-Flaschen "1,5 Liter", Meßfehler, ….

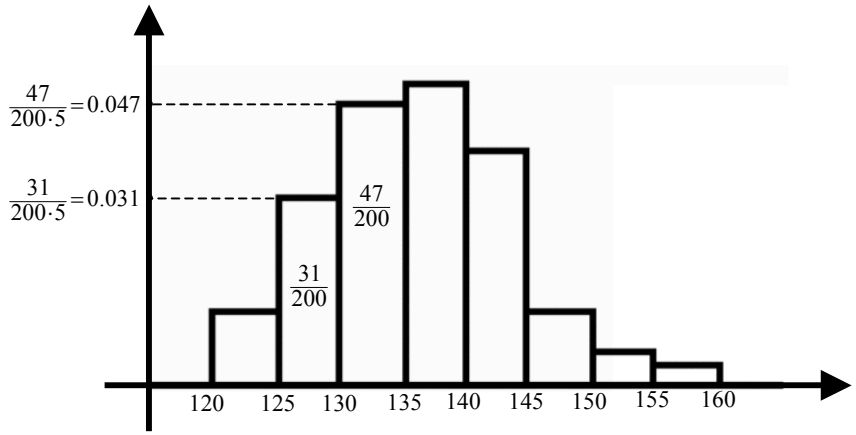
3. Die Normalverteilung

Wir betrachten wieder das Histogramm der Größe der Kinder.



Es sind 31 Wahrnehmungen im Intervall $[125,130[$ und 47 Wahrnehmungen im Intervall $[130,135[$. Auf der *y*-Achse lesen wir die *Häufigkeit* ab .

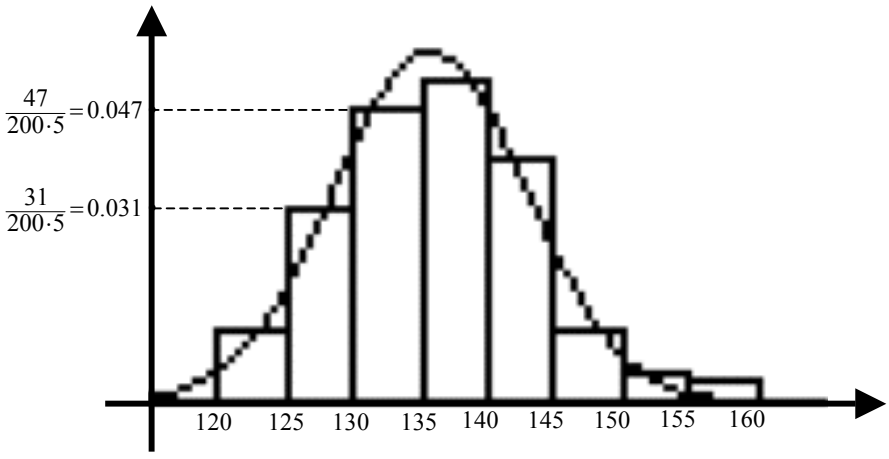
Wir passen die Einheit auf der *y*-Achse jetzt so an, dass die *Fläche* eines Rechtecks (einer Säule) zur Maßzahl der *relativen Häufigkeit* der entsprechenden Klasse wird:



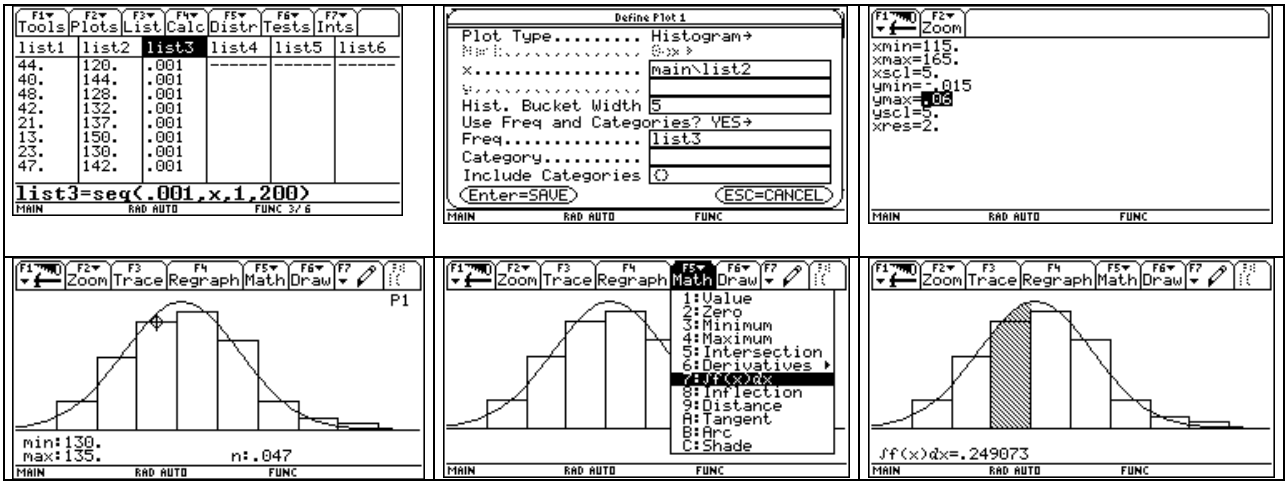
Auf der *y*-Achse lesen wir jetzt die *relative Häufigkeitsdichte* ab, oder die relative Häufigkeit geteilt durch das Produkt der Anzahl (200) der Daten und die Klassenbreite (5).

So sehen wir dass 31/200 = 15.5% der Kinder eine Länge im Intervall $[125,130[$ haben. Die Gesamtfläche unter dem Histogramm ist 1, als Summe der *relativen Häufigkeiten.*

Als m*athematisches Modell* für dieses Histogramm wird eine fließende glockenförmige Kurve oder *Normalverteilung* verwendet, Diese Funktion heißt *Normale Dichte.* Die Gesamtfläche unter der Kurve ist noch immer 1. Der *Mittelwert* dieser idealisierte Verteilung wird mit μ und die *Standardabweichung* mit σ notiert (Griechische Buchstaben !) .

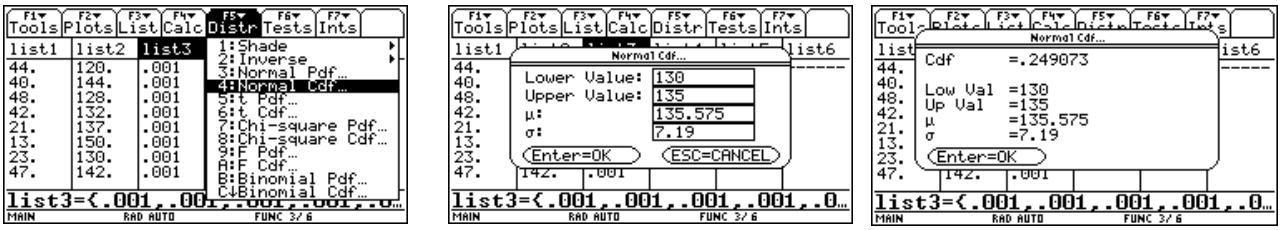


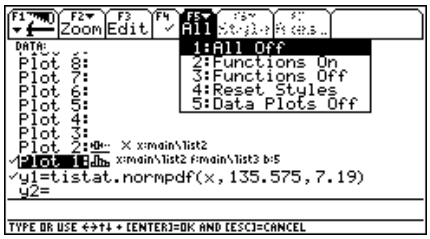
Für unseres Modell setzen wir $\mu = \bar{x} = 135.575$ und $\sigma = s = 7.19$.

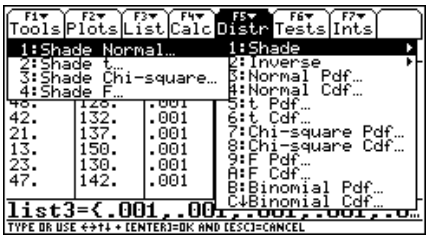


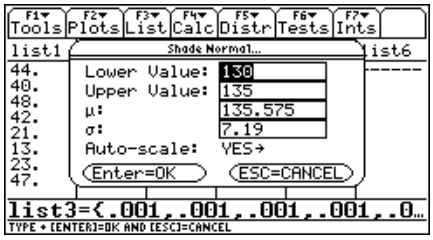
Natürlich ist das mathematische Modell *keine perfekte Wiedergabe* der Verteilung. Der Zweck eines mathematischen Modells liegt in *einer guten Beschreibung* der Realität.

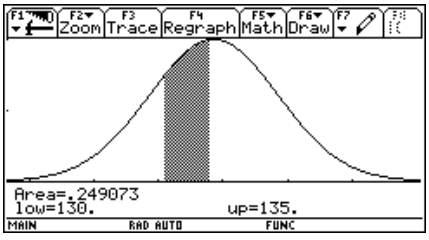
Mit der Normalverteilung als Modell können wir den *Anteil* oder die *relative Häufigkeit von Daten in einem bestimmten Intervall* (annähernd) bestimmen als Fläche unter der Kurve und über dem Intervall. So finden wir 24.9% der Daten zwischen 130 und 135 (zu vergleichen mit 47/200 = 23.5% beim Histogramm).











Für die Beschreibung der Variablen oder Größen, die wir studieren, werden Großbuchstaben verwendet. Wenn *X* die Größe eines Kindes darstellt, dann lesen wir " $X < 130$ " als " Die Größe *X* ist kleiner als 130(cm)". Wir verwenden den entsprechenden Kleinbuchstaben *x* für jeden spezifischen Wert von *X*. Nehmen wir zum Beispiel ein Kind aus der Gruppe mit einer Körperlänge von 127 cm, dann nimmt die Variable *X* hier den konkreten Wert $x = 127$ an.

Eine normalverteilte Variable *X* mit Mittelwert μ und *Standardabweichung* σ wird notiert als $X \sim N(\mu, \sigma)$.

Für den Anteil der Daten mit einer Länge *X* zwischen 130 und 135 schreibt man

$$P(130 < X < 135) = P(130 \le X \le 135) = P(130 \le X < 135) = P(130 < X \le 135) = 24.9\%$$

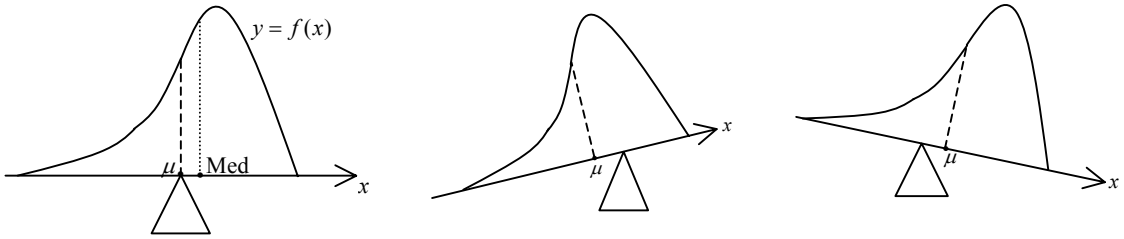
Für unseres Modell gilt ja $P(X = 125) = P(X = 130) = 0$ da die "Fläche" eines Segments 0 ist.

4. Eigenschaften von Dichten

4.1 Allgemein

Eine Dichte *f* ist ein mathematisches Modell für das Histogramm der relative Häufigkeitsdichten der gemessenen Daten einer Größe *X* mit den folgenden Eigenschaften:

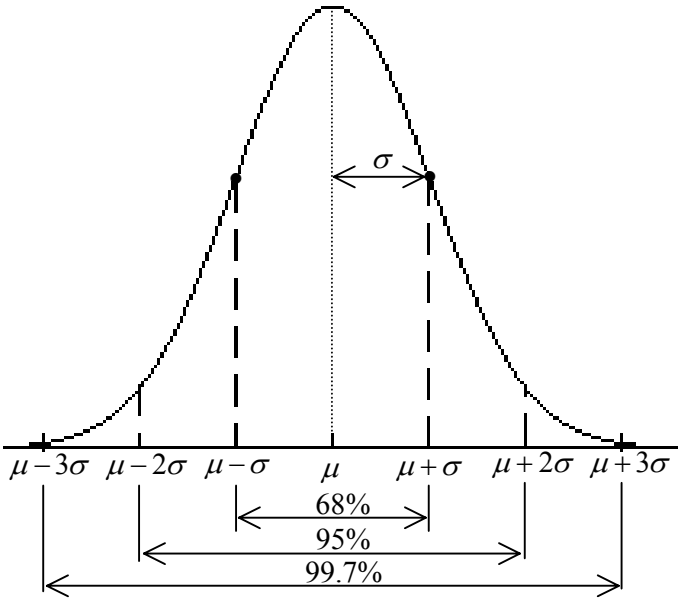
- $f(x) \ge 0$ für alle *x*
- *f* schließt mit der *x*-Achse die Gesamtfläche 1 ein oder $\int_{-\infty}^{\infty} f(x)dx = 1$.
- Der Mittelwert μ des Modells ist der "Gleichgewichtpunkt" der Verteilung:



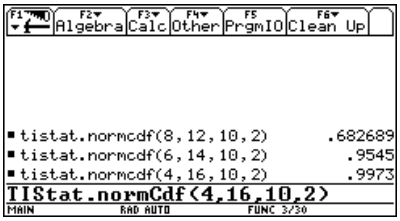
4.2 Die Normalverteilung

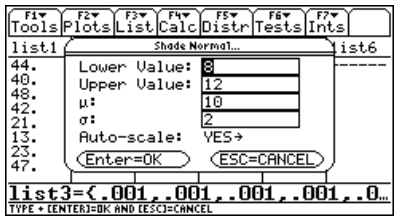
Für eine normaleverteilte Dichte *f* mit Mittelwert μ und *Standardabweichung* σ gilt außerdem:

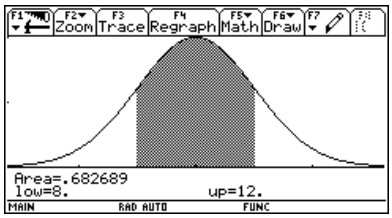
- $f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, mit $e \simeq 2.718$.
- Die Kurve ist glockenförmig und symmetrisch bezüglich der Geraden $x = \mu$.
- $\lim_{x \to \pm\infty} f(x) = 0$.
- Auch die Standardabweichung hat eine geometrische Bedeutung: Die Punkte auf der Kurve mit den *x*-Koordinaten $x = \mu \pm \sigma$ sind die Wendepunkte der Glockenkurve.
- Die 68 - 95 - 99.7 Regel:
  - 68% der Daten liegen im Intervall $[\mu - \sigma, \mu + \sigma]$
  - 95% der Daten liegen im Intervall $[\mu - 2\sigma, \mu + 2\sigma]$
  - 99.7% der Daten liegen im Intervall $[\mu - 3\sigma, \mu + 3\sigma]$



Wir verifizieren diese Regel für $\mu = 10$ und $\sigma = 2$





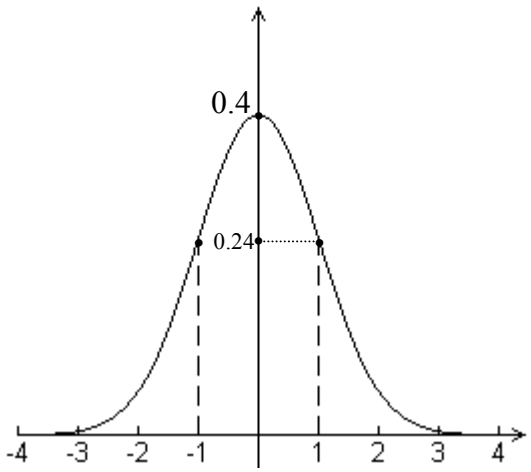


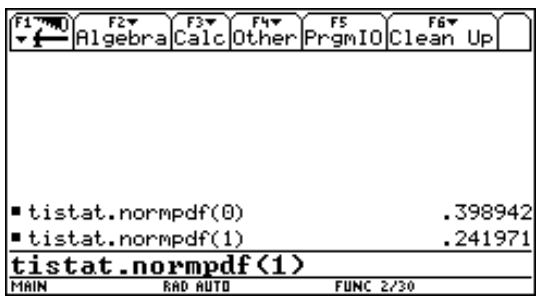
5. Normale Dichten und die Standardnormaldichte

Für die standardnormale (standardisierte) Dichte *f* gilt $\mu = 0$ und $\sigma = 1$:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} .$$

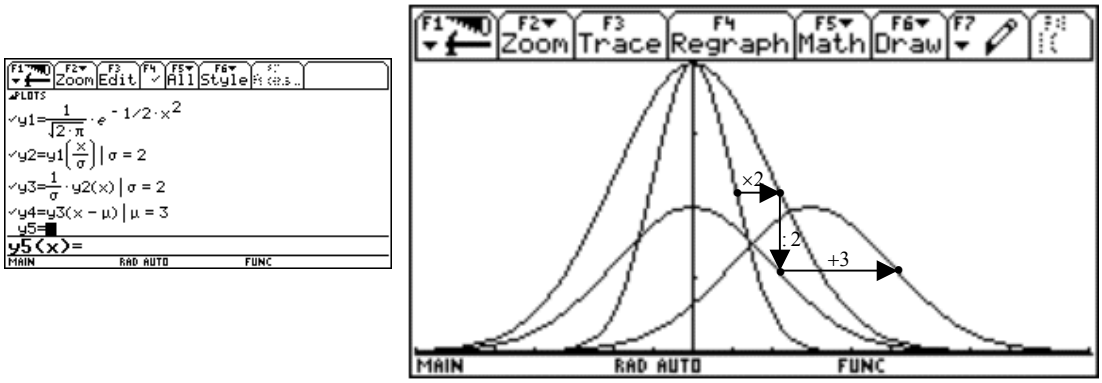
Wir illustrieren wie jede Normale Dichte graphisch durch eine flächentreue Transformation der standardnormalen Dichte erhalten werden kann.



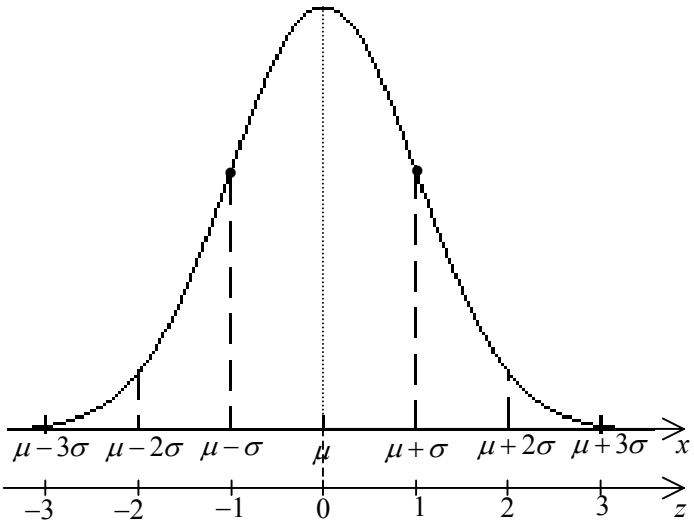


Das Maximum der Standardnormaldichte is 0.4 , die Wendepunkte befinden sich auf Höhe 0.24 oder 60% des Maximalwerts.

Wie wird nun die standardnormale Dichte transformiert zur Dichte mit $\mu = 3$ und $\sigma = 2$?



Umgekehrt, wenn $X \sim N(\mu, \sigma)$, dann ist $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$. Die Variable Z heißt standardisierte Zufallsvariable. Durch diese Transformation wird ein konkreter Wert x von X transformiert in den Wert $z = \frac{x - \mu}{\sigma}$ von Z . Die Zahl z lehrt uns, um wieviele Standardabweichungen σ die Zahl x vom Mittelwert μ entfernt ist: $x = \mu + z \cdot \sigma$.



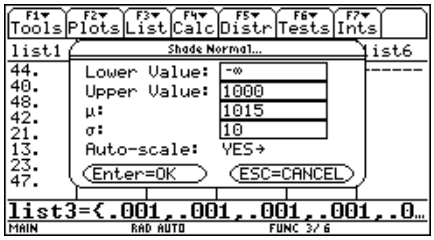
6. Übungen

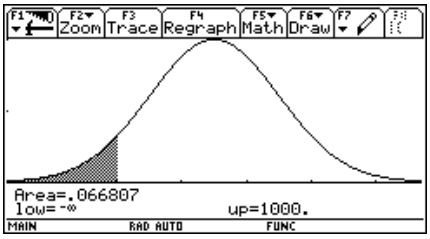
1. Eine Maschine füllt Pakete mit Zucker. Die Masse *X* der Pakete ist normalverteilt mit Mittelwert $\mu = 1015\,\text{g}$ und Standardabweichung $\sigma = 10\,\text{g}$.

   a) Welcher Anteil der Pakete enthält weniger als 1kg ?
   b) Wie muss man die Maschineneinstellung - das heißt μ - ändern (σ bleibt konstant), damit nur 1% der Pakete eine Masse unter 1kg haben würden?

Lösung:

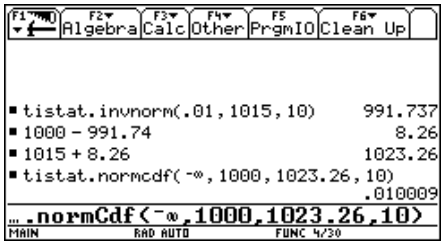
a) `normcdf(`$-\infty$`,1000,1015,10)` = 6.68%



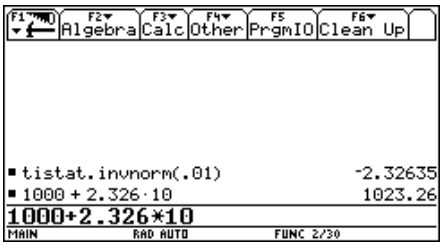


a) Methode 1:
Suche mit der Umkehrfunktion "`invNorm`" der Verteilungsfunktion den Wert *x*, für den gilt $P(X \le x) = 0.01$ und verschiebe die Kurve um den richtigen Abstand nach rechts:

Antwort:
Wähle $\mu = 1023.26$



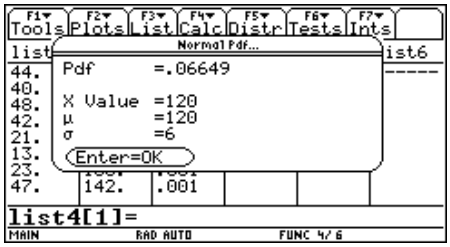
b) Methode 2: Standardisieren

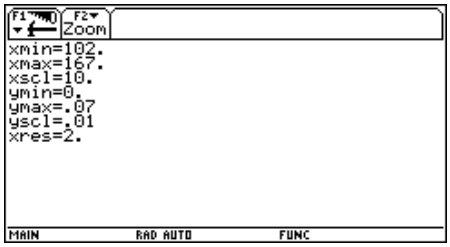


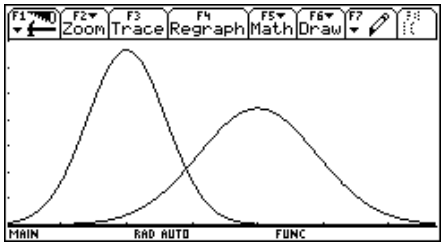
2. a) Zeichne die Dichtefunktionen zu $N(120,6)$ und $N(140,9)$ in einer geeigneten Skalierung.
   b) Finde das erste Quartil der $N(120,6)$-Verteilung.
   c) Durch welche geometrische Transformationen geht die $N(120,6)$-Verteilung über in die $N(140,9)$-Verteilung?
   d) Bestimme mit den Resultaten von (a) und (b) das erste Quartil der $N(140,9)$-Verteilung.

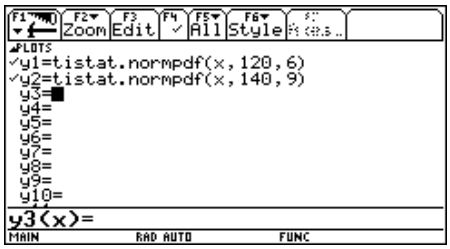
Lösung:

a) $120-3\cdot 6=102$ und $120+3\cdot 6=138$
$140-3\cdot 9=113$ und $140+3\cdot 9=167$
Also `xmin` = 102 und `xmax` = 167
`normpdf(120,120,6)` = 0.066 , wir stellen `ymax` = 0.07

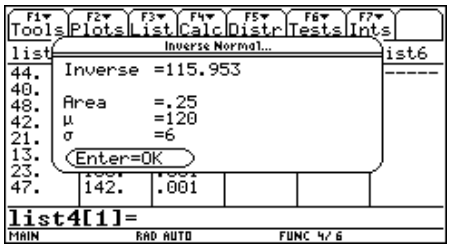


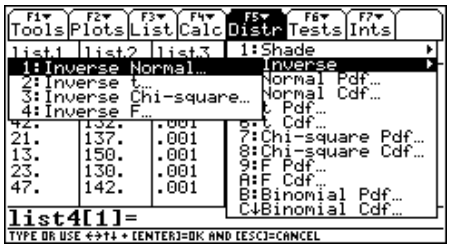






b) `invnormpdf(.25,120,6)` = 115.95





c) Verschiebung nach rechts um den Abstand 20 , horizontale Streckung gegenüber der Symmetrieachse $x=140$ mit dem Faktor 9/6 = 1.5 , vertikale Stauchung mit dem Faktor 1.5 .

d)