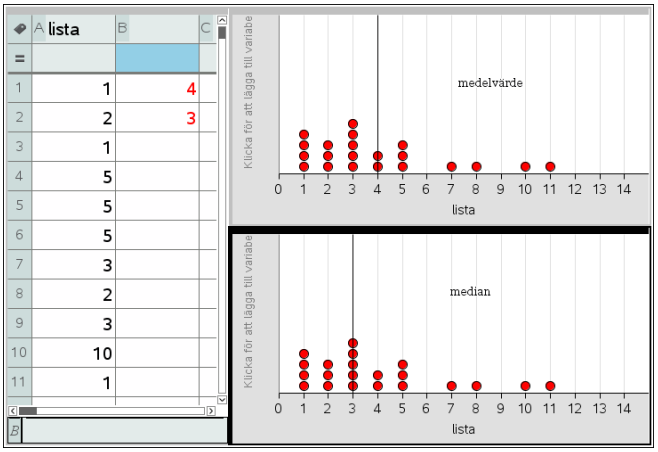
Medelvärde, median och standardavvikelse

Detta är en enkel aktivitet där vi på ett dynamiskt sätt ska titta på hur de statistiska måtten, t.ex. median och medelvärde ändras när man ändar ett värde i en data-uppsättning.

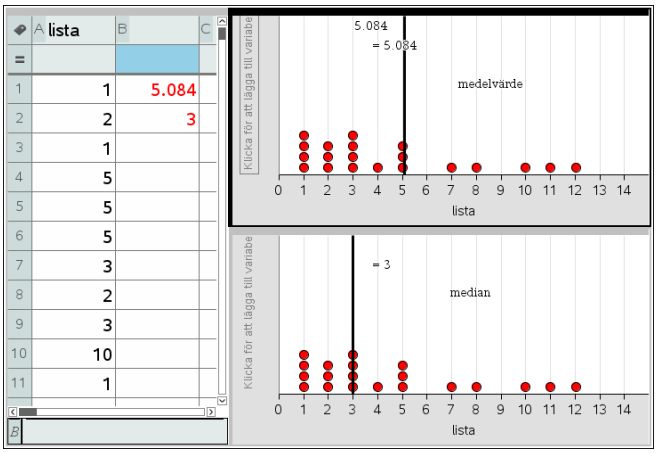
Problem 1:

På sid 1 har vi en samling data och vi har sedan plottat datauppsättningen som punktdiagram i två likadana diagram. Sedan har vi lagt till linjer som visar medel-värde och median. Det gör man med verktyget ”Rita värde” som finns under Analysera i verktygsmenyn. Man skriver då enligt syntaxen
medelvärde:=mean(lista)
eftersom våra värden finns i variabeln lista.

Vi ser att medelvärdet är 4 och medianen 3 i den ursprungliga listan. Det kan du se om du klickar på linjerna.

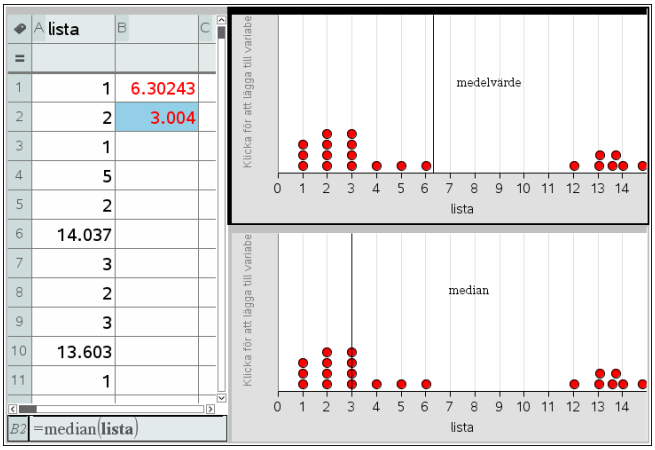


Dra nu i en punkt i något av diagrammen. Hur förändras då medelvärde och median?



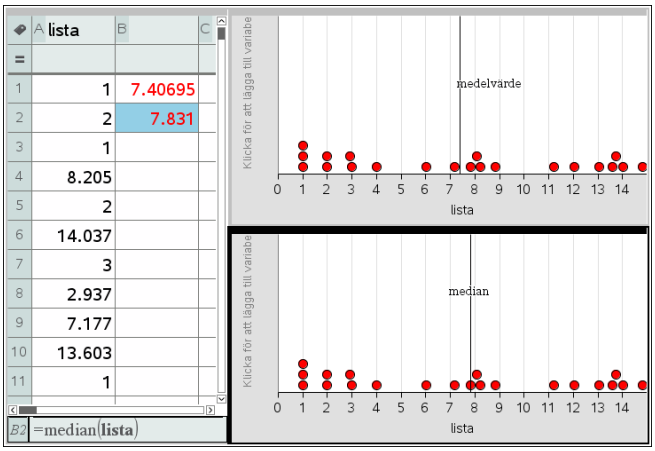
Det kan se ut så här. Vilka iakttagelser kan man då göra? Jo, medianen förändras inte alls men däremot har medelvärdet ökat till drygt 5.

Vi prövar med några dragningar till.



Vi ser att medianen knappast förändras alls. När vi drar datapunkterna så kan man inte placera dem direkt som heltal så därför få vi nu ett decimaltal som medianvärde. Alternativet är då att gå in i kalkylarket och ändra.

För att få en mer drastisk ändring av medianen så får man flytta lite fler datapunkter. Här bli nu medianen större än medelvärdet.



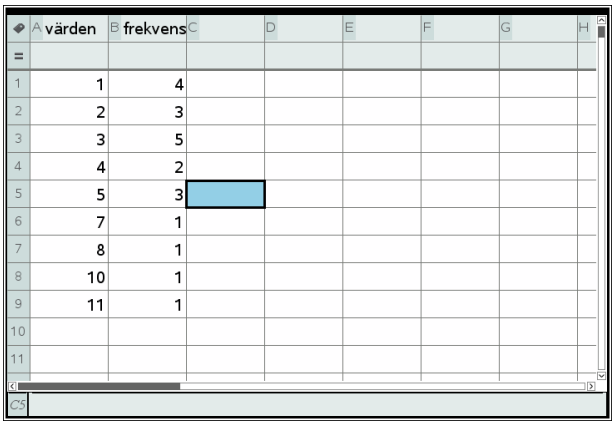
Vi har alltså konstaterat att medianen är mer okänslig för förändringar av enskilda värden.

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

Vi kan konstatera att medelvärde och median till-sammans inte alltid ger en bra och sammanfattande beskrivning av ett datamaterial. Det ska vi undersöka närmare. Först kommer ett lite tips hur man gör om de data man har redan är ordnade i en frekvenstabell.

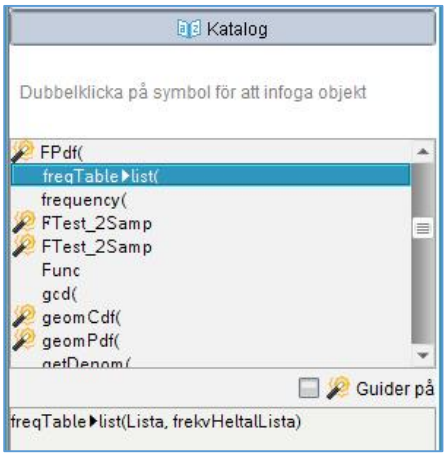
Problem 2

Ibland har man sina data i frekvenstabeller. Vi har här nu samma data som vi ursprungligen hade i problem 1 men nu är de ordnade i en frekvenstabell. Man kan fortfarande beräkna medelvärde och median. Då kan man i en ledig cell skriva *=mean(värden,frekvens)* för att beräkna medelvärdet.

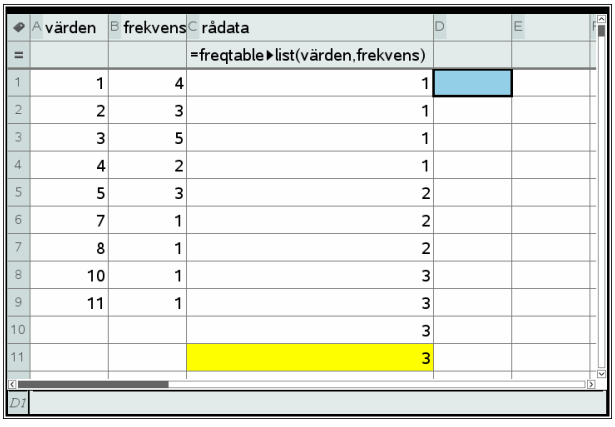


	A värden	B frekvens
1	1	4
2	2	3
3	3	5
4	4	2
5	5	3
6	7	1
7	8	1
8	10	1
9	11	1
10		
11		

Utifrån frekvenstabellen kan man inte rita punktdiagram men det finns ett knep att fixa detta. Man använder då en speciell instruktion som visas nedan. Instruktionen finns i katalogen som du når från Dokumentverktygslådan:



Så här blir nu listan med värden. Den är sorterad i stigande ordning.



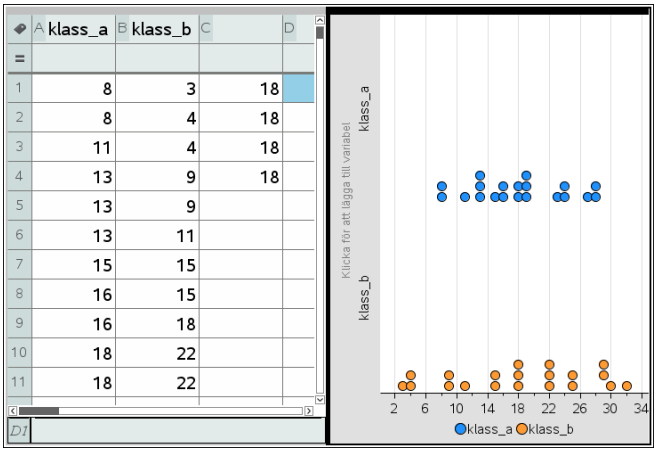
	A värden	B frekvens	C rådata
			=freqtable▶list(värden,frekvens)
1	1	4	1
2	2	3	1
3	3	5	1
4	4	2	1
5	5	3	2
6	7	1	2
7	8	1	2
8	10	1	3
9	11	1	3
10			3
11			3

Problem 3

Nu ska titta lite närmare på ett begrepp som heter *spridning*. Ofta så räcker ju inte medelvärde och median till för att ge en bra bild hur ett datamaterial ser ut.

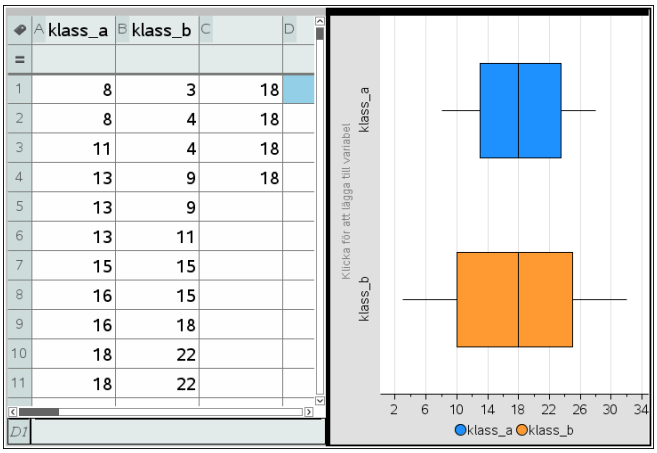
I kalkylarket nedan visar vi resultatet på ett prov för två klasser. Både medelvärde och median är lika (18 poäng) i båda klasserna.

Vi har här plottat punktdiagram för båda klasserna och vi ser att spridningen är större i klass b. kan man få ett bra mått på detta. Vi ser att differensen mellan största och minsta värdet i de två klasserna är 20 poäng i klass a och 29 poäng i klass b. Detta kallas variationsbredd och ger ett mått på spridningen.



	A klass_a	B klass_b	C	D
1	8	3	18	
2	8	4	18	
3	11	4	18	
4	13	9	18	
5	13	9		
6	13	11		
7	15	15		
8	16	15		
9	16	18		
10	18	22		
11	18	22		

Ett sätt att visa detta är att rita lådagram. Klicka i diagrammet och välj då denna diagramform.



	A klass_a	B klass_b	C	D
1	8	3	18	
2	8	4	18	
3	11	4	18	
4	13	9	18	
5	13	9		
6	13	11		
7	15	15		
8	16	15		
9	16	18		
10	18	22		
11	18	22		

Här få man också en bra bild av spridningen. Man ser ju inte alla enskilda datavärden som i ett punktdiagram utan får en sammanfattande bild. Strecket mitt i lådan är medianen och lådans kanter är undre respektive övre kvartilen.

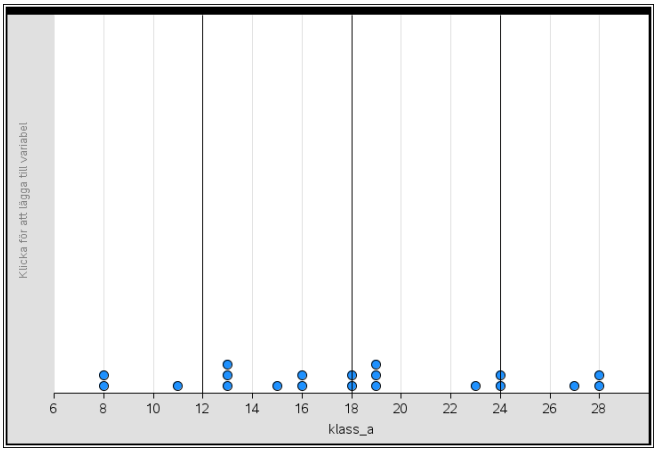
Ett mått som nästan alltid används som mått på spridningen är *standardavvikelsen*. Man kan säga att standardavvikelsen ger ett mått på den genomsnittliga avvikelsen från medelvärdet.

Här har vi nu öppnat en ny sida och plottat värdena för klass a i ett punktdiagram. Vi har också plottat linjer enligt formeln

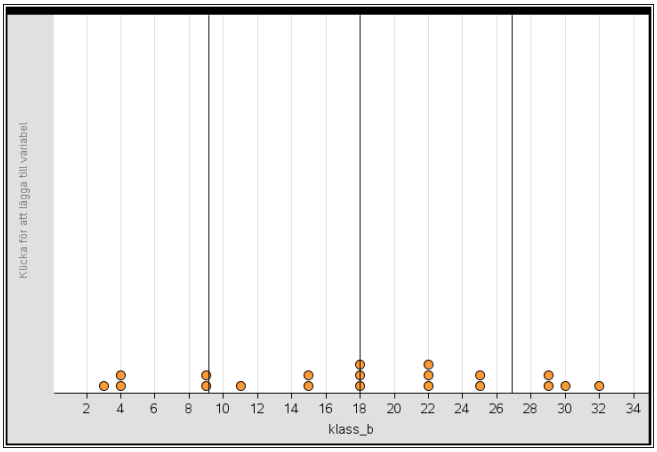
Mean(klass_a)+ stDevPop(klass_a)
respektive
Mean(klass_a) - stDevPop(klass_a)

Instruktionen för standardavvikelsen finns i katalogen som du når från Dokumentverktygslådan.

Då inhägnar vi ett område inom en standardavvikelse från medelvärdet. Vi ser att standardavvikelsen är nästan precis 6.

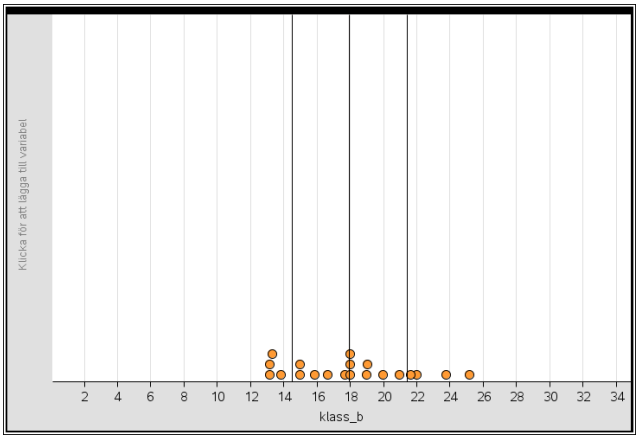


Vi gör likadant med klass b:



Standardavvikelsen är nu ungefär 9, betydligt större än den för klass a alltså. Sammantaget ger nu dessa två diagram ger alltså en väldigt bra bild av resultaten på provet.

Pröva nu att dra i punkter i diagrammen för att öka eller minska standardavvikelsen.



Här har vi packat ihop observationerna för klass b. Vi har samma medelvärde som förut men standardavvikelsen är betydligt mindre. De lodräta linjerna hänger ju med när du flyttar punkterna. Elegant!

Problem 4:

Hur beräknas standardavvikelsen egentligen? TI-Nspire har ju en inbyggd funktion för att direkt beräkna standardavvikelsen från ett datamaterial. Det gäller båda data som ligger i en lista med värden och data som finns i en frekvenstabell.

Vi visar nu i steg hur det går till.

- I kolumn a har vi våra data.
- I kolumn b har vi för varje värde avvikelsen från medelvärdet.
- I kolumn c kvadrerar vi avvikelserna för att summan av avvikelserna i de fortsatta beräkningarna inte ska bli noll. Vissa värden är ju negativa och summerar man värdena i kolumn b blir det ju 0 som resultat.
- I kolumn d på rad 1 så har vi till sist beräknat standardavvikelsen som

$$\sqrt{\frac{\text{summan av avvikelserna i kvadrat}}{20}}$$

Markera cellen så ser du formeln i inmatningsfältet.

	A klass_b	B avvikelse	C i_kvadrat	D	E	F
=		=mean(klass_b)−klass_b	=avvikelse^2			
1	3.	15.	225.	8.8713		
2	4.	14.	196.			
3	4.	14.	196.			
4	9.	9.	81.			
5	9.	9.	81.			
6	11.	7.	49.			
7	15.	3.	9.			
8	15.	3.	9.			
9	18.	0.	0.			
10	22.	-4.	16.			
11	22.	-4.	16.			

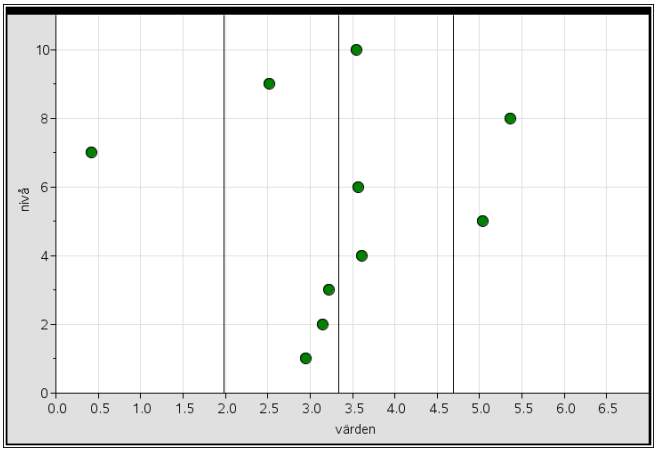
Problem 5:

I detta problem har vi i kalkylarket två listor. Det är datapunkter i kolumnen värden vi ska titta på. Den andra kolumnen har vi bara med för att vi ska kunna plotta datavärden på ett tydligt sätt i ett diagram.

	A värden	B nivå	C	D	E	F	G
=							
1	2.944	1.	medelvärde	3.33682			
2	3.145	2.	standardavv	1.35639			
3	3.218	3.					
4	3.608	4.					
5	5.041	5.					
6	3.564	6.					
7	0.4212	7.					
8	5.365	8.					
9	2.52	9.					
10	3.542	10.					
11							

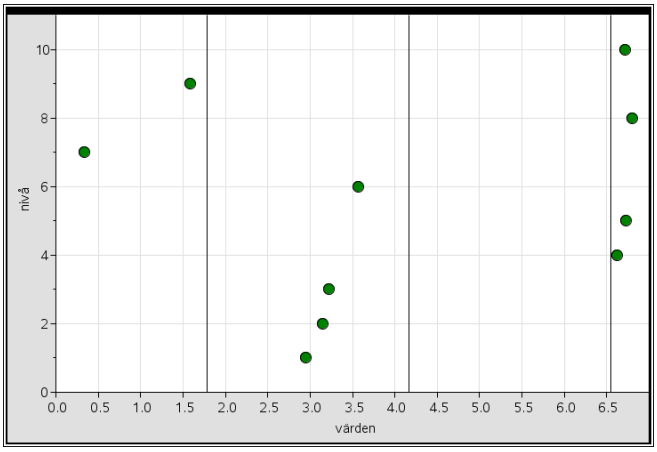
B3 =3

Här har vi plottat datapunkterna på olika nivåer i ett xy-diagram. Vi har också lagt in linjer för medelvärdet och linjer på avståndet en standardavvikelse från medelvärdet. När vi flyttar punkter så flyttar sig också dessa tre linjer. I diagrammet ser vi att tre datapunkter ligger utanför gränsen för en standardavvikelse.

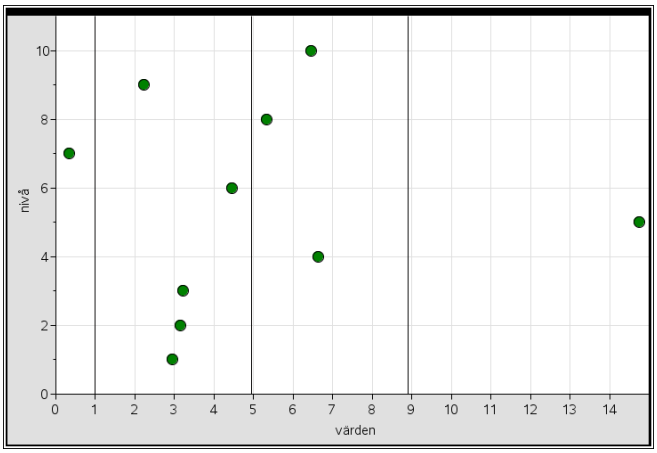


Försök nu att flytta en punkt i taget så att t.ex. 6 punkter ligger utanför dessa gränser. Vad händer då med linjerna? Jo, linjerna flyttas och andra punkter som ligger utanför gränserna riskerar att hamna innanför och vi får börja om igen.

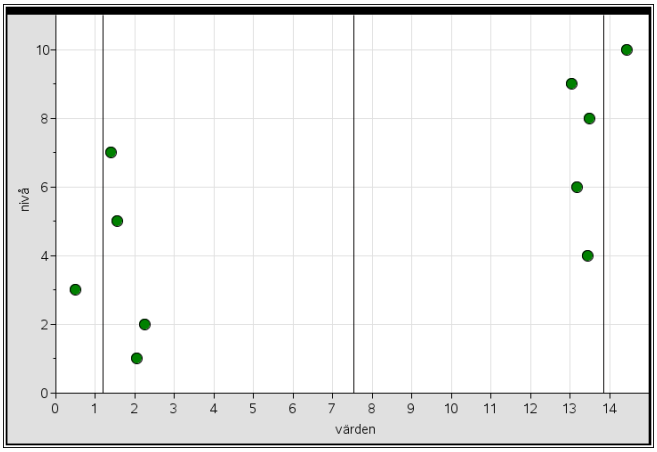
Detta säger något om vad standardavvikelse är. Vi har ju ”bara” 10 punkter och en *stor* förflyttning av bara en punkt från centrum i diagrammet gör att medelvärde och standardavvikelse ändras tillräckligt mycket för att åter de flesta punkterna ska ligga innanför gränsen för en standardavvikelse.



Här har vi dragit iväg med en punkt och vi ser att standardavvikelsen ändras mycket. Den är ca 4 nu. 8 punkter finns i intervallet medelvärde $\pm$ en standardavvikelse.

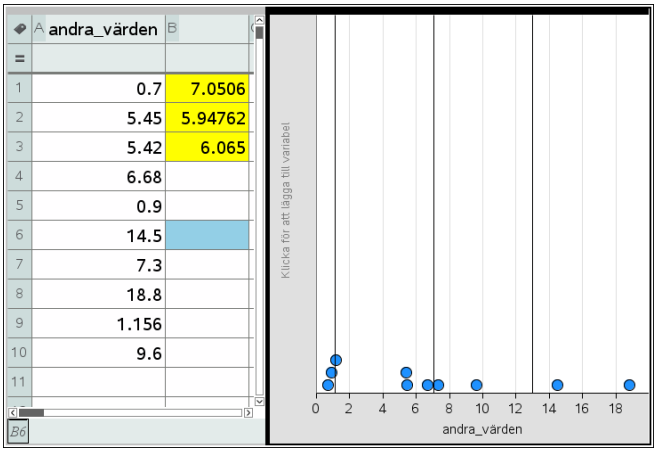


Här en extremsituation. Både medelvärde och median ligger ungefär på 7,5 och nästan alla punkter ligger inom gränserna. Standardavvikelsen är drygt 6.



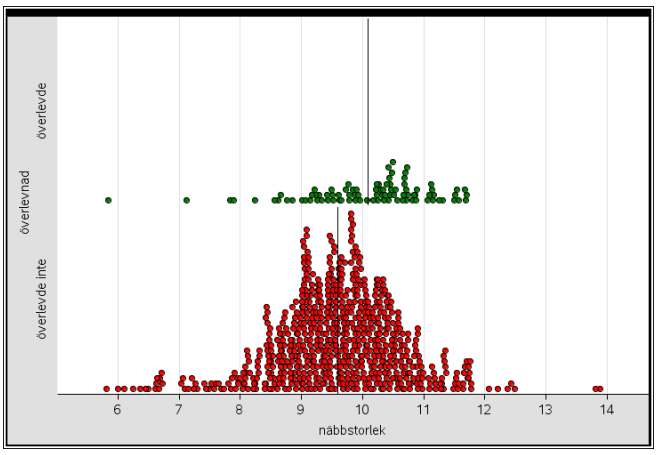
Här en annan datauppsättning som ser helt annorlunda ut trots att de statistiska måtten är ungefär lika. 4 datapunkter ligger nära medelvärdet. Medelvärde, median och standardavvikelse säger alltså inte allt!

I många datamaterial är ofta fördelningen ungefär *normalfördelad* och då ligger ca 2/3 av datapunkterna inom en standardavvikelse från medelvärdet.



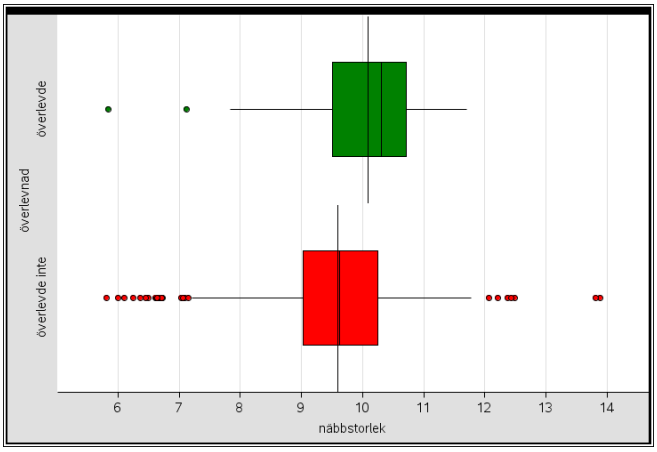
Problem 6:

Vi avslutar denna aktivitet med att visa ett kalkylark med insamlade data från den verkliga världen. Det handlar om näbbstorlekar i mm hos en koloni av finkar på Galapagosöarna och vilka som överlevde en svår torka. Man kan konstatera att finkar med kraftigare näbbar överlevde i högre grad.



I punktdiagrammen har vi lagt in linjer för medelvärdena. Vi har också delat upp variabeln näbbstorlek efter *kategorier*, i detta fall överlevnad. Vi ser den symmetriska fördelningen för de som inte överlevde. Fördelningen för de som överlevde är inte lika symmetrisk.

Om vi klickar i diagrammet kan vi ändra till visning av lådagram också.



Här ser vi både medianer och medelvärden för respektive population. Punkterna till vänster och höger är s.k. *utliggare*, värden som ligger långt ifrån lådans kanter.