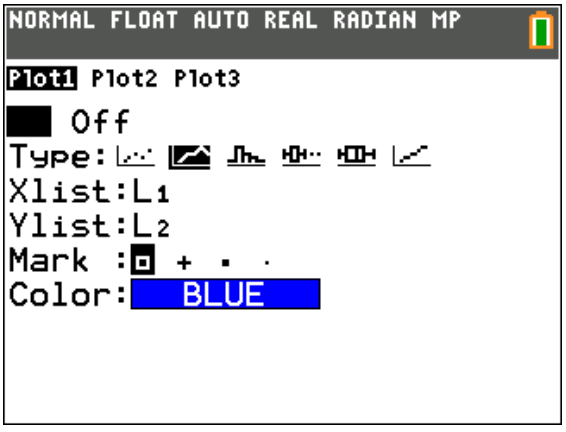
Beskrivande statistik

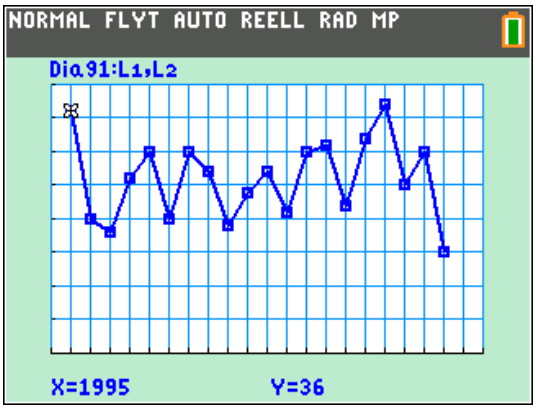
Tabellen ovan visar antalet allvarliga olyckor på en vägsträcka under 15 år.

år	Antal olyckor
1995	36
1996	20
1997	18
1998	26
1999	30
2000	20
2001	30
2002	27
2003	19
2004	24
2005	27
2006	21
2007	30
2008	31
2009	22
2010	32
2011	37
2012	25
2003	30
2014	15

Ett sätt att få en bild av hur antalet olyckor har förändrats är naturligtvis att rita ett *tidseriediagram*. Man matar då in data i två listor. Sedan går man till diagraminställningen och väljer följande inställning:



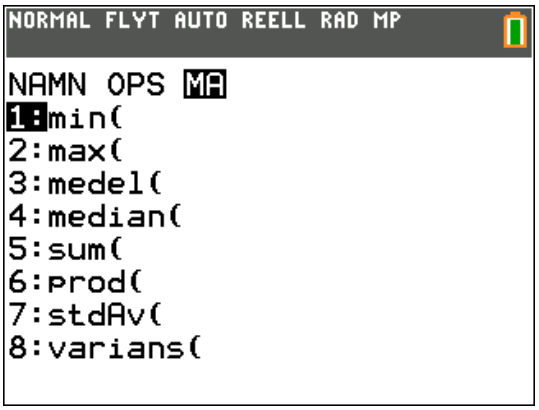
Nu kan vi plotta diagrammet. Man får tänka på att ha en bra fönsterinställning som passar våra data. Gå till [WINDOW] och ställ in. Om man trycker på [ZOOM] och väljer Zoomstat kan man få en grov inställning som man sedan kan justera.



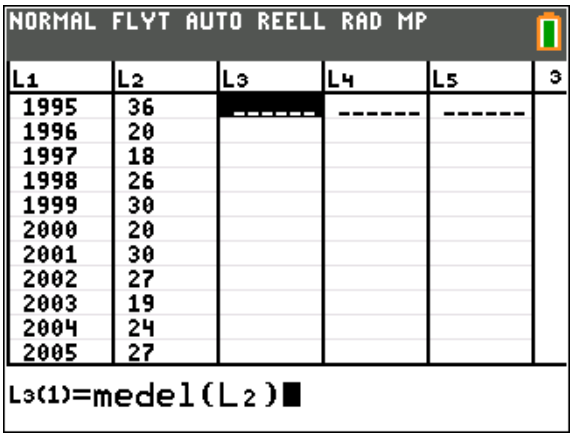
Vi ser att antalet olyckor hoppar lite upp och ner. Vi har några värden som avviker. För att få en samlad bild av utvecklingen kan man beräkna *medelvärdet*.

Medelvärde

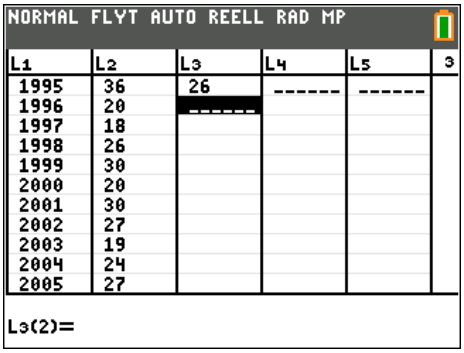
Om vi går tillbaka till statistikeditorn kan vi enkelt göra den beräkningen. Placera då markören i första raden i kolumn L3 t.ex., och tryck på [2nd][LIST]. Välj MA på översta raden.



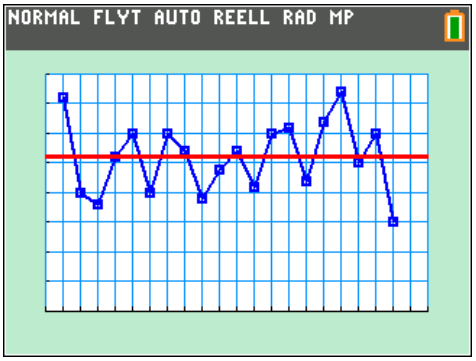
Då får du uppsättning på beräkningar du kan göra på listor. Välj nu 3:medel och skriv så här:



Tryck på [ENTER] och du får en beräkning av medelvärdet.



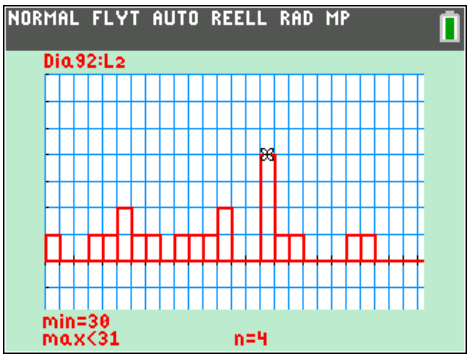
Nu kan vi lägga in medelvärdet i diagrammet genom att i editorn för funktioner (tryck på [Y=]) skriva Y1=26.



Vi ser att vi har ett ungefär lika många värden över som under medelvärdet. Medelvärdet säger ju bara hur många olyckor det har varit i *genomsnitt*. Vi återkommer till detta.

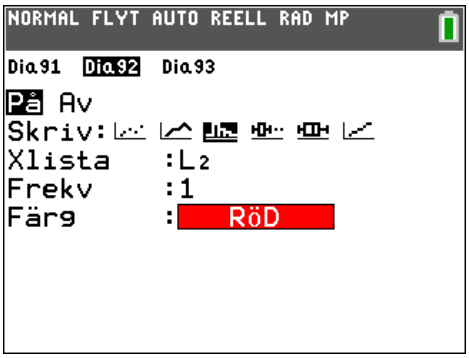
Ett sätt att sammanställa våra data är att göra en *frekvensindelning*, dvs. titta på under hur många år som det t.ex. var mellan 15 och 20 olyckor. Man kan då visa frekvensindelningen genom att rita *histogram*.

Om vi ritar histogrammet så att varje förekommande värde ska visas så blir det så här:



Vi ser t.ex. att värdet 30 förekommer 4 gånger. Då det är många staplar och nästan alla värden har frekvensen 1 så delar vi in data i grupper eller *klasser*. Vi gör en klassindelning.

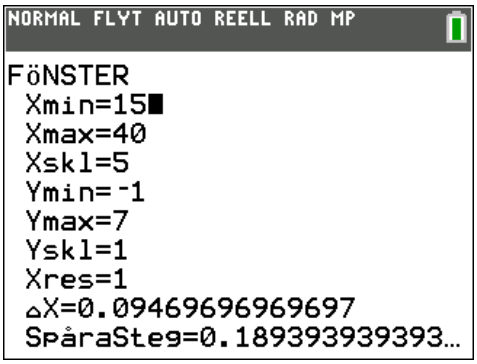
Diagraminställningen ser ut så här:



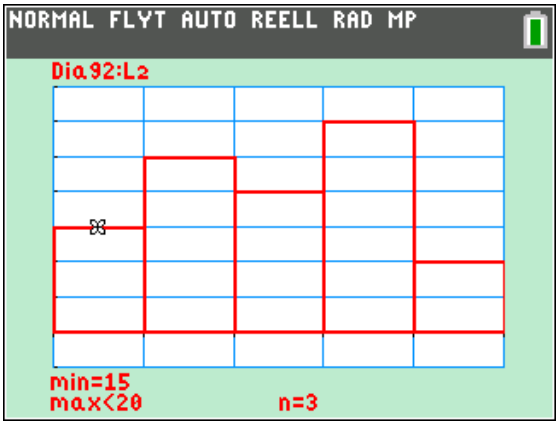
Vi ska alltså rita histogram. Data ligger i lista L2 och varje värde förekommer en gång.

Nu kommer vi till en viktig sak: vi måste ställa in bredden på klasserna. Säg att vi vill räkna hur många gånger vi hade 15-20 olyckor. Det betyder att vi ska ha en klassbredd på 5.

Då ställer vi in vårt fönster så här:

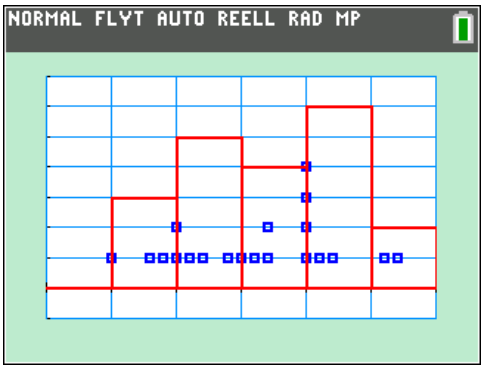


Det viktiga här är att vi anger **Xskl** till 5.
Nu kan vi plotta histogrammet.



Vi få tre värden i intervallet 15-20, dvs. värdena 15, 16, 17, 18 eller 19 förekommer 3 gånger. 20 tillhör nästa klass.

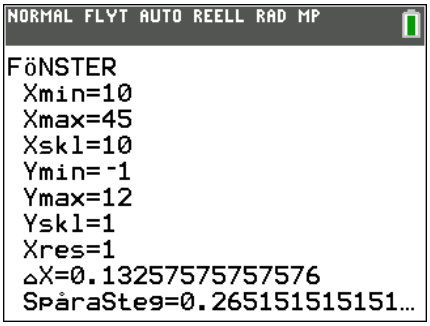
Om vi lägger in alla data som punkter i diagrammet så ser det ut enligt nedan. Lika värden staplas *på varandra*. Sådana diagram kallas *punktdiagram*. Räknaren kan inte plotta punktdiagram utan vi har använt ett speciellt räknarprogram för detta.



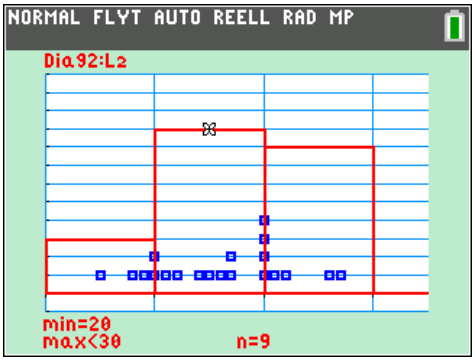
Titta på första stapeln. Där har vi tre värden. De värden som ligger på gränsen tillhör nästa klass. Där ska vi ha värden 20, 21, 22, 23 och 24. Enligt stapeln så ska det vara 5 st och om du räknar punkterna fram till den högra gränsen så blir det också 5 st.

Vi kan ställa om klassbredden till 10 och rita staplar för intervallen 10-20, 20-30, 30-40

Då blir fönsterinställningen så här:



Diagrammet blir så här:



Du har 9 värden i intervallet 20-30. Observera att värdet 30 (förekommer 4 gånger) hör till nästa klass.

En frekvenstabell skulle då se ut så här:

olyckor	frekvens
$10 \le x < 20$	3
$20 \le x < 30$	9
$30 \le x < 40$	8

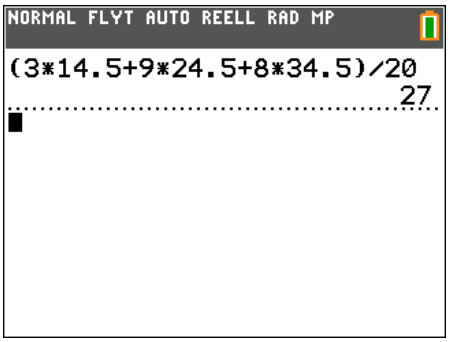
Histogrammet är ju en grafisk frekvenstabell.

Hur gör man om vi bara har tillgång till frekvenstabellen ovan och ska beräkna medelvärdet?

Då får man använda något som kallas *klassmitten*. Man tänker sig att alla inom en viss klass ligger i mitten. I klasserna ovan skulle det då bli 14, 24 och 34. Första klassen innehåller ju värdena
10, 11, 12, 13, **14**, **15**, 16, 17, 18, 19

och det mittersta värdet är ju medelvärdet av de två i mitten, dvs. 14,5.

Nu beräknar vi medelvärdet



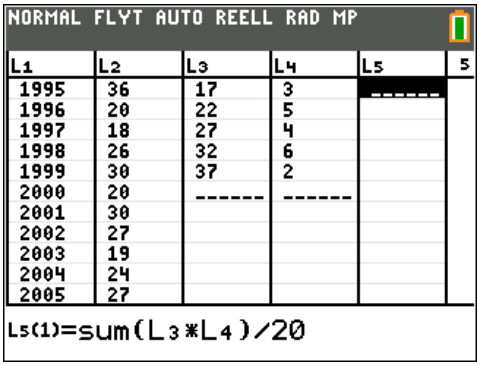
Vi får ett värde som ligger lite högre än det verkliga medelvärdet, som är 26.

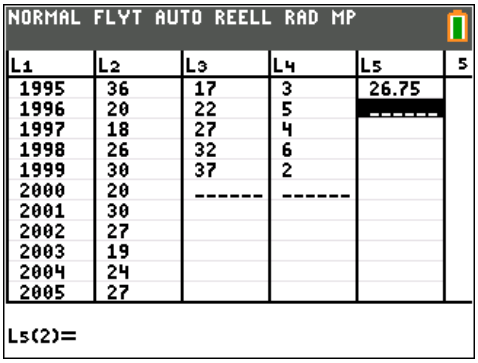
Om vi istället har 5 klasser enligt diagrammet i vänstra spalten så får vi följande frekvenstabell:

$15 \le x < 20$	3
$20 \le x < 25$	5
$25 \le x < 30$	4
$30 \le x < 35$	6
$35 \le x < 40$	2

Första klassen innehåller värdena 15, 16, **17**, 18, 19. 17 ligger alltså i mitten.

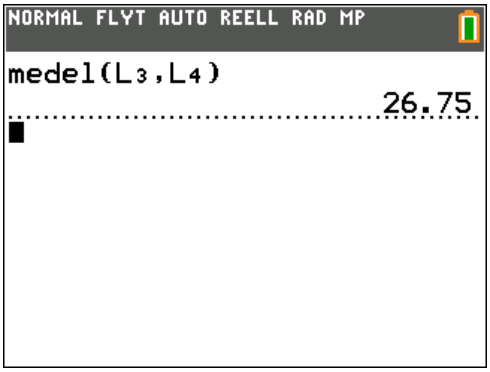
Vi kan räkna ut det i statistikeditorn så här:
Vi har klassmitten i kolumn L3 och frekvenserna i kolumn L4. På rad L5(1) skriver vi sedan





Vi får resultatet 26,75.

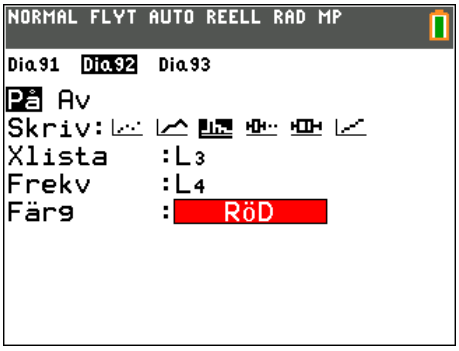
Man kan också *direkt* göra en beräkning av medelvärdet om man har en frekvenstabell:



Man skriver alltså in i vilken lista man har klassmitten och frekvensen av data i intervallet.

Vanligtvis tänker man sig att man har data som kan anta *alla* värden inom ett intervall och då brukar man använda klassmitten som medelvärdet i intervallet. I vårt exempel har vi heltalsvärden. Antalet olyckor kan ju bara vara hela tal. Intervallet $15 \leq x < 20$ har då klassmitten 17,5.

Man gör på motsvarande sätt om man ska rita ett histogram. Då blir diagraminställningen så här:

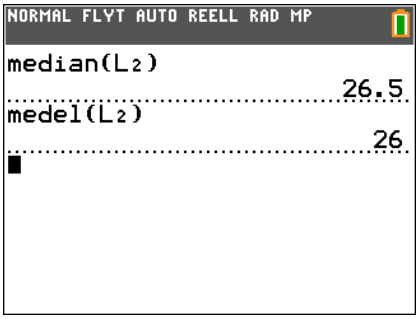


Pröva gärna detta!

Median

Nu har vi beräknat medelvärdet på olika sätt. Det finns ett annat mått, *median*, som är det värde som ligger i mitten om man sorterar alla värden. Om det är ett jämnt antal värden så tar man medelvärdet av de två talen i mitten.

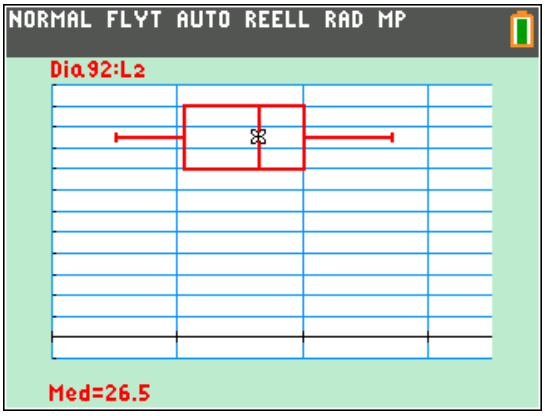
Medianen räknar man ut så här på räknaren:



Medelvärde och median är ganska lika.

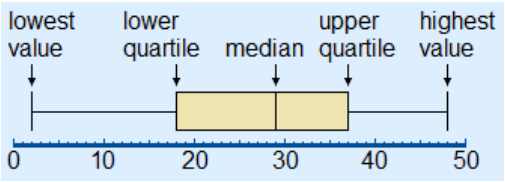
Det finns en särskild diagramform, *lådagram*, som visar medianen på ett mycket tydligt sätt. Vi ställer alltså om diagramplottningen till lådagram.

Dags att plotta! Vi ser en låda med två utstickare. Ytterkanterna på dessa visar det minsta resp. största värdet. Lådans kanter visar övre gränsen för de 25 % minsta värdena och undre gränsen för de 25 % största. Dessa värden kallas *kvartiler*. Se nästa sida!



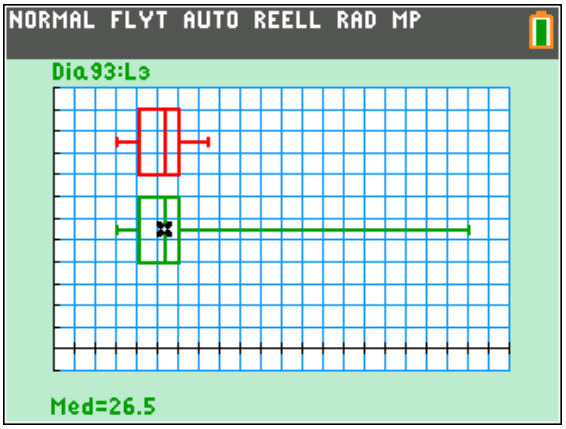
Med TRACE kan man spåra i diagrammet och avläsa minsta värde, den undre kvartilen, medianen, den övre kvartilen och det största värdet.

Medianen är strecket inne i lådan. Till vänster och höger om detta streck ligger 50 % av observationerna. Se bilden nedan.

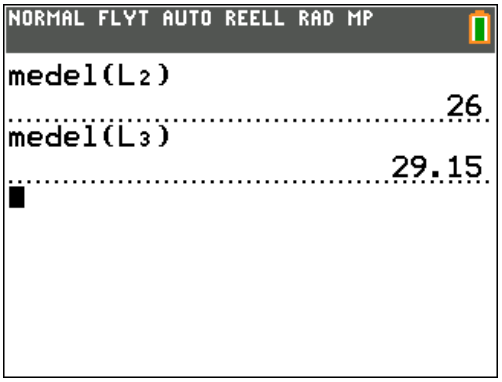


Lådan i mitten är alltså de 50 % som ligger i mitten.

Tänk er nu att vi ändrar det största värdet (37) till 100. Vi lägger då in en ny lista i lista L3 som innehåller samma värden utom för år 2011. Nu ritar vi lådagram för båda listorna.



Vi ser att medianen inte ändras. Den är fortfarande 26,5. Vad händer med medelvärdet? Vi kontrollerar.



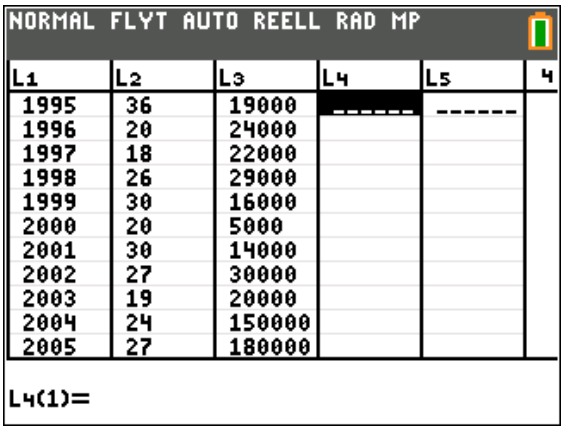
Medelvärdet har höjts med drygt 3.

Både medelvärde och median är mått som används för att visa det genomsnittliga värdet. I vissa fall är medelvärde att föredra och i andra fall medianvärdet.

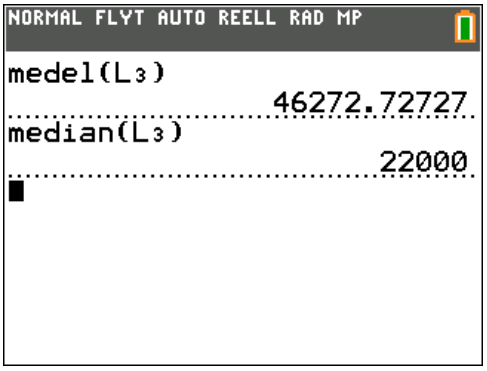
Medianen kan vara ett lämpligt mått om data har en sned fördelning med många höga eller låga värden. I motsats till medelvärdet påverkas *inte* medianen av sådana extremvärden.

Ett exempel när där medianen är ett bättre mått är t.ex. inkomster. Många har låga eller medelhöga inkomster och ett fåtal har höga eller mycket höga inkomster. Ett medelvärde skulle i detta fall ge en missvisande bild eftersom de med höga inkomster drar upp medelvärdet.

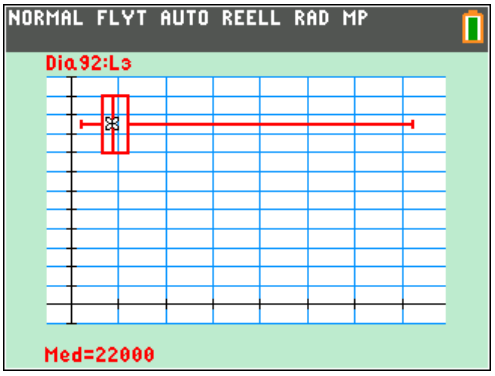
Vi tar ett exempel. Vi tänker oss att en grupp människor har följande inkomster. Se lista L3 nedan. De två nedersta inkomsterna är alltså väldigt höga.



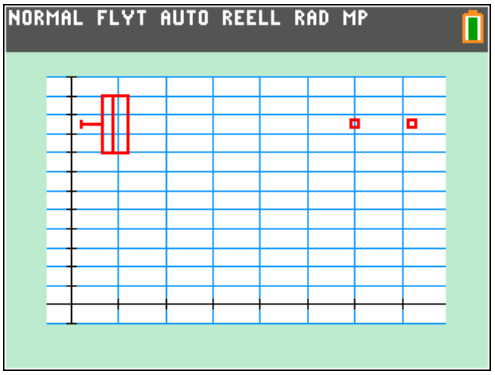
Vi beräknar nu medelvärde och median.



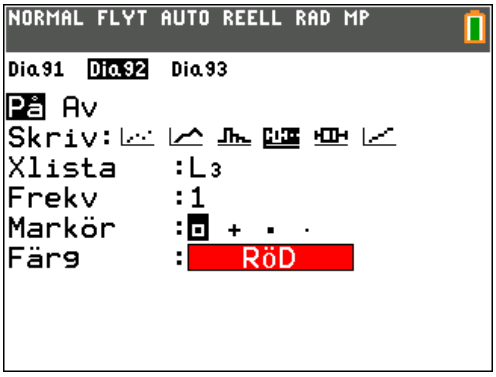
Att ange medelinkomsten är ju här missvisande. Medianen är ett bättre mått.



När man har en sådan här sned fördelning kan man plotta diagrammet så här:



Man ställer då in diagramvisningen så här:



De två yttervärdena visas då som två punkter och är värden som ligger långt ifrån lådans kanter. Sådana värden kallas *utliggare*.

Spridning

Vi har nu gått igenom hur man med räknaren kan beräkna medelvärden och medianer. De är båda s.k. *lägesmått*, dvs. mått som på något sätt visar var tyngdpunkten ligger.

Nu är man ju också intresserad av att beräkna ett mått på *spridningen* i datamaterialet. Ta t.ex. följande enkla exempel.

A: 1000 2000 3000

B: 1000 1500 2000 2500 3000

Båda datauppsättningarna har medelvärdet och medianen 2000 men är spridningen lika? Då måste vi först definiera vad spridning är. Vi gick ju egentligen igenom spridning när vi ritade lådagram. Lådans kanter talar om var 50 % av datamaterialet ligger och de vågräta strecken anger hur långt det är mellan det lägsta till högsta värdet. Det kallas för variationsbredd.

Nu vill vi på något sätt skaffa oss ett mått på den typiska eller genomsnittliga *avvikelsen* är från medelvärdet. Ett sådant värde kallas för *standardavvikelse*.

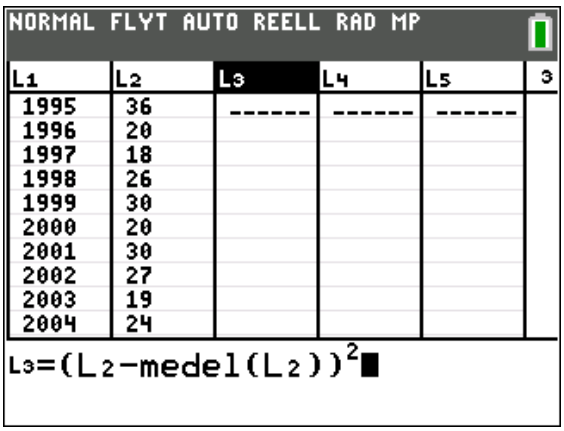
Definition

För värden $x_1, x_2, \ldots, x_n$ med medelvärde $\bar{x}$ är standardavvikelsen

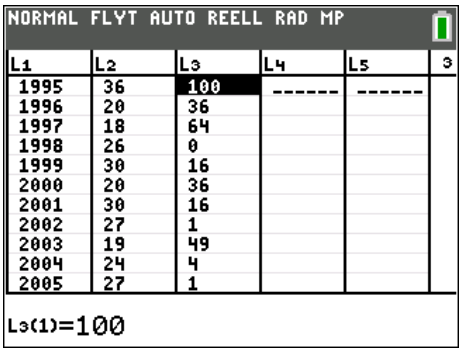
$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

Det här ser ju krångligt ut men vi faktiskt ganska enkelt göra beräkningen på räknaren.

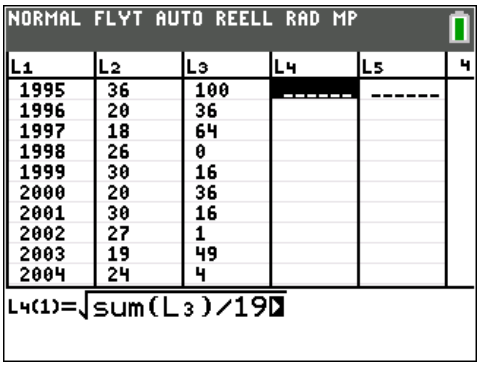
Först beräkningarna avvikelsen frånmedelvärdet i kvadrat. Se formeln.



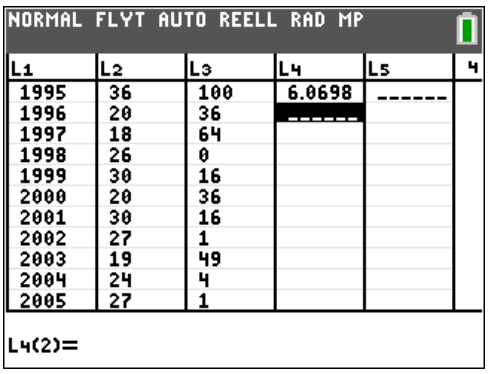
Det blir så här:



Nu tar vi roten ur summan av avvikelserna och dividerar med antalet värden -1.

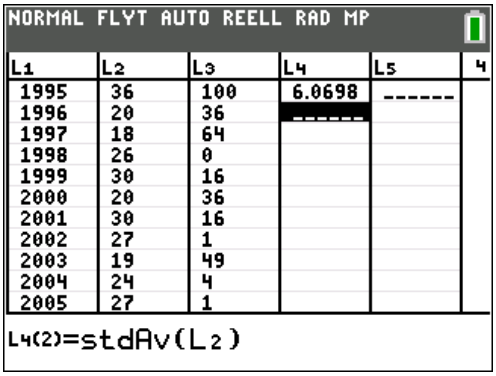


Tryck på ENTER.



Vi får att standardavvikelsen blir 6,1.

Nu kan vi ju direkt beräkna detta med räknarens inbyggda funktion. Skriv då enligt nedan.



Vi får då samma värde.

Ändra nu värdet nu det största värdet i lista L2 igen från 37 till 100. Vilken standardavvikelse får du då?

Medelvärde och standardavvikelse är viktiga mått när man studerar s.k. normalfördelningar och när man ska dra statistiska slutsatser från ett slumpmässigt urval av data.